Definition and integration of new insights for improving disease surveillance systems

Mehtab Alam Syed

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale Information, Structures, Systèmes (I2S)

Unité de recherche TETIS, CIRAD

Definition and integration of new insights for improving disease surveillance systems

Présentée par Mehtab Alam SYED

le 08/12/2023

Sous la direction de Mathieu Roche, Maguelonne Teisseire et Elena Arsevska

Devant le jury composé de

Antoine Doucet, Professeur, Université de la Rochelle, France	Rapporteur
Natalia Grabar, Chargée de Recherche CNRS - HDR, Université de Lille, France	Rapportrice
Bruno Martins, Associate Professor, University of Lisbon, Portugal	Examinateur
Isabelle Mougenot, Professeure, Université de Montpellier, France	Examinatrice
Mathieu Roche, Chercheur CIRAD - HDR, France	Directeur de thèse
Maguelonne Teisseire, Directrice de Recherche INRAE, France	Co-Directrice de thèse
Elena Arsevska, Chercheuse CIRAD, France	Encadrante - Invité

UNIVERSITÉ DE MONTPELLIER

PUBLICATIONS

Journals

Syed MA, Arsevska E, Roche M, Teisseire M. GeospatRE: Extraction and Geocoding of spatial relation entities in textual documents. *Cartography and Geographic Information Science*, Taylor & Francis, in press, 2023 [Q1, IF 2.5]
https://doi.org/10.1080/15230406.2023.2264753

Arınık N, Van Bortel W, Boudoua B, Busani L, Decoupes R, Interdonato R, Kafando R, van Kleef E, Roche M, **Syed MA**, Teisseire M. An annotated dataset for event-based surveillance of antimicrobial resistance. *Data in Brief*. Elsevier, 108870, 2023 [Q2, IF 1.2]
https://doi.org/10.1016/j.dib.2022.108870

Syed MA, Decoupes R, Arsevska E, Roche M, Teisseire M. Spatial opinion mining from COVID-19 twitter data. *International Journal of Infectious Diseases*. 116, suppl.: 527. International Meeting on Emerging Diseases and Surveillance (IMED 2021), 2022 (Extended abstract) [Q1, IF 8.4]
https://doi.org/10.1016/j.ijid.2021.12.065

Conference Proceedings

Syed MA, Arsevska E, Roche M, Teissere M. A metadata approach to classify domain-specific documents for Event-based Surveillance Systems. In Proceedings of *Conference on Communication, Computing and Digital Systems (C-CODE)*, 2023, pp. 1-5
https://doi.org/10.1109/C-CODE58145.2023.10139883

Syed MA, Arsevska E, Roche M, Teisseire M. GeoXTag: Relative spatial information extraction and tagging of unstructured text. Proceedings of *AGILE conference on geographic information science* - AGILE GIScience Ser., Volume 3, 16, 2022 [Best paper candidate]
https://doi.org/10.5194/agile-giss-3-16-2022

Syed MA, Arsevska E, Roche M, Teisseire M. Feature Selection for Sentiment Classification of COVID-19 Tweets: H-TFIDF Featuring BERT. In Proceedings of the *15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022)* - Volume 5: HEALTHINF, 2022, pp. 648-656
https://doi.org/10.5220/0010887800003123

Syed MA, Arsevska E, Roche M, Teisseire M. A data-driven score model to assess online news articles in Event-based surveillance system. In Proceedings of "*International Conference on Information Management and Big Data*", Springer CCIS, Volume 1577, 2021, pp. 264–280
https://doi.org/10.1007/978-3-031-04447-2_18

RÉSUMÉ

Une recrudescence des maladies infectieuses a conduit à une augmentation significative des menaces sanitaires signalées à partir de diverses sources en ligne. Les systèmes de surveillance basés sur les événements (EBS) détectent les menaces sanitaires ou les événements en utilisant des approches automatisées pour aider les parties prenantes à prendre des mesures préventives en temps opportun. Il existe un important potentiel d'amélioration dans l'extraction d'événements pour renforcer l'efficacité des EBS. Dans cette thèse, nous proposons d'améliorer l'extraction automatique de données pour les événements et fournir des informations plus précises. Et tout particulièrement, nous nous intéressons à la qualité des données, l'amélioration de la précision géographique et l'analyse de sentiment. Ce travail est soutenu par le projet MOOD qui vise à améliorer la surveillance en épidémiologie des systèmes de type EBS.

Pour surveiller efficacement les maladies infectieuses à partir de sources de données en ligne, il est impératif de mettre en œuvre des mesures d'évaluation de la qualité des données afin d'obtenir des informations fiables et dignes de confiance. Dans notre travail visant à améliorer la qualité des données, nous introduisons une approche basée sur les données pour classer les articles de presse comme pertinents ou non pertinents en enrichissant le contexte épidémiologique. Nous explorons également les caractéristiques des métadonnées des actualités en appliquant une approche d'apprentissage automatique pour identifier les métadonnées importantes. De plus, nous explorons également l'amélioration des attributs de qualité de la source d'actualités en proposant l'identification de la spécialisation de la source et l'identification de la couverture géographique.

Pour extraire des informations sur les événements l'exactitude géographique joue un rôle essentiel en épidémiologie. Nous proposons une approche de reconnaissance d'entités nommées (NER) basée sur des règles pour extraire les relations spatiales liées aux emplacements mentionnés dans les données textuelles, qui est évaluée à l'aide d'un ensemble d'articles de presse couvrant diverses maladies. De plus, nous présentons un algorithme pour calculer les coordonnées géographiques sous forme de polygones pour les emplacements de relations spatiales identifiées, avec des évaluations qualitatives impliquant les utilisateurs finaux.

Extraire des sentiments des médias sociaux, par exemple des tweets géolocalisés offre des aperçus en temps réel pour évaluer la gravité d'un événement. Nous avons effectué une analyse de sentiment en utilisant des mesures basées sur la hiérarchie spatiale pour l'analyse des tweets (H-TFIDF) afin de comprendre les sentiments locaux pendant l'épidémie de COVID-19. Cette analyse a été évaluée sur un jeu de données liés au COVID-19 catégorisé en groupes spatiaux. De plus, diverses fonctionnalités, y compris celles basées sur le modèle de langue Bidirectional Encoder Representations from Transformers (BERT), H-TFIDF, la fréquence des termes-inverse de la fréquence du document (TF-IDF) et le sac de mots (BOW), ont été évaluées pour mesurer leur importance dans la classification des sentiments.

Mots-clés: Fouille de textes, Extraction d'événement, One Health, Traitement automatique du langage naturel (TALN)

ABSTRACT

An escalation in infectious diseases has led to a significant increase in health threats reported across diverse online sources. Event-based surveillance (EBS) systems detect health threats or events by utilizing automated approaches to assist stakeholders in taking timely preventive measures. There is significant room for improvement across various aspects of the event to enhance the effectiveness of EBS. In this thesis, we improve several aspects of the event to provide more precise information by ensuring prior data quality assessment, geographical accuracy enhancement, and post-situational awareness. This work is supported by the MOOD [1] project, which aims to enhance the utility of EBS.

To effectively monitor infectious diseases reported from online sources, it is imperative to implement data quality assessment measures in order to obtain trustworthy and reliable information. In our work to improve data quality, we introduce a data-driven approach to classify news articles as relevant or irrelevant by enriching the epidemiological context. We also explore metadata features of online news by applying a machine learning approach to identify important metadata features. Moreover, we also explore enhancing news source quality attributes, proposing the identification of source specialization and geographical coverage identification for improved classification performance.

To extract event information, the geographical accuracy of events plays a pivotal role in epidemiology allows precise tracking, containment thereby significantly impacting public health outcomes. Secondly, in our work to improve geographical accuracy, we propose a rule-based Named Entity Recognition (NER) approach to extract spatial relations related to locations mentioned in text data, evaluated using a diverse news article dataset covering various diseases. Additionally, we present an algorithm to compute geographical coordinates in the form of polygons for identified spatial relation locations, with qualitative assessments involving end-users to ensure their quality and utility.

Extracting situational awareness from social media e.g. geotagged tweets of geographically accurate event region are offering real-time insights to gauge severity of event. Finally, for situational awareness, we performed sentiment analysis using Hierarchy-based measures for tweet analysis (H-TFIDF) to understand local sentiments during the COVID-19 epidemic, evaluated with early COVID-19-related tweets from the E.Chen dataset categorized into spatial groups. Furthermore, various features including Bidirectional Encoder Representations from Transformers (BERT), H-TFIDF, term frequency-inverse document frequency (TF-IDF), and bag-of-words (BOW), were employed in spatial opinion mining to assess their significance in sentiment classification.

Keywords: Text mining, Event Extraction, One Health, Natural language processing (NLP)

[1] https://mood-h2020.eu/

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all the individuals and institutions that have supported me during my doctoral studies. Their contributions have been instrumental in the successful completion of my PhD thesis. Specially, I am truly grateful to the thesis jury for their commitment to academic excellence. Your expertise and thorough evaluation have been instrumental in enhancing the quality and shaping of this thesis.

I would like to express my deep gratitude to my thesis supervisors, Mathieu Roche, Maguelonne Teisseire and Elena Arsevska for their continuous guidance, support, and invaluable insights throughout my journey in the fields of Text Mining, Epidemic Intelligence and Data Science. Your expertise, dedication, and mentorship have been instrumental in shaping my research and academic growth. I am deeply appreciative of the opportunities you have provided me and the knowledge you have shared. Your encouragement and constructive feedback have been the driving force behind the successful completion of this thesis. Moreover, I am profoundly grateful for your continued belief in my abilities. I would say thank you for your significant impact on my academic and professional development.

I am deeply grateful to the MOOD (Monitoring Outbreaks for Disease surveillance in a data science context) project for their generous funding and support throughout my academic journey. This project's financial assistance has not only enabled me to conduct this research but has also provided valuable resources and opportunities for professional growth. I extend my sincere appreciation to the entire MOOD team for their unwavering encouragement and belief in the significance of this work. Your support has been instrumental in the successful completion of this thesis, and I look forward to contributing further to the project objectives in the future.

I extend my thanks to the researchers and fellow PhD students who have contributed to my academic journey. Their commitment, constructive criticism, and stimulating discussions in the laboratory, seminars, and conferences have greatly influenced the progress and refinement of my research. I am also grateful to CIRAD (UMR TETIS) for providing the necessary resources, infrastructure, and laboratory equipment that have facilitated my research. The institutional support and academic environment have been crucial in enabling me to carry out my work. Specially, I would like to extend my appreciation to Annie Huguet for her outstanding secretariat and administrative support. Her dedication and efficiency have been instrumental in the smooth progress of this thesis project, and her assistance is deeply valued.

I would like to express my deepest gratitude to my family, specially to my wife Hina and my adorable daughter Merha, for their unwavering support and understanding throughout my academic journey. Your love and encouragement have been my greatest motivation, and I am truly thankful for your patience and sacrifices.

List Of Abbreviations

AI	Avian Influenza
AH	Animal health
AMR	Antimicrobial Resistance
BERT	Bidirectional Encoder Representations from Transformers
BOW	Bag-Of-Words
BH-TFIDF	BERT combined with Hierarchy-based measure for tweet analysis
COVID-19	Coronavirus Disease 2019
EBS	Event-based Surveillance
EWS	Early Warning System
FAO	Food and Agriculture Organization
GIS	Geographical Information System
GPHIN	Global Public Health Intelligence Network
H-TFIDF	Hierarchy-based measure for tweet analysis
IBS	Indicator-Based surveillance
IR	Information Retrieval
MedISys	Medical Information System
NER	Named Entity Recognition
NLP	Natural Language Processing
PADI-web	Platform for Automated extraction of animal Disease Information from the web
POS	Part-Of-Speech
ProMED	Program for Monitoring Emerging Diseases
PH	Public health
spatRE	spatial relation entity
TBE	Tick-borne Encephalitis
TF-IDF	Term Frequency-Inverse Document Frequency
UMLS	Unified Medical Language System
WHO	World Health Organization
WAHIS	World Animal Health Information Database Interface

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

Introduction

CHAPTER 1

INTRODUCTION

This chapter provides an understanding of event-based surveillance systems as part of the epedimic intelligence. Furthermore, it underlines the research objectives of the thesis and the summary of contributions associated with these research objectives.

Outbreaks of infectious diseases pose serious threats to public, animal, and plant health (one health) (Kim et al., 2020). Moreover, infectious disease outbreaks affect not only one health but also the national and international economy and trade (Rees et al., 2019). Therefore, it is important to implement health surveillance methods to recognize potential infectious disease outbreaks and to minimize their associated devastating effects on affected population and indirectly on the society. In the existing literature (WHO, 2008; Rees et al., 2019), there are two main types of surveillance: 1) event-based surveillance (EBS) and 2) indicator-based surveillance (IBS). IBS uses official sources to detect important disease outbreaks (Runge-Ranzinger et al., 2008). They produce structured and reliable data, offering an extensive range of information regarding the pathogen, outbreak source, species, clinical signs, etc. As a result of official procedures, the declaration of outbreaks experiences a considerable time delay. Whereas, EBS refers to the collection of information regarding events that hold the potential to pose risks to public health reported in unstructured data sources like textual data, i.e., news articles, social media updates (WHO, 2020). As per the World Health Organization (WHO), approximately 60% of all outbreaks are identified through informal sources (Abbood et al., 2020). Both of these surveillance strategies complement one another in terms of benefits due to their unique data collection, verification, assessment, and data interpretation processes (WHO, 2008) and are treated as fundamental in constructing a comprehensive surveillance system (Balajee et al., 2021). Our research was mainly focused on EBS, whereas IBS was out of scope.

EBS is the organized process of detecting and reporting information (i.e., represented as events) to public health authorities by rapidly capturing information from different unstructured data sources (Balajee et al., 2021). It enables concerned authorities to be better prepared for endemic and pandemic disease outbreaks by functioning as a key component of an effective early warning system (WHO, 2008; Balajee et al., 2021). For information acquisition, online information sources (e.g., news articles, blogs, rumours, social media (such as Twitter, etc.), and other ad hoc reports, etc.) have gained great attention in implementing 'web-based' or 'internet-based' EBS systems (Valentin, 2020).

There are three types of EBS systems: moderated, partially moderated, and fully automated (Linge et al., 2009). The way of flow of information in these EBS systems from online data sources, e.g. news aggregators, depicts its level of automation. The final output of all these types of EBS systems is to identify and extract signals or potential events (health threat from potential disease in a certain region over time) from heterogenous data sources (Arsevska et al., 2018). In every type of EBS systems mentioned, there are certain advantages and disadvantages dependent on the level of priorities of certain factors. For instance, the Program for Monitoring Emerging Diseases (ProMED) is an example of a moderated system in which experts identify news articles, validate the content and report events (Carrion and Madoff, 2017; Yu and Madoff, 2004). The main advantage of this system is less signal-to-noise ratio (low false outbreak detection rate) due to human validation of content, with disadvantages of resource limits (experts), situational awareness and expert biases towards the events. Similarly, the Global Public Health Intelligence Network (GPHIN) (Blench, 2008) is a partially moderated system that automatically identifies a stream of thousands of news articles per day and group of experts and further moderated by group of experts to identify events. It has the advantage of automated data collection method but with the same disadvantages as ProMED. Fully automated systems include the European Commission Medical Information System (MedISys)

(Linge et al., 2010), Platform for Automated extraction of Disease Information (PADI-web) (Valentin et al., 2021) from the web and HealthMap (Freifeld et al., 2008). Unlike moderated systems, fully automated systems are faster at processing data and cost-efficient as compared to moderated systems. However, the main weakness of such systems is the higher signal-to-noise ratio as compared to moderated systems, as well as less accuracy of event information i.e., there is significantly higher rate of identifying false health threats or information associated with health threats (Cacciatore, 2021). Our research focuses only on the EBS component because there is a need for a better understanding how to better detect, extract, analyse unstructured information from EBS sources. The final output of such systems are signals or potential outbreak events.

1.1 Motivation

An event refers to a specific occurrence, pattern, or cluster of health-related incidents in a geographical region or community population that are monitored and investigated to assess their significance in terms of public health (PH)/animal health (AH) (Shakeri Hossein Abad et al., 2021). These events include outbreaks of infectious diseases, unusual increases in disease cases, the emergence of new diseases, or any other health-related incident that require high attention, investigation, and response from public health authorities (Welby-Everard, Quantick, and Green, 2020). There are five key aspects in the text, i.e., *when*, *where*, *which*, *who* and *why* that defines an event in the text (Ibrahim, 2020; Sims, Park, and Bamman, 2019; Yu et al., 2020). These aspects are important in epidemiology because these are the potential features of an outbreak. For instance, *when* is the start date of event, *where* represents the affected region, *which* is the disease, pathogen, *why* represents the cause of the disease and *who* is affected (host e.g. humans, animal) and stakeholders. There are several approaches to extract events from the textual data sources that include, i.e., rule-based approaches, supervised machine learning, semi-supervised learning, unsupervised learning, deep learning models and ontology-based approaches, etc (Mujtaba et al., 2019). Hence, it is important to enhance different aspects to obtain precise information about events, accurately reflecting the real on-ground situation.

Data quality is the foundation of effective disease surveillance, as it ensures that the information gathered is relevant, reliable, trustworthy, and up-to-date (Kington et al., 2021). Inaccurate or unreliable information can lead to misguided PH/AH responses and after assessment by epidemiologists. Therefore, data quality assurance is essential to filter out noise, filter incredible sources, to reduce bias in order to achieve appropriate decision-making related to control and surveillance.

The precision of outbreak locations, which involves identifying the exact geographical region of an event, significantly influences automated event extraction methods within EBS (Valentin et al., 2020). Achieving this precision can be challenging, as it requires discerning specific location details mentioned in the text, including complex spatial relations such as 'North of Paris' or 'near the France border'. Subsequently, after extracting the location, it is equally crucial to pinpoint the precise geographical area (point or polygon) or geospatial representation of the event. Without this level of accuracy in geographic information, surveillance efforts may lack the spatial context needed to target interventions effectively (Friis and Sellers, 2020) and may also affect the spatial epidemiology

studies (Kirby, Delmelle, and Eberth, 2017). This gap could potentially lead to the rapid spread of diseases both within and across regions because of lack of information or less accurate information on potential disease outbreaks to be assessed by risk assessors.

Spatial opinion mining on social media like Twitter provides a dynamic and real-time source of information for situational awareness during an event (Vernier, Farinosi, and Foresti, 2019). Spatial opinion mining involves analysing the content of geotagged tweets to extract opinions, sentiments, and information related to the health events (Steiger, De Albuquerque, and Zipf, 2015). Natural language processing (NLP) techniques are applied to understand the context of tweets. In the context of an event, it can reveal public sentiment regarding the severity of the outbreak, symptoms, prevention measures, personal experiences and more. Moreover, continuous monitoring and evaluation of the spatial opinion mining of geotagged tweets of event region can help in improving the situation awareness effectiveness over time.

This thesis is mainly funded by the '**MOnitoring Outbreaks for Disease surveillance in a data science context (MOOD[1])**' project. The MOOD project aims at taking advantage of data mining, analysis and visualization of health, environmental and other data to **enhance the utility of EBS**. Ultimately, MOOD is supporting the work of European and global public and veterinary health agencies and surveillance practitioners by providing existing monitoring platforms with novel features, and methodological and practical support adapted to their needs.

1.2 Objectives

Our primary research question is how can event extraction be improved within the framework of fully automated EBS Systems? The subsidiary research questions to support the main question are as follows:

1. Which strategies can be employed to enrich the data quality of news sources for ensuring reliable information for event extraction? Moreover, to which extent does enhancing the data quality of news sources contribute to more accurate event extraction within EBS?

2. How does ensuring the geographical accuracy of events impact the effectiveness of automated event extraction in EBS? Moreover, which techniques can be employed to accurately identify the location of events, e.g., spatial relation associated with locations?

3. In which ways can spatial opinion mining of social media platforms like Twitter enhance situational awareness of events over time in EBS?

[1] https://mood-h2020.eu/

1.3 Contributions

To address the objectives, we proposed three main contributions in the thesis, i.e., 1) different approaches to assess the data quality, 2) an approach to improve geographical accuracy of event, and 3) an approach for situational awareness of event as shown in Figure 1.1. The overview of these contributions are as follows:

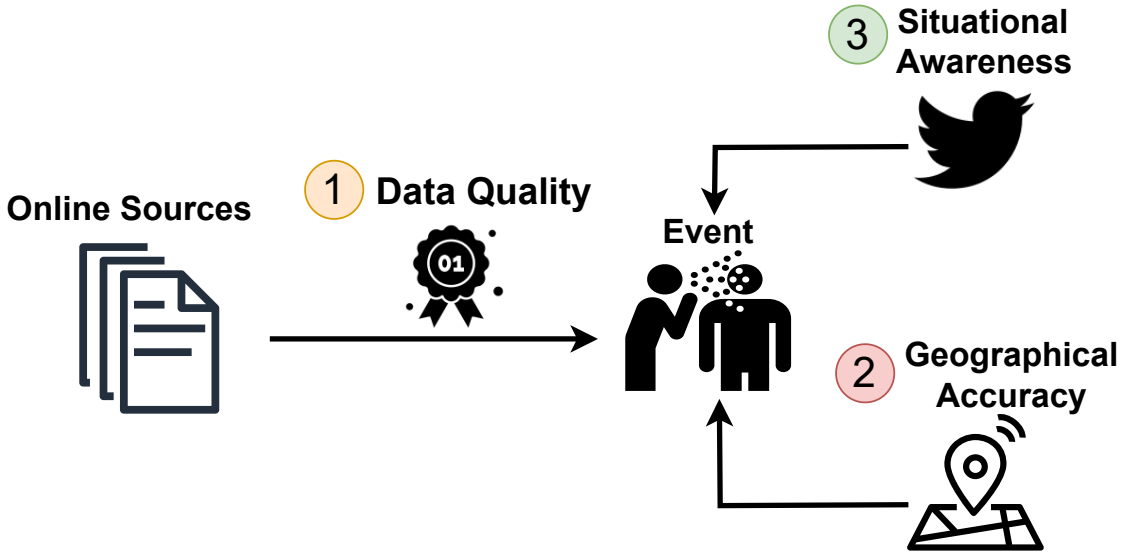


Figure 1.1: Research Focus

1. In step 1, we addressed the data quality challenges by introducing a data-driven, score-based approach as an extension of the baseline method proposed by Alomar et al. (Alomar et al., 2016). Our approach involves enriching the epidemiological context to classify news articles as either relevant or irrelevant. To assess the effectiveness of our approach, we conducted evaluations using the avian influenza news article dataset extracted from PADI-web[2]. Additionally, we introduced a machine learning-based approach that leverages news article metadata features to classify news articles as relevant or irrelevant. We conducted evaluations using the same dataset as in the previous approach. Lastly, we explored a novel perspective focused on enhancing the quality attributes of news sources themselves, as opposed to individual news articles. We proposed a text mining approach to determine whether an online source is specialized or not. Furthermore, we introduced an algorithm for identifying the geographical coverage of these news sources. In essence, we introduced several news source metadata features with the aim of improving the classification pipeline performance. The results were assessed using a news source dataset derived from the same dataset discussed in the score-based approach.

