Prédiction et caractérisation des biais textuels basés sur le discours

Nicolas Devatine

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE
Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Nicolas DEVATINE

Le 23 octobre 2023

Discourse-Driven Prediction and Characterization of Textual Bias

---

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Informatique et Télécommunications**

Unité de recherche :
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Philippe MULLER et Chloé BRAUD

Jury

M. Alexandre ALLAUZEN, Rapporteur
Mme Karën FORT, Rapporteure
M. Maxime AMBLARD, Examinateur
M. Philippe MULLER, Directeur de thèse
Mme Chloé BRAUD, Co-directrice de thèse
M. Rufin VANRULLEN, Président

Abstract

In an expanding information-based society, where public opinion is influenced by a plurality of sources and discourses, assessing the presence and extent of textual bias is of paramount importance. Therefore, the research undertaken in this thesis revolves around the detection and characterization of such biases, by placing a particular focus on political biases in news articles. What distinguishes this research from prior work on the subject lies in its shift beyond mere lexical analysis of documents. Instead, it integrates argumentative and rhetorical dimensions by considering the structure of the documents. To do so, we draw upon methodologies derived from the field of discourse analysis in Natural Language Processing (NLP). We latently induce a document structure by relying on elementary discourse units, which are sub-components of sentences and constitute the smallest textual unit capable of expressing a coherent proposition or idea. From an extensive set of experiments on the prediction of political leanings in news articles, we not only reveal the effectiveness of the proposed discourse-driven method, but also highlight several noteworthy findings that hold potential implications for further research. However, the ambition of this thesis goes beyond simply predicting biases, we aim to characterize them by getting some insights into the model's decisions. We therefore delve into the growing field of explainability in NLP, by making a particular focus on model-agnostic and perturbation-based explanation methods for text classification. While such methods have previously demonstrated their effectiveness across a wide range of tasks, they are not without their limitations, especially in terms of their computational cost and their ability to process long documents. To address these shortcomings, we propose a series of new strategies based on different levels of granularity. These include the development of explanation methods centered on discourse units, on specific vocabularies of interest, or on the document structure induced by the model. Following on from the experiments carried out on the prediction of political leanings in news articles, we evaluate both quantitatively and qualitatively the explanations generated for this task using our approach and demonstrate the benefits of the proposed strategies over existing methods. Thus, this work introduces a new perspective to the analysis of textual biases in NLP by proposing an integrated discourse-driven method for both predicting and characterizing biases.

Résumé

Dans une société de l'information en pleine expansion, où l'opinion publique est influencée par une pluralité de sources et de discours, l'étude de la présence et de l'étendue des biais dans les textes se révèle être d'une importance capitale. Ainsi, la recherche menée dans cette thèse s'articule autour de la détection et de la caractérisation de ces biais, en mettant un accent particulier sur les biais politiques dans les articles de presse. Ce qui distingue notre étude des travaux existants sur le sujet est que nous allons au-delà de la simple analyse lexicale des documents. En effet, nous intégrons également les dimensions argumentatives et rhétoriques en prenant en compte la structure du texte. Pour ce faire, nous nous appuyons sur des méthodes dérivées du domaine de l'analyse du discours en Traitement Automatique des Langues (TAL). Nous induisons de manière latente une structure du document basée sur les unités élémentaires de discours, qui sont des sous-composants des phrases et qui constituent les plus petites unités textuelles capables d'exprimer une proposition ou une idée cohérente. À partir d'un ensemble d'expériences sur la prédiction des biais politiques dans les articles de presse, nous démontrons à la fois l'efficacité de la méthode proposée basée sur le discours et soulignons également plusieurs résultats notables ayant de potentielles implications pour de futures recherches. Cependant, l'ambition de cette thèse dépasse la simple prédiction des biais, nous cherchons aussi à les caractériser en examinant les décisions du modèle. Nous nous intéressons ainsi au domaine de l'explicabilité en TAL, en nous concentrant plus particulièrement sur les méthodes d'explication agnostiques au modèle et basées sur des perturbations pour la classification de texte. Bien que ces méthodes aient démontré leur efficacité sur un grand nombre de tâches, elles présentent certaines limites, notamment en ce qui concerne leur coût de calcul et leur capacité à traiter les documents longs. Afin de remédier à ces problèmes, nous proposons plusieurs nouvelles stratégies basées sur différents niveaux de granularité, parmi lesquelles le développement de méthodes d'explication basées sur les unités discursives, sur des vocabulaires spécifiques d'intérêt ou sur la structure du document induite par le modèle. Dans la continuité des expériences menées sur la prédiction des biais politiques dans les articles de presse, nous évaluons quantitativement et qualitativement les explications générées à l'aide de notre approche pour cette tâche et démontrons les bénéfices des stratégies proposées par rapport aux méthodes existantes. Ainsi, ce travail apporte une nouvelle perspective à l'analyse des biais textuels en TAL en proposant une méthode intégrée basée sur le discours permettant à la fois de prédire et de caractériser les biais.

Remerciements

Je tiens à exprimer ma profonde gratitude à toutes les personnes qui ont contribué à la réalisation de cette thèse et qui m'ont soutenu durant ces trois années.

En premier lieu, je souhaite remercier mes directeurs de thèse, Philippe Muller et Chloé Braud, pour leur soutien indéfectible et leur accompagnement précieux. Je leur suis profondément reconnaissant pour tout ce qu'ils m'ont apporté tant sur le plan scientifique qu'humain. Merci d'avoir cru en moi, de m'avoir toujours encouragé, en particulier durant la période du Covid-19, et pour tout le temps que vous m'avez consacré. J'ai adoré travailler avec vous et nos rendez-vous hebdomadaires me manqueront beaucoup.

Je souhaite également remercier les rapporteurs de cette thèse d'avoir accepté de relire et d'évaluer mon travail, Alexandre Allauzen et Karën Fort, dont les retours constructifs ont été très appréciés et ont participé à l'amélioration de ce manuscrit. Ma reconnaissance va aussi aux autres membres du jury et aux membres invités, Maxime Amblard, Rufin VanRullen et Tatjana Scheffler, pour l'intérêt porté à mon travail et leur présence durant la soutenance. Les échanges passionnants que nous avons eus m'ont permis d'enrichir ma réflexion et d'envisager de nouvelles perspectives pour le futur.

Un remerciement tout particulier à l'IRIT et aux membres de l'équipe MELODI, pour m'avoir accueilli et offert un environnement de travail stimulant dans lequel j'ai pu pleinement trouver ma place. Les échanges et les moments de partage que j'ai pu avoir avec vous ont largement contribué à mon épanouissement scientifique. Je n'oublie pas mes collègues de bureau pour l'ambiance conviviale, les instants d'entraide, et les pauses gourmandes autour d'un thé ou d'un café. Nos discussions, tant professionnelles que personnelles, et votre soutien constant ont été un pilier essentiel dans mon parcours de doctorant.

Je remercie également l'ANR pour le financement de cette thèse ainsi que le projet SLANT, dans lequel s'inscrit mon travail, pour m'avoir offert cette opportunité de recherche passionnante.

Je suis infiniment reconnaissant envers ma famille et mes amis. Votre soutien sans faille durant ces années, et particulièrement durant la période difficile du Covid-19, a été ma force. Merci à mes parents pour leur amour, leurs encouragements, pour avoir toujours cru en moi, et pour m'avoir permis d'en arriver là. Vous m'avez tant apporté et je vous dois beaucoup. À mon frère et ma sœur, pour votre compréhension, pour m'avoir supporté et avoir su me changer les idées dans les moments difficiles. Mes remerciements vont également à ma famille de Toulouse : mes grands-parents, ma tante, mon oncle et mes cousines qui ont été d'un grand soutien pour moi dans cette nouvelle ville. Votre accueil chaleureux, votre présence et nos repas partagés ont été pour moi un réconfort précieux. Je souhaite aussi remercier ma famille de Paris : ma grand-mère, ma tante, mes cousins. Malgré la distance et le manque d'occasion pour se voir, vous m'avez toujours entouré d'affection et de compréhension. Je n'oublierai pas nos nombreuses soirées passées sur Discord durant lesquelles je pouvais me vider la tête.

Un merci particulier à Jason, avec qui j'ai pu développer mon intérêt pour l'informatique à travers nos nombreux projets, ce qui m'a mené à poursuivre mes études dans ce domaine. C'est en partie grâce à toi si je suis arrivé à ce stade de ma vie professionnelle.

Enfin, je dédie une partie importante de ces remerciements à ma moitié. Son soutien inconditionnel, sa patience et son amour ont été ma plus grande force dans l'achèvement de cette thèse. Anne Lise, ta présence à mes côtés, tes encouragements et ta compréhension dans les moments les plus intenses ont été primordiaux. Merci pour ta tendresse et ta bienveillance, merci pour tes conseils avisés et ton écoute attentive, merci d'avoir supporté mes doutes et mes longues heures de travail. Merci d'avoir partagé ces moments avec moi, et pour tous ceux à venir.

À vous tous, mes sincères remerciements.

Contents

List of Tables

List of Figures

Chapter 1

Getting to the Slant

In the age of information, we are surrounded by a vast amount of textual data, shaping our perspectives, opinions, and understanding of the world. While the analysis and processing of such data have significantly improved over the years, the study of the biases inherent in these texts has remained a challenge. Such biases, whether conscious or unconscious, may influence our cognition and decision-making processes, and consequently, have a profound impact on society at large. We explore the intersection of natural language processing (NLP) and textual bias, identifying the various factors that contribute to the propagation of biases in texts, and the consequences that arise from them.[1]

This first chapter lays the groundwork for our exploration by contextualizing the study and framing the key notions that will be discussed throughout the thesis. This chapter also provides an up-to-date review of the literature on the subject, as well as an illustration of the different dimensions of textual bias. We begin by establishing the contextual background and providing a clear definition of textual bias, before delving into its manifestations within the field of NLP. In the subsequent sections, we narrow our focus to the political biases present in news articles, as these forms of bias are particularly pervasive and influential in shaping public opinion. To further illustrate the concepts introduced, we present a case study that examines political bias in news articles. This real-world example serves to shed light on the potential influence of textual bias on public opinion, political discourse, and decision-making processes.

By examining the multifaceted nature of biases in texts, we aim to provide a deeper understanding of the current advances and remaining challenges that lie ahead in the pursuit of more ethical, equitable, and transparent sharing of information from an NLP perspective.

[1]This thesis was funded and supported by the ANR (ANR-19-CE23-0022) and the SLANT project: https://www.irit.fr/Slant/.

1.1 Contextual Background and Definition

Textual bias is a widespread issue impacting the credibility and fairness of information sharing across numerous domains, such as media, academia, or public discussions. It involves systematic deviations from objectivity or neutrality in the presentation, interpretation, or selection of information in written content, including news articles, research papers, political speeches, or social media posts. One of the fundamental factors contributing to its manifestation is the inherent subjectivity of human authors who, intentionally or not, inject their personal beliefs, preferences, and values into their writing. Furthermore, cultural, social, and historical contexts can shape the framing and perception of information, perpetuating certain biases (Eberhardt, 2019). Textual bias manifests in several forms, each with a unique impact on the reader's interpretation. Among the most prevalent biases are:

- **Selection bias:** The systematic over- or under-representation of particular topics, sources, or perspectives in a text, occurring when authors focus on subjects or viewpoints that align with their beliefs or interests while ignoring or downplaying alternative perspectives.

  An example of selection bias

  A book covering the history of a country may place excessive emphasis on the narrative of victory, leading to a biased perspective: "In the glorious revolution, our forefathers heroically defended the homeland against the invaders, resulting in a decisive victory."

- **Framing bias:** Presenting information in a manner that emphasizes specific aspects or interpretations, often excluding alternative perspectives (Tversky and Kahneman, 1981; Entman, 1993). This can involve using loaded language, rhetorical devices, and narrative structures to influence the reader's understanding of a topic or issue. In the field of psychology, this notion is known as the "framing effect", and Tversky and Kahneman (1981) demonstrated that the way choices are framed, in terms of gains or losses, can significantly influence decision-making outcomes.

  An example of framing bias

  These two sentences relay the same information but framed differently, leading to different reader perceptions: "Immigrants take away our jobs." vs. "Immigrants contribute to our economy."

- **Language bias:** The use of biased, prejudiced, or discriminatory language, manifesting as slanted word choices or subtle linguistic cues that convey negative or positive judgments of specific groups or individuals.

An example of language bias

"The aggressive protesters demanded an end to the regulations." The word "aggressive" has a negative connotation, which influences readers' perceptions of the demonstrators.

- **Confirmation bias:** The tendency to seek or interpret information that validates pre-existing beliefs or expectations (Nickerson, 1998). This results in selectively presenting evidence or arguments supporting the author's viewpoint while disregarding or downplaying contradictory information.

An example of confirmation bias

A climate change denier may selectively cite studies that question the extent of human contribution to global warming, ignoring the vast body of literature that affirms it: "Though some claim that human activities are leading to climate change, a study found no significant correlation between CO_2 emissions and global warming."

These types of textual bias are not mutually exclusive, and a single text may exhibit several of them. The effects of textual bias can be far-reaching, impacting not only individual readers, but also society and knowledge dissemination as a whole. At the individual level, exposure to biased texts can lead to the formation of distorted beliefs, attitudes or perceptions, influencing decision-making or support for specific policies or actions (Gentzkow and Shapiro, 2010). At the societal level, textual bias can contribute to the fragmentation and polarization of public discourse, as people are increasingly exposed to information that confirms their pre-existing beliefs while isolating themselves from opposing viewpoints (Flaxman et al., 2016; Marie et al., 2023). This polarization can intensify social divisions and undermine democratic processes, as individuals become less willing to engage in constructive dialogue with those who hold different views. In the context of knowledge dissemination, textual bias can hinder the development of a comprehensive and accurate understanding of complex issues by privileging certain perspectives, theories or findings over others (O'Connor et al., 2023). This is particularly problematic in scientific research, where biased reporting or interpretation of results can lead to the dissemination

of false or misleading information and erode the public trust in science (Ioannidis, 2005).

The growing prevalence of digital media and social networks has amplified the importance of understanding and addressing textual bias. The internet allows for rapid information dissemination, often without oversight or verification, potentially exacerbating textual biases. Additionally, the algorithms used by search engines and social media platforms can inadvertently reinforce bias by prioritizing content that matches user's beliefs or interests, creating "filter bubbles" (Pariser, 2011). As a result, the study of textual bias has moved beyond the traditional areas of media and communication studies to encompass the fields of artificial intelligence (AI) and NLP (Blodgett et al., 2020). It is important to recognize the growing role of AI and algorithmic systems in shaping the dissemination of information, and the potential for these technologies to both perpetuate and mitigate textual bias. Future research should continue examining the intersection of textual bias and AI and develop strategies to promote fairness, accountability, and transparency in information production and consumption.

Textual bias is a multifaceted and complex phenomenon that involves various types of deviations from objectivity or neutrality in the presentation, interpretation, or selection of information. As textual bias affects multiple aspects of society and public life, it is essential to acknowledge its existence and work towards mitigating its effects.

1.2 Textual Bias in Natural Language Processing

In the rapidly evolving field of NLP, the analysis of textual bias has emerged as an important area of inquiry. By examining how human biases manifest in textual data, researchers can develop a deeper understanding of the pervasive influence these biases exert on our interpretation of language. This section explores the intricacies of textual bias in NLP, beginning with a comprehensive framing of the study. By defining and positioning ourselves in relation to the complex notion of bias, we build upon previous state-of-the-art work to establish a solid foundation for our analysis.

A particularly concerning aspect of textual bias lies in the intersection of political and media landscapes. The influence of political and media biases on public opinion has been widely acknowledged (McCombs and Shaw, 1972), and we will examine the role NLP can play in detecting and understanding these biases. By conducting a focused analysis of political and media biases through the lens of NLP, we aim to contribute to the growing body of knowledge in this domain.

In this section, we synthesize the state-of-the-art findings in NLP research on textual bias, providing a comprehensive overview of the methodologies, techniques, and insights for a deeper understanding of the subject. Through this exploration, we aim to shed light on the multiple aspects of textual bias analysis in NLP.

1.2.1 Framing the Study

The analysis of textual bias in NLP has garnered significant attention in recent years, as there is a growing awareness about the importance of understanding and mitigating biases in human texts. In the context of NLP, biases can be broadly categorized into two types: model-based biases and textual biases. Model-based biases refer to biases that arise due to the underlying architecture or training processes of NLP models, such as the presence of gender bias in pre-trained word embeddings (Bolukbasi et al., 2016). Whereas textual biases are those that stem from the data itself, often reflecting societal and cultural biases present in the texts (Caliskan et al., 2017; Bender et al., 2021), as presented in Section 1.1. While model-based biases have been extensively studied and addressed in various ways (Bolukbasi et al., 2016; Zhao et al., 2018), the focus of this thesis is on the analysis and characterization of textual biases in human texts (Recasens et al., 2013). By investigating textual biases, we aim to provide insights that can encourage fairness, accountability, transparency, and ethics in the production or consumption of written content, as well as inform the development of more robust NLP techniques and tools that are less likely to perpetuate or exacerbate these biases.

The study of textual bias in NLP has been approached from various perspectives, and several tasks have been proposed to tackle different aspects of it, such as sentiment analysis, stance detection, or hate speech detection, among others. Sentiment analysis, or opinion mining, involves determining the sentiment expressed in a piece of text, such as positive, negative, or neutral (Pang and Lee, 2008). Recent work in this area has acknowledged the role of textual bias in shaping sentiment analysis results. For instance, Kiritchenko and Mohammad (2018) showed that gender and racial biases can influence sentiment classification, demonstrating the need for considering these biases during model development and evaluation. Stance detection is another relevant task that seeks to determine the author's position or viewpoint on a given topic within a text (Mohammad et al., 2016). This task is particularly sensitive to textual biases, as it directly deals with the expression of subjective opinions. Recent studies have explored methods for predicting these biases and mitigating their effects, such as adversarial training (Allaway et al., 2021) and conditional encoding (Augenstein et al., 2016). Hate speech detection is a subfield of NLP that deals with the identification and classification of offensive or hateful language in text (Fortuna and Nunes, 2018), and is closely related to sexism detection

(Stanczak and Augenstein, 2021). Textual biases are of great concern in this area, as they can manifest in the form of stereotypes and prejudiced language. Recent work has focused on developing more robust and generalizable hate speech detection models by leveraging transfer learning (Mozafari et al., 2019) and exploring methods to reduce bias in training data, such as data augmentation and re-sampling techniques (Park et al., 2018). Regarding knowledge dissemination, Recasens et al. (2013) examined framing and epistemological bias from Wikipedia articles using logistic regression and a list of handmade features. They identified common linguistic cues for detecting these biases, including subjectivity, presuppositions and entailments. Braud and Søgaard (2017) were interested in writing style for the detection of scientific fraud using logistic regression and a set of features including: word features, syntactic features and discourse features. From their experiments, they identified a number of fraud markers such as the absence of comparison, as well as the presence of different types of hedging and ways of presenting logical reasoning.

At the intersection of all these tasks and applications, fairness in NLP has emerged as a crucial topic and gained popularity in recent years. It refers to the development of models and techniques that ensure equitable treatment of diverse groups and perspectives in the processing and analysis of natural language data (Hovy and Spruit, 2016; Chang et al., 2019). As NLP applications increasingly permeate various aspects of daily life, such as social media, healthcare, and education, it is imperative to ensure that these systems do not perpetuate or exacerbate existing societal biases. The study of fairness in NLP is closely related to the analysis of textual biases, as addressing and understanding such biases is a prerequisite for developing fair and inclusive NLP systems. Recent research in this area has focused on identifying and quantifying biases in NLP models, developing methods to mitigate biases in these systems, and creating benchmarks to assess the fairness of NLP applications (Sun et al., 2019; Basta et al., 2019; Mitchell et al., 2019).

Despite the significant progress made in the analysis and understanding of textual biases in NLP, several challenges remain to be addressed. A notable limitation of most existing approaches is their exclusive reliance on lexico-syntactic information for analyzing textual biases. While these features are undoubtedly important, they may not fully capture the subtler aspects of bias that can manifest in the structural aspects of texts, such as rhetoric, argumentation, and discourse structure (Kiesel et al., 2015). Consequently, there is a growing need for research that examines the role of these structural aspects in contributing to textual biases. In parallel, the recent emergence of explainability techniques for understanding the decisions made by machine learning models has opened up new avenues for studying textual biases in NLP (Ribeiro et al., 2016). By providing human-interpretable explanations of model predictions, these techniques can

help to identify potential sources of bias and to guide efforts to address them (Guidotti et al., 2018). In addition, explainable AI methods can facilitate a deeper understanding of the complex interplay between lexical, syntactic, and structural features that contribute to textual biases, which may, in turn, lead to the development of more effective debiasing techniques (Jain and Wallace, 2019). However, the effective application of explainability techniques in the context of textual bias analysis is not without its challenges, as it requires careful consideration of factors such as the trade-offs between model complexity and interpretability. Addressing these challenges will be critical for deepening our understanding of textual biases in NLP, and it is within this framework that the work of this thesis is situated.

1.2.2 Political and Media Bias: A Focused Analysis

In this section, we look at the relationship between political and media bias by examining their manifestations in various textual sources through the lens of NLP. The motivation behind such a focus stems from the growing societal and academic interest in understanding and mitigating the impact of these biases on public opinion, democratic processes, and decision-making. In this context, the application of NLP techniques in the study of political and media bias offers valuable insights into how these biases manifest and propagate, thus contributing to the development of more transparent and unbiased information dissemination.

To that end, this section is organized into three distinct yet interconnected subsections. First, we focus on political bias analysis in NLP, exploring the methodologies employed in detecting and quantifying bias in political texts and discourses. Second, we turn our attention to media bias analysis, discussing the role NLP plays in identifying how various media outlets may exhibit and propagate biases through language use, framing, and other linguistic mechanisms. Finally, the third subsection bridges the gap between the previous two, investigating the interplay between political and media biases, which allows for a deeper comprehension of how these biases manifest and interact within the media landscape. By examining these three dimensions, this section aims to provide a comprehensive understanding of political and media bias in the context of NLP, highlighting current work and challenges in moving towards a more balanced and transparent information society.

Political Bias

Political bias refers to the slant or favoritism that individuals or groups exhibit towards specific political ideologies, parties, or policies. Analyzing political bias is essential for understanding how they may influence language and communication, as well as the impli-

cations they hold for policymaking, public opinion, social behavior, and democracy. It may contribute to uncover hidden agendas, promote transparency and encourage critical thinking among citizens. Several studies have demonstrated the impact of political biases on individual's perceptions and decision-making processes (Druckman, 2001; Taber and Lodge, 2006). A large field of research has emerged in NLP, investigating the presence and impact of political bias in different types of textual content, such as congressional speeches, legislative debates, or social media posts (Gentzkow et al., 2019; Grimmer and Stewart, 2013; Johnson and Goldwasser, 2016). Table 1.1 presents the main existing datasets in this field. We leave aside here the work on the analysis of political bias in the media, which will be discussed in a later section.

Name	**Lang.**	**#Data**	**Source**	**Annotation**
Convote (Thomas et al., 2006)	English	3,857	Congressional Speeches	Liberal, Conservative
Political Blogs (Yano et al., 2009)	English	8,818	American politics blogs	Liberal, Conservative
Twitter Ideology (Preoţiuc-Pietro et al., 2017)	English	4.8M	Tweets	7-point scale (Left to Right)
Politifact (Rashkin et al., 2017)	English	10,483	Political statements	6-point scale (True to False)
LIAR (Wang, 2017)	English	12,836	Political statements	6-point scale (True to False)
YouTube Political Discussion (Wu and Resnick, 2021)	English	134M	YouTube user comments (US political channels)	Left, Right
US Politics (Pujari and Goldwasser, 2021)	English	188K	Wikipedia, Tweets, Press statements, News articles	Republican, Democrat, Other
Reddit Political (Alkiek et al., 2022)	English	527K	Reddit comments from liberal and conservative groups	Liberal, Conservative
PoliTweet (Kawintiranon and Singh, 2022)	English	10,000	Political tweets (US 2020 election)	None

Table 1.1: Summary of existing popular datasets for the analysis of political bias in texts. Newspaper article datasets are excluded here and are referenced in Table 1.3.

Early work on the analysis of political bias in NLP focused on analyzing ideologically slanted language in political texts, such as speeches and manifestos. Laver et al. (2003) introduced an innovative technique, called *Wordscores*, for extracting policy positions from legislative speeches and party manifestos using keyword analysis. This approach was later refined by Slapin and Proksch (2008), who developed the *WORDFISH* algorithm, which

is based on word frequencies and allowed for the scaling of party manifestos by comparing them to reference texts with known positions. Another notable work is Monroe et al.'s (2017) development of a method called *Fightin' Words*, based on Bayesian shrinkage and regularization, which identifies politically charged words in partisan speeches in the U.S. Senate. More recently, supervised machine learning methods, such as naive Bayes classifiers (NB) and support vector machines (SVM), have been extensively applied to detect political bias in textual data. Thomas et al. (2006) used SVM to determine from the transcripts of U.S. Congressional floor debates whether a speaker supports a proposal or not, by taking into account the relationships between speech segments. Similarly, Yu et al. (2008) applied SVM and NB to classify U.S. Congressional speeches based on the political party of the speaker. Their approach utilized various feature representations, including bag-of-words and tf–idf (term frequency-inverse document frequency). Conover et al. (2011) proposed to predict the political alignment of Twitter users from the texts and hashtags of their tweets using a SVM on the tf-idf. Peterson and Spirling (2018) used logistic regression to predict the party of the members of parliament from records of British parliamentary debates. Sapiro-Gheiler (2019) compared NB and SVM, as well as decision trees (DT) and lasso-penalty regression (LR) using a bag-of-words to predict senators' party from U.S. Congressional records, and found SVM and LR to be the most efficient. Sim et al. (2013) focused on measuring political candidate's ideological positioning from their speeches. They deduced ideological cues from a corpus of political writings and used them to infer the proportions of ideologies each candidate uses in election campaigns by applying a domain-informed Bayesian Hidden Markov Model (HMM). The analysis of political bias has also been addressed from the perspective of topic modeling. Lin et al. (2008) proposed a novel probabilistic model that simultaneously learns topics and perspectives from a corpus of political speeches. Later, Ahmed and Xing (2010) introduced multi-view topic models, based on Latent Dirichlet Allocation (LDA) to identify ideological perspectives on a topical level from political blog data. Finally, Nguyen et al. (2013) proposed supervised hierarchical LDA which captures both the topic structure and the perspectives on those topics, and conducted experiments using the famous U.S. Congressional floor debates transcripts dataset.

With the emergence of deep learning techniques, researchers have increasingly turned to neural network-based models for political bias analysis. The development of word embeddings, such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), has further facilitated the identification of semantic relationships and biases in text. Iyyer et al. (2014) were among the first to employ Recursive Neural Networks (RvNN) and pre-trained word embeddings to classify sentences from U.S. Congressional floor debate transcripts according to their political ideology (Liberal, Conservative or Neutral). They

demonstrated that RvNN outperformed the baseline methods, which included logistic regression models. Preoţiuc-Pietro et al. (2017) examines users' political ideology on Twitter using a seven-point scale and a broad range of language features including unigrams, word clusters (word2vec) and emotions. Also based on Twitter data, Demszky et al. (2019) proposed the clustering of tweet embeddings from 4.4M tweets on 21 mass shootings to uncover the topical and framing dimensions of political polarization. Long short-term memory networks (Hochreiter and Schmidhuber, 1997, LSTM), have also been employed in the study of political bias. Rashkin et al. (2017) used LSTM to predict the truthfulness of individual statements made by public figures. Wu and Resnick (2021) were interested in the political analysis of cross-partisan discussions posted from political videos on Youtube. They trained a hierarchical attention network (HAN) from the comments to predict the user's political leaning (left or right).

The introduction of transformer-based models and pre-trained language models (PLMs), such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019b), has made it possible to further improve and deepen the analysis of political bias in texts. Davoodi et al. (2020) were interested in predicting the passage or failure of a bill using data on the text of all bills introduced in Indiana, Oregon, and Wisconsin between 2011 and 2018. They relied on BERT and Relational Graph Convolutional Network (RGCN) to model the interactions between the text of a bill and the legislative context in which it is presented. Guo et al. (2020) examined the impact of political ideology biases on social topic detection from Twitter data. They trained BERT and LSTM models to predict whether a tweet is about gun control or immigration and distinguish between two datasets, one containing right-leaning tweets and the other left-leaning tweets. They showed that BERT is more likely to propagate the bias seen during training. Pujari and Goldwasser (2021) collected political texts about 455 members of the U.S. Congress from press statements, Wikipedia articles and tweets. They proposed a Compositional Reader model using BERT and LSTM to generate representations for political figures and evaluate their model on politician's grade prediction. The *National Rifle Association* (NRA) assigns letter grades to politicians based on their gun-related voting. They showed that the representations they learn effectively capture nuanced political information. Alkiek et al. (2022) were interested in the political users of the social network Reddit. From 574K political users on Reddit they trained a RoBERTa model over their comments to infer political affiliation, and showed that there are heterogeneous types of political users. Some work has also focused on mitigating these biases in language models, Liu et al. (2021a) suggested metrics for measuring political bias in generative language models and propose a reinforcement learning (RL) framework for mitigating them.

Given the growing interest in the study of political bias in NLP and the widespread use of PLMs, several works have proposed the training of large-scale specialized language models for the analysis of political bias. Kawintiranon and Singh (2022) introduced PoliBERTweet, an English pre-trained language model for analyzing political content on Twitter. They collected over 83M unique politics-related tweets during the U.S. 2020 presidential election period and fine-tuned BERTweet, a pre-trained RoBERTa model fine-tuned on Twitter data. They evaluated PoliBERTweet on several NLP political tasks and showed its dominance over general-purpose language models in domain-specific contexts.

Although much progress has been made in recent years, the analysis of political bias remains a complex task by its very nature. The rating of the bias, which is essential for its analysis and for supervised approaches, can be categorized along numerous ideological dimensions such as left-wing versus right-wing, liberal versus conservative, or even more nuanced ideological distinctions, including the continuous characterization of the political spectrum (see Table 1.1). While most previous studies have favored discrete labels, a few have tried to infer continuous values (Preoţiuc-Pietro et al., 2017), but there remains a serious lack of consistency in the annotation schemes for political bias that hinders the comparison of results. This complexity is further compounded by the fact that political bias is not a static construct; it is influenced by cultural backgrounds and evolves over time, often differing significantly between regions or countries. For example, a conservative political stance in Sweden, where there is broad support for social welfare, may differ significantly from a conservative stance in the United States. As a result, the process of evaluating and annotating political bias in texts is often ambiguous and prone to subjectivity, making it a challenging task for both human annotators and NLP models.

Subsequently, several limitations to existing work on this task can be identified. The vast majority of existing research on political bias analysis in NLP has primarily focused on English-language data and English politics, resulting in a lack of multilingual work on the subject (see Table 1.1). Only a few works have proposed the analysis of corpora of other languages, as for example Lehmann and Derczynski (2019), who were interested in detecting political stances using quotes from Danish politicians. This limitation not only hinders the generalizability of the results, but also the ability to better understand the role that linguistic and cultural nuances play in the formation of political biases across languages and countries. In addition, there is a crucial need for explanation and interpretability of model decisions. Most state-of-the-art models are black-box systems that do not provide straightforward explanations for their predictions. Therefore, it is hard to understand the underlying factors that influence these decisions and, subsequently, to

draw reliable and generalizable conclusions about the nature and extent of political biases. While certain linguistic patterns and word usage may be indicative of political bias, the opacity of the model makes it challenging to discern the precise linguistic features driving these biases. Addressing the issue of explainability is therefore essential for understanding political biases and assessing the reliability and validity of methods. Finally, working on political bias analysis raises ethical considerations, such as the potential misuse of NLP models to manipulate public opinion or amplify existing biases (Hovy and Spruit, 2016). Researchers must remain vigilant in addressing these concerns by developing guidelines and best practices to ensure that their work does not inadvertently contribute to the proliferation of misinformation or the spread of harmful ideologies.

Media Bias

The proliferation of digital media and the widespread availability of information have transformed the way people consume news and engage with the world around them. With a plethora of news sources available, it is becoming increasingly difficult for individuals to separate fact from fiction, particularly in the context of news reporting. The rise of fake news and alternative facts has led to a growing concern about the impact of biased information on public opinion, political discourse, and democracy (Scheufele and Tewksbury, 2007). Textual bias, defined as the distortion of information by a writer to serve a particular agenda, whether intentional or not, is a pervasive problem in contemporary media. A recent study by Allcott and Gentzkow (2017) found that false news stories spread much more quickly and widely than true stories, and that the spread of false news was largely driven by people's preference for information that confirmed their pre-existing beliefs. This suggests that bias is not only a problem in the way news is reported, but also in the way it is consumed and shared by the public.

One prominent initiative that aims to provide a comprehensive resource for assessing media bias is Media Bias/Fact Check (MBFC). The MBFC[2] website is an independent online platform that rates and categorizes news sources based on their perceived bias, from left-leaning to right-leaning, as well as their factual reporting, from high to low, conspiracy, pseudoscience and satire. MBFC employs a team of reviewers who apply a rigorous methodology to review and evaluate news outlets, taking into account factors such as the use of loaded language, factual reporting, story selection, and political affiliation, among others. The platform also provides details about the ownership, funding, and history of each news source, offering users a more comprehensive understanding of the context in which these media operate. In the context of media bias analysis in NLP, MBFC serves as a valuable resource for researchers, providing ground truth labels and a means to

[2] https://mediabiasfactcheck.com/

evaluate the performance of models in detecting and classifying media bias.

The problem of biased information is further compounded by the use of algorithms and machine learning models in automated news aggregation and recommendation systems. While these systems are designed to deliver personalized news to users based on their preferences, they can also amplify and reinforce biases in the news that users receive (Pariser, 2011; O'Neil, 2016; Agan et al., 2023). As a result, there is a growing need for methods and tools to detect, characterize, and mitigate bias in news articles (Groseclose and Milyo, 2005; Hamborg et al., 2019). The current state-of-the-art in this field can be categorized into several key research areas, and Table 1.2 presents the main existing datasets for each of them. As in the previous section, we leave aside here the work on political bias in the media, which will be discussed in the next section.

> [WHERE THE JOBS ARE]$_{\textbf{Economic}}$ [Critics of illegal immigration can make many cogent arguments to support the position that the U.S. Congress and the Colorado legislature must develop effective and well-enforced immigration policies that will restrict the number of people who migrate here legally and illegally.]$_{\textbf{Policy Prescription}}$ [It's true that all forms of [immigration exert influence over our economic and cultural make-up.]$_{\textbf{Cultural Identity}}$ In some ways, immigration improves our economy by adding laborers, taxpayers and consumers, and in other ways immigration detracts from our economy by increasing the number of students, health care recipients and other beneficiaries of public services.]$_{\textbf{Economic}}$ [Some economists say that immigrants, legal and illegal, produce a net economic gain, while others say that they create a net loss]$_{\textbf{Economic}}$

Figure 1.1: Example of an annotated article (a 2006 editorial in the *Denver Post*) in terms of media framing from the MFC corpus, taken from Figure 2 in Card et al. (2015).

The study of framing bias in news was one of the first subjects of media bias analysis in NLP. Framing refers to the way information is presented in the media, which can affect the audience's perception of the issue (Entman, 1993). In NLP, several studies have focused on detecting and analyzing framing in media texts (Ali and Hassan, 2022). For instance, Card et al. (2015) proposed the Media Frame Corpus (MFC), containing several thousand news articles annotated in terms of media framing using 15 framing dimensions, such as *Economic*, *Health* or *Cultural identity* (Boydstun et al., 2014). An example of an annotated article from the MFC corpus is shown in Figure 1.1. Afterward, Card et al. (2016) proposed to use logistic regression classifiers to predict the framing of an article from the MFC dataset. Ji and Smith (2017) improved on these results by introducing an approach based on a recursive neural network and a new attention mechanism that computes a discourse-aware representation of the text. Similarly, Field et al. (2018) compared their lexicon approach, based on pointwise mutual information, and

obtained equivalent results. Media framing bias has also been studied from the perspective of gun violence. Liu et al. (2019a) introduced a new dataset consisting of headlines from U.S. news articles about gun violence and annotated them in terms of their framing on the subject (politics, public opinion, society/culture, economic consequences, guns rights, gun control/regulation, mental health, school/public space safety and race/ethnicity). They proposed several models to predict the frames of headlines, including: BERT, Bi-LSTM and Bi-GRU (Gated Recurrent Unit). BERT has proven to be the most efficient for this task. Lastly, Lee et al. (2022) proposed to mitigate framing bias in news articles through neutral multi-news summarization. They generate a framing-bias-free summary from news articles with varying degrees of bias using deep neural models for multi-document summarization.

Another important task in this domain is fake news detection, or fact-checking. Fake news refers to false or misleading information that is deliberately disseminated (or not) by the media and which may have a far-reaching impact on public opinion (Wardle and Derakhshan, 2017). Several worldwide projects have been launched recently to fight mis- and dis-information online, such as FirstDraftNews[3] or FullFact,[4] due to the potential societal consequences. An illustrative example is the widespread misinformation regarding the recent COVID-19 pandemic, which led to confusion and mistrust in public health measures, ultimately resulting in serious consequences for the spread of the virus (Cinelli et al., 2020; Zhou et al., 2020a). Detecting fake news in textual data has thus become a major field of research in NLP (Zhou and Zafarani, 2020; Nakov and Da San Martino, 2020). Many techniques have been explored to tackle this issue, including machine learning algorithms, deep learning models, and linguistic feature analysis. Early approaches to fake news detection focused on the use of handcrafted linguistic features, such as writing style and sentiment analysis, to identify deceptive content (Pérez-Rosas et al., 2017; Horne and Adali, 2017). Baly et al. (2018a) studied the factuality of reporting and bias of news media and proposed a rich set of features from which they trained an SVM classifier. More recent approaches have leveraged deep learning techniques, such as CNN, RNN, and transformer-based models, to capture more complex and context-dependent patterns in the textual data (Rashkin et al., 2017; Zhou et al., 2020b; Wang, 2017; Oshikawa et al., 2020). With an alternative approach, Karimi and Tang (2019) proposed "Hierarchical Discourse-level Structure for Fake news detection" (HDSF), a model that learns and constructs a discourse-level structure for fake news detection. They showed that real/fake news present substantial differences in their structures and that it improves performance over standard approaches. Other studies have also investigated

[3] https://firstdraftnews.org/
[4] https://fullfact.org/

discourse structures in newspaper articles (Choubey et al., 2020), further supporting the importance of considering this aspect of textual bias. Shu et al. (2019) and Zellers et al. (2019) worked towards mitigating these biases with the introduction of methods to explain model decisions and identify potential threats using fake news generation models.

Name	Lang.	#Data	Source	Annotation
BiasedSents (Lim et al., 2020)	English	46	U.S. news articles	4-point scale (not biased to very biased)
Framing				
Media Frames Corpus (Card et al., 2015)	English	20,037	U.S. newspapers between 1990 and 2012	15 frames (Boydstun et al., 2014)
Gun Violence Frame Corpus (Liu et al., 2019a)	English	1,300	News headlines from 21 U.S. news media	9 frames on gun violence
Fake News Detection				
Fake News Net (Shu et al., 2020)	English	23,921	Newspapers articles	Fake or Real
ReCOVery (Zhou et al., 2020a)	English	2,029	Newspapers articles on COVID-19	Reliable, Unreliable
FACTOID (Sakketou et al., 2022)	English	3.4M	Reddit posts on political discussions (2020)	Real, Fake, Unlabeled
NELA-GT-2022 (Gruppi et al., 2023)	English	1.8M	News articles from 361 media	6-point scale veracity labels
Stance Detection				
Emergent (Ferreira and Vlachos, 2016)	English	2,595	U.S. Newspapers articles	For, Against, Observing
Arabic News Stance (Baly et al., 2018b)	Arabic	3,042	Arabic News articles	Agree, Disagree, Discuss, Unrelated

Table 1.2: Summary of existing popular datasets for the analysis of media bias. Datasets on political bias in news are excluded here and are referenced in Table 1.3.