2. In step 2 referring to geographical accuracy of events, our focus was on the identification of spatial relations related to locations mentioned in text data, such as those found in online news articles, especially with regard to the event's location. Initially, we proposed a rule-based Named Entity Recognition (NER) approach to identify and extract spatial relations within the

[2] https://padi-web.cirad.fr/en/

text. These spatial relations were linked to various locations. To evaluate the effectiveness of this extraction phase, we employed a diverse news article dataset covering different diseases. Subsequently, we introduced an algorithm to calculate the geographical coordinates, represented as polygons, for the locations identified in the first step. These polygons were generated for 19 different spatial relations across nine distinct European cities. To ensure the accuracy and utility of these polygons, qualitative evaluations were conducted with groups of end-users.

3. In step 3 concerning situational awareness of event, we focused on the sentiment analysis performed on terms selected by Hierarchy-based measure for tweet analysis (H-TFIDF) (Decoupes et al., 2021) for spatial tweets groups to know local situations during the ongoing epidemic COVID-19 over different time frames. To evaluate the effectiveness of our approach, we utilized early-stage COVID-19-related tweets sourced from the E.Chen dataset (Chen, Lerman, and Ferrara, 2020). These tweets were categorized into spatial groups representing regions. Secondly, we performed spatial opinion mining utilizing diverse features, which encompassed Bidirectional Encoder Representations from Transformers (BERT), H-TFIDF, term frequency-inverse document frequency (TF-IDF), and bag-of-words (BOW). Our objective was to thoroughly examine the significance of these features in the context of sentiment classification. The outcomes were assessed using the same E.Chen dataset (Chen, Lerman, and Ferrara, 2020).

The remainder of the thesis is structured as follows: concerning contribution **1**, Chapter 2 outlines the proposed methodologies to ensure the data quality of news sources for improved event extraction. Subsequently, referring to contribution **2**, Chapter 3 describes the methodologies to improve geographical accuracy of the event. Afterwards, with reference to contribution **3**, Chapter 4 details the approaches to understand the situation of affected regions, allowing to understand the event impact. Finally, Chapter 5 summarizes the contributions and perspective of this research.

Part II

Ensuring Data Quality of Online News Sources

CHAPTER 2

ENSURING DATA QUALITY OF ONLINE NEWS SOURCES

This chapter examines various dimensions of assessing data quality in online news sources. Additionally, it elaborates the assessment of data quality of news articles and news sources using different text mining and machine learning approaches.

2.1 Introduction

EBS is the organized process of detecting and reporting information on potential health threats and hazards (i.e. represented as events), most commonly as outbreak/cases, to healthcare authorities by rapid capturing of information from different unstructured data sources (Balajee et al., 2021). It enables health authorities to be better prepared for different disease outbreaks by functioning as a key component of an effective early warning system (WHO, 2008; Balajee et al., 2021). For information acquisition, online information sources (e.g. news articles, blogs, social media e.g., Twitter), and other ad-hoc reports (e.g., access to laboratory reports, electronic health records, expert networks and exchanges) are preferred in EBS systems (Freifeld et al., 2008; Yu and Madoff, 2004; Blench, 2008; Abdelmalik et al., 2018) as compared to traditional data collection methods which are labor-intensive (Cato, Cohen, and Larson, 2015; Lin et al., 2010).

More than 60% of the first signals on new outbreaks come from news sources (Abbas et al., 2022). This finding underscores the role of news sources as early indicators of potential disease outbreaks, highlighting the need for data quality assurance in EBS systems. Online news information is diverse and collected from heterogeneous online data sources, it gets crucial to verify this unstructured information that can pose serious threats to public health to avoid misinformation (i.e. a piece of information that is false having no evidence) (Zhou et al., 2021) and disinformation (i.e. intentionally generated false information) (Bastick, 2021). These sources are preferred as they improve timeliness and detection of outbreak-related information. To avoid false information on potential outbreaks, it is important to verify the information associated with the online news at two levels, i.e., news article, and news source in order to get accurate and reliable information.

In this work, we will discuss different approaches and the data quality attributes in order to ensure relevant and reliable information detection in EBS systems. These approaches will ensure the data quality at news article level (direct information) and news source level (authorization and credibility) of the information provider. The detail objectives of the work are discussed in the subsequent Section 2.3.

2.2 State-of-the-art

News articles of online news sources serve as vital data streams in disease surveillance systems, enabling real-time outbreak detection and informed public health interventions (Wilson and Brownstein, 2009). The aim is to leverage existing research in order to ensure the quality of data by evaluating the reliability, relevance, and accuracy of news articles and news sources, which serve as the primary information source for Early Warning Systems (EWS).

Research on data quality, which is crucial for evaluating online news sources and constructing EBS systems, began in the 1990s. Wang and Strong (Wang and Strong, 1996) defined data quality as “the information which is fit for use". Moreover, the dimensions for assessing data quality are a set of attributes representing single or multiple aspects of data, including the currency, accuracy,

relevance, authority, and purpose of information (Batini, Scannapieco, et al., 2016). Data quality of news sources in the context of an EBS system refers to the reliability, relevance, timeliness, bias and accuracy of the data collected and used for surveillance purposes (Craig et al., 2016). Therefore, it is important to ensure these dimensions to avoid reducing the false alerts for disease surveillance. Reliability refers to the trustworthiness and credibility of the information being reported by news sources. Therefore, reputable news sources are bounded with strict fact-checking and verification processes before publishing information (Conroy, Rubin, and Chen, 2015). In the existing literature, there is a degree of overlap identified among the data quality dimensions and their assessment methods. For instance, Mandalios and Jane (Mandalios, 2013) used the following assessment criteria to evaluate online sources: purpose, authority and credibility, accuracy and reliability, currency and timeliness and objectivity. In addition, Zhu and Gauch (Zhu and Gauch, 2000) proposed six quality metrics, including currency, availability, information-to-noise ratio, authority, popularity and cohesiveness, for investigating the assessment of online sources. Additionally, Nozato and Yoshiko (Nozato, 2002) stated that the timeliness, depth, reputation, and accuracy of online sources are the most important data quality dimensions. Another study (Bachmann, Eisenegger, and Ingenhoff, 2021) used the quality attributes of the respondents and general perception of the news sources for news classification. Moreover, another study (Bhuiyan et al., 2020) investigated news credibility assessments by comparing crowds and expert opinions to understand the differentiation in the rating of the source. Relevant news sources offer information about better understanding of disease, mode of transmission, symptoms, level of risks in timely manner (Slaughter et al., 2005). Accuracy of information from news sources depends on several factors, i.e., source credibility, transparency of information, cross-referencing across credible sources, bias and experts input, etc (Di Domenico et al., 2021). Moreover, these quality dimensions are dependent on multiple sub factors which are needed to be evaluated.

In addition to the data quality dimensions and their assessment methods as described above, there exist different studies that employ various state-of-the-art techniques (Cato, Cohen, and Larson, 2015) based on text mining, information retrieval, machine learning, deep learning and knowledge representation graphs for assessing the relevance of news sources. For example, Essam and Elsayed (Essam and Elsayed, 2020) defined a specialized information retrieval technique by assessing the topics and subtopics of the news to identify highly relevant background articles. Elhadad et al. (Elhadad, Li, and Gebali, 2019) adopted a machine learning technique for extracting features from the news content and prepared a complex set of metadata for identifying the credibility of the news sources. Another study (Islam et al., 2020) proposed a method based on deep learning techniques to find patterns in news sources to avoid false information, rumours, spam, fake news, and disinformation. Moreover, Hu et al. (Hu et al., 2006) analysed the visual layout information of news homepages to utilize the mutual relationship that exists between news articles and news sources using a semi-supervised learning algorithm. However, this approach is not only based on a computationally expensive learning model to establish a relationship between new articles and sources but is also limited to small news corpora. To address this limitation, a system named MediaRank was designed (Ye and Skiena, 2019) to incorporate large datasets for measuring the quality of news sources by a mix of computational signals reflecting peer reputation, reporting bias, bottom-line pressure, and popularity. A study employing the application of knowledge graphs by Rudnik et al. (Rudnik et al., 2019) implemented a method using a Wikidata knowledge base for generating the semantic annotation of news articles to filter

relevant news articles.

Metadata refers to structured information that provides details about various forms of data such as images, multimedia, books, and scientific articles (Vellucci, 1998; Riley, 2017). In a research study, a metadata approach is used for categorization of historic newspaper collection. This metadata was collected by analysis of fined-grained search patterns within the newspaper collection (Bogaard et al., 2019). In another research study, two primary types of metadata are introduced for digital news article archives (Khan et al., 2016). These metadata categories consist of explicit metadata (linked to news articles) and implicit metadata (crucial metadata embedded within the content), which serve the purpose of searching for news articles within archives. In another research, Bidirectional Encoder Representations from Transformers (BERT) model is proposed for the merging of text representations with metadata and knowledge graph embeddings, specifically encoding author-related information for book classification task (Ostendorff et al., 2019). In another research study, the DANIEL system, which is a text genre-based Information Extraction (IE) system, was proposed to efficiently filter out irrelevant documents in epidemic surveillance, particularly for low-resourced languages (Lejeune et al., 2015). The benefit of this method was to increase the coverage across a variety of languages at a low cost, rather than focusing solely on optimizing results for a specific language. In another research study, a framework is proposed for the profiling of cities through automatic extraction and analysis of metadata of news articles using data mining and machine learning techniques (Cascone, Ducange, and Marcelloni, 2019). The cities profiles were characterized in terms of criminality, events, services, urban problems, decay and accidents. Another research proposed a neural network based approach for multi-label document classification, in which two heterogeneous graphs are constructed i.e., metadata heterogeneous graph for modelling various types of metadata and their topological relations and label heterogeneous graph constructed based on labels hierarchy and their statistical dependencies (Ye et al., 2021). In another research study, an approach was proposed for multi-label document classification using the available metadata for evaluating the performance of metadata-based features compared to content-based methods (Sajid et al., 2023). The proposed technique has been assessed for two diverse datasets, namely, from the Journal of universal computer science (JUCS) dataset and dataset of the articles published by the Association for computing machinery (ACM). Another research proposed transformer-based models for finding the documents that contain epidemic events and event extraction, with focus on high-resource and low-resource languages (Mutuvi et al., 2020).

A study (Alomar et al., 2016) that is based on a domain-oriented news article classification problem, is taken as fundamental to this research and used to develop the proposed approach.The research discussed a direct method (i.e., identification, review, and evaluation of known sources to find relevant information sources) and an indirect method that assesses quality attributes of news content and metadata. To fill the research gap, we proposed the assessment of the data quality at two levels, i.e., 1) News article, 2) News Source. In the first step, we proposed automatically extraction of quality attributes from the metadata and content of the news article and assesses the quality of news articles. The second step is to compliment the first step, we additionally identify quality attributes associated with news source and the possible one to extract automatically and evaluate the quality at source level.

2.3 Objectives and Contributions

The main objective is to assess quality attributes to ensure the relevant information in EBS systems. The subsidiary research questions to support the main question are as follows:

1. What are the quality attributes of the news article identified from metadata and content, and how to evaluate these attributes of news article in EBS systems?
2. What are the external quality attributes to ensure the quality of news source, and evaluate these attributes to verify news source in EBS systems?
3. How the combination of attributes at two levels ensures the quality of information provided to EBS systems?

The following contributions are included to address the objectives are as follows:

1. We proposed a data-driven score approach to assess the quality of online news articles in EBS. The online news articles are assessed through metadata and content of the news articles in order to filter relevant news articles.
2. We proposed machine learning approaches to classify relevant news articles in EBS through metadata features.
3. We proposed several quality attributes (source metadata) of news sources and their automated extraction. Moreover, we analysed the impact of news article metadata and source metadata towards classification of relevant news in EBS.

2.4 Datasets

In order to evaluate our contributions, we proposed two datasets in this chapter. The details of these datasets are discussed in the subsequent sections.

2.4.1 EBS News Classification Dataset

This dataset consists of news articles related to Avian-Influenza (AI) events extract from PADI-web, with 317 articles classified as relevant and 374 articles classified as irrelevant. This dataset is manually labelled by epidemiologists for the news classification task. The dataset is limited in size because it involved manual annotation, which required human effort. Each entry in the dataset includes the following information: ID, title, text, URL, language, source language, creation date and class (label) of the news article. *Relevant* class contains AI outbreaks. While *irrelevant*, does not contain an event of AI outbreaks.

2.4.2 EBS Source Relevance Classification Dataset

The dataset contains the news sources detected by the PADI-web relevant (detected articles for avian-influenza outbreaks) or irrelevant with no event. This dataset is derived from the news article dataset in Section 2.4.1. Each entry in the dataset includes the following information: news_source, source_description, relevant_frequency, irrelevant_frequency, annotated_category, annotated_geographical_coverage and confidence (label) on the news source. For the experiments, confidence is a binary variable employed to assess the source classification task. The different approaches to assess the data quality in the context of EBS are discussed in the subsequent sections.

2.5 A data-driven score-based approach to assess the data quality of online news articles

Data quality measures (DQM) are the metrics to rank elements based on their quality, facilitating the identification of reliable news sources in terms of relevance, accuracy, and reputation (Vaziri and Mohsenzadeh, 2012). Various criteria exist for computing the data quality of online news sources, including metadata attributes and attributes extracted from the content of the news article. Alomar et al. (Alomar et al., 2016) proposed the measures of metadata score (MS) derived from extracted metadata, content score (CS) computed from extraction of various attributes inside the content of news article. Our approach proposed a new measure called epidemiological entity extraction score (E3S) calculated using weighted named-entities, specifically spatio-temporal entities related to epidemiology. We choose the unsupervised data-driven score approach for news classification because it helps us understand how different factors affect the results. Moreover, the dataset size and its focus on specific information can also be reasons for choosing this approach. The overall process pipeline, including all the components, is depicted in Figure 2.1.

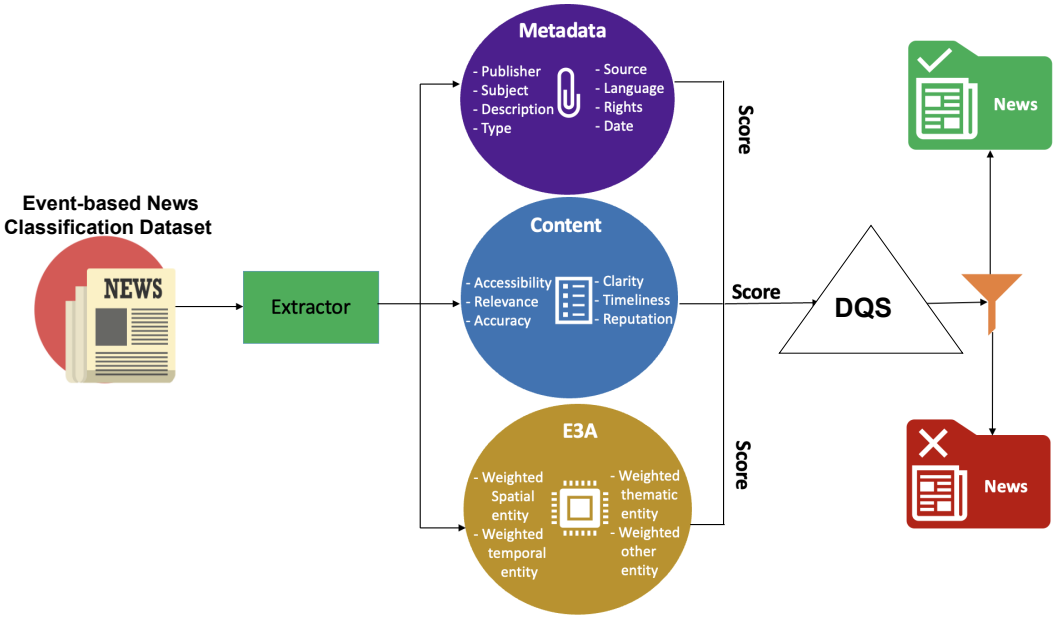


Figure 2.1: Process Pipeline: A data-driven score-based approach to assess data quality

The details of these measures are as follows:

1. **Metadata Score (MS):** Metadata plays a crucial role in rapidly retrieving information (Chan et al., 2001). Search engines rely on metadata to quickly identify relevant results from countless online sources (Chan et al., 2001). Reliable online sources define specific tags that are analysed and retrieved by search engines to deliver ranked results.

   When assessing the metadata of news sources, various metadata attributes are taken into account (Alomar et al., 2016). The different kinds of metadata of the news article are as follows:

   (a) **Title:** The title of the news article, which provides a brief summary or description of the story.

   (b) **Author:** The name of the individual or organization responsible for creating or reporting the news.

   (c) **Publication Date:** The date when the news article was published or made available to the public.

   (d) **Source:** The name of the news organization or publication that produced the article.

   (e) **Description:** A brief summary or abstract that provides an overview of the article's content.

   (f) **Keywords/Tags:** Relevant keywords or tags assigned to the article to describe its topic, subject, or themes. These help in categorization and searchability.

   (g) **Language:** The language in which the news article is written.

   (h) **Type:** Type of the news article, e.g. topic or category.

   (i) **Rights:** Copyright of the source.

We have identified and selected metadata attributes that are both relevant and widely accessible in most news articles, e.g., title (subject), description, author, date (publication date), type (topic of news article), rights (copyrights), source (URL of news article) and language. The extraction of each metadata attribute from the online news article allows for the computation of the metadata score (MS) using the formula proposed by (Alomar et al., 2016). MS is defined based on 8 metadata attributes, which is as follows:

$$MS = \sum_{n=1}^{8} Presence(attribute_n) \tag{2.1}$$

$$Presence(attribute_n) = \begin{cases} 1, attribute \notin metadata \\ 2, attribute \in metadata \end{cases}$$

We used the same scoring mechanism as discussed in the baseline approach. The scoring mechanism for the metadata attributes in this approach is, i.e., a score of 2 is assigned for the presence of an attribute, and a score of 1 is given for its absence.

2. **Content Score (CS):** Online news content comprises information about one or more events presented in the form of electronically available information for the public (Westerman, Spence, and Van Der Heide, 2014). The quality of news articles is ensured through considerations of currency, timeliness, relevance, accuracy, and impact (Mandalios, 2013). Therefore, it is important to analyse the content of news sources and extract the quality attributes to quantify the data quality score.