Stance detection is the last major task that has been the subject of much work related to textual bias analysis in media. It is the task of determining the attitude or position of a text towards a target (Mohammad et al., 2016). In the context of media bias, stance detection can be used to analyze the alignment of news articles with particular viewpoints or ideologies. Ferreira and Vlachos (2016) proposed a novel dataset for stance detection of news articles and used a logistic regression classifier on a bag-of-words to predict the

stance. Baly et al. (2018b) compared neural network-based approaches on Arabic news stance detection.

In addition to the tasks mentioned above, several other NLP tasks have been employed to study media bias. These include sentiment analysis (Pang and Lee, 2008; Carvalho et al., 2017), which can be used to assess the emotional tone of news articles, and argument mining (Lippi and Torroni, 2016; Vecchi et al., 2021), which aims to extract and analyze the structure and content of arguments in order to reveal potential biases in the way issues are presented or argued.

One of the major challenges of these tasks is the processing of long documents, as most NLP models, including transformers, have limitations in handling long input sequences (Beltagy et al., 2020). This constraint limits the analysis of media bias, as news articles are often long documents, where the context is essential for understanding the underlying bias. Most existing approaches to media bias analysis focus on shorter texts, such as headlines, abstracts, or the beginning of the article. However, bias can manifest in various ways throughout an article, and analyzing the complete content is crucial for a comprehensive understanding of media bias. Researchers have proposed various methods to address this issue, such as hierarchical attention networks (Yang et al., 2016) and sliding window approaches (Liu et al., 2022). Recent work on techniques like Longformer (Beltagy et al., 2020) has started to address this issue by enabling the processing of longer texts using transformers, but their applicability and effectiveness in the context of media bias analysis still need to be explored and evaluated.

Another limitation is the lack of comprehensive and diverse datasets that cover various media sources, languages, and cultural contexts. Most existing studies rely on datasets from a limited number of English-language news sources, which may not adequately represent the diversity of media outlets and their respective biases and generalize to other languages or cultural contexts (Hamborg et al., 2019). To address this issue, future research should prioritize the development of multilingual and multicultural datasets for media bias analysis.

Media bias is a dynamic process, that can evolve over time and across different topics (Groseclose and Milyo, 2005). Current NLP models often struggle to capture these temporal variations and topic-dependent biases, limiting their applicability to real-world scenarios. More attention needs to be given to methods that can adapt to changing biases and incorporate temporal information into the analysis. Moreover, the potential influence of confounding factors on the performance of models, such as topic similarity or article length, is an important challenge in media bias analysis (Baly et al., 2020a). Accounting for these

factors is crucial for improving the validity and reliability of media bias detection models.

Political Bias in the Media

As we have explored political bias and media bias in NLP separately in the previous parts, we now shift our focus towards understanding the interplay between these two notions. This section delves into the study of political bias in news articles from an NLP perspective, examining the methods that have been proposed to identify, analyze and quantify such bias. Understanding political bias in news articles has gained substantial interest recently due to the increasing polarization of media, the rise of social media platforms, and the potential impact of biased reporting on public opinion and democracy (Allcott and Gentzkow, 2017). Political bias in media refers to the slanting or distortion of news reporting, influenced by the political ideology of a journalist, editor, or publisher. This bias can manifest in various forms such as the choice of words, framing of issues, argumentative processes, or the selection of stories covered. Given the significant role of media in shaping public opinion, it is crucial to investigate political bias in news articles to ensure a balanced and transparent sharing of information.

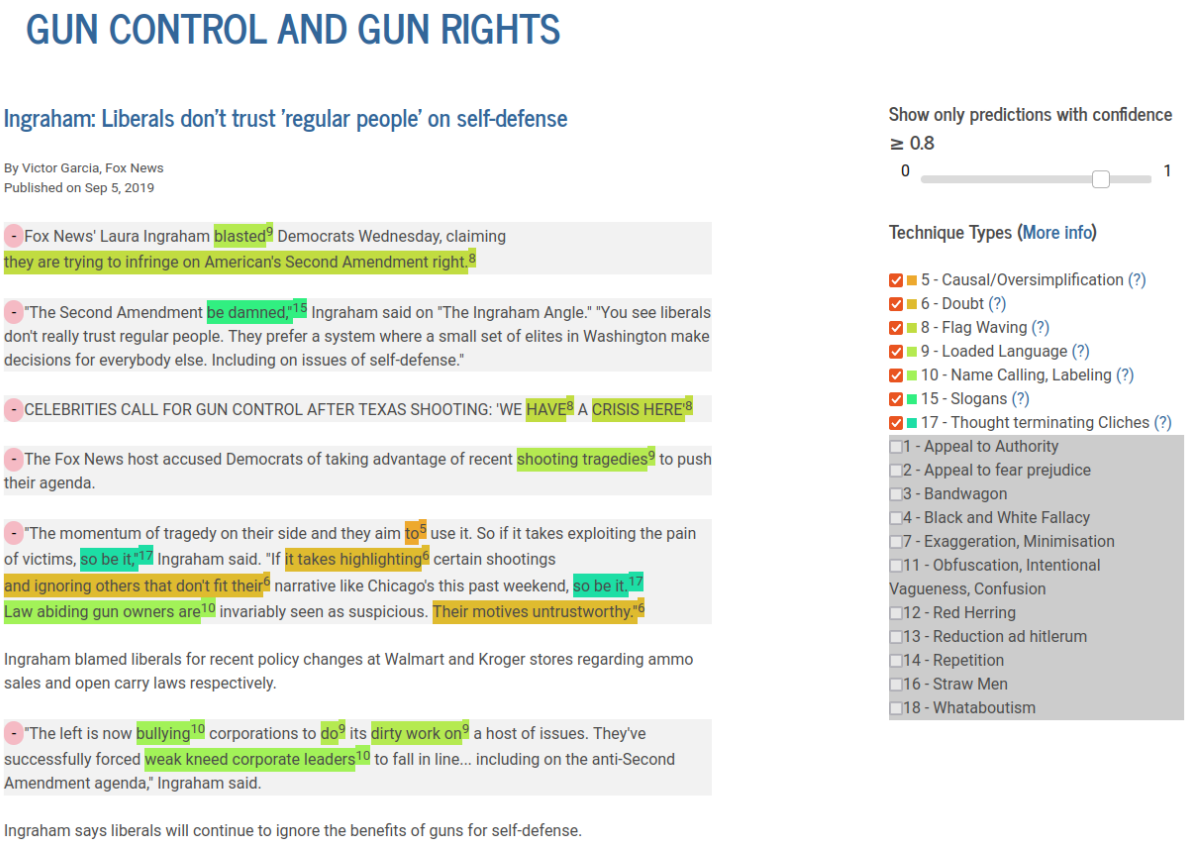


Figure 1.2: Example of an article annotated in terms of persuasion techniques from the Propaganda Persuasion Techniques Analyzer (Prta) tool (Da San Martino et al., 2020b).

As a result, several platforms specialized in the analysis and rating of political bias in newspaper articles have emerged and are receiving increasing attention. These platforms employ different methodologies and rating systems to evaluate the political leanings of news media and their content:

- AllSides[5] uses a multidimensional approach that combines human input, community feedback, and algorithmic analysis to rate news sources and individual articles on a five-point scale: left, left-center, center, right-center, and right. The AllSides team first assigns an initial rating based on a comprehensive review of the source's content and editorial stance, which is then subject to revision based on user feedback and third-party reviews.[6]
- Media Bias / Fact Check[7] follows a similar approach, relying on human reviewers and measures to analyze news sources and articles. Their ratings span across seven categories: least biased, left-center, left, right-center, right, extreme right, and extreme left. The reviewers assess bias through various factors, such as the use of loaded words, the selection of stories, and the political affiliation of the source. In addition, they provide scores for factuality, and credibility of media sources.
- ad fontes media,[8] on the other hand, developed the Media Bias Chart (Figure 1.3), which visually represents news sources on a two-dimensional plane, with the x-axis indicating political bias (left to right) and the y-axis representing reliability and quality (high to low). AdFontes Media uses a team of analysts with balanced right, left, and center self-reported political viewpoints to rate articles.
- GROUND News[9] takes a distinct approach by comparing how a particular news story is covered across various sources with different political biases. By analyzing the wording, story choices and political affiliations in each article, Ground.news provides users with a comprehensive view of the different perspectives and biases present in the coverage of a story.

The emergence of these platforms and the growing interest in political bias analysis have led to the development of numerous NLP projects focused on this subject (Nakov and Da San Martino, 2020). One such prominent project is the Propaganda project[10]

[5] https://www.allsides.com
[6] https://www.allsides.com/media-bias/media-bias-rating-methods
[7] https://mediabiasfactcheck.com
[8] https://adfontesmedia.com
[9] https://ground.news
[10] https://propaganda.qcri.org/

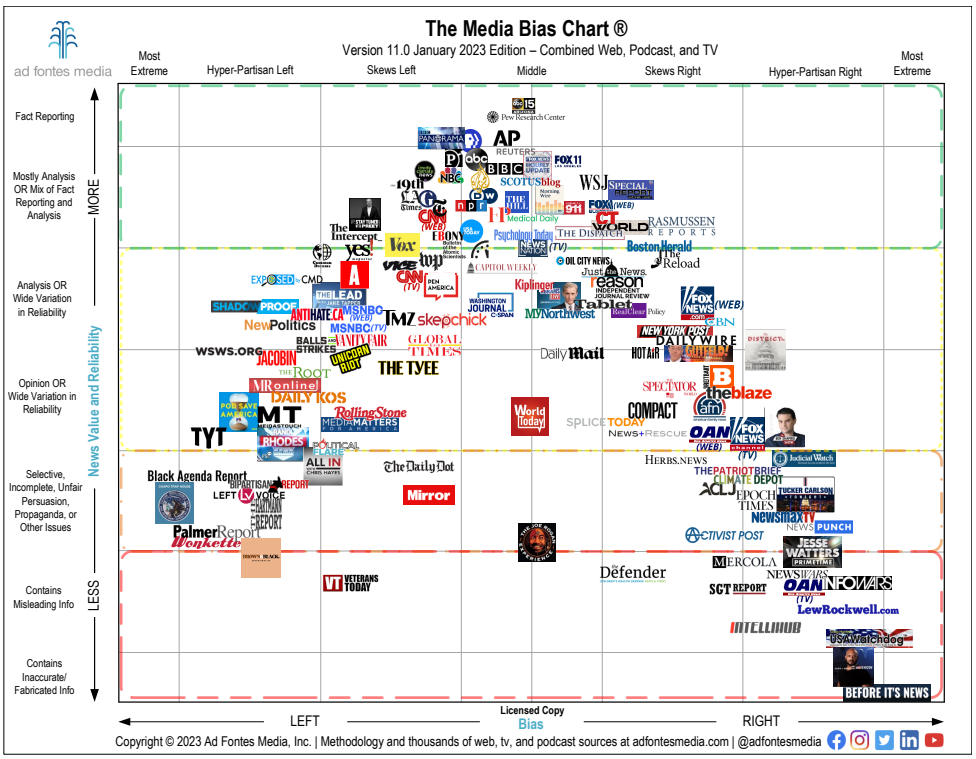


Figure 1.3: Media Bias Chart[12] (Ad Fontes Media): A visual representation of media bias in the United States plotted along two dimensions - horizontal axis for political bias (left, center, right) and vertical axis for reliability.

(Barrón-Cedeño et al., 2019), which aims to develop NLP techniques for detecting propaganda and manipulation in news articles. The project encompasses various tasks, such as identifying specific propaganda or persuasion techniques, determining the overall level of propagandistic content, and analyzing the impact of propaganda on the perceived reliability of news sources (Da San Martino et al., 2019; Yu et al., 2021). In particular, they proposed a tool for analyzing persuasion techniques, for which an example of an analyzed article is given in Figure 1.2. This work is part of a larger project, the TANBIH project,[11] which focuses on the identification and aggregation of various aspects of media bias, including framing, slant, and propaganda, to create a more comprehensive understanding of bias in news. By leveraging a multi-task learning approach, TANBIH's framework effectively detects different types of biases, as well as fake news and clickbait, providing a valuable resource for researchers and news consumers.

The significance of analyzing political bias in news articles is further reflected in the shared tasks organized within the NLP community. *SemEval-2019 Task 4: Hyperpartisan News Detection* (Kiesel et al., 2019) aimed to identify hyperpartisan news articles, which exhibit extreme political biases, using a dataset of 754K articles collected from the web.

[11] https://tanbih.qcri.org/

The best-performing systems in this task combined sentence-level embeddings with a convolutional neural network to achieve high accuracy in detecting hyperpartisan articles (Jiang et al., 2019). *SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles* (Da San Martino et al., 2020a) focused on detecting specific propaganda techniques used in news articles, such as loaded language, name-calling, and appeals to authority. The participating systems used a range of approaches, including machine learning and rule-based methods, with transformer-based models achieving the best performance (Morio et al., 2020).

The focus on political bias in news articles has led to various studies and methods in NLP and Table 1.3 references the main existing datasets. One of the early work in this field was conducted by Gentzkow and Shapiro (2010), who developed a method to quantify media slant by analyzing the frequency of politically charged sentences in news articles. More recently, Chen et al. (2018) were interested in the headlines of news articles, as these often carry strong political biases. Given an article headline with a particular political leaning (left or right), they trained a bias flipper model based on autoencoders to generate a headline on the same topic but of the opposite political leaning. Kulkarni et al. (2018) proposed an attention based multi-view model for political ideology detection of news articles. From a set of multi-modal features consisting of the title, the link structure (to other news media and sources), and the article content, they trained a neural network using stochastic attention and demonstrated the effectiveness of their approach. Another noteworthy approach to detect political orientation and hyperpartisanship in news articles has suggested the study and comparison of stylometry between several political orientations (Potthast et al., 2018). Using multiple modeling approaches on a set of stylometric features, including readability scores, dictionary features, word frequency and paragraph length, they showed that there is a distinction between the writing style of hyperpartisan and non-hyperpartisan articles. Fan et al. (2019) studied political bias in media by looking at informational bias, or factual reporting. Rather than focusing only on lexical biases, their claim is that political biases in the news can also be characterized by the decisions made about content selection and organization within articles. Based on this assumption, they created a new dataset, BASIL, annotated at the level of informational bias spans, and carried out a first experiment for informational bias prediction using BERT.

Another line of research has focused on leveraging social information and external knowledge to improve the detection of political bias in news articles. Li and Goldwasser (2019) took advantage of the social context of the articles to predict its political perspective. From an article and its social information, consisting of Twitter users who share links of the article and follow political users, they construct a socially-infused textual repre-

Name	Lang.	#Data	Source	Annotation
BuzzFeed-Webis (Potthast et al., 2018)	English	1,627	News articles from 9 U.S. media	Mainstream, Left, Right
Webis Bias Flipper (Chen et al., 2018)	English	6,447	News articles from 5 U.S. media	Left, Center, Right
Telugu (Gangula et al., 2019)	Telugu (India)	1,329	News articles from Telugu newspapers	5 political parties in India
Proppy (Barrón-Cedeño et al., 2019)	English	52,000	News articles from 100+ news outlets	Propagandist, Non-propagandist
BASIL (Fan et al., 2019)	English	300	News articles from Fox News, NY Times, Huffington Post	Lexical/Informational, Direct/Indirect, Positive/Negative
Hyperpartisan (Kiesel et al., 2019)	English	754K	News articles from hyperpartisan and mainstream websites	Hyperpartisan, Non-hyperpartisan
Allsides (Baly et al., 2020a)	English	34,737	News articles from 73 news media, covering 109 topics	Left, Center Right
NLPCSS-20 (Chen et al., 2020c)	English	6,964	News articles from U.S. news outlets	Political Bias, Unfairness, Non-objectivity
PTC-SemEval20 (Da San Martino et al., 2020a)	English	536	News articles from 49 U.S. media	14 propaganda techniques
NewB (Wei, 2020)	English	264K	Sentences from from 11 news media regarding Donald Trump	Liberal, Conservative
Politik (Aksenov et al., 2021)	German	47,362	News articles from 34 German news outlets	7-point scale (Left to Right)
PIP22 (Sinno et al., 2022)	English	175	News articles from 5 U.S. media	At the paragraph level: Liberal, Conservative, Neutral – Social, Economic, Foreign
BIGNEWS (Liu et al., 2022)	English	3.6M	News articles from 11 U.S. media	Left, Center, Right

Table 1.3: Summary of existing popular datasets for the analysis of political bias in media.

sentation using Graph Convolutional Networks (GCN) and demonstrate its effectiveness for political orientation detection. Still using information from Twitter users, Stefanov et al. (2020) proposed to use tweets and retweets from users about media outlets to identify the media's political leaning. Another approach based on social context analyzed

media readers on Facebook, Twitter, and YouTube, as well as what was written about the media on Wikipedia (Baly et al., 2020b). Using these features and the text of the article, they fine-tuned a pre-trained BERT model for the prediction of political bias and showed that the social media context, as complementary information, allows for significant improvements in the results. Baly et al. (2020a) released a large dataset of news articles for the prediction of political ideology, and propose for their classification model to incorporate, along with the content of the article, information describing the target media, including Twitter and Wikipedia descriptions of the media. Since the majority of articles from a same media share the same political annotation, the authors paid particular attention to diversifying the topics covered in the articles, selecting topics of interest, and ensuring that the model could not use lexical cues specific to the media, unrelated to the political bias, which would allow it to easily achieve high performance without looking at politically charged content. In particular, they propose a media-based split where the media present in the training set are excluded from the test set, in order to ensure that the model is actually modeling the political ideology and not the media from which the articles originate. Furthermore, they propose to remove media bias through triplet loss pre-training and adversarial adaptation, in order to encourage the model to focus on politically charged content, and compare two approaches for the classification model, BERT and LSTM, resulting in BERT performing better. Using external knowledge as well, Li and Goldwasser (2021) were interested in the entities mentioned in news articles, and tried to predict the political bias of the article with respect to those entities. The proposed framework consists in extracting the "person" and "organization" entities from the article in order to learn a representation of these entities and their relationship from external knowledge sources and text corpus, such as Wikipedia. A classification model is then trained on this representation and the text of the article to predict its political ideology. Similarly, Zhang et al. (2022) relied on knowledge graphs to model entity cues in news article and performed political perspective detection using relational graph neural networks.

Among the most recent work, Chen et al. (2020b) studied how secondary information about politically biased spans in an article, such as their frequency, positions, and sequential order, helps to improve the effectiveness of political bias detection. Using the probability distributions of these measures in a Gaussian mixture model, they showed that incorporating this information is beneficial for detecting political bias. Chen et al. (2020c) proposed not only to predict the political leaning of news articles, but also to explore how political bias and unfairness are manifested at different levels of granularity. One of their main findings is that the last quarter of the article seems to be the most biased part. Since the process of annotating corpora on political bias is a challenging task and annotations of existing corpora are not always very robust, Lazaridou et al. (2020) propose

a new corpus annotated by experts, and compare them to crowdsourced and automatic annotations. They concluded that the use of automatically generated annotations is not suitable for this task, and that expert knowledge can be used to boost the classification performance. Using the BASIL dataset, Lei et al. (2022) studied discourse structures for sentence-level political bias analysis in news articles. They extracted the discourse roles and relations for each sentence in the article (Choubey et al., 2020) and used them to inform the RoBERTa-based bias classification model. They showed that using discourse structure information yields to improvements on political bias prediction. Sinno et al. (2022) considered the multidimensional aspect of political ideology and polarization in media. In particular, they introduce a new dataset for which trained political scientists and linguists annotated the articles at paragraph level along three political levels (liberal, conservative, neutral) according to three political dimensions: social, economic and foreign (article about foreign issues). Finally, Liu et al. (2022) proposed POLITICS, a pre-trained language model fine-tuned from RoBERTa on news articles for political ideology prediction and stance detection. From BIGNEWS, a large-scale dataset of more than 3.6M political news articles, POLITICS was fined-tuned using a novel ideology-driven pretraining objective based on the comparison of articles on the same story. POLITICS surpassed both strong baseline methods and state-of-the-art techniques in various political ideology prediction and stance detection tasks, as demonstrated by their experiments.

While most existing work focuses on English-language datasets, which limits the analysis to a narrow view of the political spectrum, some work has focused on the study of political bias in newspaper articles from other countries and languages, where the political context might not be the same. For instance, Gangula et al. (2019) published a dataset of news articles from various newspapers in Telegu, a language spoken in the Indian state of Andhra Pradesh and Telangana. They proposed to predict the political bias using a headline attention network, based on the assumption that the headline of the article often reflects its political bias. Similarly, Agrawal et al. (2022) created and annotated a dataset of Hindi news articles on which they fine-tuned various pre-trained language models to predict the political orientation, with XLM-RoBERTa being most successful. Kameswari and Mamidi (2021) built from the Telegu dataset and proposed a new fine-grained annotation scheme for it, introducing 10 labels to capture various aspects of political bias in news. Turning to another language, Han et al. (2019) analyzed the political slants of user comments from Korean partisan news articles. They fine-tuned KorBERT, a pre-trained BERT model for the Korean language, to detect the political leaning of both comments and articles. For German, Aksenov et al. (2021) introduced a dataset of news articles labeled for political bias on a five-point scale, and placed particular attention on data cleaning and balancing. They experimented with different classification

methods, and showed that political bias classification is particularly challenging when using fine-grained labels. From another perspective, Padó et al. (2019) were interested in the construction of discourse networks for political debates based on the identification of political claims and actors in German news articles.

Based on current advances in the field, several challenges remain to be addressed in order to effectively analyze political bias in news articles. One such challenge is the cross-target or cross-topic performance; when trained on a specific target or topic, the performance tends to degrade considerably when applied to new topics, which is often the case due to the heterogeneity of the topics covered, and the different time windows considered between the numerous datasets. Additionally, it is crucial to address the potential for models to exploit lexical cues related to the media source as a shortcut for predicting orientation, since most articles from the same media outlet tend to share the same annotation. The construction of high-quality datasets with the necessary characteristics and properties to mitigate these issues remains an ongoing endeavor (Baly et al., 2020a). Another significant challenge is annotation consistency, as political bias can be defined in multiple ways and lacks a consensus on how it should be annotated in news articles, resulting in numerous datasets with varied annotations on the subject (see Table 1.3). This inconsistency complicates the comparison of different methods and approaches. Finally, the field must place greater emphasis on the explainability of methods employed to characterize political bias, moving beyond mere prediction and toward a deeper understanding of the underlying factors.

1.3 Political Bias in News Articles: A Case Study

In this case study, we will illustrate the notions introduced above by examining the presence of political bias in the article "*The Coronavirus Hoax*" published in *The New American* media on March 16, 2020, by Ron Paul, about the COVID-19 pandemic in the United States (`https://thenewamerican.com/the-coronavirus-hoax/`).

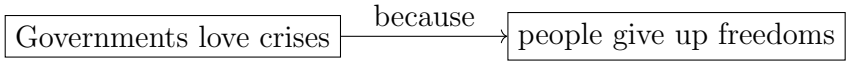
Before diving into the case study, it is important to provide a brief context of the author and the media outlet. Ron Paul, the author of this article, is an American author and retired politician who served as the U.S. Representative for Texas's 22nd congressional district from 1976 to 1985 and again from 1997 to 2013. He is affiliated with the Libertarian and Republican parties. *The New American* is an American media which presents news and opinion from a conservative, constitutionalist, and libertarian perspective. According to the evaluation made by the Media Bias/Fact Check platform, *The New American* is

classified as a "Right Biased" source. We will now analyze the text of the article paragraph by paragraph, identifying and discussing various types of biases found within, starting with the first paragraph.

Paragraph 1

Governments love crises because when the people are fearful they are more willing to give up freedoms for promises that the government will take care of them. After 9/11, for example, Americans accepted the near-total destruction of their civil liberties in the PATRIOT Act's hollow promises of security.

In this first paragraph, we can observe the presence of framing bias and language bias. The author frames the relationship between governments and crises negatively, implying that governments take advantage of crises to increase their power and control over citizens. This framing bias is supported by the use of emotionally charged words, such as "fearful", "give up freedoms", "near-total destruction of their civil liberties", and "hollow promises of security", which contribute to the language bias. By using these words, the author subtly influences the reader's perception of the government's actions and intentions. In the rhetorical structure of this paragraph, we can observe the following biased relation:



The author makes the assumption that governments loves crises because it leads to people giving up freedoms as they are fearful. Here, "because" is a discourse connective that establishes a causal relationship between the two propositions, with the second proposition providing the cause or reason for the first.

Paragraph 2

It is ironic to see the same Democrats who tried to impeach President Trump last month for abuse of power demanding that the Administration grab more power and authority in the name of fighting a virus that thus far has killed less than 100 Americans.

In the second paragraph, the author uses of the word "ironic" to point out the perceived inconsistency in the Democrats' behavior, it adds a negative connotation to the Democrats' actions. The phrasing "the same Democrats who tried to impeach President Trump last month" can be seen as an example of framing bias, as it reminds readers of the impeachment trial and frames Democrats as adversaries of the current administration. Furthermore, the statement "in the name of fighting a virus that thus far has killed less than 100 Americans" shows selection bias, as it minimizes the threat posed by the virus by focusing on the number of deaths at the time of the article, ignoring other relevant information about its potential impact.

Paragraph 3

Declaring a pandemic emergency on Friday, President Trump now claims the power to quarantine individuals suspected of being infected by the virus and, as Politico writes, "stop and seize any plane, train or automobile to stymie the spread of contagious disease." He can even call out the military to cordon off a US city or state.

We can observe the presence of language bias. The author uses strong and authoritative language to describe President Trump's actions in response to the pandemic emergency, such as "claims the power", "stop and seize", and "call out the military". These word choices contribute to a negative portrayal of the President's actions, suggesting that he is overreaching in his authority to deal with the pandemic.

Paragraph 4

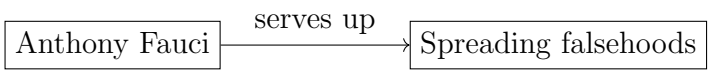
State and local authoritarians love panic as well. The mayor of Champaign, Illinois, signed an executive order declaring the power to ban the sale of guns and alcohol and cut off gas, water, or electricity to any citizen. The governor of Ohio just essentially closed his entire state.

The author uses framing bias and language bias by referring to state and local officials as "authoritarians" who "love panic" which has a negative connotation and frames them as individuals having bad intentions. It negatively describes these officials and suggests they are exploiting the situation for their benefit. Describing the actions of the mayor of Champaign and the governor of Ohio using terms like "ban the sale of guns and alcohol", "cut off gas, water, or electricity" and "closed his entire state" can be seen as selection bias, as they focus on a few instances without considering the broader context or the reasoning behind these decisions.

Paragraph 5

The chief fearmonger of the Trump Administration is without a doubt Anthony Fauci, head of the National Institute of Allergy and Infectious Diseases at the National Institutes of Health. Fauci is all over the media, serving up outright falsehoods to stir up even more panic. He testified to Congress that the death rate for the coronavirus is ten times that of the seasonal flu, a claim without any scientific basis.

Anthony Fauci is labelled as the "chief fearmonger" of the Trump Administration, exhibiting language bias by using a negative term to describe Fauci's role. The statement that Fauci is "serving up outright falsehoods" and the claim that his testimony to Congress lacks "any scientific basis" can be considered framing bias, as they imply that Fauci is intentionally spreading misinformation to manipulate public opinion.



We can observe a rhetorical bias, as the author implies that Anthony Fauci is intentionally spreading falsehoods to create panic, casting doubt on the credibility of his statements.

Paragraph 6

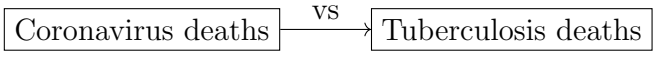
On Face the Nation, Fauci did his best to further damage an already tanking economy by stating, "Right now, personally, myself, I wouldn't go to a restaurant." He has pushed for closing the entire country down for 14 days.

In the sixth paragraph, the author highlights Fauci's statement on Face the Nation, accusing him of trying to "further damage an already tanking economy". This assertion can be seen as an example of framing bias, as it insinuates that Fauci's advice is driven by ulterior motives, rather than genuine concern for public health. Additionally, the mention of Fauci's recommendation to close the entire country for 14 days demonstrates selection bias, as it emphasizes a single aspect of his suggestions without considering the context of his recommendations.

Paragraph 7

Over what? A virus that has thus far killed just over 5,000 worldwide and less than 100 in the United States? By contrast, tuberculosis, an old disease not much discussed these days, killed nearly 1.6 million people in 2017. Where's the panic over this?

Here, the author downplays the severity of COVID-19 by comparing it to tuberculosis, questioning why there is no panic over the latter. It creates a misleading context by comparing two diseases with different transmission rates, global impacts, and preventive measures. The author's use of rhetorical questions serves to reinforce his perspective that the government's response to COVID-19 is excessive.



Paragraph 8

If anything, what people like Fauci and the other fearmongers are demanding will likely make the disease worse. The martial law they dream about will leave people hunkered down inside their homes instead of going outdoors or to the beach where the sunshine and fresh air would help boost immunity. The panic produced by these fearmongers is likely helping spread the disease, as massive crowds rush into Walmart and Costco for that last roll of toilet paper.

The author refers to Fauci and others as "fearmongers" describing them negatively, and argues that their actions will make the disease worse. The use of the terms "panic" and "fearmongers" adds to the language bias. Additionally, the assertion that "martial law" is their ultimate goal demonstrates framing bias, as it implies that their intentions are harmful to the public. The author also makes speculative statements about the consequences of these measures, such as increased panic and disease spread, without providing concrete evidence to support his claims.

Paragraph 9

> The madness over the coronavirus is not limited to politicians and the medical community. The head of the neoconservative Atlantic Council wrote an editorial this week urging NATO to pass an Article 5 declaration of war against the COVID-19 virus! Are they going to send in tanks and drones to wipe out these microscopic enemies?

In this paragraph, the author extends his critique of the COVID-19 response to include a broader range of actors, such as the Atlantic Council. The use of the word "madness" to describe the situation illustrates language bias, as it negatively portrays the actions and views of those involved, implying irrationality and overreaction. The author's sarcastic rhetorical question about sending tanks and drones to combat the virus mocks the idea of a collective response to the pandemic and implies that the proposed actions are irrational.

Paragraph 10

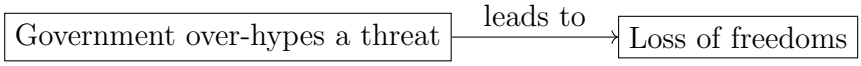
> People should ask themselves whether this coronavirus "pandemic" could be a big hoax, with the actual danger of the disease massively exaggerated by those who seek to profit — financially or politically — from the ensuing panic.

The author suggests that the coronavirus pandemic could be a "big hoax", with the danger of the disease being "massively exaggerated", suggesting that the entire situation may be fabricated for personal gain. By using quotation marks around the word "pandemic", the author further undermines the legitimacy of the situation and casts doubt on the seriousness of the crisis.

Paragraph 11

That is not to say the disease is harmless. Without question people will die from coronavirus. Those in vulnerable categories should take precautions to limit their risk of exposure. But we have seen this movie before. Government over-hypes a threat as an excuse to grab more of our freedoms. When the "threat" is over, however, they never give us our freedoms back.

In this last paragraph, the author acknowledges that the coronavirus is not harmless and that vulnerable populations should take precautions. This concession helps to establish a more balanced perspective. However, the author quickly shifts back to the main argument, asserting that the government exaggerates threats to seize freedoms. The phrase "we have seen this movie before" is an example of framing bias, as it implies that the situation is repetitive and predictable. Again, by using quotation marks around the word "threat", the author downplays the seriousness of the pandemic and reinforces the idea that the government's response is primarily driven by a desire for control. The author argues that the government's response to crises typically results in a loss of individual freedoms.



The article's overall argumentative structure follows a pattern of criticizing the government and medical community's response to the COVID-19 pandemic. The author consistently employs language bias by using emotionally charged and negative language to describe those involved in the response, such as "fearmongers." This language bias serves to discredit the actions and motivations of those responding to the pandemic. Additionally, the author employs framing bias throughout the article, often using comparisons or rhetorical questions to present his perspective as more reasonable and rational than that of the government or medical community. This bias contributes to the overall argument that the pandemic response is exaggerated and driven by a desire for power and control.

In summary, this case study reveals a range of political biases present in the article, including language bias, framing bias, and selection bias, among others. The author employs these biases to advance his argument that the government and medical community are exaggerating the threat of COVID-19 for personal and political gain. The consistent use of these biases throughout the article contributes to an overall argumentative structure that is heavily influenced by the author's political stance and skepticism toward the authorities' response to the pandemic. From this case study, we can clearly understand

the far-reaching impacts of textual bias and why it is necessary to work toward mitigating its effects. Given the multiple dimensions of textual bias, and while we will not cover all the aspects introduced in this chapter, our work will not only focus on the lexical effects of bias, whose impacts have been widely acknowledged in NLP, but also examine the less recognized discursive processes that contribute to the manifestation of bias. By keeping a focus on political biases in news articles, our intention is to build tools that can provide insights into textual biases not only through their prediction, but also and more importantly through their characterization.

Chapter 2

Beyond Words: Investigating Discourse Processes in Biased Texts

Building upon the foundation established in the first chapter, where the concepts of textual bias and their study in NLP were introduced, this section seeks to expand the understanding of textual bias beyond the lexical level. While bias has traditionally been associated with mere lexical cues, we argue that it encompasses a broader range of discursive processes, such as rhetorical, argumentative, and structural ones. Our focus here goes beyond the surface-level of textual biases, to explore the complex discourse structures that characterize biased texts.

2.1 Motivation and Background

The study of textual bias has predominantly focused on the analysis of lexical cues and linguistic patterns. While these approaches have certainly contributed to the detection and understanding of biases in texts, it is important to recognize that bias can manifest itself in more subtle ways, beyond the mere choice of words. One such avenue that merits further investigation is the analysis of discourse processes in biased texts, which includes the examination of rhetorical and argumentative structures, among other aspects. Consider the following two fabricated excerpts on the topic of environmental regulations:

Excerpt A

Environmental regulations are necessary to protect our planet. Recently, several studies have supported this claim, showing a direct correlation between reduced pollution levels and strict regulations. This body of evidence proves that we must continue to regulate industries to ensure a sustainable future for the next generation.

Excerpt B

While many argue that environmental regulations are beneficial, the economic burden they place on businesses cannot be ignored. Several industries have reported slowed growth and decreased job opportunities as a result of these strict rules. We must reconsider the true cost of such regulations before pushing for more.

In excerpt A, the argumentative structure is mainly in favor of environmental regulation, based on studies that show a positive impact on the environment. The discourse begins with a claim: "Environmental regulations are necessary to protect our planet" which is then immediately backed up by a reference to empirical evidence: "Recently, several studies have supported this claim, showing a direct correlation between reduced pollution levels and strict regulations". It then points to future consequences to persuade readers: "ensure a sustainable future for the next generation". In contrast, excerpt B emphasizes the economic implications of these regulations, suggesting a potential downside, with a different rhetorical strategy. It starts with a concession "While many argue that environmental regulations are beneficial", that turns into a counter-argument emphasizing the economic repercussions "the economic burden they place on businesses cannot be ignored", supported by an evidence "Several industries have reported slowed growth and decreased job opportunities". It then makes a call to reconsider these regulations based on perceived negative impacts, making a direct appeal to the reader: "We must reconsider the true cost of such regulations before pushing for more". This side-by-side comparison illustrates how, beyond words, discourse processes can shape and reveal various underlying biases, in particular when discussing the same topic.

Discourse, as per Schiffrin (1998), refers to the connected, coherent, and purposeful use of language, which manifests itself in the structure of the text through the organization of language beyond the level of sentences. Fundamentally, the discourse structure can be seen as a reflection of the argumentative strategies that authors employ to persuade or inform their readers. The strategic placement of claims and evidence, the choice of rhetorical moves, the establishment of coherence and cohesion, and the use of pragmatic markers, all contribute to a structured discourse that, when analyzed, can unveil potential biases. This view is supported by van Dijk's theory of news discourse, which proposes that discourse structure is one of the primary areas where biases can emerge (van Dijk, 1988). van Dijk (1998) further emphasized the role of discourse structures in shaping public opinion and perpetuating biases. His theory of news discourse elucidates how news stories are not mere factual presentations but are shaped by a complex interplay of societal norms, power relations, and implicit biases. According to van Dijk, every text is a reflection of these

underlying ideologies, and a careful study of its discourse structure can reveal them. This perspective establishes a clear motivation for the study of discourse processes in biased texts.

Discourse processes in biased texts encompass all aspects that contribute to influencing the reader beyond individual words or sentences. Several manifestations of bias in discourse can be identified. First, the organization of a text can in itself be biased. This is because the structure of a text determines what information is presented first, what is emphasized, and what is omitted, which in turn influences how readers perceive the text (van Dijk, 1998). Second, bias can manifest in the argumentative structures used in a text. Some authors may employ fallacious reasoning or biased argumentation strategies, subtly manipulating the reader's understanding and perspective (Toulmin, 2003). Third, bias can be found in how coherence and cohesion are established in a text. Authors can create a biased representation of reality by connecting ideas in a way that supports their own perspective (Halliday and Hasan, 2014). Finally, rhetorical strategies can also be a source of bias. Rhetorical devices such as metaphors, analogies, and loaded language can be used to influence the reader's opinion (Lakoff and Johnson, 2008). Given these considerations, the analysis of discourse processes represents a promising direction for improving the understanding of bias in text. It provides a means to uncover subtle forms of bias that can be overlooked in a purely lexical analysis and hence, allows for a more comprehensive and nuanced understanding of bias in texts. However, it is a challenging task, which requires models capable of handling not only lexical features, but also the complex relationships and dependencies between different parts of a text. While some previous work has already investigated the discursive aspects of textual bias in the media and politics (see Section 1.2.2), it remains an under-explored dimension.

2.2 Discourse Analysis in Natural Language Processing

Having introduced the notion of discourse, which we are interested in for the analysis of textual bias, we now turn our attention to the study of this concept from the perspective of NLP. Discourse analysis in NLP is a field that examines the structure and organization of texts beyond individual sentences. It aims to uncover the relationships and coherence between sentences and to understand how meaning is constructed in a text. One key aspect of discourse analysis is discourse parsing, which involves analyzing and representing the hierarchical structure of a text. By introducing the concept of discourse parsing, we explore different formalisms used to analyze discourse and discuss the remaining challenges in this field.

Moving beyond theory, we examine the practical applications of discourse parsing in various downstream tasks, showing that, while it holds great promise for improving the model's capabilities, leveraging parsed trees comes with many downsides. Building upon the motivation established in the previous section, our aim is to provide a comprehensive overview of discourse analysis in NLP and its implications for the study of textual bias.

2.2.1 Discourse Parsing

Discourse parsing is a crucial aspect of discourse analysis in NLP that involves structuring a text into segments, and analyzing the semantic and rhetorical relationships between them. It aims to understand texts beyond the sentence-level, focusing on inter-sentence relationships to infer meaning and intention from the larger discourse. Discourse parsing has been applied to multiple fields in NLP, such as machine translation (Chen et al., 2020a), question answering (Jansen et al., 2014), summarization (Christensen et al., 2013), and sentiment analysis (Bhatia et al., 2015), where understanding the text's overall structure and relationships between segments can improve the system's performance significantly. This process is also particularly important in understanding and interpreting the meaning and structure of news articles, as they typically involve complex and interrelated ideas. In this section, we will discuss the main approaches to discourse parsing, the key challenges faced, and the recent developments in the field.

Several discourse formalisms have been proposed in the literature, each having its unique way of representing discourse structure. Among the most widely adopted are the Rhetorical Structure Theory (RST), the Segmented Discourse Representation Theory (SDRT), and the Penn Discourse Treebank (PDTB). These theories, while fundamentally attempting to solve the same problem of representing discourse structure, propose a different perspective on how discourse can be parsed and represented. One of the earliest and most influential is Rhetorical Structure Theory (RST), which was introduced by Mann and Thompson (1988). RST is a functional theory of text organization that aims to capture how parts of a text are connected to each other to form a coherent whole. In RST, a text is parsed into a tree structure where each node represents a text segment and the relationships between these nodes capture the rhetorical relations between the segments. RST distinguishes between nucleus and satellite roles in these relations, where the nucleus is the more important segment and the satellite provides additional information about the nucleus. First, the text is segmented into Elementary Discourse Units (EDUs), which are typically clauses or phrases that express a single proposition. The idea is to identify the minimal units of discourse that still convey a complete thought. For example, consider the following sentence: "Although it was raining, she went to the store because she

needed milk." This sentence can be broken down into three EDUs: (i) "Although it was raining," (ii) "she went to the store," (iii) "because she needed milk." Each of these units expresses a single proposition. A discourse segmenter, which is a specialized model for this task, is employed to identify these units and segment the text by recognizing linguistic cues such as punctuation, syntactic structures, and conjunctions among others. After segmenting the text into EDUs, the next step is to identify the rhetorical relationships between these EDUs. This involves first identifying the attachment, i.e. which pairs of EDUs are linked (with adjacency constraint in RST), then the type of relation (e.g., elaboration, contrast, cause, etc.) and finally the direction of the relation (i.e., which EDU is the nucleus and which is the satellite). Rhetorical relations can be broadly classified as either multinuclear or mononuclear. Multinuclear relations are those in which the linked EDUs have equal status, such as in the case of a list or a sequence. On the other hand, mononuclear relations consist of a nucleus and a satellite. The classification of rhetorical relations is a challenging task due to the large number of potential relations (78 in the RST-DT corpus, Carlson et al., 2001) and the lack of explicit markers for many of these relations. Once the rhetorical relations have been identified, the discourse structure can be represented as a tree, with the EDUs as leaves and the rhetorical relations as edges. The tree is typically rooted at the most general or global rhetorical relation, with the other relations nested within it. Figure 2.1 shows an example of RST tree.

SDRT (Asher and Lascarides, 2003), on the other hand, is a formalism that extends Discourse Representation Theory (Kamp, 1981). In SDRT, a discourse is segmented into discourse units that are linked by discourse relations. However, unlike RST, SDRT represents discourse structure as a graph rather than a tree, allowing for more complex and flexible discourse structures. SDRT is particularly useful for tasks that involve understanding the temporal or causal structures in texts, or to analyze dialogues (Asher and Lascarides, 2003; Asher et al., 2016; Li et al., 2020a).

The PDTB-style (Prasad et al., 2008), is another widely used formalism in discourse parsing. PDTB is not a theory of discourse structure per se, but rather a large-scale, corpus-based resource that provides rich annotations of discourse relations in a text. Unlike RST and SDRT, whose purpose is to obtain a complete and coherent representation of discourse structure, PDTB focuses on identifying and annotating discourse connectives and their arguments, without attempting to build a global structure of the discourse. The primary components of PDTB-style annotation are discourse connectives, along with their arguments, and senses. Discourse connectives are the explicit words or phrases (e.g., "because", "however") that signal a discourse relation, the arguments of a connective are text spans that the connective relates, and the sense of a connective refers to the

type of relation it signals (e.g., Comparison, Contingency). In PDTB, a distinction is made between explicit and implicit discourse relations. Explicit discourse relations are signaled by discourse connectives (e.g., "however", "therefore"), while implicit discourse relations have no explicit connectives present in the text, but can be inferred from the context.

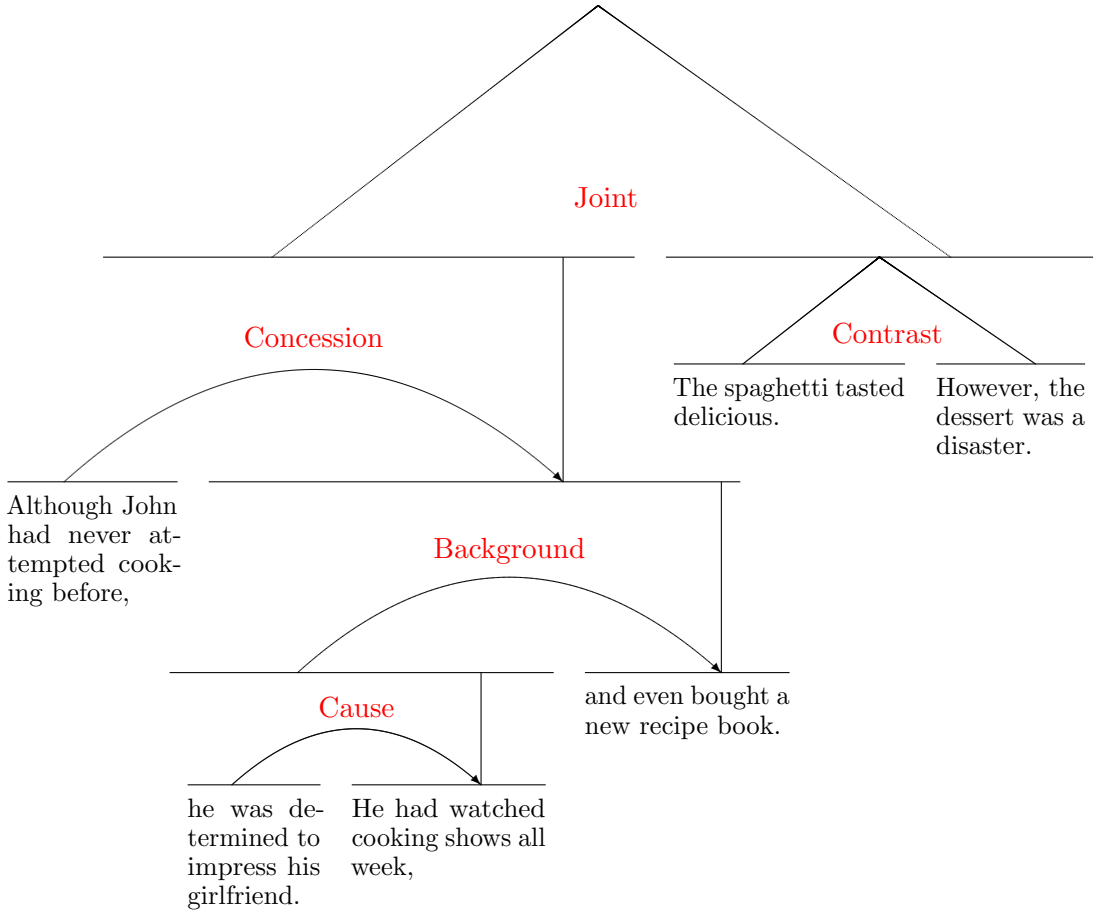


Figure 2.1: Fabricated example of RST discourse tree. The text is segmented into EDUs, which are linked via rhetorical relations (shown as a directed edge for mononuclear relation and a simple edge for multinuclear relations). The "satellite" of each mononuclear relation is pointing to the "nucleus".

Most studies on discourse have predominantly focused on RST (Feng and Hirst, 2012; Joty et al., 2012; Ji and Eisenstein, 2014; Li et al., 2016; Yu et al., 2022) and PDTB (Lin et al., 2010; Biran and McKeown, 2015; Xue et al., 2015), due to the availability of large annotated corpora in English, i.e., RST Discourse Treebank (Carlson et al., 2001) and Penn Discourse Treebank (Prasad et al., 2008), respectively, but also on other languages (da Cunha et al., 2011; Toldova et al., 2017; Cao et al., 2018; Peng et al., 2022). Despite the success of these models, discourse parsing is still a challenging task, with many areas for improvement. One such area is the handling of implicit discourse relations, which are relations not marked by explicit connectives. For example, take the following two

sentences: "John put on his heavy jacket. It was cold outside". In these sentences, there is no explicit connective such as "because" or "therefore". However, an implicit relation can be inferred: John put on his heavy jacket because it was cold outside. The cause-and-effect relationship between the cold weather and John's action is implied without being explicitly stated. Implicit relations make up a large part of the discourse (about 50% in PDTB) and are often more difficult to identify and classify than explicit ones (Braud et al., 2023). Another challenging aspect of discourse parsing is dealing with long-range dependencies, where relations hold between sentences far apart in the text. Finally, and perhaps one of the main difficulties, is the lack of large-scale, high-quality annotated corpora for training and evaluating parsers. While corpora such as the RST-DT and PDTB exist, they are relatively small, are limited to a few languages, mainly English, and do not cover the full range of discourse domains and phenomena found in natural language as they are mostly based on news articles. Moreover, the manual annotation of discourse structure is a time-consuming and complex task, further limiting the availability of training data.

2.2.2 Applications and Limitations

One of the promising applications of discourse parsing in NLP is text classification. Ji and Smith (2017) demonstrated that integrating RST discourse structures can significantly improve the performance of text categorization tasks. They proposed an attention mechanism that learns to weight the importance of sentences in a document according to their positions and relations in an RST discourse tree, as given by a discourse parser (Ji and Eisenstein, 2014). They evaluated their approach on five datasets corresponding to a wide range of classification tasks, including sentiment analysis, framing of news articles, political vote prediction, movie reviews and congressional bill survival prediction. We can note that several of these tasks involve political or media datasets, for which the incorporation of discursive information is of particular interest, further supporting the motivations of this thesis. The results of the study showed that this model outperformed prior work on four out of five tasks considered, demonstrating the usefulness of discourse parsing for downstream classification tasks. Furthermore, they found that the application of discourse parsing to text classification not only resulted in better performance, but also allowed the extraction of more meaningful insights from the text. Bhatia et al. (2015) further confirm these results by showing that RST discourse structures and recursive neural networks improve performance for sentiment classification. Another interesting work is *Pathos*, a framework that performs document sentiment analysis based on a document's discourse structure, using RST to classify the text's polarity (Heerschop et al., 2011). Given a sentence-level RST structure, the text is split into important and less important text spans, and the sentiment conveyed by distinct text spans is weighted according to their importance within

the RST structure. They showed that the performance on sentiment classification can be improved compared to a baseline not considering discourse structure. Another application of discourse parsing is text summarization, and in particular abstractive summarization. Gerani et al. (2014) explored the use of RST discourse structure in generating abstractive summaries of product reviews. They proposed a model that uses parsed trees (Joty et al., 2013) to obtain a discursive representation of the review, which is then used to generate the summary. The study shows that incorporating discourse structure into the summarization process can significantly improve the quality of the generated summaries. In the field of machine translation, Chen et al. (2020a) proposed a method to improve document-level translation by incorporating RST discourse structure information into the model. The proposed approach uses a path encoder to embed the discourse structure path of each word and combines it with the corresponding word embedding. Experimental results show that their approach outperforms competitive baselines on the English-to-German translation task.

In relation to the analysis of political and media bias, for which we do a particular focus in this thesis (discussed in Section 1.2.2), several works have proposed the examination of discourse structures. Choubey et al. (2020) proposed an approach for understanding the discourse structure of news articles. They introduced a new annotation scheme that categorizes sentences in news articles into different content types based on their roles in the discourse, following the news discourse theory proposed by van Dijk (1988). The scheme is designed to capture the discourse structure around the main event reported in a news article. The authors also develop a classifier to automatically assign these content type labels to sentences in a new article by using hierarchical neural networks. Their results show that the distributions of content types vary considerably depending on either domains or media sources. Lei et al. (2022) worked on detecting biased sentences in news articles, considering the discourse role of a sentence in telling a news story. The authors proposed to use a news discourse structure model (Choubey et al., 2020) and the PDTB discourse relations through a knowledge distillation model to identify biased sentences. Their experimental results show that incorporating both the global discourse structure and local rhetorical discourse relations can effectively increase the performance of biased sentence identification. More recently, Hong et al. (2023) proposed an innovative approach to detect political bias in news articles by considering both sentence-level semantics and document-level rhetorical structure. Their method uses a novel multi-head hierarchical attention model that encodes the structure of long documents. This method demonstrated its robustness and accuracy in detecting political bias in news articles, outperforming previous approaches highly focalized on lexical biases.

While discourse analysis has demonstrated its effectiveness in many applications, it also has a number of limitations that should be considered. A primary concern lies in the accuracy of the existing discourse parsers. Despite recent progress (Guz and Carenini, 2020; Liu et al., 2021b), parsers are not yet perfect, often struggling with ambiguity and complex sentence structures, which can ultimately lead to error propagation and erroneous downstream results, thus reflecting the expression "Garbage in, garbage out". Furthermore, the training of discourse parsers is usually done in a supervised fashion, which requires a large amount of data. Annotating this data is often costly and requires domain expertise, making the task difficult and limiting the amount of training data available. Another issue identified by Ji and Smith (2017), is the fact that discourse conventions can vary greatly across different genres, which may lead to underperformance in certain tasks, and in particular their work, political tasks. In this case, the model underperformed in making predictions about legislative bills, a genre in which discourse conventions are quite different from those in the training data of the discourse parser. This suggests that the effectiveness of a discourse parser can be heavily influenced by the type of text it is applied to, and that their performance tends to be domain-specific. Finally, the computational complexity of discourse parsing also poses an important limitation to its use. The task of parsing involves substantial computational resources and time, especially for large corpora, which may represent a constraint that is too strong for certain applications.

To sum up, discourse parsing has been applied to a variety of NLP tasks, with notable contributions to text classification (Ji and Smith, 2017), abstractive summarization (Gerani et al., 2014), and machine translation (Sim Smith, 2017). It offers a deeper understanding of text beyond the lexical level, capturing the relationships and interactions between different parts of the text. However, the limitations surrounding the performance of current discourse parsers, the lack of generalizable approaches, and the computational complexity of discourse parsing are considerable challenges to its adoption as an additional resource for downstream tasks. In the next section, we explore an alternative approach to discourse parsing for discourse-driven text classification based on the learning of latent structures.

2.3 Discrete Latent Structure as an Alternative to Discourse Parsing

Following our exploration of discourse processes in biased texts, we have acknowledged the role that discourse structure plays in understanding discourse phenomena. Discourse parsing has been the conventional way to extract such structures, however, we are now going to focus on an alternative approach which presents several notable advantages and

provides a new way of incorporating structured representations into models: discrete latent structures (Martins et al., 2019; Wu, 2022; Niculae et al., 2023).

Discrete latent structures allow the model to be fed with structural knowledge not through the usual parsed trees, but through a more adaptable method that does not require additional resources and can be trained in an end-to-end fashion driven solely by the downstream objective. While this method may not reflect typical discourse structures, it aims to produce something comparable, by capturing the task-specific structural biases of the input document. As we discuss the motivations for considering this approach, as well as the existing formalisms and challenges of discrete latent structure learning, we are making a shift in the way we analyze discourse.

2.3.1 Motivation

As discussed in Section 2.2, one of the most widely recognized approach to discourse analysis in NLP is discourse parsing, which focuses on creating structured representations of the input data from a discourse formalism. However, while discourse parsing has indeed proven to be an effective approach to understanding and modeling the discursive dimensions of language, it does not come without its challenges, as discussed in Section 2.2.2. Among these, perhaps the most critical one is the propagation of errors in NLP systems, which are often structured as pipelines. Such systems frequently include off-the-shelf components or analyzers that produce structured representations of the input data, which are then used as features in subsequent steps of the pipeline. However, as these analyzers may not have been designed with the ultimate goal in mind, the pipelines are susceptible to error propagation. Moreover, these pipelines require the availability of high performing parsers or the data to train them, which remains one of the limitations of discourse parsing. Nevertheless, one notable advantage of pipeline architectures is that they are transparent and allow for interpretability, the predicted structures can be directly inspected and used to interpret downstream predictions. This transparency is not only essential for understanding the inner workings of the system, but also crucial for identifying and fixing errors.

An interesting alternative that has gained attention in recent years is the use of discrete latent structure models in deep learning (Wu, 2022; Niculae et al., 2023). Discrete latent structures refer to structures which are not directly observed in the input data but are inferred by the model during the training process. This approach combines the strengths of both pipeline architectures and deep neural networks. Deep learning models are known for their ability to learn dense, continuous representations of data, driven solely by the downstream objective. In contrast to pipeline architectures, these models learn the structure of the data directly from the input, without the need for pre-constructed

analyzers. In essence, discrete latent structure models are a powerful tool for learning to extract representations that offer a way to incorporate structural bias and discover insight about the data that are not immediately apparent, opening the door to novel interpretations and understanding. Recent applications have employed discrete latent structures and demonstrated their potential in tasks such as machine translation (Bisk and Tran, 2018), question answering (Bogin et al., 2021), semantic parsing (Yin et al., 2018) and text summarization (Balachandran et al., 2021; Qiu and Cohen, 2022).

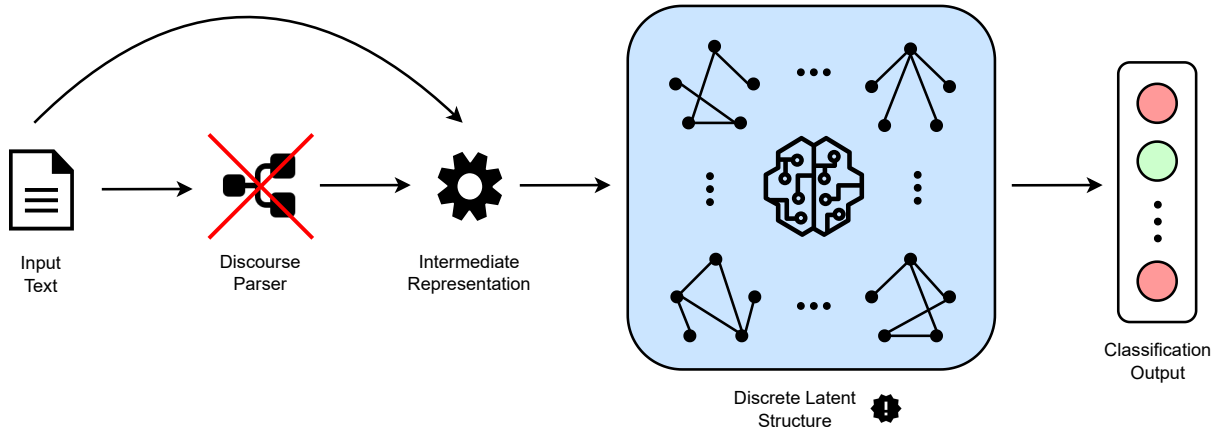


Figure 2.2: Overview of the discrete latent structure model we propose as an alternative to discourse parsing for textual bias prediction. The pipeline approach, which uses an external parser to extract the discourse structure from the input text, is replaced by an end-to-end model (in blue) where a discrete structure is latently induced for the downstream classification task.

In the context of biased texts, we propose to use latent structures to uncover and understand the hidden, sometimes implicit, biases that permeate the discourse. Figure 2.2 illustrates the proposed framework, which relies on a discrete latent structure model rather than a discourse parser to generate discourse structures for downstream tasks. This can be achieved through the careful analysis of the inferred structures, which can unveil the implicit connections and relationships that underline the bias. Consequently, this provides a more nuanced understanding of the text and its inherent biases, which traditional discourse parsing may not fully capture due to its susceptibility to error propagation. Moreover, the interpretability of the discrete latent structures provides a unique advantage in the context of biased texts. The latent structures induced by these models can be directly inspected, enabling us to interpret the decisions made by the model. This interpretability not only aids in understanding the bias inherent in the discourse, but also provides a means to identify and rectify any biases in the model itself (Ribeiro et al., 2020).

However, it is important to note that while discrete latent structure models offer promising benefits, they are not without their challenges. The induction of discrete

structures in a deep learning model can be a complex task, particularly due to the non-differentiability of discrete operations, which can hinder the application of gradient-based optimization techniques commonly used in deep learning. Moreover, the interpretability of these models, while a significant advantage, can also present its own set of challenges. The structures produced by these models can be complex and potentially difficult to interpret, particularly in the context of long and complex texts, such as those typically encountered when studying discourse processes. Therefore, future research should focus on improving the interpretability of these models and developing techniques for effectively analyzing and understanding the induced structures.

2.3.2 Discrete Latent Structure

As we have just explained, we aim to latently infer a structured representation of the input document in order to capture discourse features relevant to the downstream task. In a multitude of tasks in NLP, data can be represented through discrete structures. These structures, such as graphs, trees or sequences, can be effectively inferred using latent structure models through neural networks (Martins et al., 2019; Wu, 2022; Niculae et al., 2023). Such models are powerful tools for inferring these structures, making it possible to incorporate structural biases into the model. However, one primary challenge that remains is that the structures we are trying to learn are discrete by nature, while neural networks are designed for continuous computation. As such, it can be difficult to learn discrete latent structures, and a range of strategies have to be proposed to address this issue.

Before diving into discrete latent structures, we will first establish the common background of supervised and structured prediction with neural networks in the context of NLP. Let $\boldsymbol{x} \in \mathcal{X}$ denote an input text, and $\boldsymbol{y} \in \mathcal{Y}$ denote its corresponding label for a given prediction task. Given a dataset $\mathcal{D} = (\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)$ consisting of N input data, a generic supervised machine learning model involves learning a function $f(\boldsymbol{x}; \theta)$, parameterized by θ, that minimize the discrepancy between the model's predictions $\hat{\boldsymbol{y}}_i = f(\boldsymbol{x}_i; \theta)$ and the actual outputs $\boldsymbol{y}_i$. This discrepancy is computed by a loss function $\mathcal{L}(\boldsymbol{y}_i, f(\boldsymbol{x}_i; \theta))$ which quantifies the error in prediction. For a given dataset, the objective is then to learn the parameters θ^* that minimize the average loss over the dataset:

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{y}_i, f(\boldsymbol{x}_i; \theta)) \tag{2.1}$$

Following the same notation, we now transition into supervised discrete structure prediction. The term "discrete structure" is abstracted away from any specific formalism, and we generically denote any discrete structure with $\boldsymbol{z} \in \mathcal{Z}$, where $\mathcal{Z}$ is the set of all possible structures. An example of such a structure could be a collection of binary parts

$\boldsymbol{z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_{n^2}] \in \{0, 1\}^{n^2}$, where each entry denotes the existence of an edge between a pair of sentences in an input text $\boldsymbol{x} = [\boldsymbol{s}_1, ..., \boldsymbol{s}_n]$ composed of n sentences $\boldsymbol{s}$:

$$\begin{array}{c} \\ \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \\ \boldsymbol{s}_3 \\ \vdots \\ \boldsymbol{s}_n \end{array} \begin{array}{c} \begin{array}{ccccc} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 & ... & \boldsymbol{s}_n \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & ... & 1 \\ 1 & 0 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & ... & 1 \end{pmatrix} \end{array}$$

The model f now predicts a structured variable $\boldsymbol{z}$, and one common approach to learn $f(\boldsymbol{x}; \theta)$ is to use probabilistic models that define a distribution over possible structured outputs $\boldsymbol{z} \in \mathcal{Z}$ given an input x, such as $p(\boldsymbol{z}|\boldsymbol{x}; \theta)$. The structure $\hat{\boldsymbol{z}} = \arg\max_{z \in \mathcal{Z}} p(\boldsymbol{z}|\boldsymbol{x}; \theta)$ that maximizes this distribution is then chosen as the prediction output. In supervised structured prediction, the challenge is that the output space $\mathcal{Z}$ is typically exponentially large in the size of the input. Finding the exact structure that maximizes this distribution could be computationally expensive, therefore efficient inference algorithms are typically required in structured prediction, making the optimization problem considerably more complex than standard prediction problems.

Now that we have established the groundwork for understanding supervised prediction and discrete structures, we can further delve into the concept of latent variable models and representations in neural networks. Latent representation refers to the hidden or intermediate representation learned by a model to capture features or relationships in the data by encoding essential information from the input data in a more compact and meaningful manner. These latent variables are not directly observed, but are inferred from the observed variables. As discussed in Section 2.3.1, it is often desirable to model discrete structures latently as an intermediate representation so that the model can leverage discursive information that will be useful for the downstream task. Latent structure learning extends the concept of latent variables to structured representations. Here, instead of inferring unstructured latent variables, the model learns to infer from the input a whole structure that is not directly observed in the data. In this context, $\boldsymbol{z} \in \mathcal{Z}$ becomes a latent unobserved discrete structure, where $\mathcal{Z}$ is the latent space. We still have $f(\boldsymbol{x}, \theta)$ with $\hat{\boldsymbol{z}} = \arg\max_{z \in \mathcal{Z}} p(\boldsymbol{z}|\boldsymbol{x}; \theta)$ as the best structure, but $\hat{\boldsymbol{z}}$ is no longer the end-point in itself but is fed to some downstream task predictor g parameterized by ϕ to predict the final output $\hat{\boldsymbol{y}}$ such that $\hat{\boldsymbol{y}} = g(\hat{\boldsymbol{z}}; \phi)$, and which is trained using some loss $\mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{y}})$.

$$\hat{\boldsymbol{y}} = g(f(\boldsymbol{x}; \theta); \phi) \tag{2.2}$$

We now have the latent structure encoder model f parametrized by θ, and the downstream model g parametrized by ϕ. We consider the case where f and g are trained jointly using only the downstream loss. Unlike unstructured latent variables, which are in a continuous space, the structures we are interested in are discrete and the best structure $\hat{\boldsymbol{z}}$ is computed using the *argmax* function. As *argmax* is not differentiable, we cannot compute the gradients of the loss to optimize the parameters of the model using the gradient descent algorithm. Various strategies have been proposed to overcome this issue (Maddison et al., 2017; Peng et al., 2018; Corro and Titov, 2019b), but we will focus on a deterministic approach to learning latent structures, namely continuous relaxation.

This approach solves the non-differentiability posed by the *argmax* operation by relaxing it with a continuous and differentiable function. Rather than relying on the *argmax* operation that selects a single latent structure $\hat{\boldsymbol{z}} = \arg\max_{z \in \mathcal{Z}} p(\boldsymbol{z}|\boldsymbol{x}; \theta)$, here we consider the expectation over the set of all possible latent structures, which is differentiable. The *argmax* is replaced by taking an expectation under the distribution given by $p(\boldsymbol{z}|\boldsymbol{x}; \theta)$:

$$\hat{\boldsymbol{z}} = \mathbb{E}_{z \sim p(z|x;\theta)}[\boldsymbol{z}] = \sum_{z \in \mathcal{Z}} p(\boldsymbol{z}|\boldsymbol{x}; \theta) \times \boldsymbol{z} \tag{2.3}$$

Instead of considering a single "best" $\boldsymbol{z}$, we incorporate a form of uncertainty by taking a weighted average of all potential $\boldsymbol{z}$. Each potential structure $\boldsymbol{z}$ is associated with its probability under the distribution $p(\boldsymbol{z}|\boldsymbol{x}; \theta)$. As the space of possible structures $\mathcal{Z}$ can be exponentially large, summing over all possible structures to compute the expectation $\mathbb{E}_{z \sim p(z|x;\theta)}[\boldsymbol{z}]$ can be computationally intractable. Nevertheless, there exist efficient algorithms to solve this problem, such as the forward-backward algorithm (Rabiner, 1989) or the inside-outside algorithm (Baker, 1979), which we will discuss in the following sections. It is also important to mention here that with continuous relaxation, the predicted structure $\hat{\boldsymbol{z}}$ is no longer strictly discrete, since it represents an expectation over discrete structures. However, various strategies have been proposed for regaining discreteness, such as the use of rounding or thresholding functions post-training. Although these methods do not recover the optimal discrete structure, they enable retaining the discrete nature of the problem while still making the optimization tractable.

2.4 Structured Attention Networks

Structured Attention Networks (SANs) have emerged as a prominent class of methods offering an appealing solution to the challenges posed by discrete latent structure learning. These networks were first introduced by Kim et al. (2017), who leveraged the attention

mechanism of neural networks and the differentiability of marginal inference to learn a structured representation of the input document via the use of dynamic programming algorithms.

We make a particular focus on a variant of SANs proposed by Liu and Lapata (2018) on which we will rely for the rest of this work. Following a similar strategy, their approach aims to learn structured representations of documents, and more specifically non-projective dependency trees, using the matrix-tree theorem. With motivations similar to the ones presented in this chapter, their intention is to capture discursive phenomena in these representations by drawing on work in discourse analysis, and thus to provide an alternative solution to traditional discourse parsers.

2.4.1 A Deep Dive into Structured Attention

In the previous section, we introduced the concept of continuous relaxation as a strategy for overcoming the non-differentiability of the *argmax* operation by relaxing it with a continuous and differentiable function, i.e. an expectation over the probabilistic model. However, it can still be computationally intractable to infer the best structure $\hat{\boldsymbol{z}}$ due to the considerable size of the latent space $\mathcal{Z}$. Structured Attention Networks (SANs), which are based on continuous relaxation, were introduced by Kim et al. (2017) as a powerful mechanism for learning latent structures while effectively managing the size and complexity of $\mathcal{Z}$, by making use of the differentiability of marginal inference. SANs build on the success of attention mechanisms in NLP (Bahdanau et al., 2015) by extending the idea of attention to structured representations.

The attention mechanism in deep neural network architectures allows the model to selectively focus on certain parts of the input, essentially assigning different degrees of relevance or "attention" to different parts of the input. More formally, given a set of input values $\boldsymbol{x} = [\boldsymbol{s}_1, \boldsymbol{s}_2, ..., \boldsymbol{s}_n]$, which in our example is a text $\boldsymbol{x}$ consisting of n sentences, with their encoded representation $\boldsymbol{s}$ in the model, and a corresponding set of attention weights $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_n]$, the attention mechanism computes a weighted sum of input values $\boldsymbol{s}_i$ based on their respective attention weights $\boldsymbol{\alpha}_i$ as follows:

$$\text{Attention}(\boldsymbol{x}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{s}_i \tag{2.4}$$

Here, the attention weights $\boldsymbol{\alpha}_i$ are computed using a *softmax* function, which normalizes them to a probability distribution, over a scoring function $F(\boldsymbol{s}_i; \theta)$, which measures the relevance of each input value $\boldsymbol{s}_i$ in producing a task-specific output. F is a function that compute unnormalized scores and is typically a multilayer perceptron parameterized by θ.

$$\begin{aligned}\boldsymbol{\alpha}_i &= \text{softmax}(F(\boldsymbol{s}_i;\theta)) \\ &= \frac{\exp(F(\boldsymbol{s}_i;\theta))}{\sum_{j=1}^{n}\exp(F(\boldsymbol{s}_j;\theta))}\end{aligned} \tag{2.5}$$

In the context of SANs, Kim et al. (2017) proposed to extend this concept of attention to incorporate structured information in the attention mechanism, and to impose structural constraints on the probability distribution computed by the attention mechanism. Instead of considering a single input value $\boldsymbol{s}_i$ for computing the score and the attention weight, thus operating in a flat space where each element $\boldsymbol{s}_i$ in the input sequence is considered independently, SANs assign attention weights $\boldsymbol{\alpha}_z$ to each possible structure $\boldsymbol{z} \in \mathcal{Z}$ in relation to the input values $\boldsymbol{s}_i \in \boldsymbol{x}$. Similarly to the flat attention introduced before, the attention weights α_z are computed using a *softmax* function, which normalizes the attention weights to a probability distribution, over a scoring function $F(\boldsymbol{x}, \boldsymbol{z}; \theta)$. The function $F(\boldsymbol{x}, \boldsymbol{z}; \theta)$ represents the overall scoring function for a structure $\boldsymbol{z}$, given the set of input values $\boldsymbol{x}$ and the parameters θ, which measures the relevance of the structure $\boldsymbol{z}$ to the inputs values $\boldsymbol{x}$ in producing a task-specific output. The overall scoring function $F(\boldsymbol{x}, \boldsymbol{z}; \theta)$ is the sum of the pairwise scores $\psi(\boldsymbol{s}_i, \boldsymbol{s}_j; \theta)$ for all pairs $(\boldsymbol{s}_i, \boldsymbol{s}_j)$ that are part of the structure $\boldsymbol{z}$. ψ is a pairwise function of the neural network that compute unnormalized scores, and which is typically a bilinear function. More formally:

$$F(\boldsymbol{x}, \boldsymbol{z}; \theta) = \sum_{(s_i, s_j) \in z} \psi(\boldsymbol{s}_i, \boldsymbol{s}_j; \theta) \tag{2.6}$$

$$\begin{aligned}\boldsymbol{\alpha}_z &= \text{softmax}(F(\boldsymbol{x}, \boldsymbol{z}; \theta)) \\ &= \frac{\exp(F(\boldsymbol{x}, \boldsymbol{z}; \theta))}{\sum_{z' \in \mathcal{Z}}\exp(F(\boldsymbol{x}, \boldsymbol{z}'; \theta))}\end{aligned} \tag{2.7}$$

$\boldsymbol{\alpha}_z$ essentially quantifies the relevance of the structure $\boldsymbol{z}$ to the input values $\boldsymbol{x}$. Therefore, instead of assigning importance to individual elements of the input, structured attention assigns importance to entire structures with respect to the input values. Since the attention weights $\boldsymbol{\alpha}_z$ have been normalized to a probability distribution through the *softmax* operation, we note that $\boldsymbol{\alpha}_z$ is the equivalent of the probability distribution $p(\boldsymbol{z}|\boldsymbol{x}; \theta)$ introduced in Section 2.3.2. Thus, through continuous relaxation, the final structured representation $\hat{\boldsymbol{z}}$ is computed by taking the expectation under the distribution given by $\boldsymbol{\alpha}_z$, i.e. the weighted sum of all possible structures $\boldsymbol{z} \in \mathcal{Z}$:

$$\hat{\boldsymbol{z}} = \mathbb{E}_{z \sim \alpha_z}[\boldsymbol{z}] = \sum_{z \in \mathcal{Z}} \boldsymbol{\alpha}_z \boldsymbol{z} \tag{2.8}$$

It solves the issues related to the non-differentiability, as the computation of $\boldsymbol{\alpha}_z$ is completely differentiable, and thus the optimization can be done through gradient-based learning techniques. However, the calculation of the attention weights involves a summation over all possible structures in the latent structure space $\mathcal{Z}$, which is generally intractable due to the exponential number of such structures. To overcome this computational challenge, Kim et al. (2017) proposed two types of structured attention mechanisms to impose structural constraints on the probability distribution computed that make use of the forward-backward (Rabiner, 1989) and inside-outside (Baker, 1979) algorithms, respectively: one based on linear-chain conditional random fields (Lafferty et al., 2001) and another based on first-order graph-based dependency parsers (Eisner, 1996).

The key distinction between these two approaches lies in the nature of the dependencies captured within the structure space $\mathcal{Z}$. In the case of linear-chain conditional random fields, Kim et al. (2017) make a simplifying assumption that the relevance of an input element $\boldsymbol{s}_j$ only depends on its immediately preceding element $\boldsymbol{s}_i$. As a result, the pairwise scores $\psi(\boldsymbol{s}_i, \boldsymbol{s}_j; \theta)$ only need to consider pairs of consecutive elements, which considerably reduces the space of latent structures, and the structured attention weights $\boldsymbol{\alpha}_z$ can be computed using a relatively simple dynamic programming algorithm, namely the forward-backward algorithm (Rabiner, 1989). In contrast, the approach based on first-order graph-based dependency parsers is designed to capture more complex dependencies within the latent space, such as hierarchical and long-range dependencies. Instead of only considering pairs of consecutive output elements, the scoring function $F(\boldsymbol{x}, \boldsymbol{z}; \theta)$ in this case considers projective tree structures. Therefore, the computation of the structured attention weights $\boldsymbol{\alpha}_z$ can be computed but involves differentiating through a more complex dynamic programming algorithm, namely the inside-outside algorithm (Baker, 1979), which is capable of handling tree structures. These algorithms exploit the specific structure of the latent space and the differentiability of marginal inference to compute the normalization term in the *softmax* function in linear or polynomial time, making the approach computationally tractable.

2.4.2 Liu and Lapata (2018): Variant Approaches and their Practical Applications

We now turn our attention to a variation of SANs, which we have chosen to consider as our latent structure model in the rest of this study. This approach, which was introduced by Liu and Lapata (2018), relies on the same principle of structured attention, but employs a different strategy for the learning of the attention weights. Following the same motivations as those presented in Section 2.3.1, they seek to learn structure-aware document representations without having recourse to an external discourse parser. Taking inspiration

from existing theories of discourse and in particular from the RST formalism, which is based on tree structures, they introduced a method to learn tree-structured document representations via the matrix-tree theorem (Koo et al., 2007; Tutte and Nash-Williams, 1984). These representations act as discourse structures, but are learned latently without supervision, driven solely by the downstream objective. The authors' claim is that these representations can substitute for discourse structures given by an external parser, allowing structural biases to be incorporated into the model in an end-to-end fashion.

Our decision to adopt the model proposed by Liu and Lapata (2018) as our latent structure model over other alternatives (Niculae et al., 2018; Corro and Titov, 2019a) is founded on several key arguments. Foremost among these is the model's superior performance, demonstrated by their experimental results, which show that the representations learned by this model achieved competitive performance against strong comparison systems (Liu and Lapata, 2018). Furthermore, the method is computationally tractable and can be parallelized, the model's ability to handle longer documents, without input size limitations, and complex structures makes it well-suited to a broad range of NLP tasks, including classification of news articles. Also, their model induces intermediate structures, which can be extracted post-hoc using the Chu-Liu-Edmonds algorithm (Chu and Liu, 1965; Edmonds, 1967), that are both interpretable and meaningful, providing insights into the structural biases of the document. Compared to the tree-based approach of Kim et al. (2017), their approach considers non-projective dependency trees, which are common in document-level discourse analysis (Lee et al., 2006; Hayashi et al., 2016) and better suited to leverage discursive phenomena, while also being necessary to represent languages with free or flexible word order. In addition, the inside-outside algorithm of Kim et al. is difficult to parallelize, making it impractical for modelling long documents, whereas the approach of Liu and Lapata can be parallelized efficiently.

Let's delve into the method proposed by Liu and Lapata (2018). As a starting point, let's consider a document $\boldsymbol{x}$ constituted by a sequence of n sentences $\boldsymbol{x} = \boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n$, where each sentence $\boldsymbol{s}_i$ is a sequence of words $\boldsymbol{s}_i = \boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{m_i}$. A document's latent structured representation is conceived as a directed, rooted and non-projective dependency tree over sentences, with each sentence, in turn, being represented as a tree over its constituent words. The construction of the trees involves finding latent variables $\boldsymbol{\alpha}_{ij}$ for all $i \neq j$, where element i is the parent node of element j, under global constraints, including the single-head constraint that ensures the structure is a rooted tree. First, the sentences $\boldsymbol{s}_i$ in the document x are transformed into sequences of static word embeddings. Then, to capture the contextual information of each word in the sentence, they used bidirectional Long Short-Term Memory (bi-LSTM) networks. The bi-LSTM takes the sentences $\boldsymbol{s}_i$ as

input and produces output vectors $[\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n]$ for each sentence, each vector $\boldsymbol{h}_t \in \mathbb{R}^k$ representing a word $\boldsymbol{w}_t$ in the context of the sentence, where k is the hidden size of the bi-LSTM. These hidden representations from the bi-LSTM represent the building block for inferring latent structure. Structural information from the structured attention mechanism will be incorporated into these representations, yielding updated representations with rich structural information in it. Following recent research showing that the traditional method of using LSTM output vectors to both compute attention and encode word semantics could lead to a drop in performance (Daniluk et al., 2017; Miller et al., 2016), the authors proposed decomposing the LSTM output vector into two parts: a semantic vector $\boldsymbol{e}_t \in \mathbb{R}^{k_e}$ that encodes task-specific semantic information, and a structure vector $\boldsymbol{d}_t \in \mathbb{R}^{k_d}$ that is used to calculate the latent structured attention, where k_e is the dimension of the semantic vector and k_d is the dimension of the structure vector:

$$\boldsymbol{h}_t = [\boldsymbol{e}_t, \boldsymbol{d}_t] \tag{2.9}$$

The next step in Liu and Lapata (2018)'s approach is to capture the structural information inherent in the sentence. The structure of each sentence $\boldsymbol{s}_i$ is induced via a structured attention mechanism, based on a variant of Kirchhoff's Matrix-Tree Theorem (Koo et al., 2007; Tutte and Nash-Williams, 1984), where pair-wise attention is forced between text units to form a non-projective dependency tree. Following the SANs proposed by Kim et al. (2017), this is done in a differentiable manner, thus enabling backpropagation and learning of the attention weights. They first calculate unnormalized pair-wise scores $\boldsymbol{\psi}_{ij}$, where word i is the parent of word j in the dependency tree, computed via a bilinear function applied to the structure vectors $\boldsymbol{d}_t$ of the words. They also calculate the root score $\boldsymbol{\psi}_i^r$ which represents the unnormalized score of the word i being the root:

$$\boldsymbol{t}_p = \tanh(\boldsymbol{W}_p \boldsymbol{d}_i) \tag{2.10}$$

$$\boldsymbol{t}_c = \tanh(\boldsymbol{W}_c \boldsymbol{d}_j) \tag{2.11}$$

$$\boldsymbol{\psi}_{ij} = \boldsymbol{t}_p^T \boldsymbol{W}_a \boldsymbol{t}_c \tag{2.12}$$

$$\boldsymbol{\psi}_i^r = \boldsymbol{W}_r \boldsymbol{d}_i \tag{2.13}$$

where $\boldsymbol{t}_p$ and $\boldsymbol{t}_c$ are the representations of parent and child nodes, respectively, with $\boldsymbol{W}_p \in \mathbb{R}^{k_d \times k_d}$ and $\boldsymbol{W}_c \in \mathbb{R}^{k_d \times k_d}$ the weights for computing their representation, and $\boldsymbol{W}_a \in \mathbb{R}^{k_d \times k_d}$ the weights for the bilinear function. Then, the attention scores are normalized and constrained to reflect a non-projective dependency tree structure by computing the marginal probabilities $\boldsymbol{\alpha}_{ij}$ and $\boldsymbol{\alpha}_i^r$ using the adjacency matrix $\boldsymbol{A}_{ij}$ and the Laplacian matrix $\boldsymbol{L}_{ij}$ based on the Matrix-Tree Theorem, which represent the probabilities of a word i being a parent of word j, and a word i being the root of the tree, respectively.