   The extraction of quality attributes from the content are achieved through an automated process. Various content attributes are taken into account for assessing the content of news sources (Alomar et al., 2016). The quality attributes selected for content assessment include accessibility, relevance, accuracy, clarity, timeliness, and reputation. **Accessibility** is the preliminary step of analysing content to access an online news source. Therefore, it is to ensure that the online news source is available and accessed without any barrier. Moreover, it is also possible that it is available but with the restricted access such that it is not possible to access with any browser or external tools. Despite, in some cases, it is also possible that the online news sources are unavailable for future use in digital form. **Relevance** in the baseline approach is proposed by identifying some epidemiological attributes i.e. affected hosts, agent that affects the host and the location of the affected host. This is not the same as the Relevance of the news article. We shortlist *spaCy* (Vasiliev, 2020) natural language processing (NLP) python library is used to perform named entity recognition (NER) to extract locations, hosts, and agents respectively. spaCy is easy to adapt and customize its models and components to suit specific NER requirements. This model can be utilized for improving the NER accuracy. Some examples of the hosts and agents of disease avian-influenza are (chicken, pigs, horse, duck, goose etc) and (H5N8, H5N1, highly pathogenic avian-influenza etc) respectively. **Accuracy** is dependent on the information provided by the news sources that are the facts that can be verified and validated. In the context of EBS, it could be that the news content provide information about any health risk, outbreak information, or it could be the number of cases respectively. Alternatively, poor relevance can have poor accuracy, but not vice versa. **Clarity** is the quality

of being logical, consistent and completely understandable in terms of content that is similarly reflected in the metadata. Clarity of the article is poor if only the title is available in the metadata, and clarity is adequate if other metadata attributes are available (Alomar et al., 2016). A good clarity is if the subject, description, type etc. are available in the metadata of the news article. **Timeliness** of the news article ensures that the content of the news article relates to the current context of the events. Otherwise, the claims may not be considered, or it may be a wrong interpretation. Timeliness is the time of an outbreak saved by detection in EBS relative to the onset of the outbreak (Jafarpour et al., 2015). Furthermore, Timeliness (days) is calculated by the following equation (Kleinman and Abrams, 2006):

$$Timeliness[days] = T_{alarm} - T_{onset} \tag{2.2}$$

where T_{alarm} is the time of the event reported in the event-based system and T_{onset} can be validated from the health information databases. Lastly, **Reputation** of news sources is extracted using *MediaRank* (Ye and Skiena, 2019) algorithm which is calculated on multiple factors i.e. popularity, peer reputation, reporting bias and breadth and bottom-line pressure. E.g. the general reputation ranking using *MediaRank* (Ye and Skiena, 2019) of *New York Times* is '1' and BBC is '5'. Therefore, the general reputation of the news source has the impact on the content quality, as it is computed by considering multiple factors. After the extraction of these attributes, the CS is computed using the following formulas in the baseline approach(Alomar et al., 2016):

$$CS = \sum_{n=1}^{6} Presence(attribute_n) \tag{2.3}$$

$$Presence(attribute_n) = \begin{cases} 1, Not\ available \\ 2, Partially\ available \\ 3, Available \end{cases}$$

We adapt the same scoring mechanism proposed by baseline approach. We shortlisted 6 attributes associated with the content. The interpretation of the CS is '1' means the attribute is not available, e.g., if the news article is not accessible. Moreover, '2' score represents the attribute is partially available, e.g., relevance is dependent on host, agent so if one of them is not available, then it is said to be partially available. Lastly, '3' score represents that attribute is completely available, e.g., if the news article is completely accessible online. Subsequently, the next step is the main contribution of this approach to extract the relevant contextual information from the online news article to enhance the baseline approach.

3. **Epidemiological Entity Extraction Score (E3S):** Event extraction and early warning detection are the key components of EBS (Organization et al., 2014). An event is a verified set of processed epidemiological information of an outbreak (Arsevska et al., 2018). It contains attributes such as location, occurrence date associated with epidemiological entities such as disease or unknown syndrome, symptoms, hosts, agents, etc. (Arsevska et al., 2018). More precisely, this information is available in text in the form of spatio-temporal entities (when,

where) and epidemiological entities (which) i.e. disease, host, agent, symptoms etc. Furthermore, these attributes are extracted from text using NLP techniques. The measure (E3S) is dependent on extracted spatial, temporal and epidemiological information within the news articles.

In this measure, the title and content of a news article are processed and then named-entities are extracted. It is not sufficient to extract epidemiological (thematic) entities from state-of-the-art Name-Entity recognition (NER) techniques. In our approach, we categorized these entities into spatial, temporal, thematic and other entities. A rule-based approach is adapted to extend spaCy NER for extracting and classifying thematic entities such as hosts (e.g. humans, birds, pigs) that are associated with the disease and pathogens or agents e.g. H5N1, H5N8, HPAI, etc. spaCy was the first choice for NER task due to its balanced nature in terms of speed, accuracy, pretrained models, and ease of use as compared to other NLP libraries like NLTK, Stanford NLP. After extracting named entities from the title and content of news articles, weights are assigned to these categories depending on availability in title and content of the news article. It results into quantifying their epidemiological context in relation with its corresponding title and content. We named the resulting spatial, temporal and thematic entities as relevant entities in the context of a particular EBS (i.e. specific to proposed work) are termed as 'Contextual Entities (CE)'. For instance, *spatial entities* (Leidner and Lieberman, 2011) are the names of the geographical or spatial location available in the text. Moreover, *temporal entities* (Pustejovsky et al., 2003) are the information of date, time and duration available in the text, *thematic entities* (Chang and Manning, 2012) are the information about health related terms in the text. whereas, the remaining identified entities are labelled as 'Non-contextual Entities (NE)'. The weights assigned to the entities are calculated by the following equations:

$$EntityWeight_n = \begin{cases} 2, CE \in Title\ Sentence \\ 1.5, CE \in Content \\ 1, NE \end{cases}$$

$$E3S = \sum_{i=1}^{n} CE\_Weight / E\_Weight \tag{2.4}$$

The weights are assigned based on two criteria i.e. 1) title and description of news articles, 2) types of entities. Double weights (i.e. 2) are assigned to these entities because of their occurrences in the title of the news articles, as title is the most important element of the news article, e.g., mostly event sentence is available in the title of the news article. Whereas, a weight of 1.5 is assigned to each CE entity based on their occurrences in the content of the news article. CE gives the contextual information of the event related information, so more weight is assigned as compared to other information in the news article. Lastly, a weight of 1 is assigned to each NE regardless of their occurrences in the title and content of the news articles. The E3S is calculated as the sum of CE_Weight to the sum of E_Weight (weight of all entities) in the news article, i.e. title and content. The resulting E3S provides weights of the news article in the epidemiological context having more chances to detect events.

The Word cloud visualization provides the most frequently used relevant words using different

Figure 2.2: Word Cloud of Extracted CEs

font sizes, indicating their occurrence frequency (Lohmann et al., 2015). Figure 2.2 visualizes the occurrences of CEs in the Event-Based News Classification Dataset. We applied these score measures to classify the AI news dataset into relevant and irrelevant news. The results of these experiments are discussed in the subsequent section.

4. **Data Quality Score (DQS):** We used DQS for classifying the news article into relevant and irrelevant. The data quality score (DQS) is computed as the average of MS, CS and E3S.

2.6 Metadata-Based Machine Learning Classification

To compliment the first approach, we proposed a machine learning approach to classify online news through metadata features. It is worthwhile to explore an additional perspective: the assessment and assurance of the external quality attributes associated with the sources of news data. This aspect is instrumental in understanding how news sources have historically contributed information to EBS systems and, in doing so, can significantly enhance the EBS system effectiveness. It serves as a mechanism to ensure that the information incorporated into these systems originates from reputable and dependable news sources.

In our contribution, we address mainly two types of metadata associated with the news: a) News article metadata and b) News source metadata. News article metadata is readily available and is directly associated with the news article, providing crucial information about the article. On the other hand, news source metadata, which can be vital for the EBS context, is not directly accessible. This kind of metadata is produced through exogenous information or induced information about the news sources. We analysed both metadata to improve the data quality for the classification task. We divided this contribution into mainly two different classification tasks, i.e., 1) Metadata-Based News Article Classification and 2) Metadata-Based News Source Classification. The details of these tasks are as follows:

2.6.1 Metadata-Based News Article Classification

Metadata in news articles helps in organizing and managing news content within the context of EBS. It enables efficient search and retrieval of articles based on specific criteria, such as date, topic, or source. In our proposed approach, we extracted the above-mentioned metadata through web scrapping, we will use this metadata as features for machine learning model.

Data preprocessing is an essential step in building machine learning models with text data. It involves cleaning and transforming the raw text data into a format that can be easily understood and processed by machine learning algorithms (Maharana, Mondal, and Nemade, 2022). Some preprocessing techniques include lowercasing, removing stop words, stemming etc. For instance, we applied stop word removal on metadata textual features i.e., title, description by removing stop words from the text. Subsequently, we applied stemming and lemmatization techniques using *spaCy* on metadata features i.e., title, description to standardize the text. This text standardization converts the word into its base form by removing prefix, suffix or reduction of the word so that it could be easier to analyse by models. Additionally, we applied URL Tokenization on specific metadata features such as 'URL' and 'Source' by converting the URL into valid tokens. These tokens of source and URL can be useful for machine learning model for the classification task. In order to tokenize, we established a regular expression to split the URL into words or tokens. For instance, the URL "http://www.sample.com/level1/index.html?id=1234" is split into valid words i.e., 'http', 'www', 'sample', 'com', 'level1', 'index', 'HTML', 'id', '1234'. At the end of this step, we have a set of features for model from the metadata attributes.

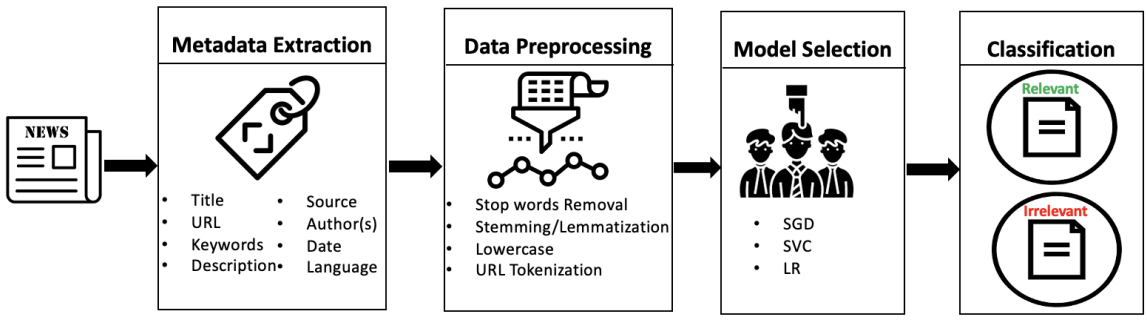


Figure 2.3: Workflow: Classification through Article Metadata

In next step, we performed experiments with several machine learning models i.e. Logistic Regression (LR), Support vector classifier (SVC) and Stochastic Gradient Descent (SGD) Classifier for the classification of relevant news articles. The idea is to see which article metadata features are important for the classification. Among the three models, SGD was performing slightly better among SVC and LR classifiers using metadata features. Stochastic Gradient Descent (SGD) Classifier is a linear classifier that is efficient and can handle sparse data well (Gite et al., 2023; Prasetijo et al., 2017). It is often used in text classification problems where the number of features is limited as compared to the size of the dataset. The goal of this approach is to classify relevant news articles in a more resource-efficient manner. The process pipeline for the approach for classification through

metadata of news article is shown in Figure 2.3.

2.6.2 Metadata-Based News Source Classification

In the second task, we identified news source metadata features for the source classification task. These source metadata include source category, e.g. specialized or generalized, geographical coverage, media bias and topic coverage. In the context of EBS, analysing news source metadata can be helpful in identifying authoritative and specialized sources (Gisondi et al., 2022) that are more likely to provide accurate and relevant information about specific events or topics (Zhang et al., 2022). The details of these source metadata features associated with the news sources are as follows:

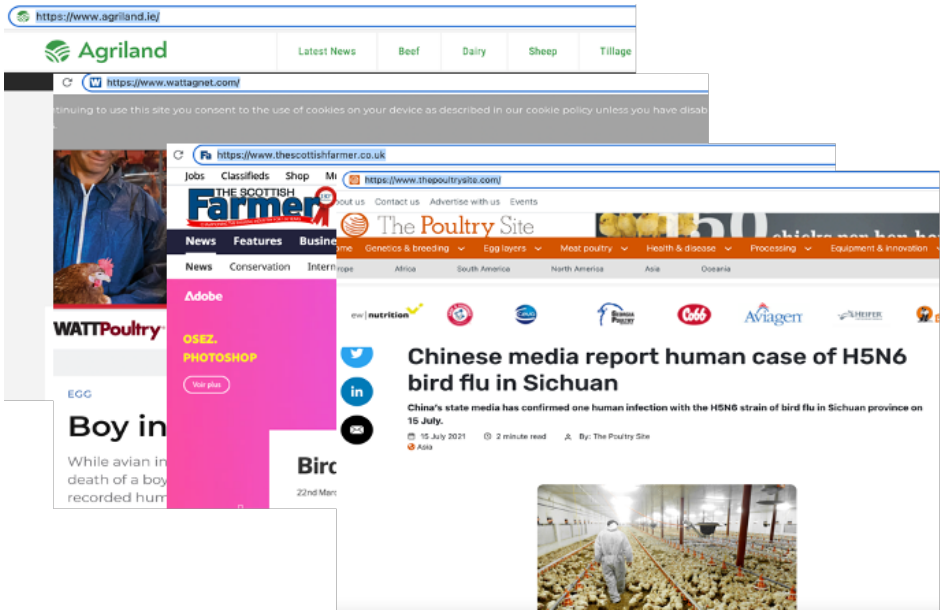


Figure 2.4: Specialized Sources of Poultry and Agriculture

1. **Source Specialization:** In the context of EBS systems, apart from government official sources, dedicated health sources mainly reports about public health, disease outbreaks and emergency preparedness. For instance, **Outbreak News Today**[1] is an online news source that reported the news about various infectious diseases and their outbreaks. In Figure 2.4, we show different specialized online news sources that mainly report news about agriculture and poultry issues.

   The automatic news source categorization is not a straight forward task. In order to achieve it, the first step is to extract the title and description information of the news source. The title and description of the news sources are extracted through web scrapping. Subsequently, *Google Translate API*[2] is utilized to check the language of the news source from the text of title and description. Afterwards, the text of the title and description is further translated using *Google Translate API* in case of non-English languages news text of title and description. Google Translate API was selected as the preferred choice due to its renowned reputation for delivering accurate translations, distinguishing it from other freely accessible libraries. In the existing

[1] https://outbreaknewstoday.com/
[2] https://py-googletrans.readthedocs.io/en/latest/

literature, clustering is one of the most famous text mining technique used for categorization of textual documents (Tandel, Jamadar, and Dudugu, 2019; Jacksi et al., 2020). Because of its unsupervised nature, it is often preferred for categorizing text documents. K-means is a straightforward and easy-to-understand clustering algorithm, preferred for simpler and quick clustering task. Therefore, we applied k-means clustering technique using textual features in title and description for source categorization, i.e., Specialized and Generalized. For clustering, we used a specific dictionary of the terms relevant in the context of avian-influenza disease. Figure 2.5 shows the flowchart to classify the news source into generalized and specialized category. For instance, specialized category news sources are livestock, agriculture etc, whereas generalized news sources report about different topic's news. The final output of the flow chart is source category (specialized/generalized).

2. **Geographical Coverage:** Local news sources excel in providing immediate, detailed information about health events at the community level, emphasizing the local impact and response. On the other hand, international source provides more global perspective with a focus on comprehensive coverage and analysis of national and international health events. Therefore, the level of information and timeliness may vary in both cases. Due to the level of information and timeliness, it is important to take into account this aspect about news sources. Figure 2.6 shows geographical coverage of three different web sources. Examples of geographical coverage of MidiLibre (1) as *local* news source of France, Farmers (2) as *national* news source of the United Kingdom (UK) and CNN (3) as *international* news source.

   To our knowledge, there is no such method to automatically extract the geographical coverage of the news source. However, by analysing different news sources, we found some pattern to identify the geographical coverage of the news source. Geographical coverage is usually available in the menu of main page of the news source websites. The menu contains geographical references e.g., world, country names, region names, city names etc. By following these patterns, we developed our custom algorithm to extract this information using web scrapping (the technique to extract the information from web pages) and NLP techniques. The steps followed to extract the geographical coverage are as follows:

   - Select the URL (home page) of the news source.
   - Analyse the webpage of the news source, and extract the *locations* from the menu items of the webpage.
   - If the *locations* are continents or countries, it is said to be an International news source.
   - If the *locations* are cities, then find its province/region using geocoding API (geopy[3] python library).
   - If the geocoded cities belong to a single region/province/state (they are often similar) can be a local news source. If it belongs to multiple regions of the same country, it can be a national news source.

[3] https://pypi.org/project/geopy/

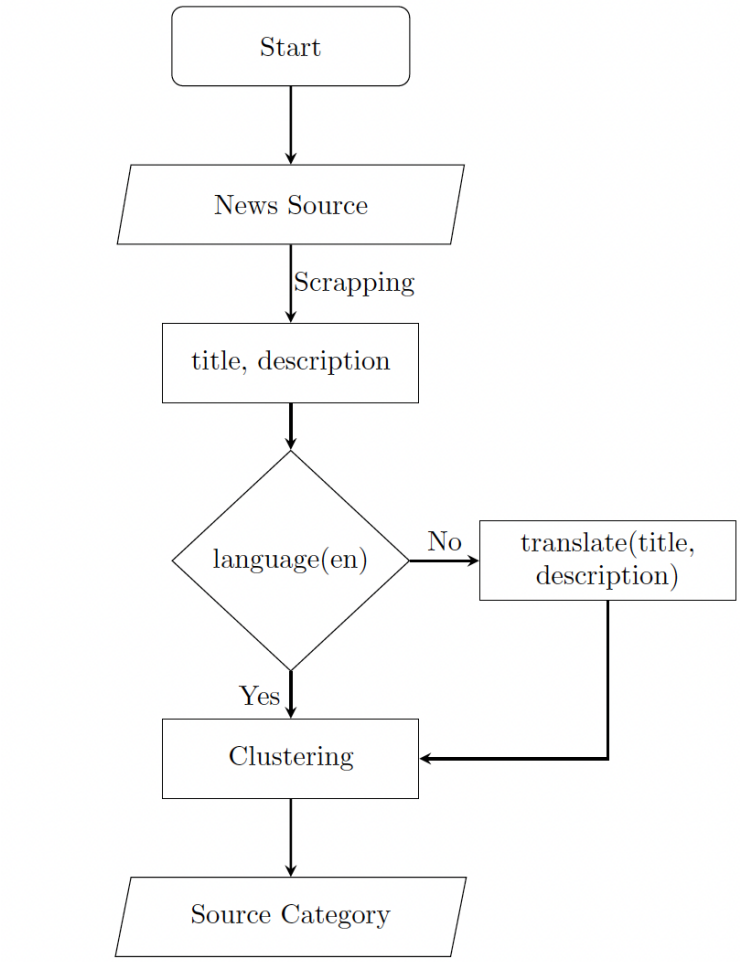


Figure 2.5: Flowchart of Source Categorization

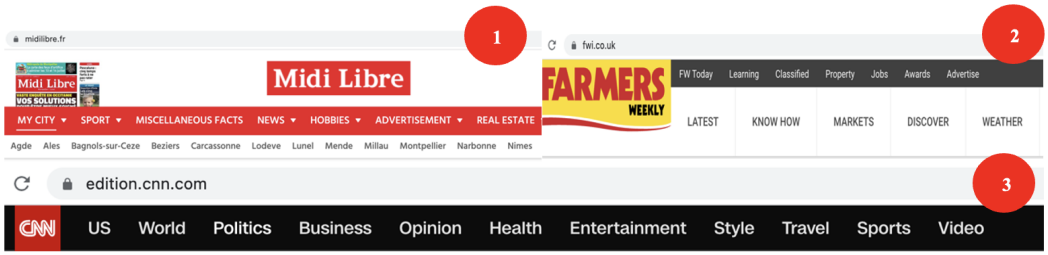


Figure 2.6: Geographical Coverage: Local, National and International News Source

3. **Media Bias:** Biased reporting may exaggerate the severity of the health situation or downplay it based on the agenda of the news source (Xu, 2023). This can lead to public confusion and panic, resulting in difficulties in effective public health responses. In order to see the media bias, World Press Freedom Index (wpfi) is an assessment measure of press freedom and the level of media independence in countries around the world (Berlinger et al., 2022). This indicator shows the fairness of media reporting. The aim is to assess media freedom, highlight countries where press freedom is restricted or violated, and the promotion of free

and independent media. Figure 2.7 shows media freedom index examples of countries, e.g., Netherlands, Norway and Nigeria.

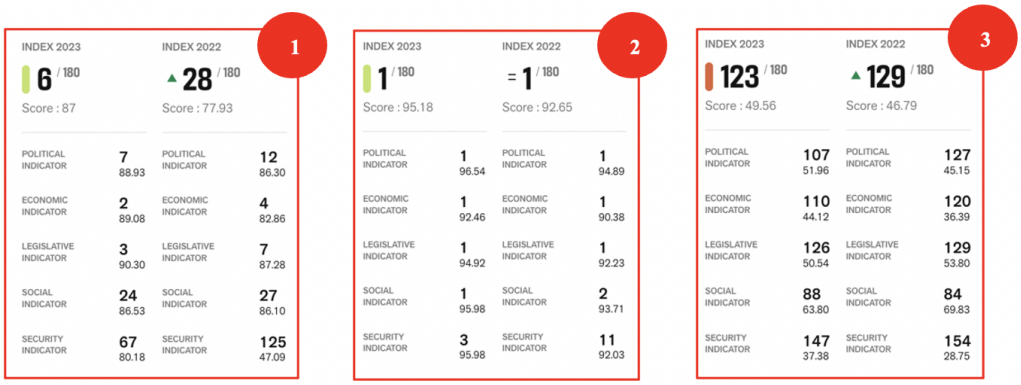


Figure 2.7: World Press Freedom Index of the Netherlands, Norway, Nigeria

The wpfi freedom index for the countries are automatically extracted through web scrapping from Reporters without border (RSF) website (*Reporters Without Borders RSF* 2012).

4. **Topic Coverage:** Topic specific news sources play significant roles in disease outbreak detection, monitoring, and response. For instance, agriculture related news sources often report on crop diseases or livestock illnesses that can serve as early warning signs of potential zoonotic diseases that can be transmitted to other humans and animals. Livestock-focused news sources report on diseases affecting animals, including those that may have zoonotic potential. For the experiments, we curated this information about the news sources from SimilarWeb [4] which is a website analytic tool. The experiments performed with results are discussed in the subsequent section.

Data and Software Availability

The whole workflow in this chapter is divided into two main approaches, i.e., 1) Data driven approach for news classification, and 2) Metadata approach for news classification. The code, datasets and results are available at GitHub repository [5].