These probabilities are used as attention weights.

$$\boldsymbol{A}_{ij} = \begin{cases} 0 & \text{if } i = j \\ exp(\boldsymbol{\psi}_{ij}) & \text{otherwise} \end{cases} \tag{2.14}$$

$$\boldsymbol{L}_{ij} = \begin{cases} \sum_{i'=1}^{n} \boldsymbol{A}_{i'j} & \text{if } i = j \\ -\boldsymbol{A}_{ij} & \text{otherwise} \end{cases} \tag{2.15}$$

$$\bar{\boldsymbol{L}}_{ij} = \begin{cases} exp(\boldsymbol{\psi}_i^r) & i = 1 \\ \boldsymbol{L}_{ij} & i > 1 \end{cases} \tag{2.16}$$

$$\begin{aligned} \boldsymbol{\alpha}_{ij} &= (1 - \delta_{1,j})\boldsymbol{A}_{ij}[\bar{\boldsymbol{L}}^{-1}]_{jj} - (1 - \delta_{i,1})\boldsymbol{A}_{ij}[\bar{\boldsymbol{L}}^{-1}]_{ji} \\ \boldsymbol{\alpha}_i^r &= exp(\boldsymbol{\psi}_r^i)[\bar{\boldsymbol{L}}^{-1}]_{i1} \end{aligned} \tag{2.17}$$

where $\bar{\boldsymbol{L}} \in \mathbb{R}^{n \times n}$ is a variant of $\boldsymbol{L}$ that takes the root node into consideration, and δ is the Kronecker delta. The Kirchhoff's Matrix-Tree Theorem states that for any graph, the sum of the weights of all directed spanning trees which are rooted at i can be calculated as the minor of the Laplacian matrix $\boldsymbol{L}$ with respect to row i and column i (determinant of the matrix you get after removing the i-th row and i-th column from the Laplacian matrix $\boldsymbol{L}$). It allows to constraint the attention weights by computing the normalized marginal probability $\boldsymbol{\alpha}_{ij}$ of the dependency edge between the i-th and j-th words and by ensuring that the attention scores $\boldsymbol{\alpha}_{ij}$ and $\boldsymbol{\alpha}_i^r$ converge to a non-projective dependency tree. The structured attention is then used to obtain the final representation by updating the semantic vectors $\boldsymbol{e}_t$ of each word as follows:

$$\boldsymbol{p}_i = \sum_{k=1}^{n} \boldsymbol{\alpha}_{ki}\boldsymbol{e}_k + \boldsymbol{\alpha}_i^r \boldsymbol{e}_{root} \tag{2.18}$$

$$\boldsymbol{c}_i = \sum_{k=1}^{n} \boldsymbol{\alpha}_{ik}\boldsymbol{e}_i \tag{2.19}$$

$$\boldsymbol{u}_i = tanh(\boldsymbol{W}_u[\boldsymbol{e}_i, \boldsymbol{p}_i, \boldsymbol{c}_i]) \tag{2.20}$$

where $\boldsymbol{p}_i$ is the vector gathered from potential parents of $\boldsymbol{w}_i$ and $\boldsymbol{c}_i$ the vector gathered from potential children. $\boldsymbol{e}_{root}$ is a special embedding for the root node. $\boldsymbol{p}_i$ and $\boldsymbol{c}_i$ are concatenated with $\boldsymbol{e}_i$ and transformed with weights $\boldsymbol{W}_u$ to obtain the updated semantic vector $\boldsymbol{u}_i$ with structural information in it.

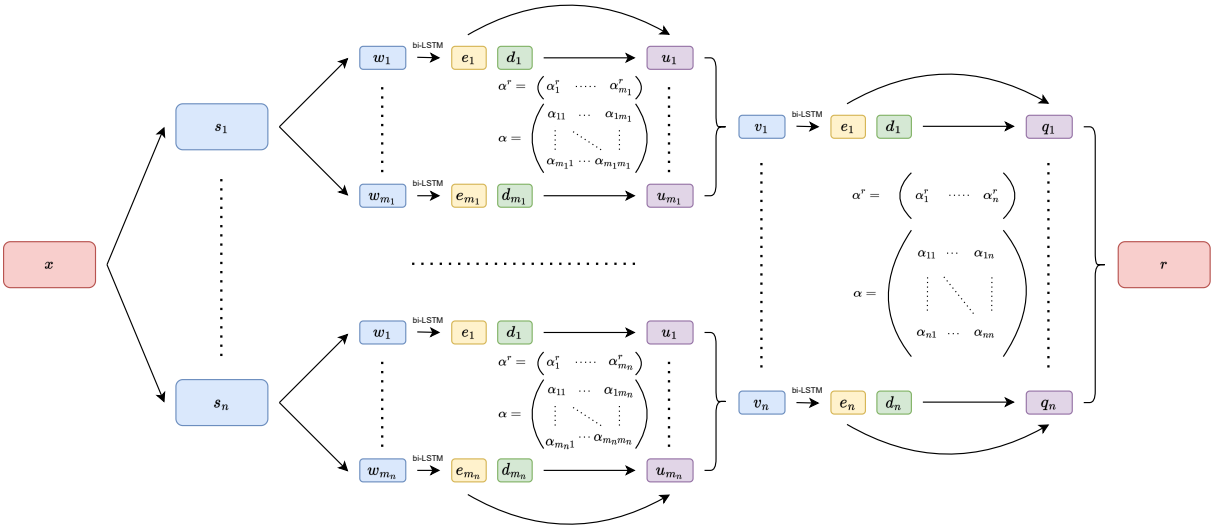


Figure 2.3: Overview of the document representation model based on latent structured attention (Liu and Lapata, 2018). A document $\boldsymbol{x}$ is composed of n sentences, where each sentence $\boldsymbol{s}_i$ is composed of m_i words $\boldsymbol{w}$. Each word embedding is fed to a bi-LSTM to obtain hidden representations $\boldsymbol{e}_i$ and $\boldsymbol{d}_i$, which are updated to representations $\boldsymbol{u}_i$ by computing the latent structured attention weight matrix $\boldsymbol{\alpha}$. Then a pooling operation produces a fixed-length vector $\boldsymbol{v_i}$ for each sentence. The same process is repeated at document level, using vectors $\boldsymbol{v}_i$ to obtain representations with structural information $\boldsymbol{q}_i$ for each sentence. A final pooling operation is used to obtain the representation of the document $\boldsymbol{r}$.

Moving on to the document-level part of the model, Liu and Lapata (2018) build document representations hierarchically, considering that sentences are composed of words and documents are composed of sentences. The same approach used for sentences is applied to obtain document representations. Given a document with n sentences $[\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n]$, where each sentence $\boldsymbol{s}_i$ consists of a sequence of word embeddings $[\boldsymbol{w}_{i1}, \boldsymbol{w}_{i2}, \cdots, \boldsymbol{w}_{im}]$ with m words, the authors feed these embeddings into a sentence-level bi-LSTM. They apply the proposed structured attention mechanism, resulting in updated semantic vectors $[\boldsymbol{u}_{i1}, \boldsymbol{u}_{i2}, \cdots, \boldsymbol{u}_{im}]$ for each word in each sentence. Then, a pooling operation is performed to obtain a fixed-length vector $\boldsymbol{v}_i$ for each sentence. Similarly, the document is viewed as a sequence of sentence vectors $[\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n]$, which are then fed into a document-level bi-LSTM. The structured attention mechanism is applied once again, yielding new semantic vectors $[\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_n]$ for each sentence. Finally, another pooling operation is performed to obtain the final document representation $\boldsymbol{r}$. An overview of the model is shown in Figure 2.3. This representation $\boldsymbol{r}$ can then be used as input for a downstream task, and the model can be trained in an end-to-end fashion, since all the operations required for computing latent structured attention are differentiable.

The authors demonstrate the effectiveness of their model through experiments on several datasets. They show that their model outperforms several state-of-the-art models

on tasks such as sentiment analysis, political vote prediction and natural language inference. They also provide an analysis of the induced structures, showing that the model can learn meaningful dependency structures without being exposed to any annotations or an external parser. The authors used the Chu-Liu-Edmonds algorithm (Chu and Liu, 1965; Edmonds, 1967) to extract the document-level dependency tree from the attention weights $\boldsymbol{\alpha}_{ij}$ and $\boldsymbol{\alpha}_i^r$. They found that, for the majority of the datasets, the trees are shallow and usually only contain nodes up to a depth of three, but most documents are relatively short, except for the political debate dataset, which contains longer documents and produces more complex trees. This finding is further supported by the fact that almost 70% of the trees induced by the model are projective trees. Still, the analysis of the document-level trees show that the model was able to learn meaningful dependency structures without any supervision.

Given the promising results demonstrated by the Liu and Lapata's approach, a number of studies have proposed adapting it to various well-known NLP tasks. One such study focuses on inducing latent dependency trees at the word level for neural machine translation (Bisk and Tran, 2018). The authors proposed a "Syntactic Attention Model" which simultaneously translates while inducing syntactic dependency trees to inform the model. The results show that the induced trees improve the translation quality. Another application is proposed by Karimi and Tang (2019), who introduced a model that constructs a discourse-level structure for fake/real news articles detection, building on the latent document representation model of Liu and Lapata. Experimental results showed that their model performed better than state-of-the-art baselines, and analysis of the induced structures according to a set of structure-related properties suggests that fake news present less coherency than real articles. Ferracane et al. (2019) explored the use of the structured attention mechanism proposed by Liu and Lapata (2018) for classification tasks, including sentiment analysis and political vote prediction, as a proxy for capturing a text's discourse structure. As their results showed that learned structures were not particularly informative, they propose several modifications to induce better structures, such as performing an additional level of percolation over the marginals to incorporate the children's children of the tree and using different pooling operations, which allowed them to obtain more complex trees and better results. In the context of abstractive summarization, Isonuma et al. (2019) proposed to leverage latent tree-structured attention for the unsupervised summarization of product reviews. This model assumes that a review can be represented as a discourse tree, with the summary as the root and child sentences explaining their parent. Their model recursively estimates a parent from its children to learn the latent discourse tree and generates a summary from the surrounding sentences of the root. They showed that their model performs better overall than other unsupervised approaches, particularly for long reviews. On the same task, Balachandran et al. (2021) proposed "StructSum", a framework

which incorporates structured document representations into summarization models using a latent structure attention module (Liu and Lapata, 2018) and an explicit structure attention module, that incorporates external linguistic structure (e.g., coreference links). This approach improves the coverage of content, generates more abstractive summaries, and incorporates interpretable sentence-level structures while performing on par with standard baselines.

2.5 From Sentences to EDUs: Changing the Building Blocks

NLP has long been established on the premise of a sentence-level representation of text (Manning and Schütze, 1999; Jurafsky and Martin, 2009). As it is the most common and natural way of segmenting a text, latent structure models have predominantly relied on this linguistic unit. Traditional sentence-level approaches, while effective at modelling local syntactic and semantic phenomena, often struggle to capture the larger, global discourse structure. This is primarily because sentences, while self-contained, are not always the best representation of the propositional content of discourse. For instance, a single sentence can contain multiple propositions, each of which may be related to different parts of the discourse (e.g. "John went to the store and Maria stayed home because she was feeling sick.").

We propose a shift towards a textual unit with a finer level of granularity, namely Elementary Discourse Units (EDUs). The concept of EDUs originates from the field of discourse parsing, where it is used as the first stage of analysis. Discourse parsing is a two-step process that begins with the segmentation of text into EDUs, followed by the construction of a discourse tree that represents the relationships between these units. An EDU is the smallest unit of discourse that can express a proposition or a coherent idea (Mann and Thompson, 1988). EDUs are more closely aligned with the propositional content of a discourse, making them a more suitable textual unit for capturing discursive phenomena. The use of EDUs as the textual unit for learning with latent structure presents several notable advantages. One of the main advantages is the ability to capture discursive phenomena at a finer granularity. Sentences, while being the basic units of syntactic analysis, often contain multiple ideas or arguments, which can be better captured and represented at the EDU level. This is particularly important in tasks that require a deep understanding of the text, such as textual bias analysis or argument mining, where the goal is to identify and extract argumentative structures from the text. By considering EDUs instead of sentences, we can identify the individual components of an argument (i.e., claim, evidence, counter-argument) more accurately, leading to a more precise and

detailed representation of the argumentative structure. Furthermore, the use of EDUs can help mitigate the problems associated with long-range dependencies in text. Sentences often contain multiple propositions that are linked to different parts of the discourse, and this can create difficulties when attempting to model the dependencies between these propositions. By focusing on EDUs, models can more easily identify and represent these dependencies, leading to more accurate predictions and better performance on downstream tasks. Finally, another advantage of using EDUs is their potential for improving the interpretability of NLP models. As EDUs represent smaller, more atomic units of meaning, they can provide more fine-grained insights into the model's decision-making process. This can be particularly useful in classification tasks, where understanding the specific parts of the text that contribute to the final decision is crucial for model interpretability. From the example below, in the sentence segmentation, the text is divided into two sentences, each expressing multiple ideas. On the other hand, the EDU segmentation divides the text into four units, each expressing a single idea, which allows for a more detailed analysis of the text.

Sentence versus EDU

Segmentation into sentences

[There's nothing abnormal about the weather this January, it's just part of the Earth's natural climate patterns.]$_1$ [The mainstream media is just pushing the idea of climate change to push their own agenda.]$_2$

Segmentation into EDUs

[There's nothing abnormal about the weather this January,]$_1$ [it's just part of the Earth's natural climate patterns.]$_2$ [The mainstream media is just pushing the idea of climate change]$_3$ [to push their own agenda.]$_4$

Moreover, a recent body of research suggests that using EDUs instead of sentences can improve the performance of NLP tasks. For instance, the work of Li et al. (2020b) leverages EDUs to improve the performance on abstractive summarization. By considering EDUs instead of sentences, their model is able to generate more coherent and informative summaries. Similarly, the work of Xu et al. (2020) introduced DISCOBERT, a discourse-aware neural summarization model that extracts EDUs instead of sentences for extractive summarization. Their model constructs structural discourse graphs based on RST trees and coreference mentions, encoded with Graph Convolutional Networks to capture long-range dependencies among discourse units. Their experiments show that DISCOBERT outperforms state-of-the-art methods on popular summarization benchmarks, further

demonstrating the potential benefits of using EDUs as the unit of analysis.

It should be noted that the process of breaking down sentences into EDUs, known as discourse segmentation, is a non-trivial task. This process requires an understanding of both the syntax and semantics of a sentence, making it significantly more complicated than sentence segmentation which can typically be achieved with simple rules or regular expressions. Therefore, it necessitates the use of a discourse segmenter to generate the segmentation in EDUs as a preprocessing step, which can add a non-negligible cost compared with sentence segmentation.

To sum up, we propose a shift in the textual unit of interest from sentences to EDUs for textual bias analysis and for the latent structure model. While approaches based on latent structure models have so far exclusively relied on sentences, this shift is motivated by the potential benefits of EDUs, including their finer granularity and their ability to capture the structure and semantics of a text more effectively. Despite the computational cost associated with the segmentation into EDUs, the potential benefits of this approach make it a promising direction for future research in NLP. The use of EDUs as the primary unit of analysis has been demonstrated to be effective in recent research, and we believe that further exploration of this approach will yield even more promising results.

Chapter 3

Experiment – Predicting Political Leanings in News Articles

Following on from the previous chapter where we introduced the framework of our approach, we present an experimental study focusing on the prediction of political leanings in news articles, which serves as a practical case study to evaluate our proposed discourse-driven method based on latent structured attention for the prediction of textual biases, thus concluding the first part of this work. We will begin by providing a description of the task at hand, emphasizing its specific objectives and challenges. We then introduce our approach, outlining the proposed classification model for predicting political leanings in news articles. Alternative approaches and baselines we considered are also presented and discussed. Subsequently, we present the datasets used for training and evaluation. Following the detailed description of the experimental setup, we proceed to the evaluation and analysis of the results obtained in order to assess the effectiveness and limitations of the proposed method with respect to the state-of-the-art and baseline approaches. This work has resulted in two scientific publications, including a preliminary workshop paper (Devatine et al., 2022) presented at *CODI*-2022 (3rd Workshop on Computational Approaches to Discourse) during the *COLING*-2022 conference, and a main paper (Devatine et al., 2023a) accepted to the Findings of the *ACL*-2023 conference.

3.1 Task Overview

Predicting political leanings in news articles is a complex task that involves analyzing the textual content of a news article and classifying it into different categories according to its political orientation. Thus, this task involves training and evaluating a classification model that takes a news article as input and predicts its political orientation as output. Here, we consider the supervised case, where we have a political annotation label for each article, which is used to train and evaluate the model (see Figure 3.1).

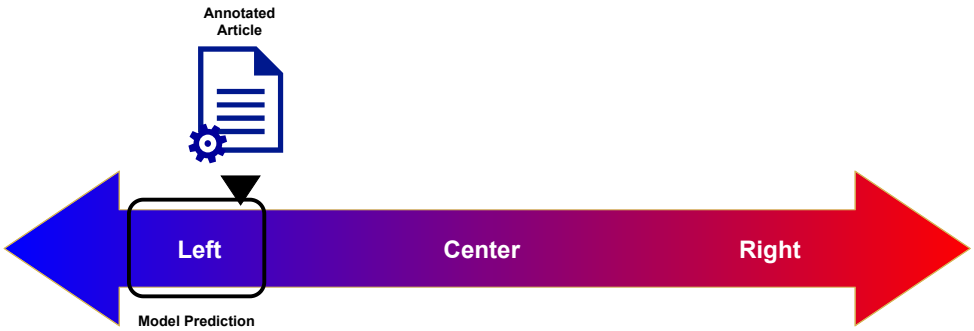


Figure 3.1: Illustration of the proposed task: predicting political leanings of news articles. Here, the classification is supervised, which means we have the gold labels given an annotation scheme on the political spectrum, in the case of this example, the gold label is 'left'.

Before discussing the specificities of the experimental setup, it is essential to define what is meant by "political orientation", "political leaning" or "political bias". Generally, all these terms refer to the positioning of beliefs or ideologies in relation to the political spectrum, which is often simplified as left-wing, center, and right-wing. However, this categorization is rather superficial, as the political landscape is actually more nuanced than that, and it is essential to recognize that this term, in itself, is multidimensional and can encompass various elements including economic policies, social issues, and more. These nuances can stem from cultural, historical, and socio-political contexts that shape the perceptions and ideologies within a region or country. Consequently, the political context of the country or region from which the articles originate plays a major role in assigning a political orientation. For instance, an article perceived as left-leaning within a predominantly conservative context may be viewed as centrist or right-leaning within a progressive one due to historical and cultural differences. In the United States, the left is often associated with Democrats, while the right is linked to Republicans. In contrast, in Europe, the political spectrum might be more diverse, with multiple parties representing various shades of the left and right, and even some that do not fit this binary view. Furthermore, it is important to note that the center class, often perceived as a neutral or unbiased stance, is not devoid of its biases, encompassing views that could be described as moderate or balanced between left and right viewpoints. Understanding that center is not equivalent to "neutral" is crucial here. Hence, the task of predicting the political leaning of a newspaper article should take into account this relative nature of political orientation, and the contextual nuances which might influence the language, tone, and content of the article. The prediction of political leanings in newspaper articles must therefore be considered for a specific political context, which will guide both the choice of dataset and the annotation scheme.

The classification of political leaning extends beyond these broad labels of left, center, and right, and extends to other divisions of the political spectrum such as extreme ideologies or other affiliations. Extremist views, for example, can be placed at the far ends of this spectrum. Furthermore, there are articles that may not fit exactly into any of these categories but may correspond to other ideologies or mixed views. For instance, an article might espouse fiscal conservatism but social liberalism, making it difficult to classify within the traditional left-center-right framework. Therefore, several variants of this task have been proposed, which are based on different annotation schemes for political bias. Some annotation schemes might use a simplified "left, center, right" version (Potthast et al., 2018; Chen et al., 2018; Baly et al., 2020a; Liu et al., 2022), while others might adopt more categories or other schemes, such as far-left, left, center-left, center, center-right, right, far-right (Wei, 2020; Aksenov et al., 2021; Sinno et al., 2022), see Section 1.2.2. However, as the number of categories increases, so does the complexity and subjectivity of the task. Notably, the political orientation can also be interpreted as a continuous variable, where articles are rated on a scale, for instance from -1 (extreme left) to +1 (extreme right). Such a representation can sometimes capture the subtleties more effectively. However, for the purposes of our experiments, we will focus here on discrete annotations. The decision to focus on discrete annotations is twofold: (i) discrete labels are more interpretable and easier for humans to understand and apply. (ii) discrete labels may align better with the way political orientations are often discussed in public discourse, with individuals and entities typically identifying with specific labels or groups. Yet, transforming political leanings into discrete categories presents its own set of challenges. Defining and annotating the data is a complex process. Political language can be implicit, and biases may be subtle. Moreover, an article might contain multiple perspectives or evolve in a manner where the political leaning is not constant throughout. This complexity makes the annotation process subjective and highly reliant on the annotator's interpretation. Moreover, working with news articles, which are long documents, introduces another layer of complexity to our task. News articles are long and complex texts, with a structure that might not be as rigid as academic papers or as predictable as novels. The content can vary widely, and the expression of political bias can be highly nuanced. Thus, analyzing the content of an article and determining a single overall political bias remains a highly complex task.

3.2 Ethical Considerations

Beyond the methodological and technical aspects, it is crucial to acknowledge the ethical considerations involved in the task of predicting the political bias of news articles. Political bias, especially in the sphere of news and media, is a controversial subject, raising numerous

ethical and moral questions. First, there is the issue of bias in the annotation itself. If the annotators have certain biases or preconceived beliefs, it could affect the labels assigned to the articles (Bender and Friedman, 2018). This, in turn, can have consequences when the model is applied in real-world scenarios where biased annotations can inadvertently strengthen or perpetuate existing biases (Dixon et al., 2018). In addition, labeling an article with a certain political leaning could inadvertently contribute to the polarization of public opinion, as news consumers could become more inclined to believe or discredit articles based on the assigned bias rather than their content. As Sunstein (2018) suggests, media fragmentation can increase polarization in society. When algorithms classify and label media sources or news articles, they might contribute to this existing fragmentation. Previous studies have also shown that explicit labels can reinforce existing beliefs and biases, even when individuals are presented with contrary evidence (Nyhan and Reifler, 2010). Another ethical aspect to consider is the potential for models to be misused. These models might be used by people to discredit news sources that don't align with their views, or to reinforce echo chambers (Pariser, 2011). Benkler et al. (2018) have shown that selective exposure to like-minded news can create a closed feedback loop, thus reinforcing pre-existing beliefs. The misuse of AI models in labeling can further contribute to the spread of disinformation, where content might be flagged or promoted based on inaccurate predictions (Diakopoulos, 2016).

It is therefore crucial to approach this task with a high degree of ethical mindfulness. Our intent is not to "label" or "classify" outlets or articles per se, but rather to build tools that can offer insights into the relationship between media and politics. Transparency in our methods, careful interpretation of our results, and open discussion about the limitations and potential misuse are crucial steps towards achieving this, and this is why an important part of this dissertation will be dedicated to the explanation of the model's decisions.

3.3 Discourse-Driven Structured Attention Network

In this section, we describe our discourse-driven structured attention model for predicting political bias in newspaper articles.[1] Building upon the latent structured attention mechanism proposed by Liu and Lapata (2018) (L&L) that we introduced in Section 2.4.2, our approach introduces several modifications and improvements to the model that we found to be beneficial for the task under consideration. In particular, these changes are threefold: (i) the level at which the structure operates, moving from sentences to EDUs, (ii) modifications to the architecture of the model proposed by L&L for text classification, (iii) adversarial adaptation to remove biases related to the media source of the article.

[1]The code is available at: `https://github.com/neops9/news_political_bias`.

Figure 3.2 presents an overview of the proposed method.

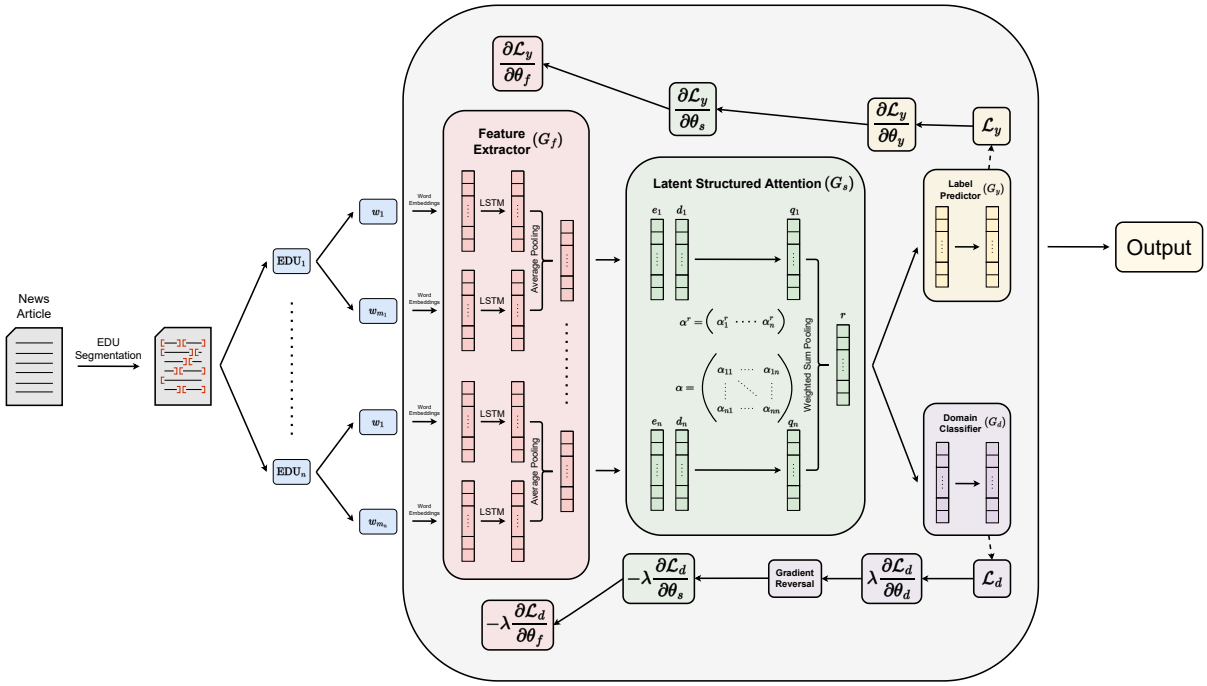


Figure 3.2: Overview of our discourse-driven structured attention model for predicting political leanings in news articles based on latent structured attention (Liu and Lapata, 2018). Taking a news article as an input, it first segments the text into EDUs based on an existing discourse segmenter (Kamaladdini Ezzabady et al., 2021). Each EDU is composed of words that are passed through an embedding layer to obtain vector representations, which are then fed to a bi-LSTM. EDU representations are obtained by aggregating word representations via an average pooling operation (red). The EDU representations are then fed into the latent structured attention network (green) to update them with structural knowledge from the latently induced non-projective dependency tree. From these updated EDU representations, a document representation r is obtained by aggregating them using a weighted sum based on the root scores. Finally, r is passed through a 2-layer perceptron to predict the distribution over class labels (yellow). We introduced an adversarial adaptation module (purple), so that the model learns to be discriminative for the main task while being media independent. The model is trained end-to-end using the loss functions $\mathcal{L}_y$ and $\mathcal{L}_d$.

Before delving into the details of these specific aspects, let us begin by describing the architecture and the general implementation of the L&L model for text classification, on which we built our model. First, the model operates at the sentence-level to create sentence representations, and then at the document-level to create a document representation from the sentence representations. Notably, for the computation of sentence representations, the process involves a first "feature extraction" module that uses pre-trained 300D GloVe embeddings (Pennington et al., 2014) for transforming the words of each sentence into vectors. These vector embeddings are then fed to a bidirectional LSTM to obtain hidden representations of words. A pooling operation aggregates the information to create the

sentence representations. Each sentence representation is then passed to the structured attention network in order to be updated with structural knowledge from the latently induced non-projective dependency tree, but we refer the reader to Section 2.4.2 for more details on this part. They thus obtain updated sentence representations with structural knowledge in it. Analogously, the same process is repeated to obtain the document representation from the sentence representations using the structured attention mechanism and a pooling operation. Finally, a two-layer perceptron with a *softmax* layer predicts the distribution over the class labels. Having introduced the approach originally proposed by L&L for text classification, we will now detail each of the modifications and additions we have made to it.

EDU Segmentation One key modification we introduce is the segmentation of texts into elementary discourse units (EDUs) rather than sentences as a preprocessing step. Instead of considering sentences as the base units for the computation of the document's latent structure, we adopt a more discourse-oriented approach using EDU segmentation as given by a discourse segmenter (see Section 2.5). We chose to use an existing discourse segmenter (Kamaladdini Ezzabady et al., 2021)[2] as it demonstrated good performance on the DISRPT 2021 shared task on segmentation (Zeldes et al., 2021), while being the only one not to require features other than tokens, making it less costly to implement. In particular, we considered the RST segmentation model on the GUM corpus (Zeldes, 2017), which achieved a F_1 score of 91.13%. Note that segmentation into EDUs as a preprocessing step represents an additional cost depending on the size of the dataset, but is performed only once.

Model Architecture Additionally, we make several improvements to the model architecture proposed by L&L based on the findings of Ferracane et al. (2019) and our own findings. First, we skip the word-level structured attention as it adds an unnecessary level of composition that was also found to have a negative empirical impact on the results, we want to focus here on the structure of the discourse. Thus, we aggregate the word representations obtained from the bi-LSTM with an average pooling operation in order to obtain the EDU representations. Secondly, following Ferracane et al. (2019), in order to capture possible children at the level of subtrees, we perform an additional level of percolation over the marginals to incorporate the children's children of the tree for the computation of the updated semantic vectors in Equation 2.20, i.e.:

[2] https://gitlab.irit.fr/melodi/andiamo/discoursesegmentation/discut

$$\boldsymbol{c}'_i = \sum_{k=1}^{n} \boldsymbol{\alpha}_{ik} \boldsymbol{u}_i$$
$$\boldsymbol{u}'_i = tanh(\boldsymbol{W}_u[\boldsymbol{u}_i, \boldsymbol{p}_i, \boldsymbol{c}'_i])$$

Finally, still following Ferracane et al. (2019) findings, in order to incorporate structural information into the pooling operation used to compute the document representation, instead of an average pooling, we aggregate over EDU representations using a sum that is weighted by the probability of a given sentence being the root, i.e., using the learned root attention scores $\boldsymbol{\alpha}_i^r$.

Adversarial Adaptation When training a model to predict the political leanings of news articles, it is expected that the model learns to recognize and understand the ideological perspectives and biases present in the text. However, in practice, this may not always be the case, and this problem, which is not often addressed in work on this task, has notably been highlighted by Baly et al. (2020a). News articles often contain specific lexical cues, such as media names, author names, email addresses, links to social media profiles, etc., which are recurrent in articles from the same source. Since articles from the same media outlet generally share the same political annotation, a model might exploit these cues for prediction rather than learning the political biases. This phenomenon raises a fundamental problem because these lexical cues have no inherent political information. Consequently, models trained this way may perform well on the training data but fail to generalize effectively to articles from different sources that lack these cues. This emphasizes the necessity to develop models that learn content-based features, rather than exploiting these lexical shortcuts. To elaborate further, let's consider an example where a model is trained to predict the political bias of articles from two hypothetical media outlets, "LiberalNews" and "ConservativePress". Suppose, "LiberalNews" always ends its articles with the slogan, "For a brighter tomorrow" and links to its social media profiles, whereas "ConservativePress" includes the author's email address at the bottom of each article. A naive model might learn to associate the presence of "For a brighter tomorrow" and social media links with liberal bias and the presence of an email address with conservative bias. Although this might achieve high accuracy on the training set, the model is not actually learning anything about political leanings and would perform poorly on data from other sources. However, removing these cues as a preprocessing step would be costly and hard to generalize, given the wide variety of forms that these lexical cues can take.

To tackle this problem, Baly et al. (2020a) suggested two approaches: Adversarial

Adaptation (Ganin et al., 2016) and Triplet Loss Pre-Training (Schroff et al., 2015). In our approach, we decided to focus on Adversarial Adaptation as it was found to be more promising from our preliminary experiments and less costly than Triplet Loss Pre-Training. This technique is inspired by the domain adaptation research that aims to make models perform well across different domains or distributions. The concept of Adversarial Adaptation is derived from the idea of training a model to be good at a primary task, predicting political leanings in this case, while being bad at a secondary task, such as identifying the source of the article. The core idea behind AA is to incorporate a media classifier into the model's architecture, and through a specialized component known as the Gradient Reversal Layer (GRL), maximize the loss of this classifier. The GRL flips the gradients during backpropagation, thereby encouraging the model to learn features that are useful for the primary task but detrimental for the secondary task of media classification.

Let the feature extractor be denoted by G_f, the label predictor by G_y, and the domain classifier by G_d. The model's parameters are updated by minimizing the loss $\mathcal{L} = \mathcal{L}_y - \lambda\mathcal{L}_d$ where $\mathcal{L}_y = \mathcal{L}(G_y(G_f(x)), y)$ is the loss for the primary task, $\mathcal{L}_d = \mathcal{L}(G_d(G_f(x)), d)$ is the loss for the domain classification task, x is the input, y is the label for the primary task, d is the domain label, and λ is a hyperparameter controlling the trade-off between the two losses. The GRL modifies the gradients such that during backpropagation, the updates encourage the feature extractor to learn representations that confuse the domain classifier, thus, L_d is maximized. Note that in order to implement this approach, we need to have access to the media label of news articles (which in practice is often the case in existing datasets), as the classification is supervised in this case. For AA training, given that the training set may contain many different media, most of which include only a few articles, we are only considering the 10 most frequent media sources for the domain classifier. Given that a dataset can contain dozens of media sources and that most of them are only represented a few times (< 5 news articles), it is less likely that these would result in the learning of any particular media-specific bias but could instead hinder the performance of the model if no relevant features can be identified for these media. Figure 3.2 illustrates the implementation of this method in our model. In essence, through this adversarial training process, the model learns to extract features that are discriminative for predicting political leanings, but are agnostic or non-informative regarding the media source. Furthermore, by not relying on source-specific cues, it increases the chances of the model generalizing well across different media sources. It is therefore crucial to acknowledge the problems associated with bias from media sources or at least to provide some form of explanation of the results in order to control this aspect of the task if nothing is done to avoid it. Otherwise, we may end up with models having misleading performances, especially if some of the media sources used in training are also present in the evaluation datasets. As this problem is still relatively

little considered and addressed in existing work on this task, it is important to emphasize it.

3.4 Alternative Approaches

In order to compare our method and to position ourselves in relation to the current best-performing models in the literature, we introduce several approaches of interest that we have considered for our experiments as a complement to the one we have proposed. We focus in particular on transformer-based approaches, which are currently the predominant class of methods in NLP due to the high performance achieved by these models. Since we have to deal with long documents (news articles) and given the difficulties posed by this type of document for transformers and NLP models in general, we also examine specialized approaches for this type of document.

3.4.1 Transformer-Based Pre-trained Language Models

Transformer models, first introduced by Vaswani et al. (2017), have become pervasive in the field of NLP with their ability to capture complex patterns in language through self-attention mechanisms. The self-attention mechanism allows the model to weigh the importance of each token in a sentence relative to all other tokens, thereby capturing dependencies between them, regardless of their distance in the text. This is a significant departure from previous models such as LSTM networks, which process sentences sequentially and often struggle with long-range dependencies. The transformer's ability to process all tokens in parallel and capture long-range dependencies has led to its widespread adoption in various tasks. While our approach doesn't rely on transformer, their efficiency and widespread use make them an inevitable reference for comparison.

One of the most notable transformer-based model is the Bidirectional Encoder Representations from Transformers (BERT). Introduced by Devlin et al. (2019), BERT is a pre-trained language model that leverages the power of unsupervised learning on large text corpora. It uses a masked language model objective to enable pre-training of deep bidirectional representations. Unlike previous models that were trained in a supervised manner on task-specific datasets, BERT is pre-trained on a large corpus of text and then fine-tuned on specific tasks. This approach allows BERT to capture general language patterns during pre-training and then adapt to specific tasks during fine-tuning, leading to state-of-the-art performance on a wide range of NLP tasks (Devlin et al., 2019; Rogers et al., 2020). A notable variant of BERT is RoBERTa (Robustly optimized BERT approach) (Liu et al., 2019b), which builds upon BERT by making modifications to the pre-training process, including training the model longer, using a larger batch size, and removing the

next sentence prediction task from the pre-training objectives. These modifications led to improved performance over BERT on a range of tasks, making RoBERTa a popular choice for many applications in NLP. In the context of our experiment, we consider RoBERTa as one of our transformer-based baselines due to its strong performance and widespread use. The power of transformer-based models like RoBERTa is further enhanced by pre-training them on massive datasets. It allows the model to learn language representations from a large amount of unlabeled data, which can then be fine-tuned with a smaller amount of labeled data for specific tasks.

In the context of predicting political leanings in news articles, a recent model that has shown promising results is POLITICS (Liu et al., 2022). POLITICS is a transformer-based model built over RoBERTa which is specifically designed for ideology prediction and stance detection in political texts. POLITICS is pre-trained on a large-scale English dataset, BIGNEWS, which contains over 3.6M news articles from 11 mainstream U.S. news outlets. The authors introduce a new pre-training objective based on the comparison of articles about the same story but written by media of different political leanings. Furthermore, they add another objective, which they call "story objective" to prevent the model from focusing on media-specific cues (as discussed in Section 3.3). From their experiments, the authors show that POLITICS outperforms strong baselines, including RoBERTa, on diverse datasets for ideology prediction and stance detection tasks, even with a limited amount of labeled samples for training.

Despite the impressive performance of transformer models, they are not without their limitations. One of the most notable is the restriction on input size. Most transformer models, including RoBERTa and POLITICS, are limited to a maximum input length of 512 tokens. This can be a significant constraint when dealing with long documents such as news articles, as it may require the text to be truncated or split, potentially leading to loss of important information (Park et al., 2022). This limitation is a result of the computational complexity of the self-attention mechanism, which scales quadratically with the sequence length. Another challenge associated with transformer models is the issue of explainability. While these models can often achieve high performance, their complex architectures and large number of parameters can make it difficult to understand why a particular prediction was made (Bibal et al., 2022). This lack of transparency can be a significant drawback in certain applications where interpretability is important. Various methods have been proposed to improve the explainability of transformer models, such as attention visualization (Vig, 2019). However, these methods often provide only a partial understanding of the model's decision-making process, and the interpretability of transformer models remains an active area of research.

3.4.2 Long Document Classification

Processing long documents, in particular for tasks such as text classification, has historically been a challenging aspect of NLP (Chung et al., 2014; Beltagy et al., 2020; Park et al., 2022). Such is the case with news articles, which are inherently lengthy and rich in content, and can easily exceed thousands of words, as they provide in-depth analysis and reporting on various issues. It is therefore essential to acknowledge the challenges associated with the processing of long sequences in NLP, particularly when it comes to transformer-based models.

Transformer-based models, including BERT and its variants, suffer from a critical limitation concerning the length of input sequences they can process. Typically, these models have a maximum input sequence length of 512 tokens. This limitation stems from the self-attention mechanism employed in transformers which computes attention weights for every token with respect to all other tokens, meaning that the memory requirement grows quadratically with the sequence length. For long documents, this becomes computationally infeasible due to inherent memory constraints. Given that news articles usually exceed the 512 tokens limit by a significant margin, as we can see in Figure 3.4 and Table 3.1 for the datasets we considered in our experiments, simply using these models as-is would lead to truncation of the input. Nonetheless, truncation may lead to a substantial loss of information, especially in our case where the goal is to extract politically relevant insights from news articles, which are long documents. As politics is often a nuanced field, cutting off parts of an article could mean missing critical context or information that could have been useful for predicting the political orientation. A solution we consider for the transformer-based baselines we are comparing against is a sliding window approach, as proposed by Wang et al. (2019) and Liu et al. (2022), with a window size of 512 (which is the maximum token limit for models like BERT) and with an overlap of 64 tokens. The overlap is used to ensure that as little information as possible is lost in between windows. Once all the windows have been processed, the outputs are then aggregated using mean pooling to form a single vector representation for the whole document. Mean pooling has the advantage of being computationally efficient while still preserving the salient features of the individual vectors. Essentially, the technique involves dividing the text into smaller parts or 'windows' that can be processed by the model. Each window is processed separately, and the resulting representations are then aggregated into a final representation for the entire document (see Figure 3.3). This method, while simple, allows us to overcome the input length limitation of transformer models without significant information loss. However, it is worth noting that there is a trade-off involved: mean pooling, used to aggregate the information from all windows, may cause some information dilution, given that it treats all windows with equal weight.

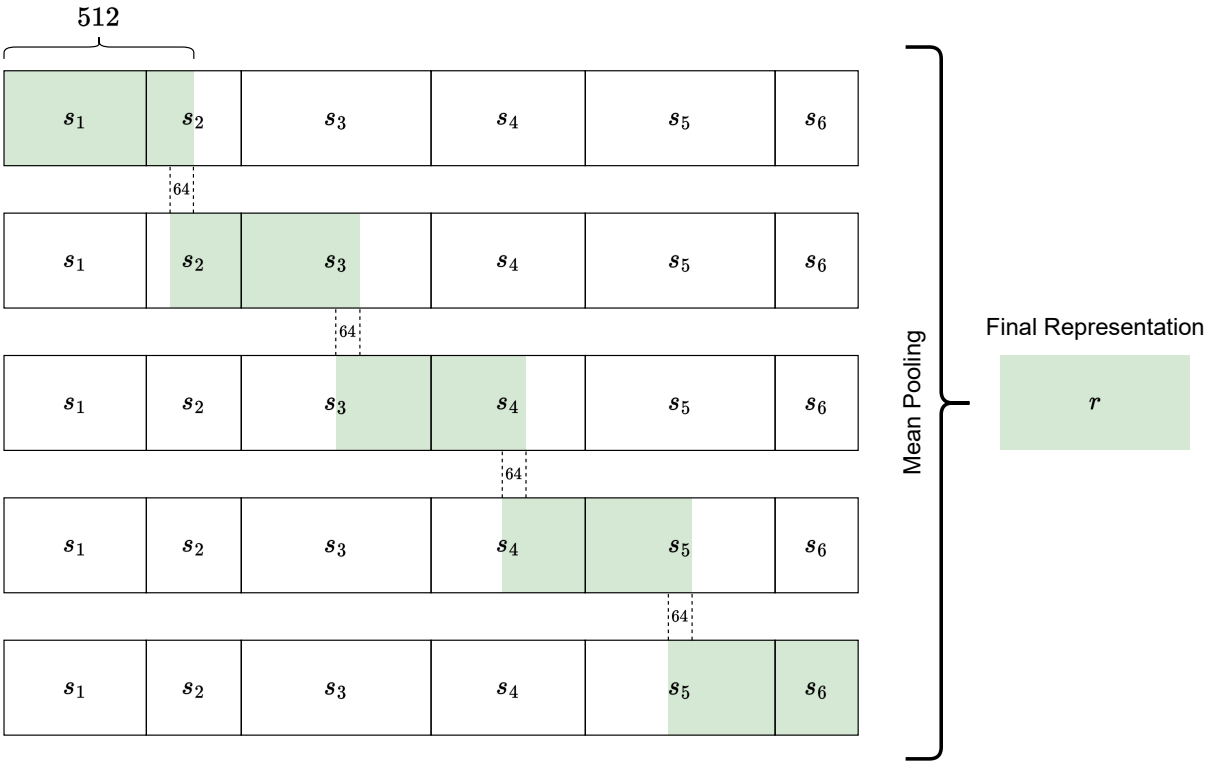


Figure 3.3: Illustration of the sliding window approach. A document x is composed of sentences s_i. We use a sliding window of size 512 with an overlap of 64, which means that we compute the representation of the first 512 tokens in the document, then the next 512 with an overlap of 64 on the first window, and so on until we reach the end of the document. All the representations obtained using this method are then aggregated using a mean pooling operation to obtain one final representation of size 512.

In light of these limitations, and to avoid having to truncate the input document, recent studies have been exploring novel architectures and techniques based on transformers that can effectively handle long sequences without significantly increasing computational costs. One such prominent approach is Longformer (Beltagy et al., 2020), a transformer model explicitly designed for long documents. The Longformer model replaces the standard self-attention mechanism with a sparse attention mechanism that scales linearly with sequence length. This allows it to process much longer sequences than a standard transformer model, up to 4096 tokens. Moreover, unlike the sliding window approach, this modification enables the model to capture dependencies across the entire document in one pass, thus reducing the computational cost. They achieve this by using a sliding window attention mechanism within the self-attention. In contrast to the standard self-attention, which attends to all previous and future tokens in the sequence, sliding window attention only attends to a certain number of previous and future tokens defined by an adaptive window size based on the token's importance. The window size is chosen such that it includes just enough context to capture relevant dependencies, but not so large as to become

computationally infeasible. Furthermore, it also includes a global attention mechanism for certain tokens, which allows these tokens to attend to all tokens in the sequence. This mechanism enables the model to retain a summary representation of the entire document. The Longformer's sliding window attention and global attention allow it to process much longer sequences than standard transformer models while still capturing the necessary dependencies. Furthermore, its low computational cost makes it suitable for large-scale tasks involving long documents. The Longformer model has been shown to perform comparably or even better than existing transformer models on a range of benchmark NLP tasks that involve long documents, such as text classification, question answering, and summarization, demonstrating its effectiveness (Beltagy et al., 2020). We therefore consider this approach in our experiments for predicting political leanings in news articles as a comparison with transformer-based approaches that use a sliding window and our latent structure method.

3.5 Datasets

As discussed in Section 3.1, predicting political leanings in news articles is a complex and multifaceted task that requires an appropriate choice of dataset and annotation scheme. It is not simply about gathering a large volume of data; it's about gathering a large volume of relevant and representative data meeting specific criteria (Wallach, 2018). Given the nature of this task, specific challenges arise concerning the annotations, topic coverage, and potential biases present in the data, which must be taken into consideration in order to build effective models and derive meaningful results.

A crucial aspect in the constitution of a good dataset is the quality of annotations. For a task such as ours, the annotation scheme and the labels assigned to data instances need to be precise, unambiguous, and reflective of the political orientations conveyed in the content. However, the lack of standardized methodologies for assigning political labels can often lead to inconsistencies in the data. In addition, political orientation is inherently subjective, and different annotators may interpret the same article differently, leading to discrepancies in the dataset. In some cases, news articles may contain nuanced positions that are difficult to categorize into distinct political labels, further complicating the annotation process. Moreover, the meaning of political labels can evolve over time and vary across geographical regions, adding another layer of complexity to the annotation task. Ensuring inter-annotator agreement becomes essential to mitigate subjectivity and maintain dataset quality (Krippendorff, 2004). Another critical consideration when constructing a dataset for political leaning prediction is the diversity of topics covered. News articles often span a wide range of subjects, and not all of them are inherently politicized. Some topics

	#BERT Tokens	#EDUs	#Sent.
Allsides	1257 ± 863	58 ± 44	32 ± 25
C-POLITICS	1008 ± 1106	100 ± 112	20 ± 24
HP	780 ± 691	81 ± 74	25 ± 24

Table 3.1: Mean and standard deviation for different levels of article length in each dataset: subtokens, EDUs, sentences.

may be neutral or have minimal political connotations (such as the result of a football match), making it challenging to determine their ideological orientation. This issue is further amplified in datasets where political bias annotations are derived from the overall bias of the media source, rather than the specific content of individual articles. Moreover, potential biases may arise from the over-representation or under-representation of certain topics in the dataset. Striking a balance between articles representing different topics and political ideologies while ensuring sufficient political content is necessary to create a comprehensive and representative dataset. Also, the geographic origin of the articles, the time period of the article publication, or even the media outlet's underlying agenda, can all contribute to potential biases in the dataset. It is therefore important to be cautious about these biases when constructing the dataset in order to minimize their impact. With these considerations in mind, we have selected three datasets[3] of interest for our experiment, which we believe best match these criteria: *Allsides* (Baly et al., 2020a), *Hyperpartisan* (Kiesel et al., 2019) and *C-POLITICS* (Liu et al., 2022). Various statistics on the length of documents are given in Table 3.1: *Allsides* and *C-POLITICS* present the longest texts. The distribution of the number of BERT tokens in the datasets is shown in Figure 3.4.

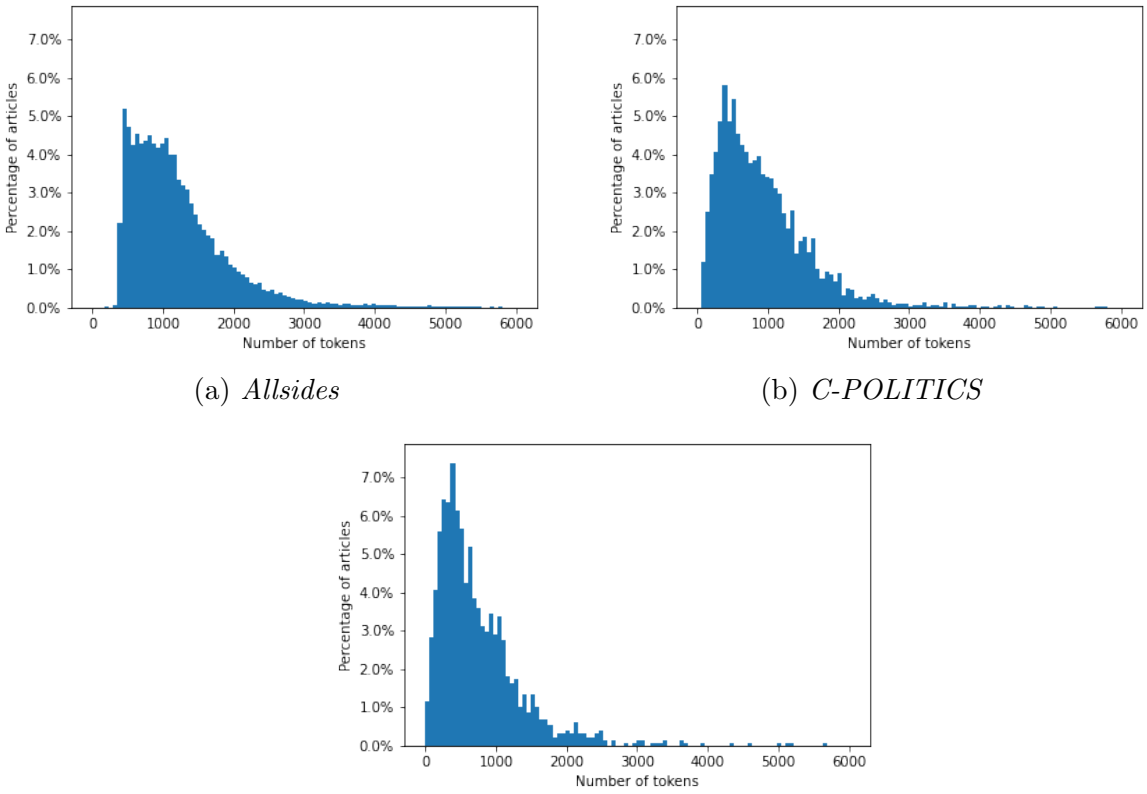
Allsides

	Left	Center	Right	Total
Train	9,618	6,683	7,189	23,490
Valid.	98	618	1,640	2,356
Test	599	299	402	1,300

Table 3.2: Number of articles per split (train, dev, test) and per class for the *Allsides* media-based dataset.

Baly et al. (2020a) were interested in the task of predicting political ideology in news articles and proposed the *Allsides* dataset. Allsides is a platform that provides an analysis

[3]Distributed under Apache License 2.0, CC BY 4.0 and CC BY-NC-SA 4.0, for *Allsides*, *Hyperpartisan* and *POLITICS* respectively.



(a) *Allsides* (b) *C-POLITICS*

(c) *Hyperpartisan*

Figure 3.4: Distribution of the number of (BERT) tokens per article for the different datasets.

of the political leaning of various English-language news articles by annotating them with 5 political classes that cover the political spectrum from the Left to the Right (see Section 1.2.2). One of the particularities of *Allsides* is that they propose an annotation of political bias at the level of individual articles, whereas the predominant method used in prior studies is distant supervision based on the annotation of the media from which the article originates (which is an oversimplification given that several articles from the same media may present different political biases). The annotation process involves a multidimensional approach that combines human input, community feedback and algorithmic analysis to rate articles.[4] Baly et al. (2020a) crawled a total of 34,737 Allsides articles from 73 news media covering 109 topics. The original Allsides data are annotated according to 5 political classes, but Baly et al. (2020a) have merged the two Left (resp. Right) classes, ending up with 3 classes for this task: Left, Center and Right. Note that the published version of the dataset[5] does not match their paper, as it includes resp. 2,817 and 119 additional articles and media. Although it complicates results comparison, we kept the published dataset, which is large and has some interesting properties.[6] In particular, it covers a wide range of political topics such as elections, immigration or

[4] https://www.allsides.com/media-bias/media-bias-rating-methods
[5] https://github.com/ramybaly/Article-Bias-Prediction
[6] Note that the original version is not available.

coronavirus, with a balanced representation of the political classes for each of these topics. Moreover, as most of the articles' political labels are aligned with those of the media from which they originate, despite the individual annotation given to the articles, the authors proposed various solutions to prevent the classifier from modelling the media instead of the political ideology. First, they preprocessed the articles in order to eliminate explicit markers such as the name of the authors or the name of the media outlet, which generally appear in the preamble of the article's content or in the content itself. Secondly, and this is one of the main features of this dataset, the authors propose two organizations for the data: random-based or media-based. In the random-based organization, articles are randomly distributed between the training, validation and test splits, which means that a given media can end up with articles in each of the different splits. Whereas in media-based organization, all articles from the same media can only appear in one of these splits at a time, which means that a media whose articles are part of the training set cannot have any articles in the test or validation sets. This ensures that the model is evaluated on articles whose sources have not been seen during training, and therefore prevents the model from predicting media rather than political ideology for evaluation, thus making the task harder. We have therefore considered this media-based split in our experiments, which contain 27,146 articles. Table 3.2 presents the distribution of articles between the different classes and splits for the *Allsides* media-based dataset.

Hyperpartisan

	Non-HP	HP	Total
Train	407	238	645
Test	314	314	628

Table 3.3: Number of articles per split (train, dev, test) and per class for the *Hyperpartisan* (HP) dataset.

The Hyperpartisan (HP) dataset, introduced by Kiesel et al. (2019), is a corpus of English news articles labeled as either hyperpartisan or not. This dataset was created as part of the shared task 4 at SemEval-2019. Hyperpartisanship refers to articles that take an extreme political standpoint. These articles were published by active hyperpartisan and mainstream websites and were all guaranteed to contain politicized content. Annotators were asked to label articles as hyperpartisan if they exhibited "extreme, hard-line views" with "unwavering and unconditional allegiance to a single party". This dataset thus adopts a binary annotation scheme for political bias that differs from the usual left/center/right scheme by focusing on extreme positions. They provided two datasets for this task: one

has 754,000 articles and is labeled in a semi-automated manner via distant supervision at the media level, while the second is smaller, with 1,273 articles labeled manually by 3 annotators. We chose the latter in order to prioritize the quality of the annotations over the quantity of data. Table 3.3 presents the distribution of articles between the different classes and splits for the *Hyperpartisan* dataset.

C-POLITICS

	Left	Center	Right	Total
Train	8,543	8,543	8,543	25,629
Valid.	890	890	890	2,670
Test	3,022	3,022	3,022	9,066

Table 3.4: Number of articles per split (train, dev, test) and per class for the *C-POLITICS* dataset.

We introduce the *C-POLITICS* dataset, a constrained version of the *BIGNEWS* dataset (Liu et al., 2022). *BIGNEWS* is a large-scale dataset of English-language news articles collected for the training of POLITICS,[7] a pre-trained language model for political ideology prediction of news articles (see Section 3.4.1). *BIGNEWS* contains 3,689,229 U.S. political news articles published between January 2000 and June 2021 from 11 media outlets ranging from far-left to far-right. They only retained news articles related to U.S. politics and annotated the political ideology into three categories, left, center and right, via distant supervision at the media level using the annotations provided by AdFontes Media[8] (see Section 1.2.2). *BIGNEWS* comes with an aligned version *BIGNEWSALIGN* where all articles discussing the same story are grouped together regardless of their political ideology. This dataset, containing 1,060,512 clusters of articles aligned on the same story (with an average of 4.29 articles per cluster), allows the comparison of articles reporting the same story but with different political ideologies. We propose a reduced and constrained version of this dataset that we call *C-POLITICS* meeting three additional desirable constraints: temporal framing, media-agnostic and class balance. First, we only kept articles published between 2020 and 2021 to ensure annotation stability, given that political labels are likely to change over time (temporal framing). Second, in the same way as in the *Allsides* dataset, we excluded the possibility of media appearing in several splits at the same time (train, validation, test), in order to avoid evaluating the model on its ability to predict media source rather than political orientation (media-agnostic). Finally, we forced each cluster to have the same number of articles of each political label, in order to guarantee

[7] https://github.com/launchnlp/POLITICS
[8] https://adfontesmedia.com

homogeneity (class balance). So, for each story in the dataset, there are as many articles representing the left, center or right class. We ended up with a dataset containing $37,365$ articles for $12,455$ clusters. Table 3.4 presents the distribution of articles between the different classes and splits for the *C-POLITICS* dataset.

3.6 Evaluation and Results Analysis

We now turn to the evaluation of the different approaches considered for the three datasets introduced. Before analyzing the results, we present our evaluation framework, starting with the evaluation metrics. We relied on the standard accuracy metric for classification, as we are dealing with well-balanced datasets. Accuracy quantifies the proportion of correct predictions among the total number of instances evaluated. For both *Allsides* and *C-POLITICS*, we also considered Mean Absolute Error (MAE) as an additional evaluation metric. The task of classifying political leanings as *left*, *center*, or *right* represents an ordinal problem. On the political spectrum, the *left* ideology is closer to the *center* ideology than to the *right* ideology (and vice versa). Thus, classifying a *left* article as *center* is less problematic than classifying it as *right*. MAE captures this cost by averaging the absolute differences between the predicted and actual values, lower MAE means better performance.

Hyperparameter	Structured Attention
# Epochs	10
Learning Rate	0.01
Batch size	8
Loss Function	Cross Entropy
Optimizer	AdamW
Weight Decay	0.01
Bi-LSTM Hidden Dim.	200
Semantic Dim.	100
Structure Dim.	100
2-layer Perceptron Dim.	200
Dropout	0.2
Adversarial Adaptation λ	0.7

Hyperparameter	RoBERTa/ POLITICS	Longformer
# Epochs	15	10
Learning Rate	$1e-4$	$2e-5$
Batch size	4	4
Max Input Length	–	4096
Loss Function	Cross Entropy	Cross Entropy
Optimizer	AdamW	AdamW
Weight Decay	0.01	0.01
Classifier # Layers	2	2
Classifier Hidden Dim.	768	768
Dropout	0.1	0.1
Sliding window size	512	–
Sliding window overlap	64	–

Table 3.5: Hyperparameters used to fine-tune the models (left table is for the structured attention models, right table is for RoBERTa, POLITICS and Longformer).

Results obtained for the different classification tasks are given in Table 3.6 (on a single run). We grouped the approaches into three categories: those from the literature, those based on pre-trained language models (PLMs) that we fine-tuned for the tasks, and those we have proposed in this thesis (see Section 3.3), based on Structured Attention (SA). For approaches taken from the literature, we have considered the state-of-the-art models proposed for the specific datasets. As these are not results that we have reproduced but that we have taken from original papers, we do not have results for all datasets for these

Model	Allsides		C-POLITICS		HP
	Accuracy	MAE	Accuracy	MAE	Accuracy
Literature					
Baly et al. (2020a)	51.4*	**0.51***	-	-	-
Jiang et al. (2019)	-	-	-	-	82.2*
PLMs					
RoBERTa	52.6	0.68	49.2	0.63	80.4
Longformer-4096	56.1	0.55	55.1	0.52	85.2
POLITICS-512	55.3	0.62	57.1	0.63	84.1
POLITICS	**60.4**	0.52	**60.5**	**0.50**	**85.8**
Structure-based models					
Structured Attention/Sent	48.8	0.67	48.6	0.57	75.6
Structured Attention/EDU	54.4	0.57	53.6	0.54	78.7

Table 3.6: Accuracy%, Mean Absolute Error (MAE, lower is better) on the test set for different versions of the model. $*$ indicates results not reproduced, taken from the original papers. "Sent"/"EDU" is for inputs segmented in sentence or discourse units. Note that POLITICS is based on RoBERTa, and already specifically fine-tuned on political texts before our own fine-tuning.

models. We performed a control experiment on the sliding window for PLMs using the POLITICS model and comparing it to the same model but without the sliding window, thus limited to the first 512 tokens of the input (POLITICS-512). We also performed a control experiment on the segmentation considered for the structured attention approach in order to evaluate the impact of the segmentation into EDUs that we proposed in comparison to a segmentation into sentences. Hyperparameters for all trained models were set using grid search (see Table 3.5). The classification model we propose (Structured Attention/EDU) contains about 120M parameters, RoBERTa and POLITICS contain about 125M parameters, and Longformer contains about 148M parameters. Training is done on an Nvidia GeForce GTX 1080 Ti GPU card.

Now that we have established the evaluation framework, we proceed to the analysis of our experimental results, which are shown in Table 3.6. Among the pre-trained language models (PLMs), the POLITICS model stands out with the highest overall accuracy and lowest MAE for all tasks. It is important to note that the POLITICS model uses extensive pre-training on a large dataset of news articles and an ideology-driven pre-training objective, which gives it significant advantages in predicting political leanings. In particular, for the *Allsides* and *C-POLITICS* tasks, the POLITICS model shows the largest gains of +4.3% and +3.4%, respectively, over the second-best model, and a low MAE of 0.52 and 0.50, respectively, among all models considered. The model's outstanding performance underlines

the benefits of specialized in-domain pre-training over more general language models such as RoBERTa or Longformer. Furthermore, the results obtained by POLITICS-512 demonstrate the benefits of considering the entire document for this task (here using the sliding window approach), and support our hypothesis that simply truncating the document would result in a significant loss of information. Indeed, the sliding window approach (POLITICS) yields a significant gain in all tasks, with an average gain of +3.4% in accuracy compared to POLITICS-512, which truncates the first 512 tokens of the input. This conclusion is further supported by the performance of the Longformer-4096 model, which has been specifically designed to handle long documents by overcoming the 512-token limit. Despite the lack of in-domain pre-training on news articles, Longformer-4096 outperforms the RoBERTa model and achieves competitive results with POLITICS models on all tasks due to its architecture's ability to process the entire document. An important finding is that the ability to process the entire article is crucial to the classification of political bias in news articles, and underlines the need to develop efficient methods for processing long documents.

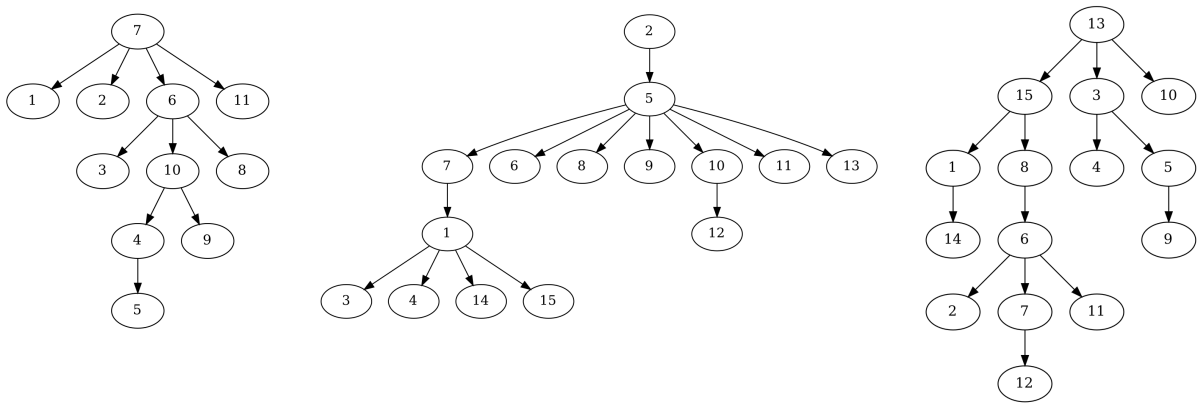


Figure 3.5: Three examples of structures induced at the EDU level by our latent structured attention model on news articles from the *Allsides* dataset and extracted using the Chu-Liu-Edmonds algorithm. The nodes correspond to the EDUs with their associated position in the document.

In contrast to the PLMs, the structured attention (SA) models we have proposed operate from a different paradigm, not relying on transformer architecture. We evaluated two versions of the SA model, one based on sentence segmentation and the other on EDU segmentation, in order to control for the impact of the finer discourse-oriented segmentation. Across all three datasets, the SA/EDU outperforms the SA/Sentence with an average of +4.6% in accuracy and −0.07 points in MAE. This shows that the discourse segmentation proves more effective in capturing the relevant features for political bias prediction when inducing the structure than sentence segmentation. Furthermore, SA/EDU outperforms by a significant margin (+3%) the state-of-the-art approach proposed by Baly et al. (2020a) for the *Allsides* dataset and the RoBERTa model (with sliding window) on

	tree height	proportion of leaf nodes	normalized arc length
Allsides			
Left	7.02	0.88	0.35
Center	6.91	0.87	0.37
Right	5.17	0.86	0.34
C-POLITICS			
Left	9.15	0.80	0.32
Center	6.20	0.76	0.33
Right	6.59	0.73	0.34
Hyperpartisan			
Non-hyperpartisan	8.42	0.81	0.33
Hyperpartisan	8.72	0.85	0.35
Global Average	7.26	0.82	0.34

Table 3.7: Average tree height, average proportion of leaf nodes and average normalized arc length of the latent trees induced by the SA model for each corpus on the test set (per class).

Allsides (+1.8%) and *C-POLITICS* (+4.4%). Nevertheless, comparing the SA models with POLITICS and Longformer, it is apparent that the SA models fall behind in terms of accuracy (−6.6% on average). While this can largely be attributed to the lack of extensive pre-training and the dominance of the transformer architecture, it shows that there is considerable room for improvement and that the structured attention approach has certain limitations. Still, despite the lower scores compared to the best-performing systems, our approach has some interesting properties that make it stand out from other methods. In particular, it can latently induce a document structure, which can then be extracted and analyzed, while being able to process long documents and not requiring massive pre-training, making it easier to adapt to a new language. Regarding the structures induced by the model, we extracted the maximum spanning trees from the attention scores using the Chu-Liu-Edmonds algorithm (Chu and Liu, 1965; Edmonds, 1967). Examples of extracted trees are shown in Figure 3.5. We report some statistics in Table 3.7 following the methodology of Ferracane et al. (2019) to examine the trees induced by the model. In particular, for each dataset and each class, we measure the average height of trees, the average proportion of leaf nodes, and the average normalized arc length (i.e. the distance between the positions in the text of the segments directly connected in the tree). Overall, the statistics obtained are similar between tasks and there are no significant differences between classes, with a global average of 7.26 for tree height, 0.82 for the proportion of leaf nodes and 0.34 for the normalized arc length. The learned trees have complex (non-flat) structures, which show that relevant information to the model has been encoded in them.

However, these results also indicate that they have marked differences from "natural" and usual structures as obtained from discourse parsers (Carlson et al., 2001), such as the presence of distant links as shown by the high average normalized arc length (direct links between EDUs that are distant in the text), and the order of nodes which tends not to follow the order of the text, in particular with the root of the tree which often does not correspond to the first segment of the document (as shown in Figure 3.5). We can note that for the *Allsides* dataset, the right-wing class have slightly more shallow trees (5.17 tree height on average), while for the *C-POLITICS* dataset the left-wing class have deeper trees (9.15 tree height on average) and the right-wing class has a lower average proportion of leaf nodes (0.73), which suggest a bias in the way documents belonging to these classes are structured in comparison with the other classes.

On a final note, the prediction of political leanings in news articles remains a challenging task, as illustrated by the overall low accuracies achieved by all models, including the state-of-the-art approaches. While the use of in-domain transformer-based models has shown superior results, there is still plenty of room for improvement. Our proposed approach based on structured attention over EDUs shows promising results by leveraging the discursive aspects of documents. However, it seems essential to push our analysis further in order to get more insight into these results and deepen our understanding of the predictions made by the models, which is not possible using these mere evaluation metrics and will be the subject of the second part of this work.

3.7 SemEval-2023 Task 3: News Genre Categorization

Before concluding this chapter, we present an auxiliary task on which we have tested our approach. SemEval2023 is a shared task on detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup (Piskorski et al., 2023). We participated in subtask 1 of "News Genre Categorization", which is a multi-class, single-label supervised classification problem that aims to determine whether a given news article is an opinion piece, an objective report, or satirical. As this task involves news articles, and is concerned with the genre of articles, where textual biases can manifest themselves, our original intention was to see how our discourse-driven model of structured attention would perform on this task. However, the competitive aspect of this shared task also led us to the exploration of other approaches of interest that proved to perform better on the leaderboard. Our system ranked among the top systems overall in most languages, and ranked 1st on the English dataset. We describe here our approach and the results obtained, which are detailed in the paper published for our participation in this

task (Devatine et al., 2023b).

Opinion Expresses the writer's opinion on a topic. Written in a persuasive style and intended to influence the reader's opinion on the subject.

Reporting Provides factual information. Aims to inform readers about the world around them and provide an objective account of events.

Satire Uses exaggeration, absurdity and obscenity to mock and ridicule people, organizations or events. Satirical pieces often mimic real articles, using irony to provide humor.

The organizers provided articles in six languages for the training phase (English, French, German, Italian, Polish, and Russian) and three surprise languages were revealed for the evaluation on the test sets (Spanish, Greek and Georgian). It is therefore a multilingual task that includes small datasets during training and zero-shot classification for surprise languages. The input data for this task are news articles in plain text format, with their title being on the first line. During the first stage, a training split and a development split were provided for each of the 6 known languages: English (*en*), French (*fr*), German (*ge*), Italian (*it*), Polish (*po*), and Russian (*ru*). Statistics for each dataset are given in Figure 3.6. The annotations were given only for the training sets at first, then also for the development sets when the test sets were released. For the evaluation on the test sets, 3 surprise languages were revealed which involves zero-shot classification (no train or dev sets): Spanish (*es*), Greek (*gr*) and Georgian (*ka*). Given the limited number of news articles available per language, this is a kind of few-shot learning task. This problem is also characterized by a strong class imbalance, especially for the *satire* genre which is represented only a few dozen times across the datasets (see Figure 3.6).

We experimented with several strategies for this task, including the structured attention approach (SA/EDU) introduced in Section 3.3 for the prediction of textual bias. SA/EDU was trained on the original data using multilingual Glove embeddings (Pennington et al., 2014). However, given that this is a competition and the aim is to achieve the best results on the leaderboard, we relied on the POLITICS model (see Section 3.4.1) as our main strategy based on the preliminary results obtained during the training phase (Liu et al., 2022). It has several advantages for this task: (i) it was massively pre-trained on more than 3.6M English news articles, (ii) it relies on the comparison of articles on the same story written by media of different ideologies, (iii) it demonstrated its robustness in few-shot learning scenarios (Liu et al., 2022). Although predicting political ideology is not the same as detecting the genre of an article, our assumption is that these two notions overlap as in both cases there is a linguistic shift in the way information is conveyed. We can also

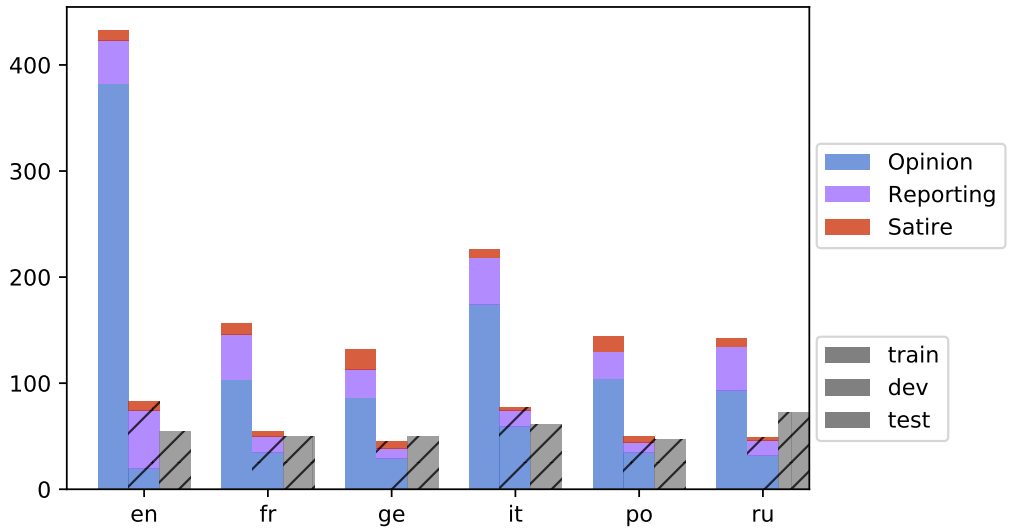


Figure 3.6: Number of news articles with associated class distribution for each known language and for each dataset split.

observe similarities between the genre reporting and articles from the center political class, or between the genre opinion and left- or right-leaning articles.

POLITICS showed much higher performance than the other models considered, and was on average better on the other languages. Thus, we have favored this model for the evaluation although a posteriori we can see that it is less efficient in certain cases (e.g. on surprise languages or on French/Russian, Tables 3.8 and 3.9). However, POLITICS was trained solely on articles from English-language media, whereas in the multilingual configuration proposed for this task, we have to classify articles in 9 languages, 3 of which were unknown during training. Because of this strong constraint and the impossibility of re-training POLITICS in a multilingual configuration, we decided to resort to translation into English. By translating all texts into English, we end up with an augmented English training set that can be used with POLITICS. This solution represents an additional cost due to the preprocessing step, and a loss of information that depends on the quality of the translation system. Several translation models were compared based on performance, language coverage, and accessibility, including GoogleTranslate,[9] DeepL[10] and OPUS-MT (Tiedemann and Thottingal, 2020), resulting in choosing GoogleTranslate as the most appropriate one. DeepL was the best performing system but had accessibility limitations due to its pricing for handling large amounts of data, we had to fall back on GoogleTranslate which was the second best performing and freely available system. We fine-tuned POLITICS on the English dataset augmented with translations. POLITICS is

[9] https://translate.google.com
[10] https://www.deepl.com/

Model	*en*	*fr*	*ge*	*it*	*po*	*ru*	Avg
Main strategy							
POLITICS	1st	7th	4th	4th	4th	5th	4th
	78.43	65.59	**77.88**	**58.66**	70.85	58.64	68.34
Baseline							
Bag-Of-Words+SVM	28.80	56.80	62.96	38.94	48.96	39.83	46.04
Control experiments							
POLITICS-512	69.43	**74.24**	76.98	53.97	66.34	58.51	66.57
Alt. approaches							
Structured Attention/EDU	61.64	54.09	62.54	57.69	53.36	51.32	56.77
XLM-RoBERTa	59.42	72.60	68.05	57.05	**79.79**	57.64	65.75
Longformer-4096	66.51	73.82	75.62	57.69	75.56	**70.57**	**69.96**

Table 3.8: Macro-F_1 scores on the test sets for each known language and different approaches. "512" means that only the first 512 subtokens of the inputs (no sliding window) were used to train the model. All results, except for the main strategy, were obtained when the submission platform reopened after the official submission deadline. We added the rank of the main system with respect to that score, according to the leaderboard published by the organizers. Note that the baseline, the structured attention model and XLM-RoBERTa are the only models that have been trained on the original data (not translated).

based on RoBERTa and, like most language models, is limited with respect to the size of the input text, here at most 512 subtokens. News articles are on average much longer, thus truncating the first 512 subtokens would result in a significant loss of information. Thus, as described previously, rather than truncating the text, we used a sliding window of size 512 with an overlap of size 64, and aggregated the information by mean pooling. Other models we have considered include, such as Longformer (see Section 3.4.2) and XLM-RoBERTa (Conneau et al., 2020). XLM-RoBERTa is a state-of-the-art multilingual language model, pre-trained on 2.5TB of CommonCrawl data containing 100 languages and including the 9 ones covered by this task. Similar to POLITICS, we used a sliding window to consider the entire article.

Tables 3.8 and 3.9 summarize the results obtained by our main strategy, the control experiments, the baseline proposed by the task organizers and the alternative approaches. The official evaluation metric used is Macro-F_1. A total of 27 teams has submitted results for subtask 1, and we ranked 1st on the English test set with a Macro-F_1 score of 78.43 (+16.8 points above the second ranked system), but were less successful on the other languages with an average of 6th place. Our main strategy (POLITICS) obtains the best results for 3 of the 9 languages, with a special distinction for English on which it particularly stands out (+9 points than the second-best approach we tested). This shows

Model	*es*	*gr*	*ka*	Avg
Main strategy				
POLITICS	5th	7th	9th	7th
	44.25	63.65	49.00	52.30
Baseline				
Bag-Of-Words+SVM	15.38	17.05	25.64	19.35
Control experiments				
POLITICS-512	48.42	**71.04**	**78.67**	**66.04**
Alt. approaches				
Structured Attention/EDU	40.71	56.08	48.75	48.51
XLM-RoBERTa	40.98	60.71	51.44	51.04
Longformer-4096	**54.81**	56.66	55.84	55.77

Table 3.9: Macro-F_1 scores on the test sets for each surprise language and different approaches. "512" means that only the first 512 subtokens of the inputs (no sliding window) were used to train the model. All results, except for the main strategy, were obtained when the submission platform reopened after the official submission deadline. We added the rank of the main system with respect to that score, according to the leaderboard published by the organizers. Note that the baseline, the structured attention model and XLM-RoBERTa are the only models that have been trained on the original data (not translated).

the benefits of large scale in-domain pre-training on English news articles, but also the limitations of translation. For French, the score is surprisingly low, which was not as pronounced in our experiments on the development split. From the control experiment, we can confirm the interest of the sliding window rather than just truncating the article, with important gains on most training languages. Interestingly, for the surprise languages, removing the sliding window leads to much better results, which shows that in a zero-shot context, introducing too much unseen language-specific information tends to confuse the model. Regarding alternative approaches, our approach based on latent structure over EDUs achieved much lower results on average than other approaches. Although this approach does not benefit from massive pre-training as with transformer-based approaches, document structure does not appear to be as relevant to this task as it is to the prediction of political bias. As for Longformer, the results confirm the previous observations on the importance of considering the whole article, with results close to or even better than POLITICS outside English and without any in-domain pre-training. Furthermore, the XLM-RoBERTa multilingual model performs well against POLITICS without the need for translation, but the lack of in-domain pre-training results in a significant performance drop, especially for English. These results should be taken with a grain of salt for the following reasons: test sets are small (30-70 instances) hence the huge variance between models and across languages. This problem is compounded by the chosen evaluation

metric: Macro-F_1 scores equalize the contributions of all classes to the final scores, meaning that one instance from the minority class classified one way or the other could swing the evaluation disproportionately. It would probably be informative to evaluate accuracy, and per class metrics, but gold labels were not provided for the test sets. This also makes it hard to estimate the distribution shift between train, development and test sets, although it is already apparent there are large differences between train and dev sets.