2.7 Results

The results are obtained for the following approaches: 1) A data-driven score-based approach, 2) Metadata-based article classification, and 3) Metadata-based source classification. The specifics of these results are outlined below:

[4] https://www.similarweb.com/
[5] https://github.com/mehtab-alam/data_quality.git

2.7.1 A data-driven score-based approach

The news classification with score-based approach is evaluated though precision, recall, and the F-Score measures (Hakala and Pyysalo, 2019; Resnik and Lin, 2010). The definitions of precision, recall, and the F-Score are as follows (Goutte and Gaussier, 2005):

$$Precision = \frac{Correctly\ Relevant\_News\_Articles\ Classified}{Total\ Relevant\_News\_Articles\ Classified} \tag{2.5}$$

$$Recall = \frac{Correctely\ Relevant\_News\_Articles\ Classified}{Total\ Relevant\_News\_Articles\ in\ Dataset} \tag{2.6}$$

$$F - Score = 2\ \mathrm{X}\ \frac{Precision\ \mathrm{X}\ Recall}{Precision\ +\ Recall} \tag{2.7}$$

Table 2.1 shows the precision, recall, and F-score for different components scores (MS, CS, E3S, DQS) used to classify news articles into relevant (outbreak-related news articles) and irrelevant (no outbreak event) categories (see Section 2.4.1). The results show varying performance across these scoring methods. The Metadata Score demonstrated moderate recall (0.7) and precision (0.44), with an F-Score of 0.54, suggesting a balanced but not highly effective performance. The Content Score, on the other hand, exhibited a better precision (0.71) but very low recall (0.1), resulting in a low F-Score of 0.14, indicating a significant problem with false negatives. Furthermore, the Epidemiological Entity Extraction Score (E3S) performed reasonably well, with a moderate F-Score of 0.53, reflecting a balanced performance in terms of precision (0.61) and recall (0.46). However, the DQS which is based on the average of all scores performed better than others in terms of F-Score of 0.8, combining exceptionally high precision (0.97) with a respectable recall (0.68).

Score Type	Precision	Recall	F-Score
MS	0.44	0.70	0.54
CS	0.71	0.10	0.14
E3S	0.61	0.46	0.53
DQS	0.97	0.68	0.80

Table 2.1: Results: Data Driven Approach

These results highlight the critical importance of selecting an appropriate scoring mechanism when classifying news articles in disease outbreak detection. While high precision is valuable to minimize false positives, a balance with recall is essential to avoid overlooking relevant articles. In this context, the DQS stands out as a robust choice, offering both high precision and reasonable recall. However, the choice of scoring method should align with specific goals, and the trade-offs between precision and recall should be carefully considered based on the desired outcomes of the classification task. This approach is valuable because it allows us to assess how different attributes influence the quality of news articles. Furthermore, in the multidisciplinary MOOD project involving end-users, attaining this degree of explainability is more significant.

2.7.2 Metadata-Based News Article Classification

We performed the classification through individual metadata feature and the combination of all metadata features in order to see the important metadata features. Table 2.2 shows the results which contain news articles extracted from PADI-web from relevant class and irrelevant class. Table 2.2 provides the evaluation results of a system that has been trained to classify through metadata features as either 'Relevant' or 'Irrelevant' article (see Section 2.4.1). The evaluation metrics used are precision, recall, and F-score. Overall, the system performs well, with the highest F-score being 0.96 for the "All" parameter, indicating that the system is able to make accurate and precise predictions.

Metadata Attributes:	**Precision**	**Recall**	**F-Score**
URL	0.86	0.87	0.86
Source	0.76	0.74	0.75
title	0.94	0.94	0.94
description	0.9	0.82	0.84
publish date	0.45	0.47	0.44
keywords	0.92	0.93	0.92
authors	0.73	0.61	0.59
language	0.31	0.5	0.38
All	0.97	0.95	0.96

Table 2.2: Results: Metadata-Based News Article Classification

The two best performing metadata features according to the table are 'Title' and 'Keywords'. 'Title' has the highest recall of 0.94 and an F-score of 0.94, meaning that the system is accurately classifying most of the metadata features as either 'Relevant' or 'Irrelevant'. 'Keywords' has a precision of 0.92, a recall of 0.93, and an F-score of 0.92, indicating that the system is making accurate predictions and finding most of the relevant results. The confusion matrix for the Table 2.2 results for the following metadata features 'Title', 'Keywords' and 'All' are shown in Figure 2.8.

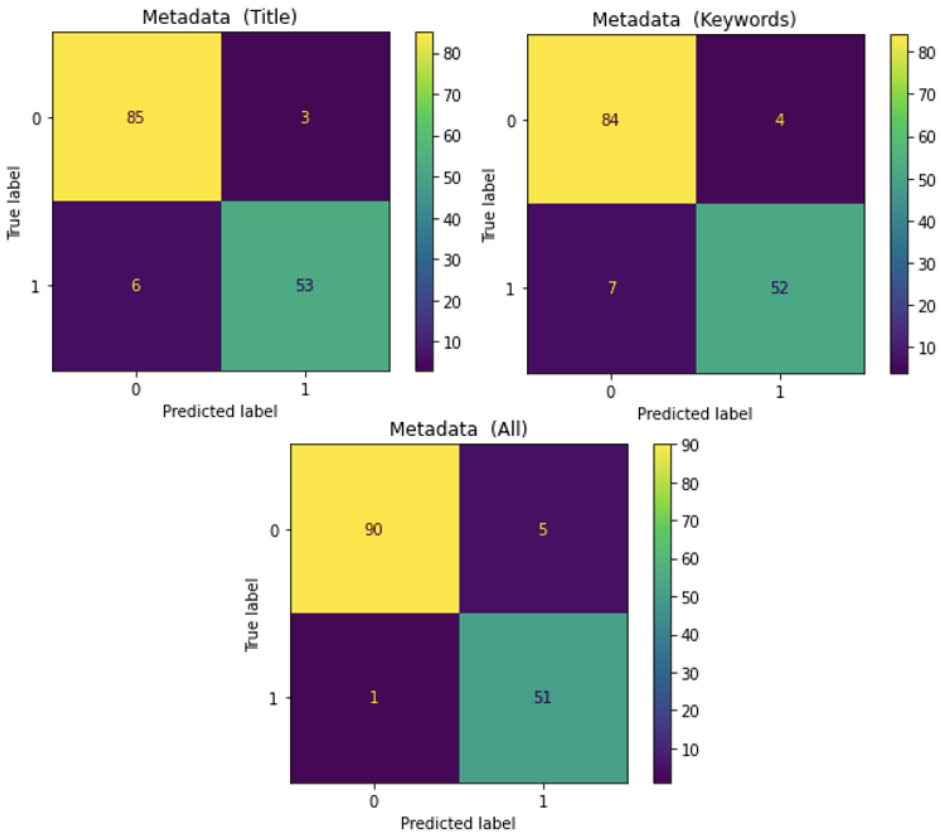


Figure 2.8: Confusion Matrices of Metadata-Based News Article Classification (Title, Keywords, All Features)

2.7.3 Metadata-Based News Source Classification

Our main objective was to identify interesting features associated with the news sources. We evaluate our results at two levels, i.e., 1) feature Classification e.g. source category and geographical coverage, and 2) Source Classification. The evaluation of source categorization is done in supervised way. In the dataset (see Section 2.4.2), we manually annotated the source category (annotated_category column) of the news source to evaluate the automated source categorization by comparing *source category* with annotated category. The clustering results into two main categories, i.e., specialized, generalized. Clustering techniques produced better results in source categorization with precision of **0.88**, recall of **0.83** and F-score of **0.86**. This category is further used for the classification of news sources. We used this source metadata as a feature for the source classification task.

In the dataset (see Section 2.4.2), we manually annotated the geographical coverage of the news source to evaluate the automated geographical coverage. Table 2.3 shows the evaluation of the geographical coverage categorization of news sources, with recall of 0.37 for local class, 0.57 for national class and we 0.87 for international class. The results show less recall for categorization of local and national sources. Therefore, the algorithm needs significant improvements for categorization of local and national sources.

Predicted Labels	True Labels				Recall
		Local	National	International	
	Local	6	4	6	0.37
	National	4	12	5	0.57
	International	6	5	78	0.88
Recall					0.76

Table 2.3: Geographical Coverage Recall Measure

We produced preliminary results for the classification of news sources for the dataset discussed in Section 2.4.2. We selected the Random Forest model as the best choice for classifying sources based on numerical (wpfi) and categorical features (Specialization, Geographical Coverage and Topic Coverage). Table 2.4 shows the results of classification of relevant and irrelevant news sources, with F-score of 0.85 for irrelevant class and 0.47 for relevant class.

Features	Class	Precision	Recall	F-Score	Accuracy
Source Category Geographical Coverage Media Bias Topic Coverage	Relevant	0.74	1	0.85	**0.76**
	Irrelevant	**1**	**0.31**	0.47	

Table 2.4: Results: Metadata-Based News Source Classification

The most important features for the random forest model for classifying news sources were 'topic coverage' and 'Media Bias (wpfi)'. The results are not promising at this stage, but these features in addition to more source metadata features can be helpful for the quality quantification of news sources.

2.8 Discussion

In the score-based approach presented in Section 2.5, we comprehensively considered contextual information attributes to ensure the relevance of news sources. These attributes encompass metadata, content, and epidemic-related aspects, including temporal, spatial, and epidemic information. The collective score derived from these weighted entities results in the contextual weight of each article. The Data Quality Score (DQS) plays a pivotal role in classifying news articles into two main categories: 'Relevant' and 'Irrelevant'. The research has contributed significantly by focusing on a dataset specifically dedicated to AI related articles. Nonetheless, certain limitations should be acknowledged. A primary limitation is the use of small dataset, which may not fully represent the diversity and complexity of articles related to different diseases. To address this, future work should involve larger datasets to explore patterns and insights across various diseases comprehensively.

Metadata extraction holds a crucial position within the EBS System pipeline. The Metadata-Based News Article Classification discussed in Section 2.6.1 has successfully highlighted key meta-

data features, including URL, source, keywords, and article titles. These features significantly contribute to the accurate classification of relevant news articles. While these findings are promising, limitations persist, particularly related to validation on smaller datasets. Tokenizing the URL and source is critical for successful article classification within the EBS System. However, it is essential to recognize that not all URLs contain relevant keywords for classification, with many news sources employing URL patterns that comprise only an ID, lacking meaningful information for the classification model.

In addition to that, we incorporated vital attributes such as source category, geographical coverage, topic coverage, and media bias (wpfi) to classify relevant news sources. Categorizing news sources based on their specializations or focus areas allows the EBS to filter and prioritize information for specific disease-related events. For instance, sources specializing in poultry and agriculture news can receive more weight when tracking AI disease outbreaks, while those focusing on finance and economics can be valuable for assessing diseases economic impacts. Additionally, understanding a news source coverage extent enables targeted monitoring of events in particular regions or countries, which is crucial for detecting and responding to geographically localized events. Moreover, the wpfi attribute into the metadata offers a quantifiable measure of a news source credibility and reputation within the media landscape.

2.9 Conclusion & Perspectives

In this work, we investigated the assessment of data quality of online news sources within the context of the EBS pipeline. We mainly proposed the three approaches to assess the quality of news sources. The first approach, named as, "a data-driven score-based approach to assess the quality of news articles" classify relevant news articles in the context of EBS. This method ensure the explainability aspect of attributes that makes it understandable for the end-users perspectives. The limitation of this work was evaluated with a small dataset. To address this, future work should involve larger datasets to explore patterns and insights across various diseases comprehensively.

The second approach, named as, "Metadata-based news article classification" is proposed to classify relevant news articles in the context of EBS. In addition to the first approach, the benefit of this approach is to classify relevant news articles with limited features (metadata) without using content features through machine learning models. The finding was that metadata features like 'title' and 'keywords' are more important for the classification task. The limitation of this work was it was evaluated with the same dataset as discussed in the first approach.

The third approach, named as, "Metadata-based news source classification" is used for the categorization of news sources. This approach can help in prioritization of news sources in the context of EBS. However, we are still investigating other attributes to enhance this approach. For instance, 'Timely Reporting' is also a potential avenue for investigation that will help in quantification of recent events reported by news source. A source frequently reporting on outdated events would be assigned a lower weight, considering the temporal relevance of its content. This approach enhances the accuracy and timeliness of assessments.

The ultimate objective is to seamlessly integrate these quality attributes into the PADI-web system, associating quality labels with both news sources and news articles. Enriched metadata features will further enhance the classification task within the pipeline, providing valuable insights and explanations for end-users. Furthermore, the research will be extended to include a two-level classification of news articles. In the case of relevant articles, it will be further classified into specific contextual classes, such as outbreak declarations, risk concerns, disease transmission, preventive and control measures. After identifying relevant news articles, the subsequent step has to improve the geographical accuracy of the event.

Part III

Geographical Accuracy of Events: The role of Spatial Relations

CHAPTER 3

GEOGRAPHICAL ACCURACY OF EVENTS: THE ROLE OF SPATIAL RELATIONS

This chapter discussed the role of spatial relations in order to improve geographical accuracy of events detected in free text. The chapter descibed NLP approaches to extract locations associated with spatial relations and an algorithm to translate the locations into geographical coordinates.

3.1 Introduction

Event-based surveillance (EBS) systems rely on three fundamental aspects i.e., spatial, temporal, and epidemiologic aspects. These aspects are used to enhance event understanding comprehensively. The spatial aspects allow understanding the geographical aspect of the event where the event happened. Moreover, the temporal aspect examines the timing and progression of the event, while the epidemiological aspect focuses on the disease patterns and spread. However, there are a lot of challenges to extract the spatial information from the informal sources (potential information provider of the EBS) and geographical representation of the extracted spatial information.

Spatial information plays a pivotal role in EBS systems, significantly enhancing the quality and depth of event understanding for epidemiologists, e.g., identification of high-risk areas, understanding disease transmission etc. The spatial information can be expressed in the media sources in both simple and complex ways, depending on the syntax and semantic of expressions, e.g., (direct mentions: "An outbreak of the flu in New York City", complex expression: "a surge in COVID-19 cases in the neighborhoods of Chicago"). This spatial information is available in the form of absolute spatial information (precise location names, e.g., Milan) and relative spatial information (spatial relations associated with the location name, e.g., North Milan) (Lesbegueries, Sallaberry, and Gaio, 2006; Zenasni et al., 2018). Both types of spatial information are equally important in the context of EBS system in order to know the geographical context of events.

In order to improve the geographical precision of event based surveillance systems, it is important to facilitates fine-grained modelling of event-spatial information (Chanlekha and Collier, 2010). Indeed, the granularity of spatial information stands as a pivotal aspect in comprehending and effectively responding to events, particularly when it comes to accurately pinpointing the location of outbreaks (Wardman, 2022). For instance, in a given text reported in a news: "The Czech Republic has found a second case of the bird flu virus, at a commercial poultry farm, an Agriculture Ministry spokesman said on Sunday. The spokesman said more details of the case, in a region east of Prague". In EBS normally, the spatial information of the reported event is "Prague". However, a careful analysis reveals that the precise location of the reported case lies in the region "east of Prague" rather than within the broader "Prague" region. The event true location might be misrepresented by risk assessor and epidemiologists or inadequately communicated if the finest details of the geographical context are not captured. Therefore, enhancing the granularity of location information becomes pivotal in improving the overall precision of spatial information within EBS systems. By refining the representation of spatial relationships, the system can better capture the event location, thereby facilitating interpretation and assessment by risk assessors and epidemiologists.

3.2 Related Work

Spatial information is important in the context of disease surveillance, disaster management practitioners and many other domains (Zeng, Cao, and Neill, 2021). In EBS, spatial information refers to data that associates geographic coordinates with events, allowing for the understanding of where

and how events are occurring in a given region (Valentin, Lancelot, and Roche, 2021). Therefore, this spatial information adds the geographical context to the event and can help authorities to deploy resources, prevention measures and rapid response where they are urgently needed (Brinks and Ibert, 2020). Moreover, it also plays a pivotal role in epidemiological analysis, particularly in public health, by identifying disease clusters, tracking the spread of infectious diseases, and discerning the factors contributing to transmission in different geographic regions (Oren and Brown, 2022). In the context of EBS systems, accurate geographical information is an important factor of a detected event (O'Shea, 2017). Therefore, it is ensured that the reported location of an event aligns closely with its actual occurrence on the map. Inaccurate or imprecise geographical information can lead to misinterpretation of event, and ultimately results into ineffectiveness of EBS systems (Li et al., 2018).

Spatial information is available in natural language through various linguistic expressions and references (Zhang et al., 2009a). The spatial information are in the following language expression types including, i.e., explicit place names, prepositions (in, at), cardinal directions, topological relations, relative distances, coordinates (Vasardani, Winter, and Richter, 2013). For instance, the most direct way is the mention of the toponyms (place names) in the text, e.g., "cases reported in **London**". Moreover, cardinal directions are words like "north", "south", "east", and "west" can indicate the direction or relative position of a place. In some cases, precise spatial information like geographical coordinates, latitude and longitude values, are available in the text to specify a location accurately (Zhang and Gelernter, 2014). However, it is a challenging task to identify and extract these complex representations of the spatial information from natural language text. Therefore, a dedicated process should be applied to identify such spatial information in unstructured text. The process of identifying locations from textual documents and translating them into geographical coordinates consists of two steps. In the existing literature, the automated process for recognizing and extracting geographic entities from natural language text is referred to as Geoparsing (Moncla et al., 2014). Whereas, Geocoding is the process of converting textual place names i.e. toponyms into geographic coordinates available in the form of latitude and longitude to locate them on geographical maps or any other geographical information systems (GIS) (Hill, 2009).

Numerous research studies have carried out with diverse approaches enhancing Geoparsing that revolves around the extraction of spatial information from unstructured text. These Geoparsing approaches include i.e., rule-based approaches, machine learning, ontology-based reasoning, geographical databases and transformer-based language models (Kokla and Guilbert, 2020; Alonso Casero, 2021). In a conducted research study, a rule-based approach was employed for geospatial relation extraction, leveraging geographical named entity recognition technology and a spatial relation annotation corpus (Zhang et al., 2009b). The rules are just limited to the specified corpus and syntactic patterns that described the geospatial relations. Furthermore, in a separate research study, a rule-based named-entity recognition method was proposed to address specific cases involving spatial named entities in textual data. This approach was validated using historical corpora (McDonough, Moncla, and Camp, 2019). However, the proposed approach did not address the complex relationship that involves other linguistic features, i.e. part-of-speech (POS), dependency parsing, word vectors etc. In another research (Zhang et al., 2009b), a rule-based approach is proposed for geospatial relation extraction based on geographical named entity recognition technology and a spatial relation-annotation corpus. However, the rules are just limited to the specified corpus and syntactic patterns

that described the geospatial relations. In another research (Chen, Vasardani, and Winter, 2017), a best-matched approach is proposed to extract geospatial relations that are referred to anchor places, gazetteered places, and non-gazetteered places. However, it is not defined in the coordinate system to be represented in geographical systems, therefore not suitable for EBS systems. Another research proposed a voting approach (SPENS) to extract place names through five different systems including Stanford NER, Polyglot NER, Edinburgh Geoparser, NER-Tagger, and spaCy (Won, Murrieta-Flores, and Martins, 2018). Another research combine multiple features that capture the similarity between candidate disambiguations, the place references, and the context where the place references occur, in order to disambiguate place among a set of places around the world (Santos, Anastácio, and Martins, 2015). Furthermore, another research (Medad et al., 2020) proposed an approach that is the combination of transfer learning and supervised learning algorithm for the identification of spatial nominal entities. However, the scope of the work was limited to the spatial entities without proper nouns e.g. conferences, bridge at the west, summit, etc. Afterwards, another research (Wu et al., 2022) proposed deep learning models i.e., CasREL and PURE in order to extract geospatial relations in the text. The proposed models were validated with two main approaches, i.e., 1) spatial entities and relations were dealt separately and jointly. The quantitative results demonstrated that pipeline approach performed better than joint approach using deep learning models. Another research (Zheng et al., 2022) proposed a knowledge-based system (GeoKG) that described geographic concepts, entities, and their relations which is used for geological problem solution and decision-making. The solution is only limited to the geological domain that contains information about geographical events, geographical relationships and concepts. Another research proposed an approach named GazPNE2 for extracting place names from tweets by combining global gazetteers (i.e., OpenStreetMap and GeoNames) to train deep learning models and pretrained transformer models, i.e. BERT (Hu et al., 2022). The extracted place names taken coarse (e.g., city) along with fine-grained (e.g., street and POI) levels and place names with abbreviations. Moreover, recent advancements have introduced the UniversalNER model with more entity types, demonstrating remarkable NER accuracy across various domains, including healthcare, biomedicine, and others (Zhou et al., 2023). To our knowledge, there is no such geoparsing method that deals to extract the toponyms associated with spatial relations in the text.

Diverse research studies have carried out for geocoding methodologies with the primary objective of transforming toponyms, which are place names or location references in text, into precise geographical coordinates (Gritta et al., 2018). Mostly, geocoding methods rely on address matching, where textual toponyms are compared to a database of known addresses to retrieve latitude and longitude information (Behr, 2010). Another research proposed a system that extracts place names from text, resolves them to their correct entries in a gazetteer, and returns structured geographic information for the resolved place name (Halterman, 2017). The system can be used for various tasks including media monitoring, improved information extraction, document annotation, and geolocating text-derived events. Another research proposed a geotagging algorithm named DBpedia-based entity recognition for places disambiguation, and Geonames gazetteer and Google Geocoder API for resolution of geographical coordinates of locations (Middleton et al., 2018). In a research study carried out, an unsupervised geocoding algorithm is proposed by taking leverage of clustering techniques to disambiguate toponyms extracted from gazetteers and estimate the spatial footprints of fine-grain toponyms that are not present in gazetteers (Moncla, 2015). Another research introduced a deep neu-

ral network that incorporates Long Short-Term Memory (LSTM) units (Fize, Moncla, and Martins, 2021). The approach was focused on modelling pairs of toponyms, where the first input toponym is geocoded based on the context provided by the second toponym. The approach effectively reduced contextual ambiguities and generated precise geographical coordinates as output. Another research proposed a representational framework that employed rules, semantic approximations, background knowledge, and fuzzy linguistic variables to geocode imprecise and ad-hoc location referents in terms of fuzzy spatial extents as opposite to atomic gazetteer toponyms (Al-Olimat et al., 2019). Additionally, geocoding services and APIs offered by technology companies and government agencies have become increasingly accessible, providing convenient and efficient solutions for geocoding tasks (Longley and Cheshire, 2017). However, there is no existing Geocoding service or method to convert the extracted toponyms associated with spatial relations into geographical coordinates.