This shared task gave us the opportunity to test our approach with another task dealing with textual bias, in this case the detection of genre in news articles, where discourse processes can also play an important role. Although this task involved the same type of data, the multilingual, few-shot and class imbalance aspects of the task enabled us to evaluate our approach in other configurations, yielding valuable insights for future work.

3.8 Conclusion and Future Directions

In concluding this first part of our exploration of textual bias and its prediction, we focused on discursive aspects and how they contribute to the manifestation of bias, particularly in the context of political bias in news articles. At the core of our investigation was the use of an approach based on the latent structured attention mechanism from EDUs, which provides a new perspective to the analysis of textual bias in NLP. Our experimental study, which focused on the prediction of political bias in news articles, demonstrated the potential and relevance of our model. We demonstrated that political bias is not just a matter of specific words or phrases, but can be identified through discursive strategies that span entire articles and are embedded in the structure of the text. Moreover, we show that the segmentation of text into EDUs is more appropriate than sentences when using latent structure models to leverage these discourse processes.

Moving forward, we see several promising directions for extending and continuing our work. Firstly, there is room for improvement in the model's performance. Among possible improvements besides further tuning of the model parameters, we could explore other representations for words, currently based on GloVe, by leveraging contextualized word embeddings, such as ELMo (Peters et al., 2018). In addition, other pooling strategies for EDU and document representations could be explored to better aggregate information. Another perspective would be to constrain latent tree induction so that it more closely aligns with existing discourse structures and formalisms, either by defining and exploring tree subspaces, or by adding a degree of supervision to the structured attention mechanism, such as the use of reference discourse trees to guide and constrain tree induction. Furthermore, segmentation into EDUs as a preprocessing step plays a crucial role in our

approach, and the quality of segmentation depends on the segmentation model considered. Although these models perform very well on this task (over 90% F_1 score), some errors are possible. Also, segmentation into EDUs depends on the formalism considered, and different formalisms (Mann and Thompson, 1988; Asher and Lascarides, 2003) may result in different segmentations, so it would be interesting to compare the impact of different discourse segmentation formalisms on our approach. Secondly, our research could benefit from extending our focus beyond just political bias in news articles. For example, other forms of bias, such as scientific fraud or fake news detection, could be addressed using similar techniques. Although the discourse processes underlying these biases may differ from those associated to political bias, our approach is domain agnostic as the structure is learned directly from the downstream objective, and can therefore be applied to any other task. Furthermore, while we focused on news articles in this study, it would be interesting to see how our model performs on other text genres, such as social media posts, political debates or scientific papers. Different types of text may exhibit bias in different ways, thus enriching our understanding of textual bias and how it operates across various communication platforms. Finally, our experiments have primarily focused on English-language datasets, but it would be interesting to consider other languages and cultural contexts with diverse political landscapes to compare how bias manifests itself in different contexts.

However, while this first part of the thesis has laid a solid foundation for understanding and predicting biases in texts, merely predicting bias isn't enough; we also aim to characterize them, not only to provide insights into textual biases, but also to address the ethical concerns raised by the task of analyzing political biases in the media (see Section 3.2). Thus, the next chapter of this dissertation will focus on the characterization of textual biases. By understanding the specifics of how bias manifests itself in text, we can gain insights into its mechanisms and perhaps even offer ways to mitigate its effects. An essential aspect of this study will therefore be the explainability of our model: not only do we want it to perform well, but we also want to understand why it makes the predictions that it does.

Chapter 4

Bias Characterization Through Explainability Techniques

Going beyond the mere prediction of textual biases, this segment of the study goes deeper into their characterization, by keeping a particular focus on political biases in news articles. To achieve this, we focus on the field of explainability in NLP, which aims to return an explanation alongside the model's prediction. By moving away from simple bias detection, we seek to understand the "why" of the biases that our models predict, thus enabling a deeper and more insightful analysis of textual biases. More specifically, we explore model-agnostic and perturbation-based explanation methods, with a focus on the LIME (Locally Interpretable Model-Agnostic Explanations) technique. While these methods have many advantages, they also pose a number of challenges, especially when dealing with longer documents. We therefore propose several strategies based on different levels of granularity (i.e. words, sentences, EDUs, structures) to address these challenges and improve the explanations generated. To assess the quality of our explanations, we employ specific evaluation measures, and perform an experimental study on the characterization of political biases in news articles.

4.1 Motivation

In the initial chapters of this thesis, the focus was primarily on predicting biases, particularly political biases in news articles. However, we are now moving beyond just identifying biases to understanding how they are formed and characterized. It leads us to the exploration of "explainability" in machine learning, a field that is driven by the need to understand and interpret the decisions made by models. Explainability serves as a bridge between sophisticated machine learning models and human comprehension (Ribeiro et al., 2016; Doshi-Velez and Kim, 2017; Samek et al., 2017). Although machine learning models are becoming increasingly effective at prediction, the explanations for their decisions

remain opaque. These models are often "black boxes", in particular neural networks, providing little insight into their decision-making processes, and this lack of transparency poses challenges, particularly when these models are used in sensitive contexts. In our case, news and information dissemination is such a domain, where the impact of biases can have far-reaching consequences on societal beliefs and public opinion. For instance, let's take the example of a news article flagged as being politically biased towards a specific political leaning. Understanding why the model made that decision could reveal specifics about the nature of that bias. Are certain keywords associated with a particular political leaning? Are there patterns in sentence structure or language use that signal a bias? Answers to these questions could inform journalists and editors about potential pitfalls to avoid when seeking to provide unbiased reporting. They could also inform readers, making them more critical consumers of news. In this way, explainability can serve as a tool to fight misinformation and promote a more informed public discourse. It may even help to uncover unexpected or counterintuitive features that the model considers important, which could potentially lead to new insights into how bias manifests itself in text. Explainability also becomes essential when considering the legal and ethical implications tied to machine learning. The European Union's General Data Protection Regulation (GDPR), for instance, includes a "right to explanation", where users can ask for clarifications on algorithmic decisions that affect them (Goodman and Flaxman, 2017). Such laws make it evident that we can't rely on "black box" models in situations where their decisions have real-world impacts. Moreover, explaining model decisions is also important for improving models and correcting errors. When we understand how a model is making decisions, we can better identify where it's going wrong and how to fix it. For instance, if a model trained to predict political leaning from news articles relies heavily on the name of the news source, we need to recognize this to adjust our model and our data accordingly.

It is worth noting that the field of explainability in AI, particularly in NLP, is still in its early stages. However, the need for explainability has grown considerably in recent years due to the increasing societal impact of AI and the decision-making processes that result from it (Gunning and Aha, 2019; Barredo Arrieta et al., 2020). Explainability helps us understand why a machine learning model makes certain decisions. It allows us to uncover what features the model relies on, whether these features are meaningful, and whether the model behaves as expected under various circumstances. By using explainability techniques in this study, we can go beyond just predicting the bias and actually begin to understand and characterize the bias.

4.2 Explainability in Natural Language Processing

Explainability in NLP, sometimes also referred to as interpretability, can be understood as the ability to understand and describe how and why an NLP model makes specific predictions or decisions (Lipton, 2018; Guidotti et al., 2018). Recent advances in NLP have led to significant improvements in the performance of many models. However, this progress has often resulted in models that are more complex and difficult to understand. Explainability therefore becomes important because it increases the trust in AI systems by making their decisions more transparent. It also provides an understanding of the model's behavior, which can be useful for identifying and rectifying potential errors or biases in the model.

One of the fundamental distinctions in the field of explainability is between local and global explanations. A local explanation provides insight into the model's decisions for a specific instance. It answers the question: why did the model make this specific prediction for this specific input? On the other hand, a global explanation is about understanding the overall behavior of the model across a wide range of inputs: how does the model generally make predictions? Another key distinction is between self-explaining and post-hoc explanations. Self-explaining models are those designed to be inherently interpretable. They integrate explainability into their architecture, meaning their inner workings are transparent, and their decision-making process can be readily understood without additional interpretation tools. These include decision trees, rule-based systems, and interpretable neural networks (Alvarez-Melis and Jaakkola, 2017). In contrast, post-hoc explanations are generated after the model makes a prediction. These are particularly useful for complex models whose internal workings are hard to interpret directly, such as deep learning models. Post-hoc explanations often involve creating a simpler, interpretable model that approximates the original model's behavior on a subset of the input space (Ribeiro et al., 2016). Furthermore, among the post-hoc methods, we distinguish model-agnostic methods (or "black box"), which are a type of post-hoc explanation that operates independently of the specific model architecture, thereby offering wide applicability. These methods generally aim to approximate the model's decision boundary around the instance of interest using a simpler, interpretable model. One of the most famous examples is LIME (Local Interpretable Model-Agnostic Explanations), which generates local explanations by fitting a simple model to the instance's neighborhood (Ribeiro et al., 2016). Moreover, we can categorize explanations based on their nature into abductive explanations and counter-factual explanations. Abductive explanations are directly derived from the model's predictions. They aim to find the most likely cause for a particular output (Miller, 2019). For instance, identifying which words in a text led the model to classify it as positive. Counter-factual explanations, however, provide insights into what could have been. They explain by showing the least amount of change needed to alter the model's decision (Wachter et al.,

2018). For example, what word changes in a text would have made the model classify it as negative instead? Table 4.1 summarizes the different categories of existing approaches.

Category	Definition	Example Methods
Local Post-Hoc	Explain a single prediction by performing additional operations after the model has made a prediction. Depend on the output, but not the internal workings of the model.	LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), Anchors (Ribeiro et al., 2018).
Local Self-Explaining	Explain a single prediction using information from within the model itself. Made as part of making the prediction and are closely tied to the model's internal workings.	Attention mechanisms in transformer models, rule-based models.
Global Post-Hoc	Involve additional operations to explain the model's overall predictive reasoning. Aim to provide an overview of the model's behavior across multiple inputs and outputs.	Partial Dependence Plots (Friedman, 2001), DeepLIFT (Shrikumar et al., 2017).
Global Self-Explaining	Use the model itself to explain its predictive reasoning across all inputs and outputs.	Decision trees, linear/logistic regression models.

Table 4.1: Overview of the different high-level categories of explanations, including examples of common methods for generating each type (inspired by Table 1 in Danilevsky et al., 2020).

When it comes to generating explanations, several techniques have been proposed, each with its own strengths and limitations. Attention mechanisms, for example, are often used in neural networks to indicate the importance of different parts of the input when making a prediction (Vaswani et al., 2017). However, as pointed out by Wiegreffe and Pinter (2019), attention scores may not always align with feature importance, calling into question their reliability as explanation methods. Gradient-based explanations, meanwhile, leverage the gradients of a model's output relative to its input (Sundararajan et al., 2017). The idea is that the value of the gradient can indicate which parts of the input affect the output the most. Other approaches, based on perturbations, are alternative methods that work by assessing the change in the model's output when the input is slightly altered. This gives an indication of which parts of the input are most influential in the model's decision (Fong and Vedaldi, 2017). Among these approaches, we distinguish surrogate models, which involve training an interpretable model (the surrogate) to approximate the predictions of the original model, thus offering a simplified, interpretable version of the model (Ribeiro et al., 2016). Alternatively, the *Anchors* method (Ribeiro et al., 2018) uses perturbations to learn decision rules. Feature interaction methods,

like Shapley Additive Explanations (SHAP), measure the contribution of each feature to the prediction by considering the interaction of features. SHAP values, based on game theory, offer consistent and locally accurate attributions (Lundberg and Lee, 2017). The choice of explanation method depends on the context, the type of model used, and the specific requirements of the task. It is essential to keep in mind that no method can offer perfect explanations, and that each has its own limitations and assumptions (Molnar, 2022).

Visualization techniques are used to make these explanations easier to understand and interpret. These techniques represent explanations visually, making them more intuitive and easier to understand. For example, saliency heatmaps can show the importance of different parts of the input, while decision trees can visually represent the decision-making process (Wattenberg et al., 2016). However, creating effective visualizations is challenging, and it is important to ensure that they accurately represent the explanation and are not misleading (Krause et al., 2016). Assessing the quality of explanations is another important aspect of explainability. It involves in particular evaluating the fidelity (how well the explanation represents the model's behavior), consistency (how stable the explanations are under slight changes in the input), and comprehensibility (how easily the explanation can be understood by a human) of the explanations (Doshi-Velez and Kim, 2017). Various methods have been proposed to evaluate these aspects, but this remains a challenging and open research question (Molnar, 2022), which we will discuss in more details in Section 4.4. Despite the progress made, there remain significant challenges in explainability. These include improving the reliability and trustworthiness of explanations, developing more effective visualization techniques, understanding the human factors in explainability (e.g., what makes an explanation understandable or useful), and bridging the gap between local and global explanations (Danilevsky et al., 2020).

4.2.1 Locally Interpretable Model-Agnostic Explanations

Among the numerous existing methods for explaining a model's decision, we chose to focus on so-called local post-hoc and model-agnostic approaches, only relying on a model's output prediction of a single instance, and not its internal representations. Model-agnostic methods are not tied to a specific model type and can be applied to any model. Since we aim to study explainability across a broad spectrum of models, this condition is important to us. Local methods, on the other hand, focus on individual predictions rather than attempting to explain the entire model's behavior. This allows us to gain insights on a case-by-case basis, which can be particularly useful when dealing with complex, non-linear models and tasks where global explanations might be challenging to provide or even misleading. Given our task, where the interactions between words can have a significant impact on the output, we believe it is more insightful to consider each document

individually. Among the most popular approaches in these categories, LIME (Ribeiro et al., 2016), Anchor (Ribeiro et al., 2018) or SHAP (Lundberg and Lee, 2017) rely on lexical features when applied to textual tasks, looking for relevant subsets of features or using perturbations by removing/switching words.

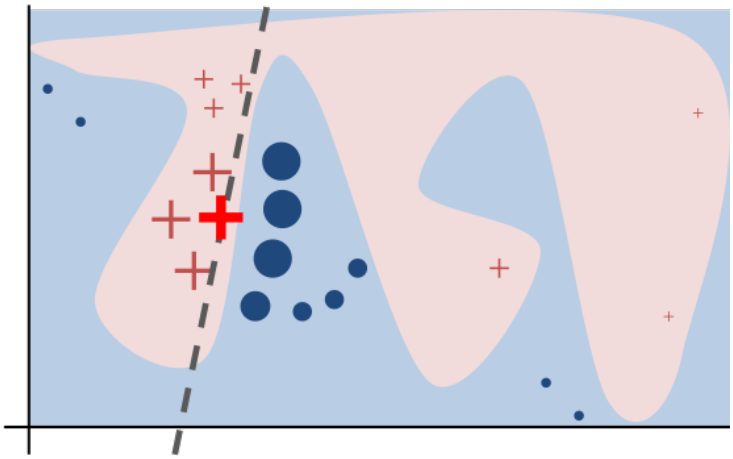


Figure 4.1: Illustration of the LIME method to generating explanations. The figure shows a two-dimensional representation of a complex decision boundary learned by a black box model, with an instance (the large red plus sign) for which we want an explanation of the model's prediction. The blue and red background areas represent the model's positive and negative prediction regions, respectively. To generate an explanation for the instance, LIME samples data around it (small crosses and circles) and weights them by their proximity to the instance (indicated by the size of the markers). LIME then fits an interpretable model (the straight dashed line) to these weighted samples, approximating the decision boundary of the black box model locally. The resulting interpretable model provides an explanation of the model's decision for the instance, showing how each feature contributes to the prediction. (Figure taken from Ribeiro et al.'s Figure 3.)

Of these methods we chose to focus on Local Interpretable Model-Agnostic Explanations, LIME (Ribeiro et al., 2016), which is a perturbation-based approach and has been shown by Atanasova et al. (2020) to have the best or near-best performance on their metrics among popular model-agnostic explanation methods while being easy to implement, making it a suitable approach for our study. LIME operates by approximating the decision boundary of the underlying model for individual instances, and then providing an explanation in terms of interpretable features, in particular words. Given a single document for which we want to generate an explanation, LIME creates a perturbed dataset by sampling perturbed instances, which are generated by removing a random subset of words from the original text, and for which it gets the model's output prediction. The number of perturbed samples to be generated is a hyperparameter in the LIME approach. Care must be taken when choosing this value, we need enough samples to create a faithful local approximation, but we must also consider the computational cost of generating these

samples and calculating predictions for them. To illustrate, consider a simple sentence, "I love this movie". A perturbed version might exclude the words "love" and "this", resulting in the sentence, "I movie". This process is repeated many times to create a dataset of perturbed instances and their corresponding model outputs. Next, LIME uses this new dataset to train an interpretable model, usually a linear model due to its inherent interpretability. This local surrogate model is trained to mimic the original model's behavior in the neighborhood of the instance under consideration. Note that this surrogate model does not aim to be a globally accurate approximation of the original model. Its objective is to capture the original model's decisions accurately around the instance we aim to explain. The coefficients of the trained linear model then serve as the explanation, indicating how much each feature (word, in the case of text data) contributes to the prediction for the selected document. We should note here that when dealing with multiclass tasks, LIME is restricted to explaining one class at a time, but we can run the LIME explanation process for each class separately if needed.

To formalize this process, let's denote the document to be explained as $\boldsymbol{x}$ and the original model's prediction function as f. For each perturbed instance, denoted $\boldsymbol{x'}$, we use a weight function w, which measures the proximity of $\boldsymbol{x'}$ to $\boldsymbol{x}$. A popular choice for w is the exponential kernel, defined as:

$$w(\boldsymbol{x'}) = \exp(-\frac{\text{distance}(\boldsymbol{x}, \boldsymbol{x'})^2}{\boldsymbol{\sigma}^2}) \tag{4.1}$$

where $\text{distance}(\boldsymbol{x}, \boldsymbol{x'})$ represents the distance between the original and perturbed instance (often computed using cosine similarity for text data), and $\boldsymbol{\sigma}$ is a parameter controlling the width of the kernel (Ribeiro et al., 2016). Once the weights are computed, a weighted linear model is trained on the perturbed dataset using the weights $w(\boldsymbol{x'})$ and the original model's predictions $f(\boldsymbol{x'})$ as targets. Let's denote the prediction function of this linear model as g. The objective of the training process is to minimize the following loss function:

$$\mathcal{L}(f, g, w) = \sum_{\boldsymbol{x'}} w(\boldsymbol{x'})(f(\boldsymbol{x'}) - g(\boldsymbol{x'}))^2 \tag{4.2}$$

This loss function encourages the surrogate model g to fit closely to the original model's predictions f for instances that are close to $\boldsymbol{x}$, while less emphasis is placed on instances that are further away from $\boldsymbol{x}$ (Ribeiro et al., 2016). Once trained, the coefficients of the linear model are used to explain the prediction made by the original model for the instance $\boldsymbol{x}$, each coefficient corresponding to the presence or absence of a specific word in the instance.

Thus, these coefficients are used to quantify the impact of each word on the output, and the explanation consists of the subset of most impactful words. An illustration of the LIME process is given in Figure 4.1 and an example of an explanation generated is given in Figure 4.2.

Let's consider a specific example to demonstrate the process. Suppose we have a sentiment analysis model that classifies text reviews into positive and negative. We select a particular review, say, "The plot was exciting, and the characters were believable". We want to understand why our model has classified this review as positive. LIME would begin by generating a number of perturbed versions of this review, such as "The plot was exciting," and "the characters were believable". The sentiment analysis model would then classify these perturbed reviews, and these predictions would serve as targets for the training of the local surrogate model. The weights for each perturbed review would be computed based on its similarity to the original review, with more similar reviews getting higher weights. Once the local surrogate model is trained, the coefficients would reveal how much each word in the review contributes to the "positive" prediction. For instance, the words "exciting" and "believable" might have large positive coefficients, suggesting that they are key contributors to the positive sentiment.

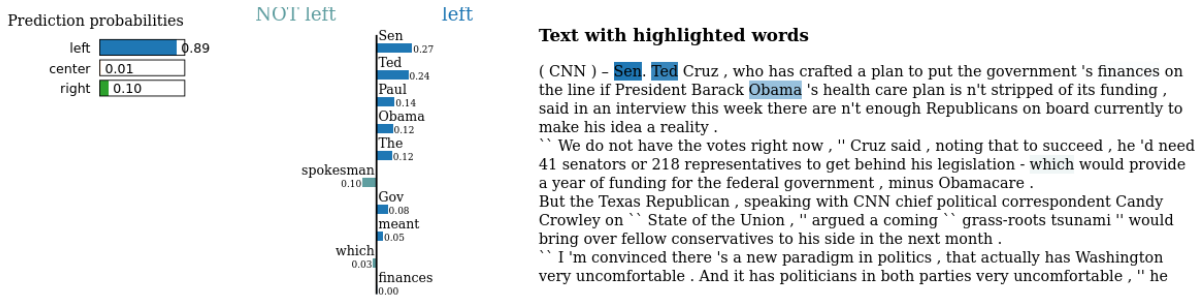


Figure 4.2: Example of an explanation obtained with the LIME implementation[1] proposed by Ribeiro et al. (2016). The explanation is generated for an *Allsides* article and for the "left" class of the experiment presented in Section 3.5 on the prediction of political bias. Here, the predicted class is "left" and the 10 most important words are highlighted and ranked according to the coefficients given by LIME. The explanation also includes words contributing negatively to the predicted class (NOT left).

However, LIME has its limitations. One such limitation is that the quality of the explanations relies heavily on the perturbation process. Perturbing textual data is not straightforward, and the way in which the perturbations are generated can have a significant impact on the explanations. For instance, in the case of text, if the perturbations result in sentences that do not make sense or are ungrammatical, the resulting explanation might not be reliable or meaningful. Therefore, the perturbation process needs to be handled carefully to ensure that the explanations are trustworthy (Ribeiro et al., 2016).

An alternative to perturbing the input by removing words would be to replace words with similar ones rather than deleting them, in order to preserve a syntactically correct sentence, but this also makes the process much more expensive (Ribeiro et al., 2018). Moreover, the explanations provided by LIME are inherently local and do not provide an understanding of the model's overall behavior. Although local explanations can offer valuable insights, they can also be potentially misleading if interpreted as a global explanation. In addition, because the surrogate model is only an approximation of the true model, there is no guarantee that the explanation provided by the surrogate model is entirely accurate or faithful to the true model's decision process. Thus, while LIME provides a practical tool for explaining individual predictions, the interpretations it provides should be considered as approximations rather than definitive explanations. Despite these limitations, LIME serves as a powerful tool for understanding model decisions (Hase and Bansal, 2020; Atanasova et al., 2020).

4.2.2 Limitations and Long Documents Explanations

Another important limitation of LIME is the cost of the sampling process (Molnar, 2022). The main issue is that the quality of the explanations highly depends on the amount of generated perturbed samples, to be representative of the model's behavior, and to avoid spurious or not robust explanations. For texts, where features are words, this can mean a high computational cost, especially for long documents where an increase in document length translates to an increase in the possible perturbations, which can affect the explanation's quality if not enough perturbed samples are generated. The more samples LIME generates, the closer the model can approximate the true decision boundary of the underlying model being explained, and thus, the more accurate the explanations. As the length of the document increases, the potential perturbations that can be generated from the document also increase, leading to a larger sample space. Given a document of $\boldsymbol{n}$ features, there are $2^{\boldsymbol{n}}$ potential binary vectors that represent possible perturbations of this document. Therefore, with long documents, there are exponentially many possible perturbations, meaning that the model would need to generate a similarly vast number of samples to approximate the decision boundary accurately. If too few samples are used: (i) the local model may not accurately reflect the behavior of the model, leading to explanations that are less reliable, and (ii) the local model may be more prone to overfitting and have higher variance. On the other hand, using a larger number of samples can help to reduce the variance of the local model, leading to explanations that are more stable and reliable. It brings us to the concept of "sampling density", which essentially refers to the number of samples generated for a given input. To accurately approximate the true decision boundary of a complex model, a high sampling density is required. However, as the document's length increases, achieving a high sampling density becomes

computationally expensive and potentially intractable. In general, it is a trade-off between quality and computational cost of the explanation. When dealing with long documents, such as the newspaper articles in our political bias prediction task (see Table 3.1), it would in practice require too many samples to generate quality explanations for these documents using this method, resulting in the time needed to generate explanations being disproportionate. On the other hand, for shorter texts, the model's behavior is likely to be more straightforward and less variable, so a smaller number of perturbed samples may be sufficient to capture the model's behavior accurately. While perturbation-based explanation methods, including LIME, can provide useful insights into model behavior, they do not offer a definitive understanding of why a model makes particular predictions. They show the direction of the influence of features, but do not necessarily explain the underlying causes.

Given these limitations, it's clear that while LIME and perturbation-based explanation methods offer valuable tools for understanding complex models, they are not without their shortcomings. Perturbation-based methods like LIME provide an accessible way to generate explanations for complex models. However, these methods face significant challenges in terms of accuracy and stability, particularly when applied to long documents. The generation of samples, which forms the foundation of these techniques, can become computationally expensive with an increase in the length of the input document. Furthermore, the instability of the generated explanations when too few samples are used for long documents reduces their reliability.

4.3 Lexical and Structural Perturbation-Based Explanations

Given the limitations of perturbation-based approaches, such as LIME, with respect to computational cost for generating explanations, especially for long documents, we propose several strategies to generate higher quality explanations while reducing computational cost by focusing on different levels of granularity. Figure 4.3 shows our complete explanation system based on LIME.

Word-level explanations The first level still operates at the word level as the original LIME approach for text by removing tokens randomly, but focusing on specific words. By generating explanations at the word level on the whole vocabulary, we are highly constrained by the length of the input as the size of the perturbation space, from which we sample, grows exponentially with it. In order to mitigate this constraint, we propose the exploration of subspaces of interest. By reducing the size of the space being sampled, we

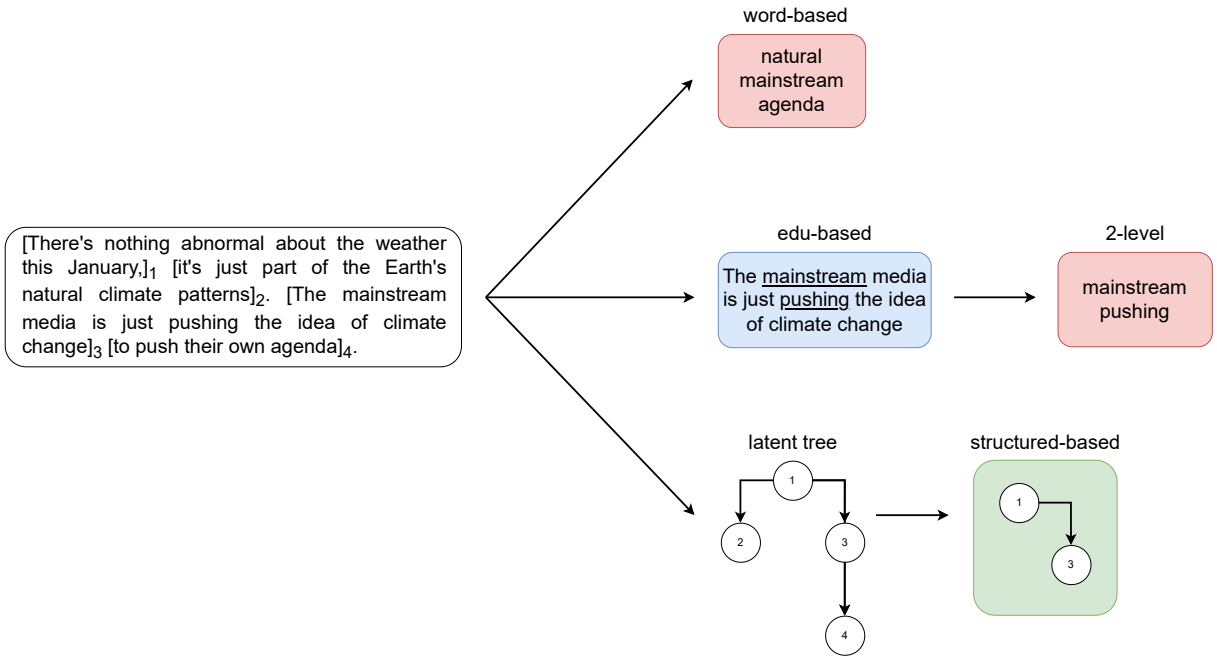


Figure 4.3: Fabricated example of explanations generated from LIME with the different strategies we have proposed based on different levels of granularity (words in red, EDUs in blue, structure in green). Structure-based explanations need the structure produced by the latent structured attention model (see Section 2.3.2). Numbers in the structure refers to EDUs.

can restrict the number of samples needed and thus generate better quality explanations at a reduced computational cost. We consider two subcases:

- **Ignoring stopwords**: The first rather obvious strategy is to ignore stopwords in the perturbation process. Stopwords are words that are commonly used in a language, but are not very meaningful and don't convey much information. The distribution of stopwords in a text follows a pattern known as Zipf's law, which states that the frequency of a word in a text is inversely proportional to its rank in the frequency table. Thus, stopwords are often over-represented in texts compared to other words, which makes them more prone to be chosen for perturbation, and thus to be found in the explanation. By excluding stopwords, the sampling process can focus on the more meaningful and informative words, which can help to reduce the number of samples needed to generate accurate explanations. We relied on the exhaustive list of common English stopwords established by the NLTK toolkit (Bird et al., 2009).

- **Focusing on specific classes of words**: Following the same principle, we can target specific vocabularies of interest and greatly reduce the sampling space by ignoring words that do not fall into these categories. Two specific vocabularies of interest are considered:

- **Named Entities**: Named entities refer to real-world objects such as persons, places, organizations, or any other specific information that can be denoted with a proper name. Named entities can significantly contribute to understanding a document's bias, in particular for detecting political bias, thereby affecting classification decisions (Li and Goldwasser, 2021). We employ spaCy,[2] a widely-used Python library, for the extraction of these entities.
- **Discourse connectives**: Discourse connectives (Webber et al., 2019) are words or phrases that exhibit the relationship between preceding and following clauses. These markers could act as shallow indicators of argumentative structures, thereby subtly influencing the model's decision-making process. We identify these discourse connectives by leveraging an extended list of 173 markers[3] proposed by Sileo et al. (2019).

EDU/Sentence-level explanations The next strategy moves beyond individual words to focus on a higher granularity: either sentences, or EDUs to take into account the general organization of the document. The process for generating explanations is then very similar to word-based ones: instead of perturbing a document by removing a random set of words, we remove a random set of EDUs/sentences. An EDU/sentence-based explanation then consists of a subset of the most impactful EDUs/sentences for the model. Given the higher granularity, this reduces drastically the perturbation space, making it more feasible and reliable to sample. However, it is important to note that using a higher granularity level for the explanation can also have some potential drawbacks. For example, although these explanations are more comprehensive than a simple list of words, higher granularity explanations are less precise and a specific local bias may be difficult to identify, making their interpretation more complex.

- **Elementary Discourse Units**: We propose to move from word-based explanations to discourse-oriented explanations of higher granularity. In the same way as the approach proposed for the classification model, we change the relevant textual units from words to EDUs, as given by a discourse segmenter (Kamaladdini Ezzabady et al., 2021). EDUs are supposed to be the atomic level of structure analysis, and thus more coherent in terms of size and content than full sentences. Using EDU-based explanations is appropriate in the context of documents containing rhetorical and argumentative processes such as articles or dialogues, allowing to capture the structure of the text more efficiently in the explanation, such as cause and effect or contrast, while being easier to understand and communicate. Furthermore, EDU-based explanations are efficient to generate and less computationally expensive as it

[2] https://spacy.io/
[3] https://github.com/sileod/Discovery/blob/master/data/markers_list.txt

drastically reduces the perturbation space with only a few dozen EDUs versus several hundred words in long texts, allowing to increase the quality of the explanations with a lower number of sample.

- **Sentences**: Sentences provide a natural division of text, encapsulating a coherent idea or piece of information, but are less specific than EDUs (see Section 2.5). For sentence boundary detection as preprocessing, we used Stanza, a Python NLP package (Qi et al., 2020).

Two-level explanations While a higher level of granularity such as EDUs may offer broad insights, it may lack the detail required for comprehensive explanations. Thus, we propose a two-level explanation strategy that combines EDU-based explanations with the classical word-based approach. In this two-stage process, we first generate EDU-based explanations and then further refine these explanations by generating word-level perturbations for words that belong to the k most impactful EDUs, k being a hyperparameter. This method, therefore, offers a balance between detail and broad analysis, enhancing the robustness of the explanations of both EDUs and words at a reduced computational cost. Indeed, the explanation at EDU-level makes it possible to generate quality explanations at a reduced cost, and by reducing the perturbation space of the explanation at the word-level to only the words from the most impactful EDUs, we also reduce the number of samples required for the word-level explanation.

Structure-level explanations Finally, we propose to generate explanations directly at the level of the structure learned by the latent structured attention model we introduced (see Section 3.3). This level of explanation is therefore specific to the structured attention approach we have considered in this work. It should be noted, however, that the same principle could be applied to any structure-based model. Here, we will perturb the entire structure extracted *via* the latent model for a given example. We chose to rely on perturbations that remove a subset of head-dependent relations in the original tree, i.e. a pair of segments. An explanation of the structure is then the subset of the most impactful relations in the tree.

Discourse relation classification We propose to augment the structure-level explanation by predicting the discourse relations between the EDUs in the tree. Discourse relations connect EDUs, conveying how different parts of the text build upon each other to form a coherent discourse. We rely on a set of 21 relations defined by the Rhetorical Structure Theory (Mann and Thompson, 1988) and use the system proposed by the DISRPT2021 (Zeldes et al., 2021) shared task winner for relations classification (Gessler et al., 2021). Note that the prediction of discourse relations is a hard task for which existing models are

still relatively poorly performing and the results obtained should be taken with caution, especially given that the relations to be predicted are not relations derived from gold discourse trees. The system we used obtained an accuracy score of 66.76 on this task. Figure 4.4 gives an example of an explanation at the structure level with the prediction of discourse relations in the tree.

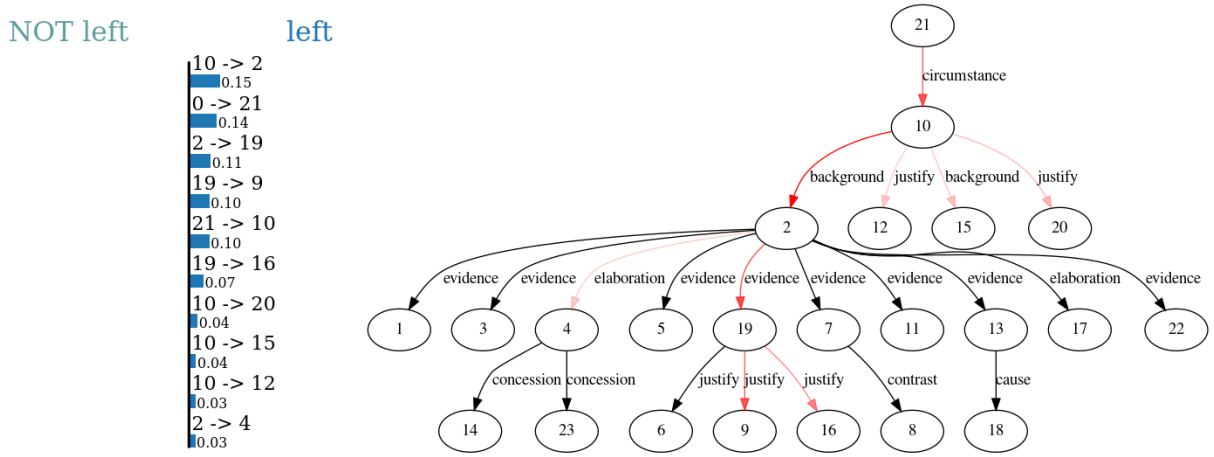


Figure 4.4: Example of a structure-level explanation with predicted discourse relations for the task of predicting political bias on a newspaper article. On the left is the explanation of the structure, with the heads → dependent relations of the tree that have the greatest impact on the prediction of the model for the "left" class. The right-hand side shows the extracted latent structure, to which the predicted discourse relations have been added. The red arrows correspond to the relations that have the greatest impact on the explanation.

By combining all these strategies, we can generate enhanced explanations that cover multiple facets of the data, thereby making the process of explanation generation more efficient and insightful. The strategies proposed here introduce a new way to handle the challenges of explanation generation for long documents in perturbation-based approaches and, in particular, LIME by focusing on different levels of granularity and incorporating both lexical and structural information.

4.4 Evaluating Explanations

Evaluating explanations is a critical and challenging task. Without a reliable method to assess the quality and accuracy of explanations, we cannot confidently use these explanations to understand our models. Therefore, it is crucial to provide quantitative evaluations of explanations, which can offer objective measures of their quality and usefulness. When it comes to evaluating explanations, the lack of standardized and widely accepted evaluation measures poses a significant challenge. Currently, common practices heavily rely on costly human judgments, which can be time-consuming and resource-

intensive (Lertvittayakumjorn and Toni, 2019; Narayanan et al., 2018). This section will address the problem of evaluating explanations, emphasizing the lack of adequate evaluation methods in current work and the importance of providing quantitative evaluations of explanations. We will introduce the method chosen for evaluating our explanations, which is based on the diagnostic properties proposed by Atanasova et al. (2020). There are few, if any, comparable studies in NLP that propose an automatic evaluation of explanations that is not based on human annotations, which motivated our interest in this approach. The authors introduced multiple metrics to evaluate explanations in the context of text classification, each targeting a different aspect of the explanation. More specifically, we are relying on three of these metrics that they have proposed for evaluating explanations.

We consider that a document is composed of a set of features, and that our explanation method generates a saliency score for each of them. Let $X = \{(\boldsymbol{x_i}, \boldsymbol{y_i})|i \in [1, N]\}$, the evaluation dataset of size N with $\boldsymbol{x_i} = \boldsymbol{w_1}...\boldsymbol{w_n}$ a document containing n features $\boldsymbol{w}$ (e.g., word, sentence, EDU, pair of segments) and $\boldsymbol{y_i}$ the predicted label. Similarly to Atanasova et al. (2020), we define $\boldsymbol{\omega}_{x_{i,j},c}$ the saliency scores of the feature $\boldsymbol{w_j}$ of $\boldsymbol{x_i}$ for the class c given by our explanation technique. We will now detail each of the metrics under consideration, each of which addresses different desirable properties of the explanations.

Confidence Indication (CI) When generating an explanation, the feature scores for each possible class can be computed. It is then expected that the feature scores for the predicted class will be significantly higher than those of the other classes. If not, this should indicate that the model is not highly confident in its prediction, and the probability of the predicted class should be low. We can then measure a confidence indication score as the predictive power of the explanation for the confidence of the model. Predicted confidence is computed from the distance between saliency scores of the different classes and then compared to actual confidence by using the Mean Absolute Error (MAE).

$$SD = \sum_{j \in [0,|x|]} D(\boldsymbol{\omega}_{x_{i,j,k}}, \boldsymbol{\omega}_{x_{i,j,K/k}}) \tag{4.3}$$

$$MAE(\boldsymbol{\omega}, X) = \sum_{i \in [1,N]} |\boldsymbol{p}_{i,k} - LR(SD)| \tag{4.4}$$

With the SD, Saliency Distance between the saliency scores, D the subtraction of the saliency value between class k and the other classes, $LR(SD)$ the predicted confidence of the class using logistic regression (LR) on SD, and $\boldsymbol{p}_{i,k}$ the output probability (confidence) of instance $\boldsymbol{x}_i$ for class k. Predicted and actual confidence are then compared by computing

the Mean Absolute Error (MAE).

Faithfulness Faithfulness is an indication that features selected in an explanation were actually useful for the model to make a prediction. It is measured by the drop in the model's performance when a percentage of the most salient features in the explanation are masked. Starting from 0%, 10%, up to 100%, we obtain the performance of the model for different thresholds. From these scores, the faithfulness is then measured by computing the area under the threshold-performance curve (AUC-TP).

Dataset Consistency (DC) DC measures if an explanation is consistent across instances of a dataset. Two instances similar in their features should receive similar explanations. Similarity between instances is obtained by comparing their activation maps, and similarity between explanations is the difference between their saliency scores. The consistency score is then the Spearman's correlation ρ between the two similarity scores. The overall dataset consistency is the average obtained for all the sampled instance pairs.

By adopting these metrics, we aim to provide a comprehensive evaluation of the explanations generated by our proposed method in order to control their effectiveness and compare them with baseline methods. However, it is essential to acknowledge the limitations of this approach. One limitation of the diagnostic approach is that it requires a sufficient number of instances in the evaluation dataset to obtain reliable estimates of the evaluation metrics. Furthermore, while the diagnostic approach provides valuable insights into various aspects of explanation quality, it may not cover all possible dimensions of explanation evaluation. Therefore, it is crucial to consider other complementary evaluation methods and explore potential improvements in the future.

4.5 Experiment – Political Bias Characterization in News Articles

We now turn to the experimental study on the characterization of political bias in newspaper articles, following on from the classification task introduced in Section 3.1. For each of the explanation strategies introduced in Section 4.3, we generate explanations for articles from the evaluation datasets of the three datasets introduced in Section 3.5 for the classification task. These explanations are generated for the model we have proposed based on latent structured attention (SA/EDU) but also for the POLITICS model (which gave the best results on all the classification tasks). For the explanations, we compare to the original version of LIME for text classification, which is based on words perturbation,

and a random explanation on the whole input.

We built on the LIME python package[4] to implement our methods (Section 4.3). We generate and evaluate explanations on 100 documents from the test set of each dataset and for 1,000 and 10,000 perturbed samples. We evaluate the different explanation methods with the metrics introduced in Section 4.4 for the structured attention model on the *Allsides* dataset. The confidence interval for the evaluation of the explanations is only given for the baseline (LIME Words) for 10 generations. Since each of the proposed improvements has a reduced perturbation space relative to the baseline, which is the impact factor of the variance, and to avoid a disproportionate computational cost, we consider that the confidence interval will be at worst equal or better, and therefore we do not give it for all explanation strategies.

Explainability technique	CI MAE ↓	F AUC-TP ↓	DC ρ ↑
Random explanation	0.053	47.45	0.010
base LIME (words)	0.036	45.78	−0.003
EDUs	**0.029**	38.80	0.075
Sentences	0.034	37.90	0.014
Structure	0.038	36.00	0.065
2-level EDUs+Words	0.034	36.40	0.131
Words w/o Stopwords	0.031	44.80	0.045
Discourse Markers	0.032	43.14	0.119
Named Entities	0.033	**35.25**	**0.176**

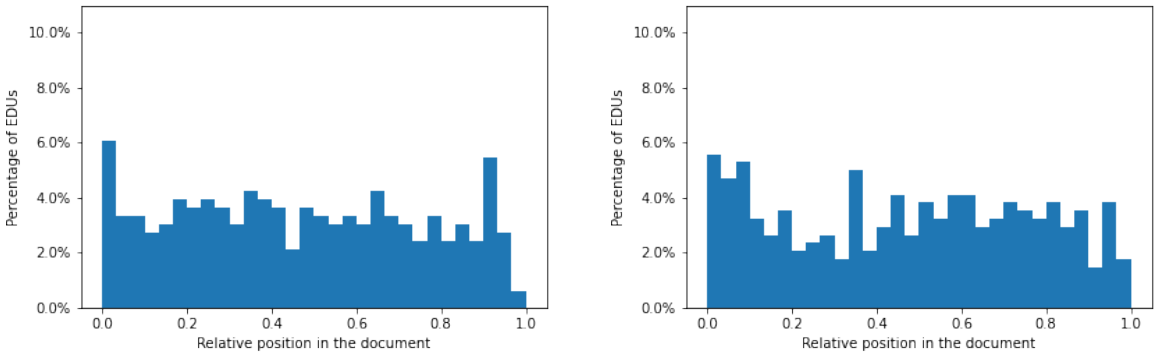
Table 4.2: Confidence Indication (CI), Faithfulness (F) and Dataset Consistency (DC) scores for the different strategies described in Section 4.3, on the *Allsides* dataset. For each document, 10,000 perturbed samples are generated. For "LIME Words", the standard deviation is ±0.002 for Confidence Indication, ±2.2 for Faithfulness, and the estimated p-value for the correlation of Dataset Consistency is 0.002.

Table 4.2 presents the evaluation metrics for each of the proposed LIME alternatives. We observe that in general, except for discourse markers and named entities, the two-level explanation performs better, obtaining strong evaluation scores for all the proposed metrics. The use of a higher level of granularity (sentences, EDUs) improves the quality of the explanations compared to the baseline; note that between EDUs and sentences, the finer segmentation into EDUs is the most accurate, further demonstrating the effectiveness of the discourse-driven approach. The higher CI score for EDUs shows that it is the appropriate level of granularity with respect to the impact of their content on the model decision, it is also the level of segmentation on which the model has been trained. Similarly,

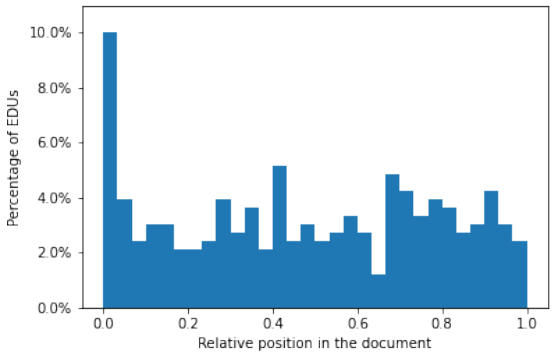
[4] https://github.com/marcotcr/lime

reducing the perturbation space by targeting classes of words generates better quality explanations, in particular for named entities, which are particularly informative for the model as already shown in the literature (Li and Goldwasser, 2021). Regarding the explanation of the structure, although the scores obtained are in the low range, we can state that they represent relevant information for the decision of the model as compared to baselines. In general, the two-level explanation seems to be the best compromise between explanation quality, computational cost, and level of detail, while the LIME baseline (words) suffers from a high perturbation space. As we are reducing the sampling space in our proposed approaches, we also made comparisons on the number of samples used to generate the explanation for these metrics, between $1,000$ and $10,000$ samples. We notice that the scores obtained by most of our approaches on $1,000$ samples remain better than those of the baseline for $10,000$ samples. This shows that it is possible to generate good explanations, and often of better quality, with a number of samples 10 times smaller, which is a major improvement over the computational cost.

By looking at the explanations generated for the different levels of granularity and properties targeted, we can gain some insights about the model's decisions. An important property that must be fulfilled by the explanation is its comprehensibility by humans, in order to characterize biases. We propose a qualitative analysis of the explanations and a comparison of the various approaches, both at the lexical and structural levels. Tables 4.3, 4.4 and 4.5 show the most recurrent and impactful words in the explanations, as given by the aggregated saliency scores of the 100 generated explanations, for each class for the *Allsides*, *C-POLITICS* and *Hyperpartisan* tasks respectively, and for each explanation method. Overall, the words that emerge seem consistent with the classes, and it is relatively straightforward to understand the possible biases that characterize them. Regarding the differences between word-based explanation approaches, we observe that two-level explanations yields more relevant information and specific lexical cues (e.g. *environmental, transgender, scientists, archbishops*), which confirms the interest of a first pass through an adapted level of granularity in order to target the most interesting parts of the text. Explanations based on discourse markers or named entities show overlap with the other methods, indicating consistency between approaches. EDU-based explanations are more comprehensive and self-sufficient, while covering information contained in word-based explanations. This seems to make it an appropriate compromise between human readability and computational cost. Furthermore, if we look at the relative position in the text of the most impactful EDUs in the explanations (Figure 4.5), we can note that while there is a slight tendency for impacting EDUs to appear at the beginning of the text, most impacting EDUs are located evenly throughout the text, which confirms the interest of keeping the entire document. Interestingly, the explanations generated for the SA/EDU



(a) *Left* (b) *Right*



(c) *Center*

Figure 4.5: Distribution of relative positions of the most impactful EDUs (as given by the explanations) for the left, center and right classes (*Allsides* dataset).

model and POLITICS show little overlap for the same dataset, with different features highlighted in the explanations. Despite the fact that POLITICS performed better on all tasks, this difference can be explained by the massive pre-training of POLITICS on newspaper articles with an ideology-driven pre-training objective, during which the model was able to learn to associate certain patterns with certain political orientations, and from which the SA/EDU model did not benefit.

Regarding structure explanations, we observe that the most impactful relationships are mainly located in the first levels of the tree, close to the root, independently of class and dataset. This suggests that the structured attention model seems to push the most important information up to the highest levels of the tree (close to the root) and less relevant information down to the leaves. A similar result is observed in work using the structured attention approach for summarization tasks, where the root of the tree corresponds to the summary returned for the document (Isonuma et al., 2019). While structure explanations (the most impactful relationships in the structure induced by the model) are more difficult for a human to interpret, the discourse relationships predicted on the tree for each relation (see Section 4.3) allow us to gain some insights into these explanations

and how bias manifests itself in the structure between different classes. Therefore, we aggregate the saliency scores of the most impactful relations according to their predicted discourse relations, as given by the explanations of the induced structures. In this way, we get a list of the most impactful discourse relations in the induced structures for each class.

We now take a closer look at the explanations generated for each of the datasets considered, and for both the POLITICS and structured attention (SA/EDU) models. By comparing the results of the different levels of explanation between the classes for each dataset, and without entering into political considerations, we can establish a first diagnosis of the biases that characterize them.

Allsides From the word-based explanations of the SA/EDU model, we observe a shift in the lexical fields between classes (*pacific, aids, percent – transgender, environmental, scientists – fired, surveillance, archbishops*), which indicates a bias in topics covered and in the way information is conveyed. Articles from the right class seem to favor negative-sounding terms, while the pitch used is more neutral for the center and left classes. We can also note the over-representation of public and political figures in the explanations, which is distinguished between each class by the political leaning and the social category of the people being mentioned. In particular, we notice that articles from the right are almost exclusively mentioning personalities from their side, with the specificity of recurrently referring to religious figures (e.g. *John Sentamu, Jerry Falwell*). While the profiles are more diversified for the left and center classes, giving a lot of attention to right-wing personalities. About discourse markers, three trends can be identified from each of the classes. The left class seems to prefer markers of certainty or uncertainty (e.g. *absolutely, maybe*), the center class focuses on markers indicating time or frequency (e.g. *then, already, frequently*), while the right class favors markers that indicate contrast or emphasis (e.g. *though, however, obviously, naturally*). Turning to the explanations of structure, in terms of discourse relations, the left's emphasis on *evidence* and *justify* indicates a possible reliance on providing proofs for claims. The center's emphasis on *preparation* and *elaboration* perhaps underlines a structured and detailed argumentation style, while the right's focus on *concession* suggests an acknowledgement of counter-arguments, but also possibly refuting them. Regarding the explanations of the POLITICS model for the same dataset, which has obtained +6% accuracy compared to SA/EDU on this dataset, we observe significant differences in the features selected compared to the explanations generated for the SA/EDU model. For the word-based explanations, the POLITICS model paints a slightly different picture, with terms such as "republican", "interview", and "conservative" emerging for the left class, possibly indicating a focus on political

Explainability technique	Left	Center	Right
LIME Words SA/EDU	obama, pacific, brass, mccain, barack, after, percent, donald, aids, with	trump, donald, continued, washington, said, ginsburg, iran, options, this, china	scalise, garnering, heard, that, anti-muslim, only, fired, president, media, surveillance
LIME Words POLITICS	trump, republican, interview, said, diary, anonymity, unlikely, international, conservative, clinton	said, aides, claim, protect, comment, fraud, some, ukrainian, might, weather	reporting, president, accused, downfall, liberal, coronavirus, muslim, mccabe, billionaire, collusion
EDUs SA/EDU	"when mainstream columnists start using words like aristocracy and kleptocracy"	"according to the american psychiatric association, not all transgender individuals suffer from gender dysphoria."	"because Stossel had done the shovel work (*cough*) of introducing fundamental concepts and breaking in nerds."
EDUs POLITICS	"some of that concern stemmed from the rise of right wing media and blogs."	"but to undermine a president..."	"this is an american disgrace!"
2-level EDUs+Words SA/EDU	media, percent, barack, columnist, worse, contrarian, sundays, interested, nationwide, watching	trump, twitter, dysphoria, manafort, donald, gender, environmental, transgender, scientists, ginsburg	stossel, scalise, president, cohen, sentamu, disgusting, nobody, media, archbishops, garnering
2-level EDUs+Words POLITICS	trump, republican, conservative, climate, right, denies, comparisons, lawmaker, nuclear, documents	said, federal, probe, guilty, correct, former, loan, examination, banks, investigation	president, leftist, liberal, accused, declared, muslim, injunctions, terror, accusation, coronavirus
Discourse Markers SA/EDU	absolutely, surely, lately, only, maybe	then, perhaps, already, frequently, still	here, though, however, obviously, naturally
Discourse Markers POLITICS	really, actually, often, meanwhile, second	first, only, however, then, later	also, previously, further, instead, immediately
Named Entities SA/EDU	Barack Obama, David Pecker, John Mccain, Preet Bharara, Hillary Clinton	Donald Trump, Paul Manafort, Bader Ginsburg, Christopher Wray, Mark Zuckerberg	Steve Scalise, John Sentamu, John Stossel, Jerry Falwell, Michael Cohen
Named Entities POLITICS	Donald Trump, Hillary Clinton, John Mccain, Tony Blair, Ryan White	Ryan White, Paul Manafort, Donald Trump, Christopher Wray, Mark Zuckerberg	Andrew Mccabe, Hillary Clinton, Zac Moffatt, Steve Scalise, John Stossel
Discourse Relations (Struct. Expl.) SA/EDU	restatement, evidence, purpose, justify, cause	preparation, elaboration, attribution, joint, circumstance	preparation, joint, attribution, concession, restatement

Table 4.3: Prototype explanations by class (*Allsides*), ordered from most to least impactful, as given by the highest aggregated saliency scores of the explanations. For context, this dataset includes articles related to various topics (e.g. elections, immigration, coronavirus, politics) published between 2012 and 2020.

ideologies and their interactions. Further, terms like "climate", "lawmaker", and "nuclear" for the left suggest thematic concerns around environmental and legislative issues. For the center class, the word "said" is emphasized, indicating a focus on reporting facts and quoting people. Meanwhile, for the right class the model focuses on words like "leftist", "liberal", "muslim", and "billionaire", with an emphasis on opposing viewpoints, religious contexts, and economic elites. Discourse markers seem less informative for the POLITICS model, with markers that don't seem to reveal any particular trend compared to the results obtained for the SA/EDU model at this level, which would indicate that the discourse-driven approach gives more importance to these words for prediction. In

terms of named entities, the POLITICS model, for the left-wing class, focuses on *Donald Trump* and *Hillary Clinton*, highlighting central figures in the political landscape. The right-wing class mentions *Andrew McCabe* and *Hillary Clinton*, highlighting individuals often criticized in right-wing media.

Explainability technique	**Left**	**Center**	**Right**
LIME Words SA/EDU	disparaging, trump, melania, pitfalls, honors, attacking, authorities, explain, which, surprising	bemoaned, reason, president, irrational, true, accomplishments, republicans, stadium, reeves, participated	president, sweeping, spokesman, chinese, surrounding, doom, lashed, caucuses, nevada, virus
LIME Words POLITICS	senate, donald, republican, escalation, congressional, sanctions, threatens, iranian, curbing, approval	said, physician, president, national, medical, economies, world, reported, ahead, china	presidential, interview, during, candidate, director, diseases, donald, infectious, anonymous, irresponsible
EDUs SA/EDU	"but trump complied,"	"whom republicans have criticized throughout the impeachment process."	"that democrats only increased the support for late-term abortion and abortion on demand."
EDUs POLITICS	"and where will the escalation end?"	"he said."	"but that's been disputed by election experts."
2-level EDUs+Words SA/EDU	contributed, e.g., repeats, replies, stance, explains, nonsense, refusing, disparaging, unhelpful	bemoaned, referencing, said, frequent, abusing, quoting, criticized, impeachment, unlike, legal	america, warn, boom, president, boycott, political, democrats, ideological, lockdown, wuhan
2-level EDUs+Words POLITICS	escalation, sounds, trump, against, forces, evicts, republican, senate, never, baghdad	said, president, tweeted, saying, national, vital, allegations, percent, reported, former	questioning, defended, national, past, respect, comments, institute, prosecutors, statement, candidate
Discourse Markers SA/EDU	honestly, increasingly, evidently, then, surprisingly	also, however, although, obviously, then	meantime, rather, absolutely, also, together
Discourse Markers POLITICS	this, well, certainly, however, otherwise	also, still, nonetheless, specifically, although	still, absolutely, initially, similarly, unfortunately
Named Entities SA/EDU	Donald Trump, Deb Riechmann, Tom Barrett, Joe Biden, Kamala Harris	Tobe Berkovitz, Devin Brosnan, Bernie Sanders, Hunter Biden, Bill Stepien	Pete Buttigieg, Donald Trump, Steven Mnuchin, Robert Unanue, Marsha Blackburn
Named Entities POLITICS	Donald Trump, Susan Paul, Eric Garner, Susan Collins, Nancy Pelosi	Tom Steyer, Joe Biden, Mitch McConnell, Roy Blunt, Devin Brosnan	Hillary Clinton, Bernie Sanders, Mike Pence, Bill Clinton, Steven Mnuchin
Discourse Relations (Struct. Expl.) SA/EDU	restatement, contrast, sequence, preparation, elaboration	joint, elaboration, restatement, contrast, sequence	restatement, elaboration, sequence, evaluation, question

Table 4.4: Prototype explanations by class (C-POLITICS), ordered from most to least impactful, as given by the highest aggregated saliency scores of the explanations. For context, this dataset includes articles related to U.S. politics published between 2020 and 2021.

C-POLITICS When analyzing the word-based explanations given by the SA/EDU model, we observe a distinctive shift in the vocabulary between the three classes. Words such as "disparaging", "trump" and "melania" associated with the left reveal an inclination towards critical assessments, targeting specific personalities, and descriptors such as

"unhelpful" and "nonsense", indicate a tone of criticism. In contrast, the center's lexicon like "bemoaned" and "reason" seems to suggest a more reflective stance. The right, however, presents terms like "doom" and "lashed", hinting at a more dramatic, perhaps confrontational tone and there is a significant focus on entities and ideologies, evident from words like "democrats", and "ideological". Comparatively, regarding the POLITICS model, the words associated with the left class, including "senate" and "escalation", appear to place emphasis on political operations and potential conflicts. For the center, the word "said" is prevalent, which might be suggestive of a more reporting style, emphasizing direct quotes and factual presentations. Interestingly, for the right class, words like "candidate" and "director" emerge, showing a focus on personalities and roles within the political landscape. At the EDU level for the SA/EDU model, the right class focuses on accusations, as shown by "that democrats only increased the support for late-term abortion and abortion on demand." Within the POLITICS model, there seems to be an emphasis on questioning for the left ("and where will the escalation end?") and direct quoting for the center ("he said."). Shifting our focus to discourse markers, the SA/EDU model for the left tends to use markers such as "honestly" and "evidently", indicating a sense of certainty. The center class uses terms like "also" and "however", suggesting a balanced view. In terms of named entities, one interesting observation is that right-wing articles seem to give particular prominence to political figures from the opposite side, with "Hillary Clinton", "Bernie Sanders" and "Pete Buttigieg", who are associated with the left-wing political side, among the personalities who have the most impact in the explanations. Lastly, in terms of discourse relations reflecting structural explanations, the differentiation between classes is subtle, with the left leaning more towards *contrast* and *preparation* and the right focusing on *evaluative* comments and *questions*. To summarize, the SA/EDU model for the *C-POLITICS* dataset appears to offer a more topic-centric and personality-focused approach, while the POLITICS model seems to emphasize political operations, confrontations, and direct reporting.

Hyperpartisan Starting with word-level explanations for the SA/EDU model, we find a pronounced divergence in the terms associated with non-hyperpartisan and hyperpartisan classes. Words such as "reported", "according", "said" and "tweeted" are dominant in the non-hyperpartisan class, emphasizing that non-hyperpartisan content is focused on presenting facts without extreme bias. On the other hand, words like "trump", "reveals", "tyranny", "racist", "treasonous" and "immigrants" dominate the hyperpartisan class. These words could possibly indicate a tendency to sensationalize events or present them with strong opinions or biases. However, a shift is observed when we consider the POLITICS model. Here, "election", "violence", and "party" are more prominent for the non-hyperpartisan class, pointing towards a more political focus, or topics that are

more public and perhaps of larger concern. The hyperpartisan class emphasizes words such as "collusion", "ridiculous", "radical", "racism" and "islamophobia", indicating a more confrontational and possibly ideological stance. Focusing on discourse markers, we observe distinct patterns. The SA/EDU model for the non-hyperpartisan class uses markers like "first", "then", and "eventually", suggesting a more sequential, fact-based narrative structure. However, for the hyperpartisan class, there doesn't seem to be a strong bias towards discourse markers, as no particular pattern emerges. Named entities also don't seem to play a crucial role in this task, as no trend seems to emerge from the explanations of named entities for the different classes and models. Finally, regarding the explanations of structures in terms of discourse relations, the non-hyperpartisan and hyperpartisan classes show relatively few differences. Nonetheless, the hyperpartisan class differs in the prevalence of *question* and *evidence* relations in the explanations, which could be indicative of an interrogative and evidential style of presenting content.

Explainability technique	**Non-hyperpartisan**	**Hyperpartisan**
LIME Words SA/EDU	reported, lewandowski, according, donald, could, news, corey, hustler, unaired, police	trump, reveals, discomfiting, reputation, controversial, hillary, politicians, immigrants, criminals, guns
LIME Words POLITICS	election, violence, party, clinton, race, isis, black, protest, celebrities, trump	collusion, allowed, election, ridiculous, radical, islamophobia, impact, clinton, trump, charges
EDUs SA/EDU	"if the 14,000 hours of unaired 'apprentice' tapes are released."	"it is an evil, oppressive ideology with governmental, judicial, educational, militaristic, and societal aspects to it"
EDUs POLITICS	"to protest gun violence in the u.s."	"of why radical muslims hate us so much?"
2-level EDUs+Words SA/EDU	said, facebook, reported, news, tweeted, lewandowski, donald, weinstein, instagram, media	tyranny, racist, chargeable, abiding, trump, treasonous, shameful, clintons, deserved, reveals
2-level EDUs+Words POLITICS	election, violence, candidates, gunman, race, clinton, republican, journalism, party, twitter	racism, cheating, hate, kremlin, administration, liberal, radical, trump, collusion, ridiculous
Discourse Markers SA/EDU	first, then, eventually, this, recently	then, perhaps, here, again, only
Discourse Markers POLITICS	often, sometimes, mostly, also, perhaps	certainly, apparently, probably, especially, finally
Named Entities SA/EDU	Harvey Weinstein, Nikki Haley, Allie Clifton, Corey Lewandowski, Jake Tapper	Donald Trump, Chrissy Teigen, Hillary Clinton, Mike Pence, Barack Obama
Named Entities POLITICS	Hillary Clinton, Lena Dunham, Harvey Weinstein, Nikki Haley, Jake Tapper	Bernie Sanders, Mike Huckabee, Harvey Weinstein, Rob Goldstone, Jimmy Kimmel
Discourse Relations (Struct. Expl.) SA/EDU	restatement, elaboration, contrast, preparation, sequence	restatement, elaboration, question, contrast, evidence

Table 4.5: Prototype explanations by class (Hyperpartisan), ordered from most to least impactful, as given by the aggregated highest saliency scores of the explanations. For context, this dataset includes articles about political news published between 2012 and 2018.

Same Story Comparison: Trump's Decision on Reopening the U.S. Economy during the Covid Pandemic

The C-POLITICS dataset contains clusters of articles aligned on the same story. We thus have access to articles that cover the same specific story but have different political biases, enabling a more focused analysis and a more relevant comparison of biases. Table 4.6 presents three articles that cover President Trump's decision on reopening the U.S. economy during the Covid pandemic. These articles originate from three sources and have different political leanings: The Washington Post (left-leaning), The Hill (center-leaning), and Breitbart News (right-leaning). For each of these articles, we predict the political bias using the POLITICS model introduced in Section 3.4, and generate the explanations associated with these decisions at different levels of granularity (2-level words, EDUs, named entities, see Section 4.3). Here, the model has predicted the correct label for all articles. The highlights in the table serve as a visual representation of the most important features given by the explanations at different levels. Our analysis will focus on comparing the explanations of these three articles based on the generated explanations.

The Washington Post (Left-leaning) This article[5] takes a clear stand in opposition to President Trump's decisions, by strongly denouncing them. The president's actions are described using negatively connoted terms to which the model seems to attach particular importance, such as "foolish" and "ill-advised". The explanation further underlines the dramatic tone of the article, which emphasizes the gravity of the situation, as shown by the EDU "This is like something out of a dystopian science fiction movie". The highlighting of "Fox News" and "Laura Ingraham's" tweet shows the importance given to criticism of right-wing media for their stance on the issue. The citation of the "study" from Imperial College gives a fact-based argument against Trump's decisions. This is further emphasized by mentioning other countries, like India and Britain, taking opposite steps as Trump.

The Hill (Center-leaning) Here, the article[6] focuses more on Joe Biden's responses to Trump's decisions, emphasizing Biden's disagreement and concern. It provides a more balanced perspective, with highlighted explanations leaning more toward the factual account of events and statements, giving preference to quotes and references, such as "the Democratic presidential candidate said on MSNBC". It primarily revolves around Former Vice President Joe Biden's comments on Trump's decision ("Biden said"), but several perspectives are confronted including health experts like Fauci ("Fauci said").

[5] https://www.washingtonpost.com/opinions/2020/03/25/trump-is-risking-terrible-tragedy-avoid-responsibility-recession/

[6] https://thehill.com/homenews/campaign/489378-biden-hits-trump-remarks-about-reopening-economy-within-weeks-he-should/

This is the most dangerous thing Trump has done yet - *The Washington Post*

If you were a passenger, you would be terrified. That is exactly how I feel as an American when I hear President Trump say he wants the country "opened up and just raring to go by Easter," i.e., April 12. The president has done many foolish and ill-advised things. But until now he has mainly been a threat to our liberties. Now he is threatening our lives. [...] The United States is about to overtake Italy as the country with the largest number of active coronavirus cases. [...] Yet what Trump is hearing from the right-wing echo chamber — and now translating into policy — is that the cure is worse than the disease. As Fox News host Laura Ingraham tweeted: "A global recession would be worse for our people than the Great Depression." [...] This is like something out of a dystopian science fiction movie ("Logan's Run," to be exact): kill our elders so that our children may enjoy a better life. I want to scream: You are not going to sacrifice my older friends and relatives on the altar of the Dow Jones industrial average! But leave aside the profound immorality of this very concept; it is also inherently impractical. [...] A study from Imperial College [...] predicted that, if left unchecked, covid-19 could kill 2.2 million Americans. The study's authors concluded that strict social distancing, along with identifying and quarantining the infected, would be necessary to substantially reduce the toll. That is why India and the Britain, both run by Trump's fellow right-wing populists, have just mandated national lockdowns. [...] In fact, it's highly doubtful that most governors will lift their shutdown orders [...] As Bill Gates said, "It's very tough to say to people, 'Hey, keep going to restaurants, go buy new houses, ignore that pile of bodies over in the corner.'" [...] It will be much harder to enforce even statewide lockdowns if the president is saying it's safe to go back to work. [...]

Biden hits Trump's remarks about reopening economy within weeks: 'He should stop talking' - *The Hill*

Former Vice President Joe Biden on Tuesday denounced President Trump for pushing to reopen the U.S. economy by Easter Sunday, saying that the president needs to "stop talking" and listen to health experts. "I would like to open up the government tomorrow if it were possible," the Democratic presidential candidate said on MSNBC just hours after Trump stated that he hoped to have the country "opened up" by April 12. [...] Health experts, including Fauci, have said that social distancing requirements could be needed for weeks, though Trump this week began floating the idea of reopening businesses, saying such a move could be necessarily to avoid severely damaging the economy. [...] "Look, if you want to ruin the economy for a long time, let's go ahead [...]" he said. "We haven't even flattened the curve. It's frustrating to hear this president speak. He should stop talking. Let the experts speak." He went on to say that the current crisis goes beyond politics and that the U.S. will be able to address economic costs later. He also cited Congress's work to pass a massive stimulus package to provide support for workers impacted by the outbreak as a positive step. "This is about how we spare this nation from a potential disaster," Biden said. Trump said during a Fox News town hall Tuesday that closing down the economy could "destroy" the U.S. and said that he'd love to have it "raring to go by Easter." Speaking at a White House briefing later that day, Fauci said that the timeline for lifting restrictions on businesses and mass gatherings was "very flexible." He said lifting restrictions would not make sense in an area like New York City, which has emerged as the epicenter of the pandemic in the U.S., and noted that the upcoming weeks would be crucial for public health officials to understand how widespread the virus is in the country.

Trump has megaphone, but states control virus shutdowns - *Breitbart News*

President Donald Trump has the biggest megaphone, but it's and local officials who will decide when to begin reopening their economies after shuttering them to try to slow the spread of the coronavirus. [...]
Q. But the president has set a in which all Americans are being urged to drastically scale back their public activities. Doesn't that amount to a national order?
A. No. [...] "When Donald Trump selects a narrative and begins to advance it, especially through his Twitter account, it has a remarkable effect on those who trust him. The more the president speaks against more robust forms of social distancing (such as shelter-in-place rules), the more noncompliance we are likely to see on the ground level from citizens sympathetic to the president," Robert Chesney, a University of Texas law professor wrote on the Lawfare blog.
Q. Still, Trump has invoked some federal laws to address the virus outbreak, hasn't he?
A. Yes, he has. The allows the expenditure of tens of billions of dollars in emergency assistance. The allows the president to direct private companies to produce goods or acquire raw materials. Trump has yet to actually order companies to do anything, over the objection of some local officials who have a desperate need for ventilators, masks and other equipment. [...] "There are real limits on the president and the federal government when it comes to domestic affairs," Berkeley law professor John Yoo said on a recent Federalist Society conference call. [...]
Q. Is it clear that state and local governments have authority to impose the severe restrictions we've seen?
A. Lawsuits already are challenging state actions on religious grounds and as seizures of property for which the government must pay compensation. But for more than 100 years, the Supreme Court has upheld states' robust use of their authority, even when it restricts people's freedoms. In 1905, the court rejected a that he should not be forced to get a smallpox vaccine or pay a fine, Malcolm noted.

Table 4.6: Explanation of the POLITICS model's prediction for articles covering the same story (President Trump reopening the U.S. economy during covid) but with different political leanings (left/center/right), taken from the *C-POLITICS* dataset. Highlighting refers to the most important features of the explanations (at EDU and word level), generated using the methods introduced in Section 4.3 (EDUs, 2-level Words). Yellow highlighting corresponds to named entity explanations. The darker it is, the more relevant it is to the model.

Breitbart News (Right-leaning) Breitbart's article[7] focuses on the legality and powers of the presidency versus state authority. The explanations emphasize the legal aspects of the situation: "Texas law professor", "Lawfare blog", "Berkeley law professor John Yoo said on a recent Federalist Society conference call". The high importance given to the highlighted EDU "to impose the severe restrictions", containing strong words such as "impose" or "severe", underlines a perception of overreach or authoritative actions with regard to the Covid pandemic for the right-wing political class.

While the explanations generated for the *Washington Post* article underscore the dangers posed by reopening the economy too soon and Trump's decisions, explanations for *The Hill* are more centered on the factual aspects of the article, giving preference to quotes and references. *Breitbart*'s explanations, on the other hand, focuses less on the health versus economy debate and more on legal aspects with the division of power between federal and state governments. All three articles also place a significant emphasis on named entities. "Trump", being the primary subject, is highlighted in all the articles. Other entities, such as "Biden", "Fauci", and "Laura Ingraham", are highlighted in the explanations, based on the context and narrative of the respective articles. Whether it's the open criticism from the left, the balanced overview from the center, or the pragmatic, legal approach from the right, each explanation shows distinct perspectives regarding the story under consideration, and a clear difference in the way information is presented.

From this qualitative analysis of the explanations generated at different levels of granularity, we can attest to the effectiveness of the different strategies proposed, particularly in the case of long documents. Using the different levels of granularity, we were able to perform an exhaustive analysis of various dimensions in which textual bias manifests itself, and to obtain valuable insights into the expression of political bias in news articles. However, it must be acknowledged that there is still considerable room for improvement in terms of the relevance and interpretability of the explanations, particularly with regard to the discursive aspects and the structures induced by the model.

4.6 Conclusion and Future Directions

In this final chapter of our exploration, we delved into the characterization of textual biases, and more specifically political bias in news articles. Our aim was not merely to predict, but to understand – to gain insights into underlying processes through which biases permeate written content, with a specific focus on the discursive dimension. We

[7] https://www.breitbart.com/news/trump-has-megaphone-but-states-control-virus-shutdowns/

explored the field of explainability in NLP, which allowed us to generate explanations alongside our model's predictions, offering valuable insights into the "why" of the biases our models identified. We were interested in model-agnostic and perturbation-based explanation methods, particularly the LIME technique, for generating explanations. The inherent limitations of these methods, particularly when dealing with long texts, led us to propose new strategies based on different levels of granularity (i.e. words, sentences, EDUs, structures). Our experimental study of political bias in news articles allowed us both to demonstrate the quality of our explanations using a series of evaluation metrics, and also to provide insights into the predicted biases from a detailed qualitative analysis of the various explanations generated. While our findings are promising, there remains a significant gap between where we are and where we aim to be in terms of truly understanding the underlying biases learned by the model, in particular regarding the discursive processes from the induced structure. Explanations generated at the level of induced structures, despite promising results, are not yet able to fully elucidate the nature of the learned structures and the potential structural biases they may highlight.

Looking ahead, we can envisage several directions for future work. One of the primary concerns remains the computational cost associated with generating explanations. In their current form, and despite the proposed improvements, generating these explanations is resource-intensive, making it challenging to produce them for large datasets in real-time scenarios. A potential direction could involve developing more efficient algorithms or techniques that maintain the quality of explanations while reducing computational costs, such as more efficient perturbation methods. Furthermore, while we have made progress in understanding the model's decisions and textual biases, the generated explanations don't yet offer a clear window into the discursive processes of the trees induced by the structured attention model. Further investigation into how the model understands and exploits the latent discourse-driven structures is essential. Such investigation might reveal more about the nature of the learned structures, how they are related to existing discourse formalisms, and any biases they might carry. Another promising direction would be to expand our approach's scope. While our current work focuses on political biases in news articles, the underlying techniques and methodologies have the potential to be applied to other genres and forms of bias. This could include the analysis of textual biases in social media posts, blogs, and even scientific literature. Lastly, our focus on English-language datasets opens the door to a broader exploration. Different languages, with their unique syntactic and semantic specificities, might exhibit bias differently. Exploring how our approach performs across diverse linguistic landscapes could offer a richer understanding of textual biases. Moreover, cultural differences play a pivotal role in how biases are perceived and manifested. Extending our analysis to various cultural contexts could lead

to a more global understanding of textual biases.

In conclusion, while this chapter has taken a significant step forward in characterizing textual biases, especially political biases in news articles, the journey is far from over. Ensuring that our systems are not just accurate, but also transparent, trustworthy, and insightful is of crucial importance.

General Conclusion

As we draw this research to a close, it becomes crucial to reflect upon the journey undertaken, the methodologies employed, the insights gained and the futur perspectives. Exploring textual bias, particularly in an information-based society, is of considerable importance. In a world where opinions are influenced by a variety of sources, understanding and quantifying textual bias can lead to a better informed and a more balanced society. By focusing on political bias in news articles, and through the automatic identification and characterization of textual bias, we aim to contribute to the large body of work on this topic with the objective to move towards a more transparent and democratic sharing of information. While previous studies have focused mainly on lexical analysis, we propose to integrate argumentative and rhetorical dimensions by considering the discourse structure of the documents. We thus introduced a discourse-driven model for the prediction of textual bias, based on structured attention networks and EDUs, that latently induces a structure over the document.

We demonstrated the effectiveness of our proposed discourse-driven approach on a series of experiments on the prediction of political bias in news articles. The results not only confirmed the effectiveness of our approach, but also led us to several key findings. Among our most important findings, we showed that, in general, segmenting documents into EDUs rather than sentences is a more appropriate level of granularity for analyzing the structure of documents. Regarding document length, we have shown that simply truncating long documents, in particular when using transformer-based approaches, results in a significant degradation of the results, and that it is necessary to move towards methods that are not constrained by document length. Finally, we found that the tree induced by our proposed discourse-driven structured attention network have non-flat complex structures, and we observe some differences in the shape of the trees between different classes when trained on the prediction of political bias in news articles. This leads us to several perspectives for the continuation of this work. In the short term, we can envisage various improvements to the structured attention model, such as leveraging contextualized word embeddings or constraining latent tree induction to learn more discourse-like structures. In the longer term, it would be interesting to develop more robust datasets for the analysis of political

bias in news articles by targeting highly politicized contents and proposing annotation schemes that are less ambiguous and more representative of the political spectrum, as well as datasets for other languages and cultural contexts. It would also be interesting to apply our approach to other forms of bias such as those present in social media posts, political debates or scientific papers.

However, our aspirations for this study were not limited to simply predicting biases. An essential aspect of our work was to characterize these biases by getting some insights into the model's decisions. Delving into the growing field of explainability in NLP, we made a particular focus on model-agnostic and perturbation-based explanation methods, more specifically LIME. While these methods have demonstrated their effectiveness, they also have their limitations, especially in terms of computational cost and their ability to process long documents. Addressing these challenges, we proposed a series of new strategies based on different levels of granularity to generate better explanations at a lower computational cost.

By evaluating and analyzing the explanations generated, we demonstrated the effectiveness of our explanation system and provided valuable insights into the underlying biases learned by the models at different levels, including the structure induced by the model, for which we could identify several biases. One important finding is that the cost of generating LIME explanations can be considerably reduced without impacting the quality of the explanations by targeting or ignoring vocabularies of interest, such as function words. Furthermore, going through a first level of high granularity explanations (sentences or EDUs) with a small number of samples in order to filter out the most important parts of the text seems to be the most efficient approach for generating quality explanations at a reasonable cost, with the advantage of having more interpretable explanations with several levels of granularity. These strategies make it possible to generate high-quality explanations for long documents in a reasonable time, which was not possible until now with LIME. Moreover, the different aspects of textual bias covered by the proposed explanation strategies (named entities, discourse markers, discourse relations, etc.) allowed us to draw several interesting conclusions about the nature of political bias in news articles. In particular, we found that the mention of political figures plays a crucial role for the model, with a tendency, for example, for certain classes to over-represent political figures from the opposite side. For the structural explanations, although their interpretation remains problematic, the predicted discourse relations showed us that there are substantial differences in the relations captured by the model in the structures induced between the different political sides, with for example some political sides favoring *evidence* and *justification* while others tend to favor *concession* or *evaluation*. While these approaches have provided us with insights

about the model's decisions and a preliminary characterization of biases, they are not yet able to fully elucidate the nature of these biases, particularly with respect to the learned structures. We have made a first step towards the automatic characterization of textual biases, but we are considering several perspectives for future work. In the very short term, we can envisage exploring other levels of explanation and subspace of perturbation, such as targeting vocabularies specific to a political topic of interest. Generating explanations on other datasets and for various languages would also make it possible to study how textual bias manifests in different cultural contexts. In the longer term, we would like to improve the interpretability of the explanations at the level of the induced structure in order to better understand how the bias manifests itself in the structure and whether this can reveal biases in the way the argument and the rhetoric of the text are constructed. Furthermore, although the strategies we have proposed can reduce the time required to generate explanations, it is still a resource-intensive and time-consuming task, and in practice it is hard to generate explanations on a large amount of data. We can therefore envisage the development of less costly methods and algorithms that preserve the quality of the explanations.