3.3 Objectives and Contribution

The research endeavours of this study are outlined as follows:

1. **Extraction of Fine-Grained Spatial Information**: The first objective of this research is to develop a robust methodology for the precise extraction of fine-grained spatial information embedded within textual content. This entails the identification and extraction of spatial relations associated with specific locations within the text.

2. **Geographical Representation of Fine-Grained Spatial Information**: The subsequent objective revolves around transforming the extracted fine-grained spatial information into a geographical representation that aligns seamlessly with established GIS standards. This involves the creation of geometric shapes that accurately reflect the extracted spatial relations.

The contributions and anticipated outcomes of this research are as follows:

1. **Novel Methodology for Fine-grained Spatial Information Extraction**: The development of a novel methodology for the extraction of fine-grained spatial information from textual sources is a substantial contribution. By providing novel NLP techniques that capture spatial relations associated with specific locations, the research aims to provide EBS systems with a more precise and comprehensive understanding of the geographical context of events.

2. **Enhanced Geographical Visualization**: The transformation of extracted spatial information into accurate and meaningful geographical shapes (geographical coordinates) contributes to enhancing the visualization of geospatial data. By adhering to established GIS standards, this outcome is expected to facilitate a more comprehensive understanding of the spatial relationships associated with locations in events.

3.4 Datasets

There are two phases of the methodology, i.e., 1) Extraction of spatial relations associated with locations in text 2) Computing coordinates (polygons) for the extracted spatial relations associated to locations. We evaluated the second phase by creating our own dataset, as there is no specific dataset for the polygon computation. The details of both datasets are discussed in the subsequent sections.

3.4.1 Geoparsing Dataset

This dataset is curated from news articles collected by PADI-web[1], an EBS system specializing in health events. The dataset encompasses news articles covering various diseases with reference to the MOOD[2] project, including: 1) *Antimicrobial Resistance (AMR)*, 2) *COVID-19*, 3) *Avian Influenza (AI)*, 4) *Lyme*, and 5) *Tick-borne Encephalitis (TBE)*. Each row is annotated in the dataset with a column (rse) contained the locations (toponyms associated with spatial relations) e.g. "south of France". These entities provide information regarding disease outbreaks and preventive measures in relation to specific locations. The dataset is structured with a separate *CSV* file for each disease category. Each row in the CSV file comprises essential columns, including 'id', 'title', 'text', 'url', 'rse', 'source_lang', and 'created_at'. For result evaluation, our geoparsing method extracts locations (toponyms associated with spatial relations) from "text" the column automatically and compares with the "rse" column.

3.4.2 Geocoding Dataset

Regarding the *Geocoding Phase*, we curated a distinct dataset encompassing *9* renowned cities across the UK and Europe with reference to the MOOD project: Paris, London, Milan, Madrid, Zagreb, Utrecht, Delft, Lyon, and Florence. This dataset contains the polygons of geospatial relations linked with these cities. Our endeavour led us to identify and define 19 geospatial relations for each city. These relations manifest in the form of singular relations and combinations thereof. The dataset, dedicated to geographical insights, comprises 19 distinct spatial relation shapes for every city. Rigorous qualitative analysis was subsequently carried out, with end-users participating to evaluate the derived shape dataset (see Section 3.7.2). Our work focused on generating shapes for 9 cities, each encompassing a combination of 19 or more geographical shapes. This comprehensive approach aims to thoroughly validate our methodology through the assessment conducted by end-users. Our algorithm facilitates the dynamic creation of shapes corresponding to various spatial relations for any given city. While the algorithm allows for the generation of shapes based on discussed spatial relations for any city, we concentrated our experiments on evaluating 19 diverse spatial relation shapes for each city as part of our evaluation process.

[1] https://padi-web.cirad.fr/en/
[2] https://mood-h2020.eu/

3.5 Spatial Relations

Spatial relations refer to descriptive terms linked with specific place names in text, indicating regions related to the mentioned place rather than its exact location. Examples of these terms include “southern Paris" and “Paris border." To our knowledge, state-of-the-art Named-entity recognition (NER) struggles to recognize such spatial relations, leading to inaccurate region identification. These relation-associated terms are known as **Spatial Relations Entities (spatRE)**. In linguistics, spatRE are represented in grammar by below expressions:

```
spatRE <- [ADVERB][NOUN] [VERB] PROPN [NOUN]
# PROPN should be a place name, which is mandatory
# At least the left or right part of PROPN is mandatory
# NOUN and ADVERB are spatial keywords, These spatial
keywords represent a lexicon of spatial relations within
natural language.
```

The expression above illustrates the linguistic representation of *spatRE*, which can be defined using a regular expression as follows:

$$\begin{aligned} \text{spatRE} \quad = \quad & ([\text{spat\_relation\_kwd}]+[\text{place\_name}]) \\ & \quad | \\ & ([\text{place\_name}][\text{spat\_relation\_kwd}]+) \end{aligned}$$

The above regular expression defines the spatRE. Furthermore, we have categorized the spatial relations into four main types. These spatial relations are defined in the subsequent sections as follows:

3.5.1 Level-1 Spatial Relations

Level-1 spatial relations consist of the *cardinal relation* and the distinct *center* relation associated with *place names* in textual documents. These cardinal relations indicate directional orientations, encompassing North, South, East, West, North-East, North-West, South-East, and South-West. Moreover, Level-1 spatial relations can also manifest through synonyms connected to the specified place name. Notable examples of spatial relation entities with Level-1 spatial relations include expressions like ‘Northern Milan’, ‘Southern Paris’, and ‘Central London’. In linguistic terms, the representation of Level-1 spatRE follows this structure:

```
level1_spatRE <- [ADVERB] [Noun] [VERB] PROPN [ADVERB]
# PROPN should be a place name, which is mandatory
# At least the left or right part of PROPN is mandatory
# ADVERB and NOUN are Level-1 keywords
```

"Level-1 keywords" are dedicated entries of lexicon e.g., north, south etc. The representation of *Level-1 spatRE* using geographical coordinates within geographical information systems is an integral aspect. Figure 3.1 illustrated an example of *Level-1 spatRE*.

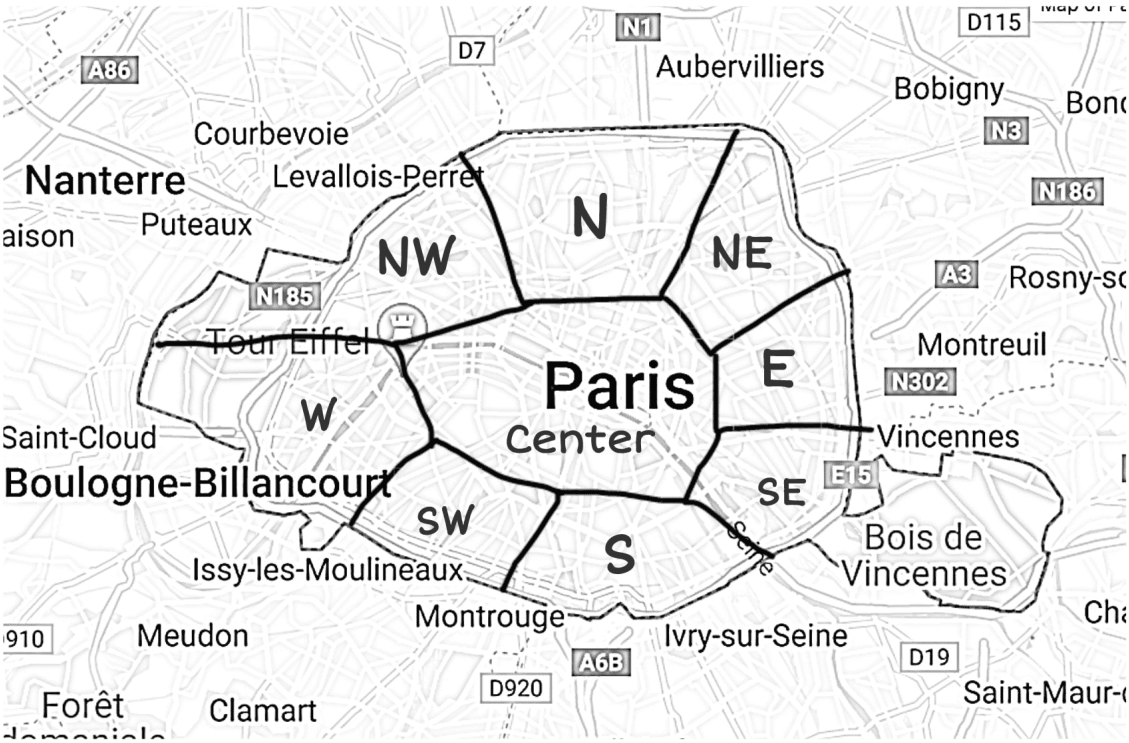


Figure 3.1: Level-1 Spatial Relations

3.5.2 Level-2 Spatial Relations

Level-2 spatial relations encompass spatial relationships coined with keywords that pertain to proximity, such as border, near, close, proximity, and their synonymous terms linked to the specified place name. These spatial relations hold significance within the context of sensitive geographical information systems, facilitating accurate geographical zone referencing. Prominent illustrations of *Level-2 spatial relations* encompass phrases like 'Milan border', 'proximity of Lyon', and 'around Bordeaux'. In the linguistic realm, the representation of Level-2 spatRE follows this structure:

```
level2_spatRE <- [NOUN] [ADVERB] [VERB] PROPN [NOUN]

# PROPN should be a place name, which is mandatory
# At least the left or right part of PROPN is mandatory
# NOUN and ADVERB are Level-2 keywords
```

"Level-2 keywords" are dedicated entries of lexicon e.g., border, nearby, close to, surrounding etc. The geographical coordinates of Level-2 spatial relations are not as straightforward. The graphical representation of the example 'Paris border' for *Level-2 spatial relations* is illustrated in 3.2.

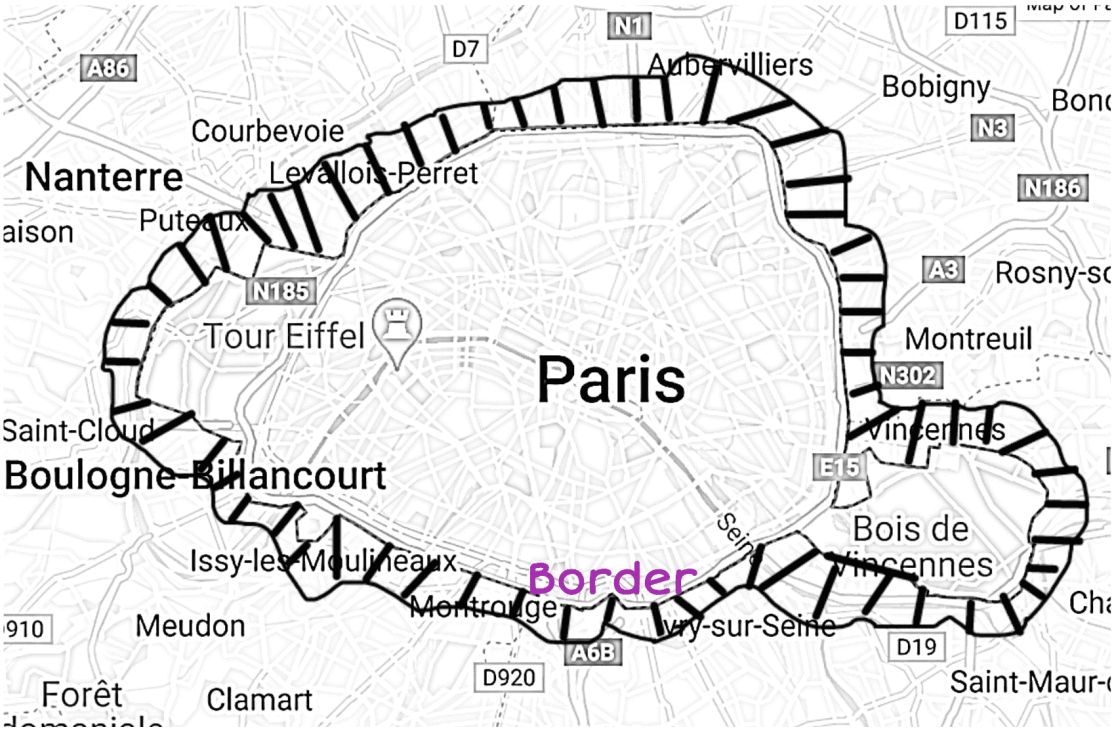


Figure 3.2: Level-2 Spatial Relations

3.5.3 Level-3 Spatial Relations

Level-3 spatial relations encompass distance associations with place names, employing compound keywords composed of a distance unit and a numerical value. These associations can be expressed in various forms, such as '2 km from', '2 km away', '2 km radius of', and so on. Additionally, the distance unit might appear in both its full-text form and as an abbreviation, for example, km, mi, ft, etc. Two examples of spatial relation entities involving Level-3 spatial relations include phrases like '2 km distance from Paris' and 'radius of 3 miles from Paris'. In linguistic terms, the representation of Level-3 spatRE follows this structure:

```
level3_spatRE <- [NUM] [NOUN] [VERB] + PROPN
# PROPN should be place Name which is mandatory
# At least left of PROPN is mandatory
# NOUN is a Level-3 keywords
# NUM is a number
```

"Level-3 keywords" are dedicated entries of lexicon e.g., 5 km radius, 3 miles away etc. Geographical coordinates pertaining to Level-3 spatial relations can be indicative of distances from the original place names. The graphical depiction of the example '1 km from Paris' within *Level-3 spatial relations* is illustrated in 3.3.

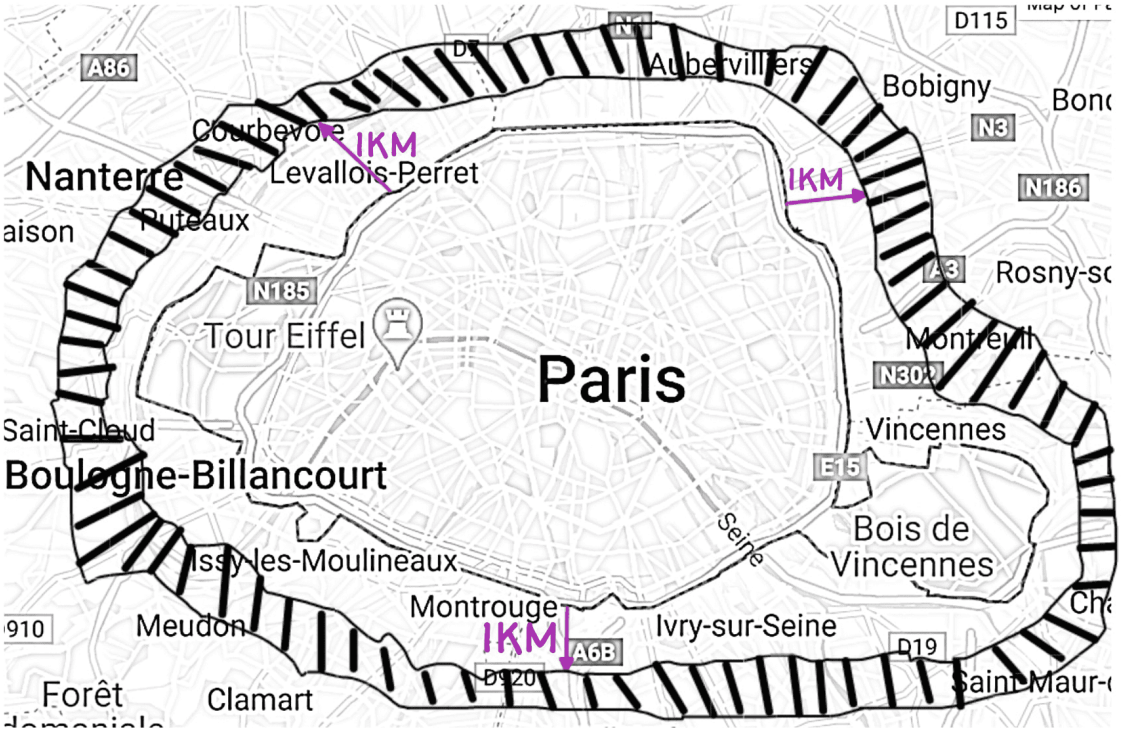


Figure 3.3: Level-3 Spatial Relations

3.5.4 Compound Spatial Relations

Compound spatial relations can also appear within text as combinations of Level-1, Level-2, and Level-3 spatial relations associated with specific place names. These spatREs are characterized by amalgamated spatial relations that are applied to the designated place names. Illustrations of *compound spatial relations* encompass phrases like '1 km from north Paris border' or '6 miles away from South Lyon', where multiple levels of spatial relationships converge. The geographical illustration of *compound spatial relations*, as seen in the 1 km from different cardinal relations, are displayed in Figure 3.4.

In order to extract such spatial relations and identify its geographical representation, a two steps methodology is proposed that are described in subsequent section.

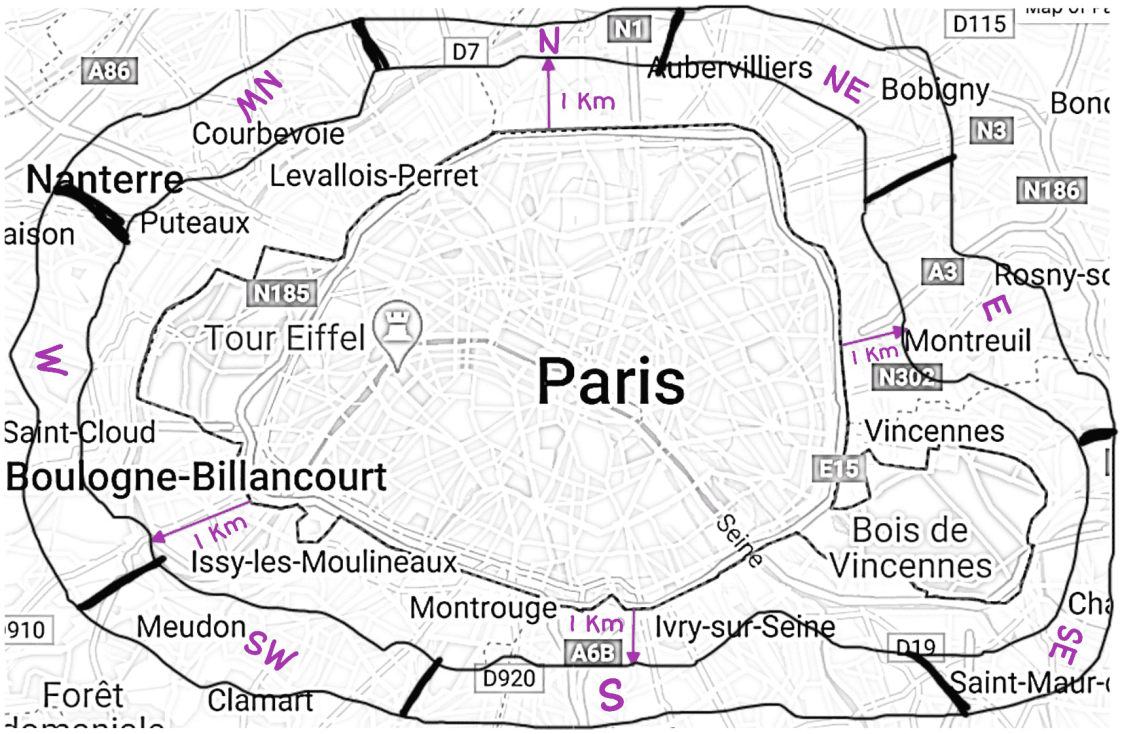


Figure 3.4: Compound Spatial Relations

3.6 Methodology

In order to deal with such spatial relations in the textual documents, our proposed methodology is divided into two main phases: 1) Extraction phase 2) Geocoding phase respectively. Extraction phase extracts spatial relation entities from the textual documents, while Geocoding phase translates the spatial relation entities into geographical coordinates compatible with GIS applications. The process workflow of the proposed methodology is shown in Figure 3.5.

The details of the two phases of the proposed methodology are explained in the subsequent sections.