Our discourse-driven integrated approach for both bias prediction and characterization of textual bias has proven to be particularly insightful, allowing for a deeper understanding of textual biases beyond the level of individual words. However, our study has only laid the foundations for a more discourse-driven characterization of textual bias, and many perspectives remain to be explored. As we move forward in a society where information shapes opinions and drives decisions, the importance of unbiased and transparent information dissemination cannot be overstated. We hope that our research will motivate future work in this area and that it will encourage the development of methods that are more transparent, and not just accurate.

Bibliography

Amanda Y Agan, Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. Automating automaticity: How the context of human choice affects the extent of algorithmic bias. Working Paper 30981, National Bureau of Economic Research, February 2023. URL http://www.nber.org/papers/w30981.
Cited on page 29.

Samyak Agrawal, Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. Towards detecting political bias in Hindi news articles. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 239–244, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-srw.17. URL https://aclanthology.org/2022.acl-srw.17.
Cited on page 39.

Amr Ahmed and Eric Xing. Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1140–1150, Cambridge, MA, October 2010. Association for Computational Linguistics. URL https://aclanthology.org/D10-1111.
Cited on page 25.

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider, and Georg Rehm. Fine-grained classification of political bias in German news: A data set and initial experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.woah-1.13. URL https://aclanthology.org/2021.woah-1.13.
Cited on pages 37, 39, and 77.

Mohammad Ali and Naeemul Hassan. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/

2022.emnlp-main.633.
Cited on page 29.

Kenan Alkiek, Bohan Zhang, and David Jurgens. Classification without (proper) representation: Political heterogeneity in social media and its implications for classification and behavioral analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 504–522, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.43. URL https://aclanthology.org/2022.findings-acl.43.
Cited on pages 24 and 26.

Emily Allaway, Malavika Srikanth, and Kathleen McKeown. Adversarial learning for zero-shot stance detection on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4756–4767, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.379. URL https://aclanthology.org/2021.naacl-main.379.
Cited on page 21.

Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, May 2017. doi: 10.1257/jep.31.2.211. URL https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211.
Cited on pages 28 and 33.

David Alvarez-Melis and Tommi Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1042. URL https://aclanthology.org/D17-1042.
Cited on page 105.

Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.
Cited on pages 53 and 102.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1432.
Cited on page 53.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.263. URL https://aclanthology.org/2020.emnlp-main.263.

Cited on pages 108, 111, and 117.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1084. URL https://aclanthology.org/D16-1084.

Cited on page 21.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

Cited on page 63.

James K. Baker. Trainable grammars for speech recognition. *Journal of the Acoustical Society of America*, 65, 1979.

Cited on pages 62 and 65.

Vidhisha Balachandran, Artidoro Pagnoni, Jay Yoon Lee, Dheeraj Rajagopal, Jaime Carbonell, and Yulia Tsvetkov. StructSum: Summarization via structured representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2575–2585, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.220. URL https://aclanthology.org/2021.eacl-main.220.

Cited on pages 59 and 70.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium, October-November 2018a. Association for Computational Linguistics. doi: 10.18653/v1/D18-1389. URL https://aclanthology.org/D18-1389.

Cited on page 30.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana, June 2018b. Association for Computational Linguistics. doi: 10.18653/v1/N18-2004. URL https://aclanthology.org/N18-2004.
Cited on pages 31 and 32.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.404. URL https://aclanthology.org/2020.emnlp-main.404.
Cited on pages 32, 37, 38, 40, 77, 81, 88, 89, 93, and 94.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3364–3374, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.308. URL https://aclanthology.org/2020.acl-main.308.
Cited on page 38.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf. Fusion*, 58(C):82–115, jun 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.12.012. URL https://doi.org/10.1016/j.inffus.2019.12.012.
Cited on page 104.

Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. Proppy: Organizing the news based on their propagandistic content. *Information Processing Management*, 56(5):1849–1864, 2019. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2019.03.005. URL https://www.sciencedirect.com/science/article/pii/S0306457318306058.
Cited on pages 35 and 37.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy, August

2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3805. URL https://aclanthology.org/W19-3805.
Cited on page 22.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020. URL https://arxiv.org/abs/2004.05150.
Cited on pages 32, 85, 86, and 87.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041.
Cited on page 78.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.
Cited on page 21.

Yochai Benkler, Robert Faris, and Hal Roberts. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press, 11 2018. ISBN 9780190923624. doi: 10.1093/oso/9780190923624.001.0001. URL https://doi.org/10.1093/oso/9780190923624.001.0001.
Cited on page 78.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1263. URL https://aclanthology.org/D15-1263.
Cited on pages 52 and 55.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.269. URL https://aclanthology.org/2022.acl-long.269.
Cited on page 84.

Or Biran and Kathleen McKeown. PDTB discourse parsing as a tagging task: The two taggers approach. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 96–104, Prague, Czech Republic, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-4612. URL https://aclanthology.org/W15-4612.

Cited on page 54.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing, 2009. ISBN 978-0-596-51649-9. doi: http://my.safaribooksonline.com/9780596516499. URL http://www.nltk.org/book.

Cited on page 113.

Yonatan Bisk and Ke Tran. Inducing grammars with and for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 25–35, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2704. URL https://aclanthology.org/W18-2704.

Cited on pages 59 and 70.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

Cited on page 20.

Ben Bogin, Sanjay Subramanian, Matt Gardner, and Jonathan Berant. Latent compositional representations improve systematic generalization in grounded question answering. *Transactions of the Association for Computational Linguistics*, 9:195–210, 2021. doi: 10.1162/tacl_a_00361. URL https://aclanthology.org/2021.tacl-1.12.

Cited on page 59.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Cited on page 21.

Amber E. Boydstun, Dallas Card, Justin Gross, Paul Resnick, and Noah A. Smith. Tracking the Development of Media Frames within and across Policy Issues. 8

2014. doi: 10.1184/R1/6473780.v1. URL https://kilthub.cmu.edu/articles/journal_contribution/Tracking_the_Development_of_Media_Frames_within_and_across_Policy_Issues/6473780.
Cited on pages 29 and 31.

Chloé Braud and Anders Søgaard. Is writing style predictive of scientific fraud? In *Proceedings of the Workshop on Stylistic Variation*, pages 37–42, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4905. URL https://aclanthology.org/W17-4905.
Cited on page 22.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada, July 2023. The Association for Computational Linguistics. doi: 10.18653/v1/2023.disrpt-1.1. URL https://aclanthology.org/2023.disrpt-1.1.
Cited on page 55.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. doi: 10.1126/science.aal4230. URL https://www.science.org/doi/abs/10.1126/science.aal4230.
Cited on page 21.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. The RST Spanish-Chinese treebank. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 156–166, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-4917.
Cited on page 54.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2072. URL https://aclanthology.org/P15-2072.
Cited on pages 29 and 31.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on*

Empirical Methods in Natural Language Processing, pages 1410–1420, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1148. URL https://aclanthology.org/D16-1148.

Cited on page 29.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001. URL https://aclanthology.org/W01-1605.

Cited on pages 53, 54, and 96.

Caio Magno Aguiar Carvalho, Hitoshi Nagano, and Allan Kardec Barros. A comparative study for sentiment analysis on election Brazilian news. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 103–111, Uberlândia, Brazil, October 2017. Sociedade Brasileira de Computação. URL https://aclanthology.org/W17-6613.

Cited on page 32.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-2004.

Cited on page 22.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.autosimtrans-1.5. URL https://aclanthology.org/2020.autosimtrans-1.5.

Cited on pages 52 and 56.

Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Learning to flip the bias of news headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 79–88, Tilburg University, The Netherlands, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6509. URL https://aclanthology.org/W18-6509.

Cited on pages 36, 37, and 77.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. Detecting media bias in news articles using Gaussian bias distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4290–4300, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.383. URL https://aclanthology.org/2020.findings-emnlp.383.
Cited on page 38.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. Analyzing political bias and unfairness in news articles at different levels of granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online, November 2020c. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcss-1.16. URL https://aclanthology.org/2020.nlpcss-1.16.
Cited on pages 37 and 38.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.478. URL https://aclanthology.org/2020.acl-main.478.
Cited on pages 31, 39, and 56.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. Towards coherent multi-document summarization. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1173, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1136.
Cited on page 52.

Yoeng-Jin Chu and Tseng-Hong Liu. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400, 1965.
Cited on pages 66, 70, and 95.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.
Cited on page 85.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucía Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The COVID-19 social media infodemic. *CoRR*, abs/2003.05004, 2020. URL

https://arxiv.org/abs/2003.05004.
Cited on page 30.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.
Cited on page 99.

Michael D. Conover, Bruno Goncalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 192–199, 2011. doi: 10.1109/PASSAT/SocialCom.2011.34.
Cited on page 25.

Caio Corro and Ivan Titov. Learning latent trees with stochastic perturbations and differentiable dynamic programming. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5508–5521, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1551. URL https://aclanthology.org/P19-1551.
Cited on page 66.

Caio Corro and Ivan Titov. Differentiable perturb-and-parse: Semi-supervised parsing with a structured variational autoencoder. In *Proceedings of Seventh International Conference on Learning Representations*, 2019b.
Cited on page 62.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. On the development of the RST Spanish treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-0401.
Cited on page 54.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China, November 2019. Association for Computational

Linguistics. doi: 10.18653/v1/D19-1565. URL https://aclanthology.org/D19-1565.
Cited on page 35.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online), December 2020a. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.186. URL https://aclanthology.org/2020.semeval-1.186.
Cited on pages 36 and 37.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.32. URL https://aclanthology.org/2020.acl-demos.32.
Cited on page 33.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.aacl-main.46.
Cited on pages 106 and 107.

Michal Daniluk, Tim Rocktäschel, Johannes Welbl, and Sebastian Riedel. Frustratingly short attention spans in neural language modeling. *CoRR*, abs/1702.04521, 2017. URL http://arxiv.org/abs/1702.04521.
Cited on page 67.

Maryam Davoodi, Eric Waltenburg, and Dan Goldwasser. Understanding the language of political agreement and disagreement in legislative texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5358–5368, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.476. URL https://aclanthology.org/2020.acl-main.476.
Cited on page 26.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. Analyzing polarization in social media: Method and application to

tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2970–3005, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1304. URL https://aclanthology.org/N19-1304.

Cited on page 26.

Nicolas Devatine, Philippe Muller, and Chloé Braud. Predicting political orientation in news with latent discourse structure to improve bias understanding. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 77–85, Gyeongju, Republic of Korea and Online, October 2022. International Conference on Computational Linguistics. URL https://aclanthology.org/2022.codi-1.10.

Cited on page 75.

Nicolas Devatine, Philippe Muller, and Chloé Braud. An integrated approach for political bias prediction and explanation based on discursive structure. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11196–11211, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.711. URL https://aclanthology.org/2023.findings-acl.711.

Cited on page 75.

Nicolas Devatine, Philippe Muller, and Chloé Braud. MELODI at SemEval-2023 task 3: In-domain pre-training for low-resource classification of news articles. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 108–113, Toronto, Canada, July 2023b. Association for Computational Linguistics. URL https://aclanthology.org/2023.semeval-1.14.

Cited on page 97.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Cited on pages 26 and 83.

Nicholas Diakopoulos. Accountability in algorithmic decision making. *Commun. ACM*, 59 (2):56–62, jan 2016. ISSN 0001-0782. doi: 10.1145/2844110. URL https://doi.org/10.1145/2844110.

Cited on page 78.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.
Cited on page 78.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.
Cited on pages 103 and 107.

James N. Druckman. On the limits of framing effects: Who can frame? *The Journal of Politics*, 63(4):1041–1066, 2001. ISSN 00223816, 14682508. URL http://www.jstor.org/stable/2691806.
Cited on page 24.

Jennifer L. Eberhardt. *Biased: Uncovering the Hidden Prejudice That Shapes What We See, Think, and Do*. Penguin, 2019. ISBN 978-0-7352-2493-3 978-0-7352-2494-0. URL http://id.lib.harvard.edu/alma/99153761411903941/catalog.
Cited on page 18.

Jack Edmonds. Optimum branchings. *Journal of Research of the national Bureau of Standards*, 71:233–240, 1967.
Cited on pages 66, 70, and 95.

Jason M. Eisner. Three new probabilistic models for dependency parsing: An exploration. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996. URL https://aclanthology.org/C96-1058.
Cited on page 65.

Robert M Entman. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58, 1993.
Cited on pages 18 and 29.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1664. URL https://aclanthology.org/D19-1664.
Cited on pages 36 and 37.

Vanessa Wei Feng and Graeme Hirst. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 60–68, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/P12-1007.

Cited on page 54.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. Evaluating discourse in structured text representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 646–653, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1062. URL https://aclanthology.org/P19-1062.

Cited on pages 70, 80, 81, and 95.

William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1138. URL https://aclanthology.org/N16-1138.

Cited on page 31.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1393. URL https://aclanthology.org/D18-1393.

Cited on page 29.

Seth Flaxman, Sharad Goel, and Justin M. Rao. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80(S1):298–320, 03 2016. ISSN 0033-362X. doi: 10.1093/poq/nfw006. URL https://doi.org/10.1093/poq/nfw006.

Cited on page 19.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3449–3457. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.371. URL https://doi.org/10.1109/ICCV.2017.371.

Cited on page 106.

Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), jul 2018. ISSN 0360-0300. doi: 10.1145/3232676. URL

https://doi.org/10.1145/3232676.
Cited on page 21.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 00905364. URL http://www.jstor.org/stable/2699986.
Cited on page 106.

Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 77–84, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4809. URL https://aclanthology.org/W19-4809.
Cited on pages 37 and 39.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL http://jmlr.org/papers/v17/15-239.html.
Cited on page 82.

Matthew Gentzkow and Jesse M. Shapiro. What drives media slant? evidence from U.S. daily newspapers. *Econometrica*, 78(1):35–71, 2010. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/25621396.
Cited on pages 19 and 36.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340, July 2019. doi: 10.3982/ECTA16566. URL https://ideas.repec.org/a/wly/emetrp/v87y2019i4p1307-1340.html.
Cited on page 24.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1168. URL https://aclanthology.org/D14-1168.
Cited on pages 56 and 57.

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on*

Discourse Relation Parsing and Treebanking (DISRPT 2021), pages 51–62, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.disrpt-1.6. URL https://aclanthology.org/2021.disrpt-1.6.
Cited on page 115.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, Oct. 2017. doi: 10.1609/aimag.v38i3.2741. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2741.
Cited on page 104.

Justin Grimmer and Brandon M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013. doi: 10.1093/pan/mps028.
Cited on page 24.

Tim Groseclose and Jeffrey Milyo. A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237, 2005. ISSN 00335533, 15314650. URL http://www.jstor.org/stable/25098770.
Cited on pages 29 and 32.

Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. Nela-gt-2022: A large multi-labelled news dataset for the study of misinformation in news articles, 2023.
Cited on page 31.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), aug 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL https://doi.org/10.1145/3236009.
Cited on pages 23 and 105.

David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58, Jun. 2019. doi: 10.1609/aimag.v40i2.2850. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2850.
Cited on page 104.

Meiqi Guo, Rebecca Hwa, Yu-Ru Lin, and Wen-Ting Chung. Inflating topic relevance with ideology: A case study of political ideology bias in social topic detection models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4873–4885, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.428. URL https:

//aclanthology.org/2020.coling-main.428.
Cited on page 26.

Grigorii Guz and Giuseppe Carenini. Coreference for discourse parsing: A neural approach. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.codi-1.17. URL https://aclanthology.org/2020.codi-1.17.
Cited on page 57.

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. *Cohesion in english*. Number 9. Routledge, 2014.
Cited on page 51.

Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles : an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415, 2019. ISSN 1432-5012. doi: 10.1007/s00799-018-0261-y.
Cited on pages 29 and 32.

Jiyoung Han, Youngin Lee, Junbum Lee, and Meeyoung Cha. The fallacy of echo chambers: Analyzing the political slants of user-generated news comments in Korean media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 370–374, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5548. URL https://aclanthology.org/D19-5548.
Cited on page 39.

Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.491. URL https://aclanthology.org/2020.acl-main.491.
Cited on page 111.

Katsuhiko Hayashi, Tsutomu Hirao, and Masaaki Nagata. Empirical comparison of dependency conversions for RST discourse trees. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 128–136, Los Angeles, September 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3616. URL https://aclanthology.org/W16-3616.
Cited on page 66.

Bas Heerschop, Frank Goossen, Alexander Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. Polarity analysis of texts using discourse structure. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*,

CIKM '11, page 1061–1070, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450307178. doi: 10.1145/2063576.2063730. URL https://doi.org/10.1145/2063576.2063730.

Cited on page 55.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Cited on page 26.

Jiwoo Hong, Yejin Cho, Jaemin Jung, Jiyoung Han, and James Thorne. Disentangling structure and style: Political bias detection in news by inducing document hierarchy. *CoRR*, abs/2304.02247, 2023. doi: 10.48550/arXiv.2304.02247. URL https://doi.org/10.48550/arXiv.2304.02247.

Cited on page 56.

Benjamin D. Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *CoRR*, abs/1703.09398, 2017. URL http://arxiv.org/abs/1703.09398.

Cited on page 30.

Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2096. URL https://aclanthology.org/P16-2096.

Cited on pages 22 and 28.

John P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2 (8):null, 08 2005. doi: 10.1371/journal.pmed.0020124. URL https://doi.org/10.1371/journal.pmed.0020124.

Cited on page 20.

Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. Unsupervised neural single-document summarization of reviews via learning latent discourse structure and its ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2142–2152, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1206. URL https://aclanthology.org/P19-1206.

Cited on pages 70 and 121.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1113–1122, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1105. URL https://aclanthology.org/P14-1105.
Cited on page 25.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357.
Cited on page 23.

Peter Jansen, Mihai Surdeanu, and Peter Clark. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 977–986, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1092. URL https://aclanthology.org/P14-1092.
Cited on page 52.

Yangfeng Ji and Jacob Eisenstein. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1002. URL https://aclanthology.org/P14-1002.
Cited on pages 54 and 55.

Yangfeng Ji and Noah A. Smith. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1092. URL https://aclanthology.org/P17-1092.
Cited on pages 29, 55, and 57.

Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. Team bertha von suttner at SemEval-2019 task 4: Hyperpartisan news detection using ELMo sentence representation convolutional network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2146. URL https://aclanthology.org/S19-2146.
Cited on pages 36 and 93.

Kristen Johnson and Dan Goldwasser. Identifying stance by analyzing political discourse on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 66–75, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5609. URL https://aclanthology.org/W16-5609.
Cited on page 24.

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 904–915, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1083.
Cited on page 54.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1048.
Cited on page 56.

Dan Jurafsky and James H. Martin. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 2009. ISBN 9780131873216 0131873210. URL https://web.stanford.edu/~jurafsky/slp3/.
Cited on page 71.

Morteza Kamaladdini Ezzabady, Philippe Muller, and Chloé Braud. Multi-lingual discourse segmentation and connective identification: MELODI at disrpt2021. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 22–32, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.disrpt-1.3. URL https://aclanthology.org/2021.disrpt-1.3.
Cited on pages 79, 80, and 114.

Lalitha Kameswari and Radhika Mamidi. Towards quantifying magnitude of political bias in news articles using a novel annotation schema. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 671–678, Held Online, September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-1.76.
Cited on page 39.

Hans Kamp. A theory of truth and semantic representation. In P. Portner and B. H. Partee, editors, *Formal Semantics - the Essential Readings*, pages 189–222. Blackwell, 1981.

Cited on page 53.

Hamid Karimi and Jiliang Tang. Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3432–3442, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1347. URL https://aclanthology.org/N19-1347.

Cited on pages 30 and 70.

Kornraphop Kawintiranon and Lisa Singh. PoliBERTweet: A pre-trained language model for analyzing political content on Twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7360–7367, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.801.

Cited on pages 24 and 27.

Johannes Kiesel, Khalid Al-Khatib, Matthias Hagen, and Benno Stein. A shared task on argumentation mining in newspaper editorials. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 35–38, Denver, CO, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-0505. URL https://aclanthology.org/W15-0505.

Cited on page 22.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2145. URL https://aclanthology.org/S19-2145.

Cited on pages 35, 37, 88, and 90.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks. *CoRR*, abs/1702.00887, 2017. URL http://arxiv.org/abs/1702.00887.

Cited on pages 62, 63, 64, 65, 66, and 67.

Svetlana Kiritchenko and Saif Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2005. URL

https://aclanthology.org/S18-2005.
Cited on page 21.

Terry Koo, Amir Globerson, Xavier Carreras, and Michael Collins. Structured prediction models via the matrix-tree theorem. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 141–150, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/D07-1015.
Cited on pages 66 and 67.

Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5686–5697, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858529. URL https://doi.org/10.1145/2858036.2858529.
Cited on page 107.

K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Content Analysis: An Introduction to Its Methodology. Sage, 2004. ISBN 9780761915454. URL https://books.google.fr/books?id=q657o3M3C8cC.
Cited on page 87.

Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1388. URL https://aclanthology.org/D18-1388.
Cited on page 36.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.
Cited on page 65.

George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008.
Cited on page 51.

Michael Laver, Kenneth Benoit, and John Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97 (2)(2):311–331, May

2003. ISSN 0003-0554. doi: 10.1017/S0003055403000698.
Cited on page 24.

Konstantina Lazaridou, Alexander Löser, Maria Mestre, and Felix Naumann. Discovering biased news articles leveraging multiple human annotations. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1268–1277, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.159.
Cited on page 38.

Alan Lee, Rashmi Prasad, Aravind K. Joshi, Nikhil Dinesh, and Bonnie Webber. Complexity of dependencies in discourse: Are dependencies in discourse more complex than in syntax? Dec 2006.
Cited on page 66.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. NeuS: Neutral multi-news summarization for mitigating framing bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.228. URL https://aclanthology.org/2022.naacl-main.228.
Cited on page 30.

Rasmus Lehmann and Leon Derczynski. Political stance in Danish. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 197–207, Turku, Finland, September–October 2019. Linköping University Electronic Press. URL https://aclanthology.org/W19-6121.
Cited on page 27.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.682.
Cited on pages 39 and 56.

Piyawat Lertvittayakumjorn and Francesca Toni. Human-grounded evaluations of explanation methods for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5195–5205, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1523.

URL https://aclanthology.org/D19-1523.
Cited on page 117.

Chang Li and Dan Goldwasser. Encoding social information with graph convolutional networks forPolitical perspective detection in news media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1247. URL https://aclanthology.org/P19-1247.
Cited on page 36.

Chang Li and Dan Goldwasser. MEAN: Multi-head entity aware attention network for political perspective detection in news media. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 66–75, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlp4if-1.10. URL https://aclanthology.org/2021.nlp4if-1.10.
Cited on pages 38, 114, and 120.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online), December 2020a. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.238. URL https://aclanthology.org/2020.coling-main.238.
Cited on page 53.

Qi Li, Tianshi Li, and Baobao Chang. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 362–371, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1035. URL https://aclanthology.org/D16-1035.
Cited on page 54.

Zhenwen Li, Wenhao Wu, and Sujian Li. Composing elementary discourse units in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.551. URL https://aclanthology.org/2020.acl-main.551.
Cited on page 72.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the*

Twelfth Language Resources and Evaluation Conference, pages 1478–1484, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.184.
Cited on page 31.

Wei-Hao Lin, Eric Xing, and Alexander Hauptmann. A joint topic and perspective model for ideological discourse. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 17–32, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-87481-2.
Cited on page 25.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A pdtb-styled end-to-end discourse parser. *CoRR*, abs/1011.0835, 2010. URL http://arxiv.org/abs/1011.0835.
Cited on page 54.

Marco Lippi and Paolo Torroni. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Technol.*, 16(2), mar 2016. ISSN 1533-5399. doi: 10.1145/2850417. URL https://doi.org/10.1145/2850417.
Cited on page 32.

Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, jun 2018. ISSN 1542-7730. doi: 10.1145/3236386.3241340. URL https://doi.org/10.1145/3236386.3241340.
Cited on page 105.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. Mitigating political bias in language models through reinforced calibration. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14857–14866. AAAI Press, 2021a. URL https://ojs.aaai.org/index.php/AAAI/article/view/17744.
Cited on page 26.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/K19-1047. URL https://aclanthology.org/K19-1047.
Cited on pages 30 and 31.

Yang Liu and Mirella Lapata. Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75, 2018. doi: 10.1162/tacl_a_00005. URL https://aclanthology.org/Q18-1005.

Cited on pages 9, 63, 65, 66, 67, 69, 70, 71, 78, and 79.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*, July 2019b. URL http://arxiv.org/abs/1907.11692.

Cited on pages 26 and 83.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.101. URL https://aclanthology.org/2022.findings-naacl.101.

Cited on pages 32, 37, 39, 77, 84, 85, 88, 91, and 97.

Zhengyuan Liu, Ke Shi, and Nancy Chen. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.codi-main.15. URL https://aclanthology.org/2021.codi-main.15.

Cited on page 57.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Cited on pages 106, 107, and 108.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=S1jE5L5gl.

Cited on page 62.

William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988. URL http:

//scholar.google.com/scholar.bib?q=info:BEw8CIWbucoJ:scholar.google.com/&output=citation&scisig=AAGBfm0AAAAAU3X_1Dq4ULnWfFzMeRsqGJcha1fReMSl&scisf=4&hl=en.

Cited on pages 52, 71, 102, and 115.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999. URL http://nlp.stanford.edu/fsnlp/.

Cited on page 71.

Antoine Marie, Sacha Altay, and Brent Strickland. Moralization and extremism robustly amplify myside sharing. *PNAS Nexus*, 2(4):pgad078, 04 2023. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgad078. URL https://doi.org/10.1093/pnasnexus/pgad078.

Cited on page 19.

André F. T. Martins, Tsvetomila Mihaylova, Nikita Nangia, and Vlad Niculae. Latent structure models for natural language processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-4001. URL https://aclanthology.org/P19-4001.

Cited on pages 58 and 60.

Maxwell E. McCombs and Donald L. Shaw. The agenda-setting function of mass media. *The Public Opinion Quarterly*, 36(2):176–187, 1972. ISSN 0033362X, 15375331. URL http://www.jstor.org/stable/2747787.

Cited on page 20.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. URL http://arxiv.org/abs/1301.3781.

Cited on page 25.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1147. URL https://aclanthology.org/D16-1147.

Cited on page 67.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2018.07.007. URL https://www.sciencedirect.com/science/article/pii/S0004370218305988.

Cited on page 105.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

Cited on page 22.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1003. URL https://aclanthology.org/S16-1003.

Cited on pages 21 and 31.

Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

Cited on pages 107 and 111.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2017. doi: 10.1093/pan/mpn018.

Cited on page 25.

Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. Hitachi at SemEval-2020 task 11: An empirical study of pre-trained transformer family for propaganda detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1739–1748, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.228. URL https://aclanthology.org/2020.semeval-1.228.

Cited on page 36.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. A bert-based transfer learning approach for hate speech detection in online social media. *CoRR*, abs/1910.12574, 2019. URL http://arxiv.org/abs/1910.12574.

Cited on page 22.

Preslav Nakov and Giovanni Da San Martino. Fact-checking, fake news, propaganda, and media bias: Truth seeking in the post-truth era. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 7–19, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-tutorials.2. URL https://aclanthology.org/2020.emnlp-tutorials.2.

Cited on pages 30 and 34.

Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *CoRR*, abs/1802.00682, 2018. URL http://arxiv.org/abs/1802.00682.

Cited on page 117.

Viet-An Nguyen, Jordan L Ying, and Philip Resnik. Lexical and hierarchical topic regression. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/f5deaeeae1538fb6c45901d524ee2f98-Paper.pdf.

Cited on page 25.

Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998. doi: 10.1037/1089-2680.2.2.175. URL https://doi.org/10.1037/1089-2680.2.2.175.

Cited on page 19.

Vlad Niculae, André F. T. Martins, and Claire Cardie. Towards dynamic computation graphs via sparse latent structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 905–911, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1108. URL https://aclanthology.org/D18-1108.

Cited on page 66.

Vlad Niculae, Caio F. Corro, Nikita Nangia, Tsvetomila Mihaylova, and André F. T. Martins. Discrete latent structure in neural networks, 2023.

Cited on pages 58 and 60.

Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, June 2010. ISSN 1573-6687. doi: 10.1007/s11109-010-9112-2. URL https://doi.org/10.1007/s11109-010-9112-2.

Cited on page 78.

Cailin O'Connor, James O Weatherall, and Aydin Mohseni. The best paper you'll read today: Media bias and the public understanding of science, Apr 2023. URL osf.io/preprints/metaarxiv/hpks9.
Cited on page 19.

Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA, 2016. ISBN 0553418815.
Cited on page 29.

Ray Oshikawa, Jing Qian, and William Yang Wang. A survey on natural language processing for fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.747.
Cited on page 30.

Sebastian Padó, Andre Blessing, Nico Blokker, Erenay Dayanik, Sebastian Haunss, and Jonas Kuhn. Who sides with whom? towards computational construction of discourse networks for political debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2841–2847, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1273. URL https://aclanthology.org/P19-1273.
Cited on page 40.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008. ISSN 1554-0669. doi: 10.1561/1500000011. URL http://dx.doi.org/10.1561/1500000011.
Cited on pages 21 and 32.

Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group , The, 2011. ISBN 1594203008.
Cited on pages 20, 29, and 78.

Hyunji Park, Yogarshi Vyas, and Kashif Shah. Efficient classification of long documents using transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 702–709, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.79. URL https://aclanthology.org/2022.acl-short.79.
Cited on pages 84 and 85.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural*

Language Processing, pages 2799–2804, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1302. URL https://aclanthology.org/D18-1302.

Cited on page 22.

Hao Peng, Sam Thomson, and Noah A. Smith. Backpropagating through structured argmax using a SPIGOT. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1863–1873, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1173. URL https://aclanthology.org/P18-1173.

Cited on page 62.

Siyao Peng, Yang Janet Liu, and Amir Zeldes. GCDT: A Chinese RST treebank for multigenre and multilingual discourse parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 382–391, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-short.47.

Cited on page 54.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Cited on pages 25, 79, and 97.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *CoRR*, abs/1708.07104, 2017. URL http://arxiv.org/abs/1708.07104.

Cited on page 30.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

Cited on page 101.

Andrew Peterson and Arthur Spirling. Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *Political Analysis*, 26(1):

120–128, 2018. doi: 10.1017/pan.2017.39.
Cited on page 25.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.semeval-1.317.
Cited on page 96.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1022. URL https://aclanthology.org/P18-1022.
Cited on pages 36, 37, and 77.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf.
Cited on pages 53 and 54.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. Beyond binary labels: Political ideology prediction of Twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1068. URL https://aclanthology.org/P17-1068.
Cited on pages 24, 26, and 27.

Rajkumar Pujari and Dan Goldwasser. Understanding politics via contextualized discourse processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1353–1367, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.102. URL https://aclanthology.org/2021.emnlp-main.102.
Cited on pages 24 and 26.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 101–108, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.14. URL https://aclanthology.org/2020.acl-demos.14.
Cited on page 115.

Yifu Qiu and Shay B. Cohen. Abstractive summarization guided by latent hierarchical document structure. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5317, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.355. URL https://aclanthology.org/2022.emnlp-main.355.
Cited on page 59.

L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi: 10.1109/5.18626.
Cited on pages 62 and 65.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1317. URL https://aclanthology.org/D17-1317.
Cited on pages 24, 26, and 30.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1162.
Cited on pages 21 and 22.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-3020. URL https://aclanthology.org/N16-3020.
Cited on pages 22, 103, 105, 106, 108, 109, and 110.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
Cited on pages 106, 108, and 111.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442.
Cited on page 59.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54.
Cited on page 83.

Flora Sakketou, Joan Plepi, Riccardo Cervero, Henri Jacques Geiss, Paolo Rosso, and Lucie Flek. FACTOID: A new dataset for identifying misinformation spreaders and political bias. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3231–3241, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.345.
Cited on page 31.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *CoRR*, abs/1708.08296, 2017. URL http://arxiv.org/abs/1708.08296.
Cited on page 103.

Eitan Sapiro-Gheiler. Examining political trustworthiness through text-based measures of ideology. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 10029–10030, Jul. 2019. doi: 10.1609/aaai.v33i01.330110029. URL https://ojs.aaai.org/index.php/AAAI/article/view/5151.
Cited on page 25.

Dietram A. Scheufele and David Tewksbury. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of Communication*, 57(1):9–20, March 2007. ISSN 0021-9916. doi: 10.1111/j.0021-9916.2007.00326.x.
Cited on page 28.

Deborah Schiffrin. Approaches to discourse. *Journal of Pragmatics*, 3(29):355–359, 1998.
Cited on page 50.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and*

Pattern Recognition (CVPR), pages 815–823, 2015. doi: 10.1109/CVPR.2015.7298682.
Cited on page 82.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org, 2017.
Cited on page 106.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '19, page 395–405, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330935. URL https://doi.org/10.1145/3292500.3330935.
Cited on page 31.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188, 2020. doi: 10.1089/big.2020.0062. URL https://doi.org/10.1089/big.2020.0062. PMID: 32491943.
Cited on page 31.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3477–3486, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1351. URL https://aclanthology.org/N19-1351.
Cited on page 114.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1010.
Cited on page 25.

Karin Sim Smith. On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4814. URL https://aclanthology.org/W17-4814.
Cited on page 57.

Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. Political ideology and polarization: A multi-dimensional approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 231–243, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.17. URL https://aclanthology.org/2022.naacl-main.17.

Cited on pages 37, 39, and 77.

Jonathan B. Slapin and Sven-Oliver Proksch. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722, 2008. ISSN 00925853, 15405907. URL http://www.jstor.org/stable/25193842.

Cited on page 24.

Karolina Stanczak and Isabelle Augenstein. A survey on gender bias in natural language processing. *CoRR*, abs/2112.14168, 2021. URL https://arxiv.org/abs/2112.14168.

Cited on page 22.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.50. URL https://aclanthology.org/2020.acl-main.50.

Cited on page 37.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159. URL https://aclanthology.org/P19-1159.

Cited on page 22.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.

Cited on page 106.

Cass R. Sunstein. *Republic: Divided Democracy in the Age of Social Media*. Princeton University Press, ned - new edition edition, 2018. ISBN 9780691180908. URL http://www.jstor.org/stable/j.ctv8xnhtd.

Cited on page 78.

Charles S. Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3):755–769, 2006. ISSN 00925853, 15405907. URL http://www.jstor.org/stable/3694247.
Cited on page 24.

Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia, July 2006. Association for Computational Linguistics. URL https://aclanthology.org/W06-1639.
Cited on pages 24 and 25.

Jörg Tiedemann and Santhosh Thottingal. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.
Cited on page 98.

Svetlana Toldova, Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. Rhetorical relations markers in Russian RST treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 29–33, Santiago de Compostela, Spain, September 2017. Association for Computational Linguistics. ISBN 978-1-945626-78-4. doi: 10.18653/v1/W17-3604. URL https://aclanthology.org/W17-3604.
Cited on page 54.

Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2 edition, 2003. doi: 10.1017/CBO9780511840005.
Cited on page 51.

W.T. Tutte and C.S.J.A. Nash-Williams. *Graph Theory*. Encyclopedia of mathematics and its applications. Addison-Wesley Publishing Company, Advanced Book Program, 1984. ISBN 9780201135206. URL https://books.google.fr/books?id=pLwdAQAAMAAJ.
Cited on pages 66 and 67.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981. doi: 10.1126/science.7455683. URL https://www.science.org/doi/abs/10.1126/science.7455683.
Cited on page 18.

T.A. van Dijk. *News as discourse*. University of Amsterdam, 1988.
Cited on pages 50 and 56.

Teun A van Dijk. *Ideology: A multidisciplinary approach*. Sage, 1998.
Cited on pages 50 and 51.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
Cited on pages 83 and 106.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.107. URL https://aclanthology.org/2021.acl-long.107.
Cited on page 32.

Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3007. URL https://aclanthology.org/P19-3007.
Cited on page 84.

S Wachter, B Mittelstadt, and C Russell. Counterfactual explanations without opening the black box: automated decisions and the gdpr. *Harvard Journal of Law and Technology*, 31(2):841–887, 2018.
Cited on page 105.

Hanna Wallach. Computational social science computer science + social data. *Commun. ACM*, 61(3):42–44, feb 2018. ISSN 0001-0782. doi: 10.1145/3132698. URL https://doi.org/10.1145/3132698.
Cited on page 87.

William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL https://aclanthology.org/P17-2067.
Cited on pages 24 and 30.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5878–5882, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1599. URL https://aclanthology.org/D19-1599.
Cited on page 85.

Claire Wardle and Hossein Derakhshan. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking*. 09 2017.
Cited on page 30.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016. doi: 10.23915/distill.00002. URL http://distill.pub/2016/misread-tsne.
Cited on page 107.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108, 2019.
Cited on page 114.

Jerry Wei. Newb: 200, 000+ sentences for political bias detection. *CoRR*, abs/2006.03051, 2020. URL https://arxiv.org/abs/2006.03051.
Cited on pages 37 and 77.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002.
Cited on page 106.

Siqi Wu and Paul Resnick. Cross-partisan discussions on youtube: Conservatives talk to liberals but liberals don't talk to conservatives. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):808–819, May 2021. doi: 10.1609/icwsm.v15i1.18105. URL https://ojs.aaai.org/index.php/ICWSM/article/view/18105.
Cited on pages 24 and 26.

Zhaofeng Wu. Learning with latent structures in natural language processing: A survey. *CoRR*, abs/2201.00490, 2022. URL https://arxiv.org/abs/2201.00490.
Cited on pages 58 and 60.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.451. URL https://aclanthology.org/2020.acl-main.451.
Cited on page 72.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/K15-2001. URL https://aclanthology.org/K15-2001.
Cited on page 54.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1174. URL https://aclanthology.org/N16-1174.
Cited on page 32.

Tae Yano, William W. Cohen, and Noah A. Smith. Predicting response to political blog posts with topic models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 477–485, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/N09-1054.
Cited on page 24.

Pengcheng Yin, Chunting Zhou, Junxian He, and Graham Neubig. StructVAE: Tree-structured latent variable models for semi-supervised semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 754–765, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1070. URL https://aclanthology.org/P18-1070.
Cited on page 59.

Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Classifying party affiliation from political speech. *Journal of Information Technology and Politics*, 5(1):33–48, 2008. ISSN 1933-1681. doi: 10.1080/19331680802149608.
Cited on page 25.

Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. RST discourse parsing with second-stage EDU-level pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.294. URL https://aclanthology.org/2022.acl-long.294.
Cited on page 54.

Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. Interpretable propaganda detection in news articles. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1597–1605, Held Online, September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-1.179.
Cited on page 35.

Amir Zeldes. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, 2017. doi: http://dx.doi.org/10.1007/s10579-016-9343-x.
Cited on page 80.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.disrpt-1.1. URL https://aclanthology.org/2021.disrpt-1.1.
Cited on pages 80 and 115.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.
Cited on page 31.

Wenqian Zhang, Shangbin Feng, Zilong Chen, Zhenyu Lei, Jundong Li, and Minnan Luo. KCD: Knowledge walks and textual cues enhanced political perspective detection in news media. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4140, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.304. URL https://aclanthology.org/2022.naacl-

main.304.
Cited on page 38.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.
Cited on page 21.

Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5), sep 2020. ISSN 0360-0300. doi: 10.1145/3395046. URL https://doi.org/10.1145/3395046.
Cited on page 30.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, page 3205–3212, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3412880. URL https://doi.org/10.1145/3340531.3412880.
Cited on pages 30 and 31.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. SAFE: similarity-aware multi-modal fake news detection. *CoRR*, abs/2003.04981, 2020b. URL https://arxiv.org/abs/2003.04981.
Cited on page 30.