3.6.1 Extraction Phase

In the extraction phase, as depicted in Figure 3.5, spatREs are extracted from the text data. For this, some language processing libraries or tools can be utilized to extract linguistic information. In our case, we chose state-of-the-art natural language processing library (NLP) *spaCy* (Honnibal and Montani, 2017) for python. In extraction phase, we used *spaCy* (see Chapter II) for NER task with the help of pre-built *spaCy* linguistic models. The steps involved to extract spatREs are as follows:

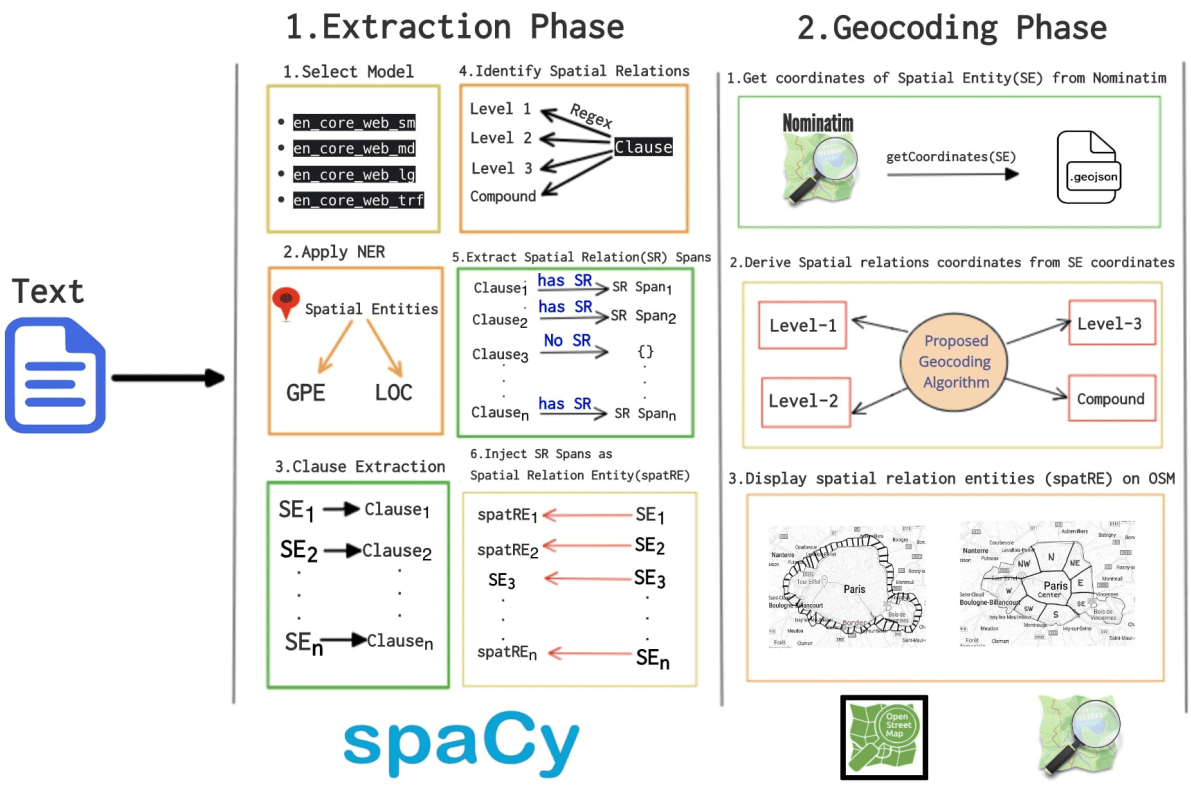


Figure 3.5: spatRE Pipeline

1. **Model Selection:** There are four linguistic models different in sizes provided by spaCy, i.e., *en_core_web_sm*, *en_core_web_md*, *en_core_web_lg* and *en_core_web_trf*. The *en_core_web_trf* model is computationally expensive compared to smaller models, and requires significant computational resources to run. However, its high performance and accuracy make it a popular choice for a wide range of NLP applications. Once the model for *spaCy* is selected, the environment is set up for the NLP tasks such NER, POS tagging, etc.

2. **Apply NER:** The next step in extraction phase is to apply NER on the textual data. For the spatial information, we only need to extract spatial entities from the textual data with the labels 'GPE' (Geopolitical entities, e.g., Paris) and 'LOC' (physical location, e.g., Safari Desert) respectively. As a result, it will identify the spatial named entity e.g. Paris, Lyon, France, Italy, etc. Once these spatial entities are extracted from the text, the next step is to identify the spatial relations that are associated with such entities. To identify the spatial relations associated with the spatial named entities, we need to extract the clauses that contain the spatial entities.

3. **Clause Extraction:** With *spaCy*, for each spatial entity we got its sentence from which the entity belongs by getting its start offset and end offset. However, in our case, we want to get its clause with which it is associated. In linguistics, technically a clause can have one spatial entity at a time. In the proposed work, we applied the rules for the clauses that are separated by conjunctions known as compound clauses. These clauses are normally separated by conjunction symbols i.e., ',', ';', ':', etc., and conjunction keywords such as 'and', ' but', 'or', 'nor', 'for' and 'yet'. In order to extract the clause of each spatial entity, every clause is separated

by the conjunctions or with some conjunction keywords i.e., and, or, etc. In our algorithm, we split the sentence into clauses and save the clause that contains the spatial entity and ignore the rest of the clauses in the sentence. Here is the example of the clause extraction as follows:

Text: Significant number of COVID-19 cases are reported close to Paris border, whereas other areas in the region are safe due to preventive safety measures.

Clause 1: Significant number of COVID-19 cases are reported close to Paris border
Clause 2: ~~whereas other areas in the region are safe due to preventive safety measures.~~

The above text which is a sentence is divided into two clauses by having conjunction to separate it. The first clause is considered as it contains the spatial entity however, the second clause is ignored because of not having spatial entity. Once the clause having spatial entities are extracted from the text, the next step is to identify spatial relations in the clauses.

4. **spatial Relations Identification:** In Section 3.5, we defined different hierarchies of spatial relations. We need to extract such spatial relations from the candidate clauses. Candidate clauses are identified in the text document as the clauses that contain spatial entities. The next step is to determine whether there are any spatial relations associated with these entities in the candidate clauses. In order to extract spatial relations in the clauses, we defined regular expressions for Level-1, Level-2, Level-3 spatial relations. The regular expressions of these spatial relations are defined using *Python* regex *re* with the help of external library *quantities*. *quantities* library is used to get the different quantity units, their abbreviations, and their interconversions. The purpose of this library is to identify the distance measurement units in the text and the interconversion of units, e.g., km to miles and miles to ft. The library also provides different representations of distance units, e.g., km, kilometres, miles and *mi*.

   If these spatial relations are identified in the clause that contained the spatial entity, then we adjust the span offset according to the spatial relation. The span offset is either adjusted from the end or the start according to the occurrence of spatial relation relative to spatial entity. In the following examples 'north of Lyon', 'Paris border'; 'north of Lyon' has the spatial relation before spatial entity i.e., 'Lyon'. Contrary to first example, 'Paris border' has the spatial relation after the spatial entity, i.e. 'Paris'. In some cases, the occurrence of the spatial relations exist in both ways. In the following examples, '2 km away from Paris border', 'south Lyon border' have the spatial relations before and after the spatial entities. '2 km away from Paris border' has the spatial relations '2 km' and 'border' with spatial entity i.e. 'Paris', 'south Lyon border' have the spatial relations 'south' and 'border' with spatial entity i.e. 'Lyon' respectively. The start offset or end offset of the spans, depending on spatial relations start offset or end offset in the text document. Once the offset of the spans are known, the next steps are to create the spans for spatial relations along with spatial entities.

5. **Geospatial Relations Spans Extraction:** Once the spans are identified in the whole text document, the next step is to save each span into the span list. In the below example, we identified the three main clauses in the whole text document. The two clauses of them are considered as spatial entities and the third clause is discarded. In the two clauses, we identified the geospatial

relations inside it and later on identified the span having `spatial_relation + spatial_entity` as shown in the below example.

Text: Significant number of COVID-19 cases are reported close to Paris border, north of Lyon whereas other areas in the closer region are safe due to preventive safety measures.

Clause 1: Significant number of COVID-19 cases are reported close to Paris border
Clause 2: north of Lyon
Clause 3: ~~whereas other areas in the region are safe due to preventive safety measures.~~
Spatial Entity 1: Paris **Spatial Entity 2:** Lyon
Spatial Entity Span 1: Paris border **Spatial Entity Span 2:** north of Lyon

After the span list is obtained, then the next step is to inject these spans as entities in the default *spaCy* NER pipeline.

6. **Spatial Relations Entities Injection:** The next step is to replace the default spatial entities in the 'Doc' (the element that contains linguistic feature information, e.g., NER, spans, and POS) element by spatRE which are identified in the geospatial relation spans. The label of the spatRE is injected in the 'Doc' element as 'spatRE'. Table 3.1 shows the example of the default NER extraction result and after incorporation of extraction phase result.

Text	**Before Extraction Phase**	**After Extraction Phase**
	DOC (Element)	**DOC (Element)**
Significant number of COVID-19 cases are reported close to Paris border, north of Lyon whereas other areas in the closer region are safe due to preventive safety measures.	Significant number of COVID-19 cases are reported close to **Paris(GPE)** border, north of **Lyon(GPE)** whereas other areas in the closer region are safe due to preventive safety measures.	Significant number of COVID-19 cases are reported close to **Paris border(spatRE)**, **north of Lyon(spatRE)** whereas other areas in the closer region are safe due to preventive safety measures.

Table 3.1: spatRE Examples

Algorithm 1 Procedure to Extract Spatial Relation Entities

Require: spaCy Linguistic Element (*Doc*)
Ensure: Spatial Relation Entities (*spatRE*)

spatEnt ← [] {Store spatial entities}
for each *ent* ∈ *Doc.ents*, **if** *ent* is 'GPE' **or** *ent* is 'LOC' **do**
    spatEnt.append(ent)
end for
spatRE ← [] {Store spatial relation entities (*spatRE*)}
for each *ent* ∈ *spatEnt* **do**
    relEntity = getLevel1(*ent*, clause(*ent*)) {Get Level 1 relation from clause}
    relEntity = getLevel2(*relEntity*, clause(*ent*)) {Get Level 2 relation from clause}
    relEntity = getLevel3(*relEntity*, clause(*ent*)) {Get Level 3 relation from clause}
    if *relEntity* is not None **then**
        spatRE.append(relEntity)
    end if
end for
Doc.ents ← *spatRE* {Inject *spatRE* into default element *Doc*}
return *Doc*

Algorithm 1 outlines the procedure for extracting spatial relation entities from a given text using the spaCy library for natural language processing. The algorithm takes a spaCy Linguistic Element, denoted as "Doc", as input and aims to identify and extract spatial relation entities, denoted as spatRE, from within the provided text. Algorithm 1 explained the steps discussed in the process workflow of Figure 3.5.

The following Algorithm 2, labelled as "Algorithm for Extraction Phase" employs the spaCy natural language processing library to extract and visualize spatial relation entities from a given input text. The steps in Algorithm 2 are as follows:

Input: The algorithm takes the input **text**, which is the text to be analysed.

Output: The algorithm produces a collection of spatRE and visualizes them using the *displacy* function.

- **Load spaCy Model**: The algorithm starts by loading the spaCy model named *en_core_web_trf*. Call Algorithm 1: The algorithm adds a custom pipeline for spatial relation extraction by calling the procedure outlined in Algorithm 1.
- **Process Text**: The algorithm processes the input text using the loaded spaCy model, resulting in the creation of a doc object that encapsulates linguistic analysis of the text.
- **Extract Spatial Relation Entities**: The Doc.ents attribute of the doc object contains the extracted entities, including spatRE.
- **Visualize Spatial Relation Entities**: The *displacy* function is used to visually represent the extracted spatREs, aiding in understanding and analysis.

In summary, this algorithm processes input text using a spaCy model to identify spatial relation entities. It then utilizes the *displacy* function to provide a graphical representation of these entities.

Algorithm 2 Algorithm for Extraction Phase

Input: text
Output: Spatial relation Entities (spatRE)

Load spaCy model: $model \leftarrow \text{load}(en\_core\_web\_trf)$
Call Algorithm 1: $model$.add_pipe$(spatial_relation_pipeline)$
$Doc \leftarrow model(text)$ {Process text}
$spatRE \leftarrow Doc.ents$ {Extract spatial relation entities}
displacy$(spatRE)$ {Visualize spatial relation entities}

3.6.2 Geocoding Phase

After the extraction phase, the next phase is *geocoding phase* as shown in Figure 3.5. The purpose of this phase is to translate the *spatRE* into geographical coordinates. The translation of geographical coordinates is derived either by slicing the polygon or by deriving using geospatial operations. In order to identify the coordinates of *spatRE*, we need to get the coordinates of the spatial entity exclusive of geospatial relations.

Algorithm 3 Algorithm for Geocoding Phase

Input: Spatial relation Entity (spatRE)
Output: coordinates

$place\_name \leftarrow getSpatialEntity(spatRE)$ {Get GPE or LOC part}
$coordinates \leftarrow getCoordinates(place\_name)$ {Get coordinates from **Nominatim**}
$coordinates \leftarrow getLevel1Coordinates(coordinates)$ {Call Algorithm 6}
$coordinates \leftarrow getLevel2Coordinates(coordinates)$ {Call Algorithm 7}
$coordinates \leftarrow getLevel3Coordinates(coordinates)$ {Call Algorithm 8}
if *output* is **GEOJSON then**

    return $geojson(coordinates)$
else if *output* is **MAP then**
    $display(coordinates)$ {Display coordinates on OSM}
end if

The Algorithm 3 facilitates the geocoding process by converting spatRE into geographical coordinates. It further provides options to visualize or return the coordinates in various formats. The steps in the Algorithm 3 are as follows:

Input: The algorithm takes a single input, which is a spatRE.

Output: The primary output is geographical coordinates of the spatRE, and the algorithm offers flexibility to choose between output formats.

1. **Get Spatial Entity**: Extracts the geographical place name part from the given spatial relation entity (spatRE).

2. **Get Coordinates**: Utilizes the extracted place name to fetch the corresponding geographical coordinates using the *Nominatim* (OpenStreetMap contributors, 2017) geocoding service.

3. **Get Level 1 Coordinates**: Refines the coordinates by invoking the procedure outlined in Algorithm 6 to accommodate Level 1 spatial relations.

4. **Get Level 2 Coordinates**: Enhances the coordinates by invoking the procedure outlined in Algorithm 7 to incorporate Level 2 spatial relations.

5. **Get Level 3 Coordinates**: Further refines the coordinates by invoking the procedure outlined in Algorithm 8 to integrate Level 3 spatial relations.

6. **Output Format Decision**: If the desired output format is GEOJSON, the algorithm generates and returns the coordinates in the GEOJSON format. If the preferred output format is MAP, the algorithm visualizes the coordinates on OpenStreetMap (OSM) (OpenStreetMap contributors, 2017).

In summary, the algorithm takes a spatial relation entity, extracts the associated place name, converts it into geographical coordinates, applies refinements based on different levels of spatial relations, and then offers a choice between returning the coordinates in GEOJSON format or displaying them on a map. The geocoding phase is further divided into different steps in order to extract different levels of spatial relations. These steps are discussed as follows:

1. **Acquire coordinates from Nominatim:** Nominatim API (Clemens, 2015) provides search by place name, feature description or free text search in OpenStreetMap (OpenStreetMap contributors, 2017) database and returns its geographical coordinates based on search queries. The API provides the *GeoJSON* which contains the geometry along with their feature attributes.

2. **Derive/Slice Geospatial Relation Coordinates:** After getting the coordinates of the place name mentioned in the spatRE, the next step is to derive the coordinates of the geospatial relations associated with the place name. This can be done depending on the type of geospatial relation. Level-1 spatial relation coordinates are acquired by slicing the main geometry of the place into **9** spatial relations geometries. For instance, The Level-1 Slicing of Paris can be sliced into 9 geographical shapes: 'Northern Paris', 'Southern Paris', 'Eastern Paris', 'Western Paris', 'North-east Paris', 'South-east Paris', 'North-west Paris', 'South-west Paris' and 'Central Paris', as shown in Figure 3.1.

Algorithm 4 Procedure to extract Cardinal Coordinates

Input: coordinates, centroid, direction
Output: Cardinal coordinates

```
cardinal_coordinates ← []
if direction is north then
    cardinal_coordinates ← getCoordinates(north)    {Coordinates having angle with centroid between 337° - 22°}
end if
if direction is east then
    cardinal_coordinates ← getCoordinates(east)     {67° - 112°}
end if
if direction is south then
    cardinal_coordinates ← getCoordinates(south)    {157° - 202°}
end if
if direction is west then
    cardinal_coordinates ← getCoordinates(west)     {247° - 292°}
end if

return cardinal_coordinates
```

The algorithm 4 is designed to retrieve specific coordinates based on cardinal directions from a given set of coordinates and a centroid, taking the direction as input. The step by step explanation of each part of the algorithm is as follows:

Input: The algorithm takes three inputs:

- **coordinates**: The set of coordinates under consideration.
- **Centroid**: The centroid coordinates for reference.
- **Direction**: The cardinal direction (east, north, west, or south) for which coordinates are to be extracted.

Output: The algorithm produces a list of cardinal coordinates based on the specified direction.

(a) **Initialization**: Initialize an empty list called *cardinal_coordinates* to store the extracted coordinates.

(b) **Extract Coordinates based on Direction**:

- If the direction is **north**, extract the coordinates with an angle (Ellefi and Drap, 2020) between 337° and 22° relative to the centroid, signifying the north direction.
- If the direction is **east**, extract the coordinates with an angle between 67° and 112° relative to the centroid, signifying the east direction.
- If the direction is **south**, extract the coordinates with an angle between 157° and 202° relative to the centroid, signifying the south direction.
- If the direction is **west**, extract the coordinates with an angle between 247° and 292° relative to the centroid, signifying the west direction.

(c) **Output**: Return the list of extracted **cardinal_coordinates** corresponding to the specified direction.

In summary, the algorithm assists in extracting coordinates corresponding to specific cardinal directions. It evaluates the input direction and fetches coordinates within the specified angular range from the centroid, ultimately providing a collection of coordinates that align with the chosen cardinal direction. The Algorithm 5 is designed to extract specific coordinates based on ordinal directions from a given set of coordinates and a centroid, considering the direction as input. The Algorithm 5 step by step operations are as follows:

Algorithm 5 Procedure to extract Ordinal Coordinates

Input: coordinates, centroid, direction
Output: Ordinal coordinates

ordinal_coordinates ← []
if *direction* is **northeast then**
    ordinal_coordinates ← *getCoordinates*(*northeast*) {22° - 67°}
end if
if *direction* is **southeast then**
    ordinal_coordinates ← *getCoordinates*(*southeast*) {112° - 157°}
end if
if *direction* is **southwest then**
    ordinal_coordinates ← *getCoordinates*(*southwest*) {202° - 247°}
end if
if *direction* is **northwest then**
    ordinal_coordinates ← *getCoordinates*(*northwest*) {292° - 337°}
end if
if *direction* is **central then**
    ordinal_coordinates ← *getCoordinates*(*central*) {Merge all cardinal extreme coordinates}
end if

return *ordinal_coordinates*

Input: The algorithm takes three inputs:

- **coordinates**: The set of coordinates under consideration.
- **Centroid**: The centroid coordinates for reference.
- **Direction**: The ordinal direction (northeast, northwest, southwest, southeast, or central) for which coordinates are to be extracted.

Output: The algorithm produces a list of ordinal coordinates based on the specified direction.

(a) **Initialization**: Initialize an empty list called **ordinal_coordinates** to store the extracted coordinates.

(b) **Extract Coordinates based on Direction**:

- If the direction is **northeast**, extract the coordinates with an angle between 22° and 67° relative to the centroid.
- If the direction is **southeast**, extract the coordinates with an angle between 112° and 157° relative to the centroid.
- If the direction is **southwest**, extract the coordinates with an angle between 202° and 247° relative to the centroid.
- If the direction is **northwest**, extract the coordinates with an angle between 292° and 337° relative to the centroid.
- If the direction is **central**, merge all extreme coordinates of the cardinal directions (east, west, north, south) to provide a central region.

(c) **Output**: Return the list of extracted **ordinal_coordinates** corresponding to the specified direction.

In summary, the algorithm facilitates the extraction of coordinates aligned with specific ordinal directions. Additionally, it provides an option to merge extreme coordinates from cardinal directions and the inside polygon become a “central" of a location.

Algorithm 6 Procedure to extract Level-1 Coordinates

Input: coordinates, centroid, direction
Output: Level-1 coordinates

$level1\_coordinates \leftarrow []$
$level1\_coordinates \leftarrow cardinals(coordinates, centroid, direction)$ {Calling Algorithm 4}
$level1\_coordinates \leftarrow ordinals(coordinates, centroid, direction)$ {Calling Algorithm 5}
if $level1\_coordinates$ is not **None then**

  return $level1\_coordinates$
end if

return $coordinates$

The algorithm 6 is designed to extract coordinates based on Level-1 spatial relations, considering both cardinal and ordinal directions. This algorithm integrates the results from Algorithm 4 and 5 to provide Level-1 coordinates. The steps to compute the coordinates of Level-1 spatial relations are as follows:

Input: The algorithm takes three inputs:

- **coordinates**: The set of coordinates under consideration.
- **Centroid**: The centroid coordinates for reference.
- **Direction**: The direction (cardinal or ordinal) for which Level-1 coordinates are extracted.

Output: The algorithm produces Level-1 coordinates based on the specified direction.

(a) **Initialization**: Initialize an empty list called **level1_coordinates** to store the extracted Level-1 coordinates.

(b) **Extract Cardinal Coordinates**: Call Algorithm 4 to extract cardinal coordinates based on the specified direction.

(c) **Extract Ordinal Coordinates**: Call Algorithm 5 to extract ordinal coordinates based on the specified direction.

(d) **Output Level-1 coordinates**: If either cardinal or ordinal coordinates were extracted, return the merged **level1_coordinates** that incorporates both. If it is not Level-1 spatial relations, return the original **coordinates**.

In contrast to Level-1 relations, Level-2 and Level-3 relations are derived by applying geospatial operations i.e. geospatial joins, geospatial unions, intersections with the help of *GeoPandas* (Jordahl et al., 2020) and *Shapely* (Gillies et al., 2007) python libraries. After the computation of geospatial coordinates of the spatRE, it is converted into compatible *GeoJSON* format, which can be utilized using any Geographical Information System (GIS) applications. Slicing is applied in order to extract cardinal and ordinal directional relations.

Algorithm 7 Procedure to extract Level-2 Coordinates

Input: coordinates, centroid, level2_keywords
Output: Level-2 coordinates

$level2\_coordinates \leftarrow []$
if $level2\_keywords$ **then**
  $polygon1 \leftarrow Polygon(coordinates)$
  $polygon2 \leftarrow polygon1.buffer(coefficient)$ {**coefficient** is buffer of external polygon}
  $level2\_coordinates \leftarrow polygon2.difference(polygon1))$

  return $level2\_coordinates$
end if

return $coordinates$

The Algorithm 7 is designed to extract Level-2 coordinates based on specified keywords e.g. border, nearby, next to, considering a centroid and utilizing polygon geometry operations. The step by step operations to compute Level-2 coordinates in Algorithm 7 are as follows:

Input: The algorithm takes three inputs:

- **coordinates**: The set of coordinates under consideration.
- **centroid**: The centroid coordinates for reference.
- **level2_keywords**: Keywords associated with Level-2 spatial relations.

Output: The algorithm produces Level-2 coordinates based on the specified keywords.

(a) **Initialization**: Initialize an empty list called level2_coordinates to store the extracted Level-2 coordinates.

(b) **Keyword Presence Check**: Check if there are **level2_keywords** specified. If not, proceed to return the original **coordinates**.

(c) **Polygon Creation**: Create a polygon (polygon1) using the input coordinates.

(d) **Buffer Operation**: Apply a buffer operation on polygon1 using a coefficient, producing an external polygon (polygon2) that extends outward from polygon1.

(e) **Difference Operation**: Extract the difference between polygon2 and polygon1. This operation yields Level-2 coordinates that corresponds to the external region added by the buffer operation.

(f) **Output Level-2 coordinates**: If Level-2 coordinates were successfully extracted, return the **level2_coordinates**. If the spatial relation is not a Level-2 spatial relation, return the original coordinates.

In summary, the algorithm enables the extraction of Level-2 coordinates by utilizing polygon geometry operations.

Algorithm 8 Procedure to extract Level-3 Coordinates

Input: coordinates, centroid, level3_keywords, distance, unit
Output: Level-3 coordinates

$level3\_coordinates \leftarrow []$
if $level3\_keywords$ **then**
  $distanceKm \leftarrow convert(distance, unit)$
  $polygon1 \leftarrow Polygon(coordinates)$
  $polygon2 \leftarrow polygon1.buffer(coefficient * distanceKm)$ {Multipy coefficient with distance in KM}
  $level3\_coordinates \leftarrow polygon2.difference(polygon1))$

  return $level3\_coordinates$
end if

return $coordinates$

The Algorithm 8 is designed to extract Level-3 coordinates based on specified keywords and a given distance, considering a centroid and utilizing polygon geometry operations. The step by step operations to extract Level-3 coordinates in the Algorithm 8 are as follows:

Input: The algorithm takes five inputs:

- **coordinates**: The set of coordinates under consideration.
- **centroid**: The centroid coordinates for reference.
- **level3_keywords**: Keywords associated with Level-3 spatial relations.
- **distance**: The numerical value representing the distance.
- **unit**: The unit of measurement for the distance (e.g., km, miles).

Output: The algorithm produces Level-3 coordinates based on the specified keywords and distance.

(a) **Initialization**: Initialize an empty list called level3_coordinates to store the extracted Level-3 coordinates.

(b) **Keyword Presence Check**: Check if there are **level3_keywords** specified. If not, proceed to return the original coordinates.

(c) **Distance Conversion**: Convert the given distance value to kilometres using the specified unit.

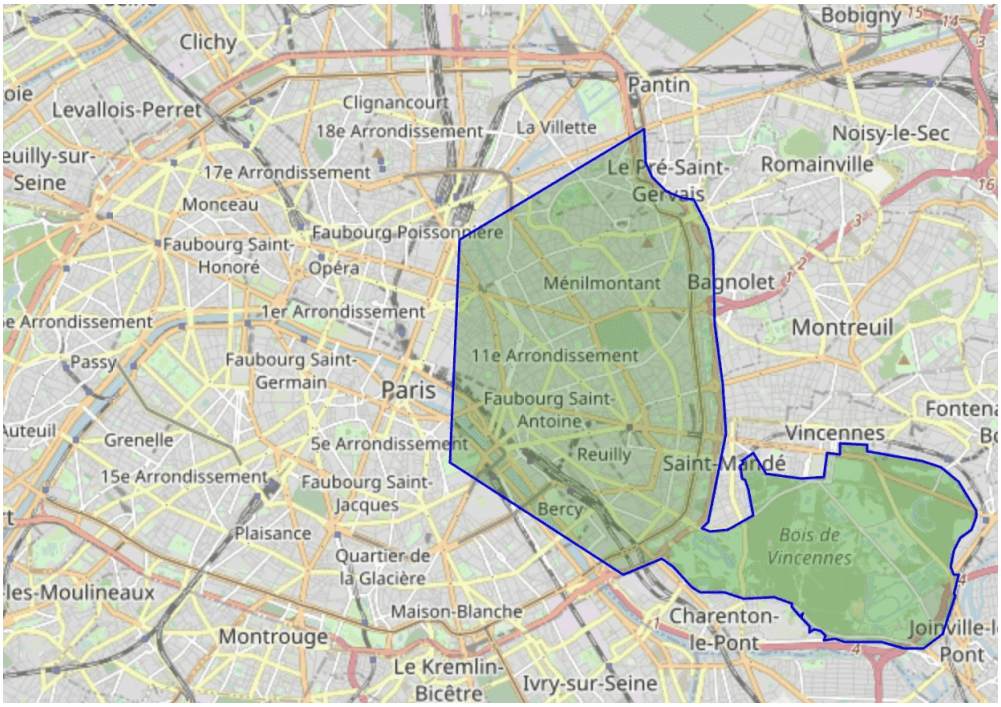
(d) **Polygon Creation**: Create a polygon (polygon1) using the input coordinates.

(e) **Buffer Operation**: Apply a buffer operation on polygon1 using a coefficient multiplied by the converted distance (Km), resulting in an expanded polygon (polygon2).

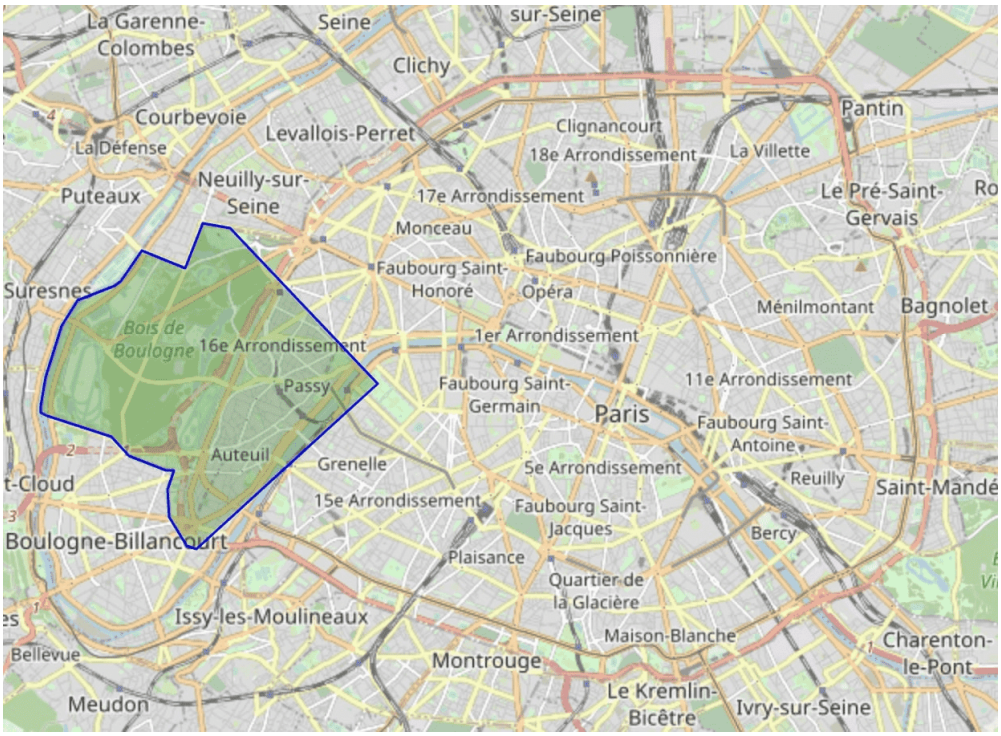
(f) **Difference Operation**: Extract the difference between polygon2 and polygon1. This operation yields Level-3 coordinates that correspond to the region added by the buffer operation, representing the specified distance.

(g) **Output Level-3 coordinates**: If Level-3 coordinates were successfully extracted, return the level3_coordinates. If the spatial keyword is not Level-3 spatial relation, then return the original coordinates.

3. **Visualization of spatRE:** The geographical coordinates of spatRE in the form of GeoJSON can be visualized using OpenStreetMap (OpenStreetMap contributors, 2017) leaflet. The OpenStreetMap leaflet is produced using the *Folium* Python library that is built on top of the Leaflet JavaScript library. The leaflet allows Python developers to create interactive maps using Leaflet.js maps in Python. This visualization can help to visualize the regions for spatRE. Moreover, in the next subsequent section, it is used to evaluate the shapes of the spatRE. Some examples of spatRE polygons visualization using OpenStreetMap Leaflet are shown in Figures 3.6, 3.7 and 3.8.



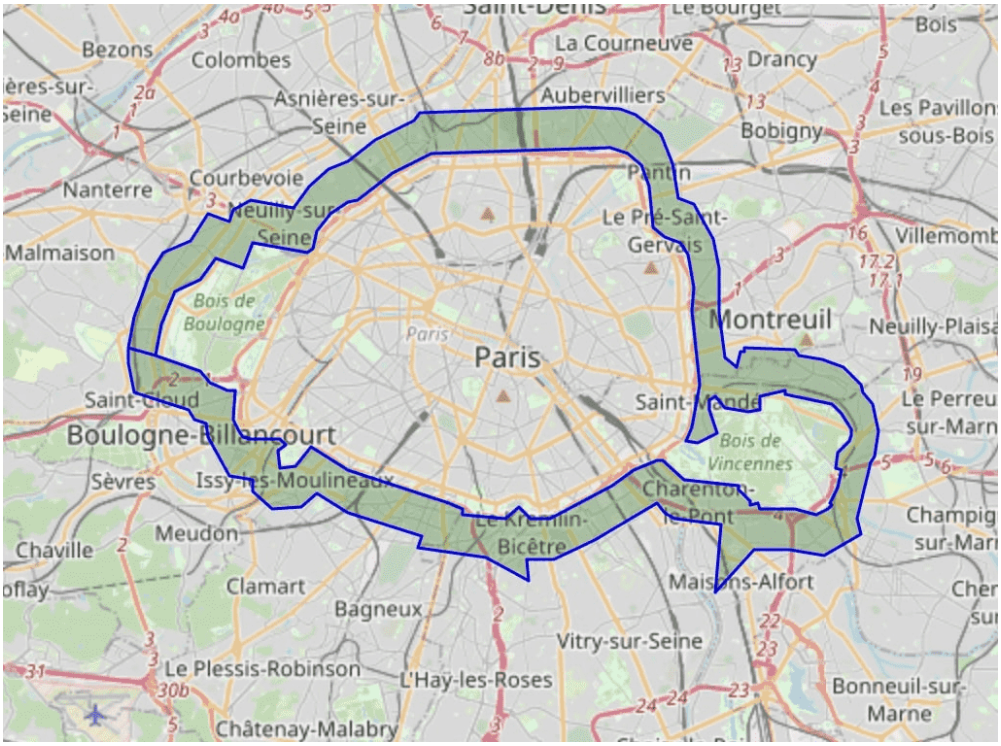
(a) Level-1 East



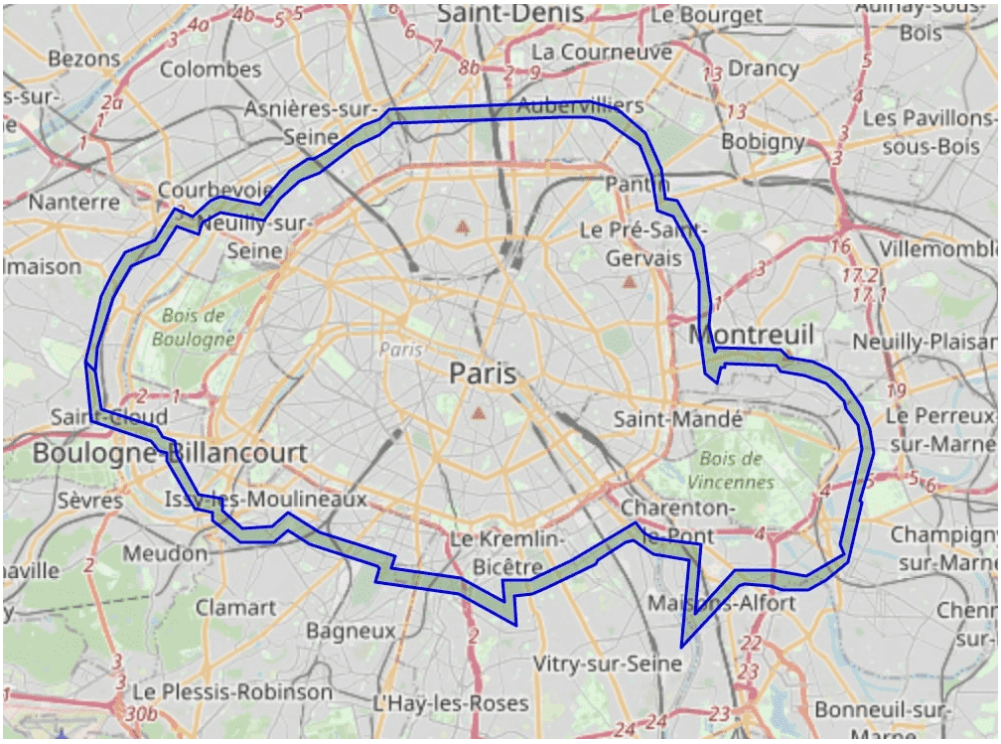
(b) Level-1 West

Figure 3.6: Level-1 spatial relations

The overarching methodology of the presented research follows a two-phase approach. Consequently, it is imperative to rigorously validate each individual phase. The assessment of both phases is expounded upon in the subsequent Section 3.7.

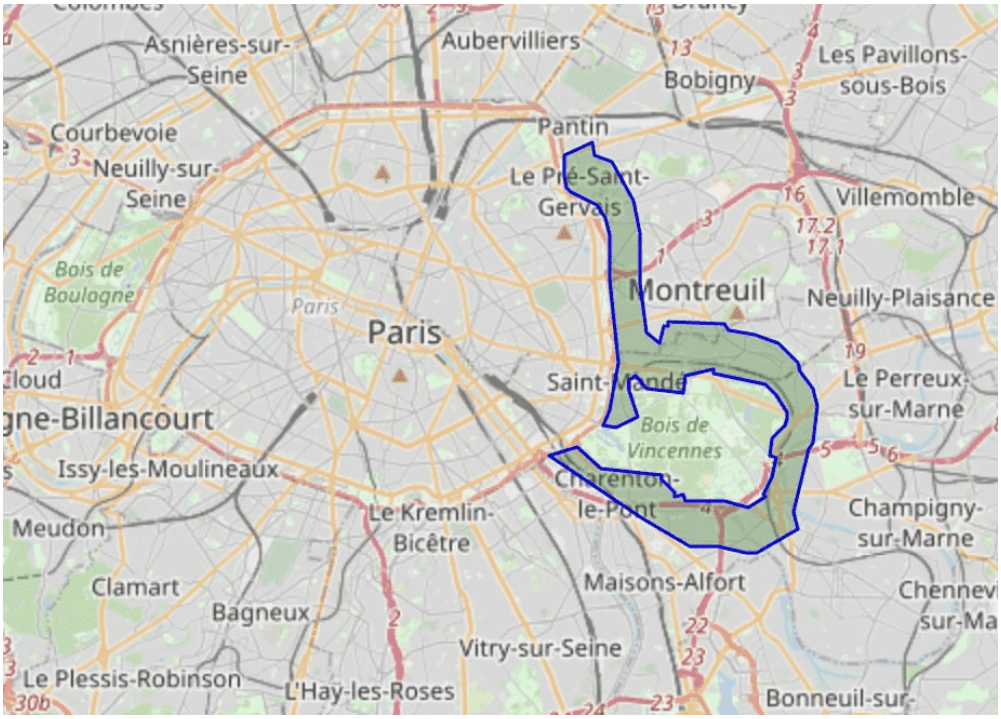


(a) Level-2 Border

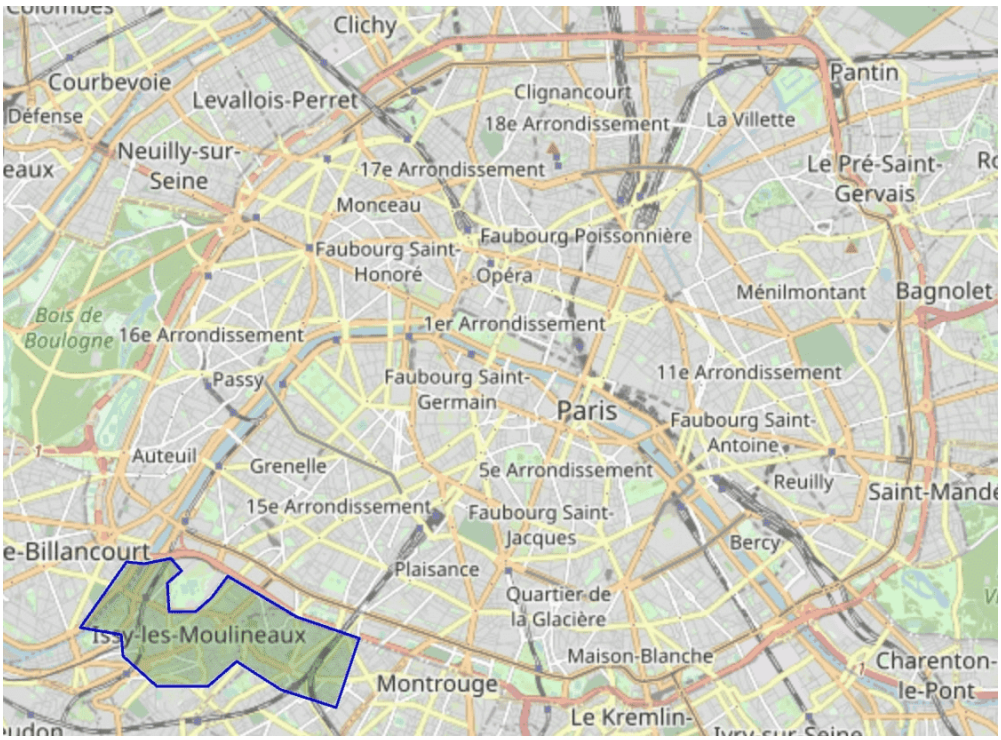


(b) Level-3 1-km

Figure 3.7: Level-2 & Level-3 spatial relations



(a) Compound East



(b) Compound South-West

Figure 3.8: Compound spatial relations

Data and Software Availability

All the contributions of this chapter were implemented in a software called *GeospaCy* (Syed et al., 2023b). Figures 3.6, 3.7 and 3.8 are generated through this software. The demonstration of this

software explains the steps involved in extracting spatREs from a free text (Figure 3.9) and geocoding them into geographical coordinates (Figure 3.10). The details of these steps, followed by software snapshots, are as follows:

1. In step 1 (Geoparsing), we can select the language model for NER task, e.g. en_core_web _sm (see Figure 3.9).
2. In step 2, we can choose the spatial entity type, e.g., RSE for spatRE, GPE for absolute spatial entity (see Figure 3.9).
3. In step 3, we can enter the text to extract the spatial entities (see Figure 3.9).
4. Step 4 shows the entered text with labelled spatial entities (see Figure 3.9).
5. In step 5, we can select the spatial entities from the list to visualize the geographical coordinates on map or GEOJson format (see Figure 3.9).
6. In step 6 (Geocoding), we can change the visualization, e.g., map or GEOJson (see Figure 3.10).
7. In step 7, we can choose to compute the polygon by midpoint or mid-midpoint (see Figure 3.10). However, this case is valid only for Level-1 spatial relations.
8. In step 8, we can export the geographical coordinates in GEOJson format on local machine (see Figure 3.10).
9. Step 9 shows the geographical representation of spatial information on map with associated information e.g., ASE is the place name, Level-1, Level-2 and Level-3 are the spatial relations (see Figure 3.10).

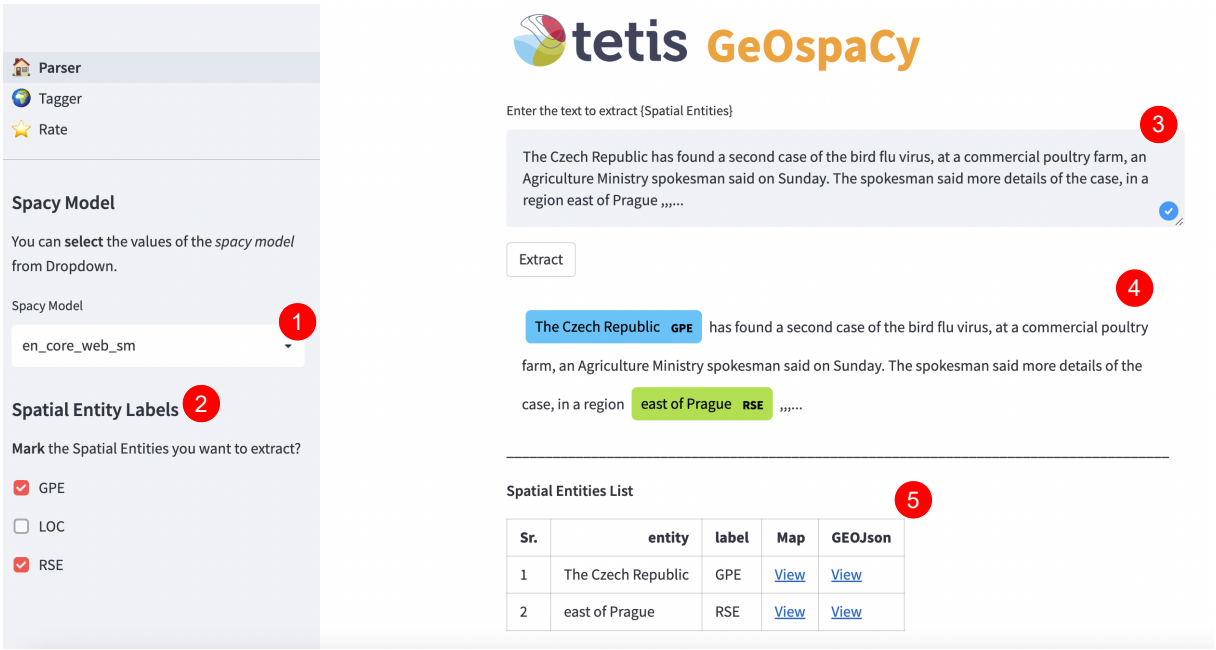


Figure 3.9: Geoparsing Method

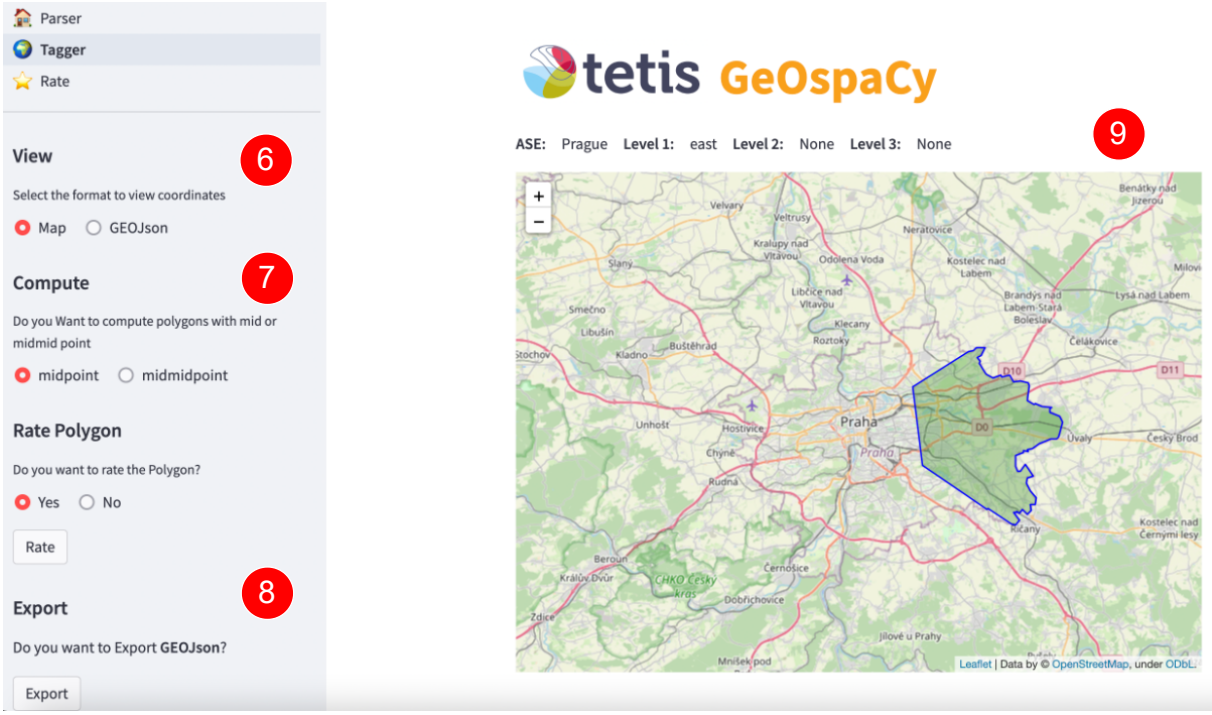


Figure 3.10: Geocoding Method

The code, datasets and results associated to this work are available in the GitHub repositories with details in Table 3.2.

Data	URL
Extraction Dataset	https://tinyurl.com/2wn7ywth
Geocoding Evaluation	https://tinyurl.com/bddu752y
Code Repository	https://github.com/mehtab-alam/GeospaCy
Shapes (Paris & London)	https://tinyurl.com/2489m4xw
Shapes (Milan & Madrid)	https://tinyurl.com/yt5m4b2m
Shapes (Zagreb & Utrecht)	https://tinyurl.com/yasy5bxz
Shapes (Delft & Lyon)	https://tinyurl.com/3z37dhu3
Shapes (Florence)	https://tinyurl.com/ycybm6r5

Table 3.2: Code, Datasets and Results

3.7 Results

The assessment of the proposed methodology involved the utilization of distinct datasets tailored to the requirements of each respective phase. The two principal phases of the study yielded distinct outcomes: 1) the extraction of spatREs from textual data, and 2) the identification of geographical coordinates associated with the spatREs. The outcomes of these two phases are expounded upon in the ensuing subsections.

3.7.1 Extraction Phase

The evaluation of the spatRE extraction from text is approached through a state-of-the-art assessment mechanism. We evaluated this phase with "Geoparsing Dataset" (see Section 3.4.1). This dataset is served as a benchmark against for performance evaluation of the extraction process. Specifically, the evaluation of NER performance revolves around key metrics, including recision, recall, and the F-Score (Hakala and Pyysalo, 2019; Resnik and Lin, 2010). The definitions of precision, recall, and the F-Score are as follows (Goutte and Gaussier, 2005):

$$Precision = \frac{Correct\ spatRE\ Recognized}{Total\ spatRE\ Recognized} \tag{3.1}$$

$$Recall = \frac{Correct\ spatRE\ Recognized}{Total\ spatRE\ in\ Corpus} \tag{3.2}$$

$$F - Score = 2\ \mathrm{X}\ \frac{Precision\ \mathrm{X}\ Recall}{Precision\ +\ Recall} \tag{3.3}$$

The results are calculated with the precision, recall and F-Score as shown in Table 3.3. "spatRE" are the extracted entities and "rse" are the annotated entities. For instance, the first row of Table 3.3 displays the results obtained from articles that contain events. In this case, 4 spatREs were extracted. The evaluation of these spatREs yielded precision, recall, and F-score performance measures of 1, 0.8, and 0.88, respectively.

Disease Name	**No. of Articles**	**spatRE**	**rse**	**Precision**	**Recall**	**F-Score**
Antimicrobial resistance (AMR)	25	4	5	1	0.80	0.88
COVID-19	100	100	92	0.87	0.94	0.90
Avian-Influenza	150	57	68	0.87	0.83	0.84
Lyme	29	10	10	0.83	1	0.90
Tick-borne Encephalitis (TBE)	73	73	81	0.93	0.83	0.87
Average	377	244	256	**0.9**	**0.88**	**0.88**

Table 3.3: Extraction Phase Results (spatRE Extraction)

Precision, recall, and F-Score metrics are individually computed for each disease dataset. The comprehensive assessment across all disease datasets culminates in an overall score characterized by a precision of ***0.9***, a recall of ***0.88***, and an F-Score of ***0.88***. Throughout the evaluation process, specific scenarios were taken into consideration with respect to spatRE instances. For example, entities such as 'North America' and 'South Korea' were deemed as spatREs. However, certain cases were identified as false positives during the evaluation process. Instances included spelling errors in spatREs, as seen in cases such as '(Yonhap)South Korea', or situations in which words were concatenated with spatREs, forming a single word that conveyed geospatial relations.

3.7.2 Geocoding Phase

We generated spatial relations polygons for 9 European cities discussed in Section 3.4.2. For each spatial relation, Level-1, Level-2, and Level-3 polygons were generated using Algorithm 3. These polygons were subsequently generated for various geographically distinct cities to evaluate the proposed methodology.

Given the absence of a standardized mechanism for evaluating the geometric accuracy of geographical coordinates, a qualitative survey was conducted to assess each shape effectiveness. A well-defined shape is characterized by a polygon boundary that closely approximates the information conveyed by the corresponding spatRE. The evaluation criteria encompass two key aspects: 1) the degree to which the shape geometry faithfully represents the intended information, and 2) the extent to which the geometry provides an accurate visualization of the geospatial relations pertinent to the associated city. This evaluation process involves active participation from end-users well-versed in GIS and geospatial information applications.

The evaluation process involves each shape being assessed by two distinct end-users, and the average score from both users is computed. Each relation shape corresponding to the selected cities undergoes evaluation by 4 groups of distinct end-users who are participants in the MOOD[3] project. Moreover, each group consists of two members. These end-users engage in the evaluation process by ranking the shapes on a scale ranging from ***1*** to ***4***. The "Rank" scale assigned by the end-users is as follows: ***1*** signifies ***unclear***, ***2*** denotes ***weak*** or not bad, ***3*** indicates ***better*** defined, and ***4*** corresponds to ***well-defined***. For the city of 'Florence', evaluation was conducted by all members of the evaluation group. The total score for each city, considering all geospatial relations, amounts to *152*, except for 'Florence', which holds a total score of *608*. The average score represents the collective scores assigned by the participants.

For instance, the first row of Table 3.4 showcases the average score computation for 'Paris' spatial relations, including north, south, east, and west based on evaluations. The cumulative score reaches 152, while the aggregated score amounts to 136. After the evaluation, the accuracy in determining the shapes of spatial relations for Paris stand at 89.5%, with an average score of 3.6 out of 4, leading to an "Excellent" designation in the remarks section. Table 3.4 displays the average score assigned by the group of end-users for each city, including 'Accuracy', 'Mean' as the fundamental unit of evaluation by the end-users, and 'Remarks' pertaining to the geometries of all geospatial relations for the city. Table 3.4, it becomes evident that the shapes of geospatial relations are better defined for 8 out of 9 cities. Notably, for the city of 'Madrid', anomalies are observed in different geospatial relations, indicating that the accuracy falls short of expectations.

[3] https://mood-h2020.eu/

City	Average Score	Total Score	Accuracy	Mean	Rank
Paris	136	152	**89.5**	3.6	**Excellent**
London	142	152	**93.4**	3.7	**Excellent**
Milan	106	152	69.7	2.8	Good
Madrid	77	152	50.7	2	Weak
Zagreb	116	152	**76.3**	3.1	**Excellent**
Utrecht	105	152	69.1	2.8	Good
Delft	121	152	**79.6**	3.2	**Excellent**
Lyon	114	152	75	3	Good
Florence	477	608	**78.5**	3.1	**Excellent**

Table 3.4: Qualitative Evaluation of Spatial Relation by City

Table 3.5 illustrates the qualitative evaluation of Level-1 spatial relations shapes. These spatial relations are north (N), south (S), east (E), west (W), northeast (NE), northwest (NW), southeast (SE), southwest (SW) and Central. The scores assigned to each geographical shape reflects the cumulative quality assessment. Higher score indicates better definition of shapes. The qualitative ranks categorize the overall quality of Level-1 shapes for each spatial relation. Overall, the analysis indicates that spatial relations i.e., E, W, NE, NW, SE, SW, and Central are "excellent" as they accurately represent the geographical region. These spatial relations are robust and accurate in the context of the evaluation. On the other hand, N and S are rated as "Good", which implies they are reasonably well-defined but are not at the same level of excellence as the others. Whereas, E stands out as the top-performing spatial relation, as its polygon in all the cases (cities) are very well represented in terms of accuracy.

	N	S	E	W	NE	NW	SE	SW	Central
Score	67	61	82	73	75	78	81	78	82
Total Score	96	96	96	96	96	96	96	96	96
Accuracy	69.8	63.5	**85.4**	**76**	**78.1**	**81.3**	**84.4**	**81.3**	**85.4**
Mean	2.9	2.4	3.5	3.1	3.1	3.4	3.3	3.2	3.4
Rank	Good	Good	**Excellent**	**Excellent**	**Excellent**	**Excellent**	**Excellent**	**Excellent**	**Excellent**

Table 3.5: Qualitative Evaluation of Level-1 Spatial Relations

Tables 3.6 and 3.7 provide a qualitative evaluation of Level-2,Level-3 and Compound spatial relations in terms of assessment of their performance. Some relations, such as Border, SW Border, E Border, and 2 Miles+Border, have notably high accuracy percentages, exceeding 80%, while others like NE Border, NW Border, SE Border, N Border, and S Border have slightly lower but still respectable accuracy levels. On the other hand, NE Border, NW Border, SE Border, N Border, and S Border scores indicating some room for improvement. Overall, the analysis demonstrates that certain spatial relations, such as Border, SW Border, E Border, and 2 Miles+Border, are particularly accurately representing the geographical regions.

	Border	NE Border	NW Border	SE Border	SW Border
Score	87	65	63	72	77
Total Score	96	96	96	96	96
Accuracy	**90.6**	67.7	65.6	75	**80.2**
Mean	3.6	2.7	2.7	3	3.3
Rank	**Excellent**	Good	Good	Good	**Excellent**

Table 3.6: Qualitative Evaluation of Level-2, Level-3 and Compound Spatial Relations

	N Border	S Border	E Border	W Border	2 Miles+Border
Score	59	61	78	71	84
Total Score	96	96	96	96	96
Accuracy	61.5	63.5	**81.3**	74	**87.5**
Mean	2.6	2.5	3.3	2.9	3.5
Rank	Good	Good	**Excellent**	Good	**Excellent**

Table 3.7: Qualitative Evaluation of Level-2, Level-3 and Compound Spatial Relations

3.8 Discussion

The proposed work is aimed at extracting geospatial relation entities from text data and at determining their geographical coordinates for visualization on GIS. This two-step process involves the extraction of spatial relation entities followed by the geocoding of these entities. The outcomes of this research have significant implications in various domains, such as disease surveillance, disaster identification, and event surveillance systems, and in approximating the region of interest in identifying geospatial relations with locations from informal sources of information. For instance, we consider the following text from a digital news article: "One outbreak was noted in Podkarpackie province (east of Poland), one in a farm in south of Warsaw in Mazowieckie province and one in Lubuskie province in the west of Poland" [4]. The locations of the outbreaks in the text are 'east of Poland' and 'south of Warsaw' rather than 'Poland' and 'Warsaw' respectively. It is crucial to identify the accurate region of the outbreak to provide appropriate information to health officials in this context. The same situation holds for other geographically sensitive alert systems. By accurately identifying spatial relation entities and their corresponding geographical coordinates, this research contributes to improving the accuracy and efficiency of such systems, the early detection of outbreaks, timely response to disasters, and effective surveillance of events in real time. However, certain limitations in the current methodology that need to be addressed. For instance, in the extraction phase, North America, South Africa, South Asia, and South Korea are spatial entities. However, they were recognized as spatREs by our approach which should be discarded. In the geocoding phase, we evaluated the proposed algorithm by generating geographical coordinates for geospatial relations specific to several European and UK

[4] https://www.agroberichtenbuitenland.nl/actueel/nieuws/2020/08/05/significant-increase-in-asf-outbreaks-in-pig-farms-in-poland

cities. However, this approach was limited to few cities, and future work should be aimed at enriching the geocoding dataset. During the evaluation, we observed some irregularities in the shapes of geospatial relations for specific cities, including Madrid. For instance, the algorithm produced geographical coordinates for "2 miles away from Madrid border". However, these coordinates are meaningless without a polygon. Similarly, some irregular polygons were generated by the algorithm, such as 'near to the south of Madrid', 'west Madrid border', and 'vicinity of northwest Madrid' were generated by the algorithm. These irregular polygons need to be further investigated to improve the accuracy of the algorithm. Upon closer examination of the compound spatial relation, which involves a combination of Level-1 and Level-2 spatial relations such as Border with North, it has become apparent that further investigation is necessary to improve the accuracy of the results. Specifically, we need to focus on enhancing the border with Level-1 spatial relations. This enhancement may involve making adjustments to the boundaries to refine the output. Overall, the results for the polygon shapes of spatial relation entities are promising, except for the shapes of Madrid. To improve the accuracy of the algorithm, further investigation is needed to address the limitations and irregularities identified in this research. The proposed research offers a unique perspective on interpreting shapes for different geospatial relations. Table 3.8 presents examples of spatial relations within the context of an avian influenza case study in news articles extracted by PADI-web.

Event sentence	**spatRE**	**polygon**
The chief of the Veterinary Directorate in the southern Zhambyl region, Erbol Zhienqulov, said on September 29 that the village of Qaratas had been locked down after cases of bird flu were confirmed at a local poultry farm on September 21.	southern Zhambyl	https://github.com/mehtab-alam/RSI_case_studies/blob/master/geojson/south_Zhambyl.geojson
France has detected a highly pathogenic strain of bird flu in a pet shop in the Yvelines region near Paris, days after an identical outbreak in one of Corsica's main cities	near Paris	https://github.com/mehtab-alam/RSI_case_studies/blob/master/geojson/near_Paris.geojson
The Czech Republic has found a second case of the bird flu virus, at a commercial poultry farm, an Agriculture Ministry spokesman said on Sunday. The spokesman said more details of the case, in a region east of Prague	east of Prague	https://github.com/mehtab-alam/RSI_case_studies/blob/master/geojson/east_Prague.geojson
In less than a week after bird flu was detected in two poultry farms in Vengeri and west Kodiyathoor in Kozhikode district	west Kodiyathoor	https://github.com/mehtab-alam/RSI_case_studies/blob/master/geojson/west_Kodiyathoor.geojson

Table 3.8: Avian-Influenza Spatial Relations Examples

3.9 Conclusion and Perspectives

The main contribution of this work focused on the extraction of spatREs from the text and its translation into geographical coordinates. We proposed a combination of NLP techniques to extract spatRE i.e., Level-1, Level-2, Level-3 and compound spatREs from the text documents. Afterwards, we proposed a tailored geocoding algorithm to translate spatREs into geographical coordinates compatible to GIS. The methodologies developed in this research have the potential to be integrated in EBS systems, such as PADI-web EBS, in collaboration with our colleagues at CIRAD. This integration could significantly enhance the capabilities of PADI-web and similar systems. Moreover, the accurate extraction and representation of spatial information can also be applied in other domains i.e., disaster management, urban planning, epidemiology, and other fields where geographical context is important. Geographical precision, especially concerning spatial relations, can be achieved through the freely available code provided (see Section 3.6.2).

In essence, the perspective of comparing the geographical aspects of official events and EBS-detected events serves as a critical mechanism for gauging the surveillance system's performance and its alignment with the actual geographical realities of events. This comparative analysis holds the potential to give valuable insights about the coverage of the surveillance system in capturing events within their true geographical context. Therefore, in case of fine-grained accurate spatial information of EBS detected event compare to official event can empower decision-makers and response teams with precise information. Therefore, in case of avian-influenza, there are different organizations reporting the official events having IBS systems i.e., WAHIS by World Organization of Animal Health (WOAH), EMPRES-i by Food and Agriculture Organization (FAO) (Lin et al., 2023). These official events are reported at different geographical granularity level, i.e., Admin level-1, Admin level-2, Admin level-3. So, it is important to compare the official events with EBS detected events at geographical level. Various disease case studies can be examined to compare the spatial granularity of officially reported events and those detected by EBS systems. After the exploration of the geographical precision of events, the next chapter explains the aspect of situational awareness of precise affected regions of events.

Part IV

Spatial Opinion Mining for Situational Awareness of Events

CHAPTER 4

SPATIAL OPINION MINING FOR SITUATIONAL AWARENESS OF EVENTS

This chapter describes situational awareness of events within the context of EBS systems, with a specific focus on social media data, particularly Twitter. Furthermore, our approach emphasizes on spatial opinion mining through various features and identifies the most suitable features for this task. Finally, we underscore the future perspectives of this research.

4.1 Introduction

In recent years in addition to online media (used in chapter 2 of this thesis), the use of data from social media, especially Twitter, has been studied for event detection, including disease surveillance and forecasting, and human health behaviours (Gupta and Katarya, 2020; Bigeard and Grabar, 2019). Therefore, the weak signals of possible outbreaks detected by online media can be verified through social media (Twitter) data specific to the region. Indeed, a significant amount of tweet topics are headline news or persistent news of online media sources (Kwak et al., 2010), thus suggesting that Twitter can be an efficient tool to access event reports in news articles, blog posts, and other sources of online media. A Twitter message referred to as a "tweet" have several attributes timestamp (data and time of tweet), username (author), location, message and hashtags (used to categorize the tweet). In the context of tweet data about events reported, we mainly cover three dimensions, i.e., spatial, temporal and epidemic dimensions.

In the beginning of March 2020, World Health Organization announced COVID-19 outbreak as a global pandemic (Dubey, 2020). The lockdown, at the beginning of the pandemic, affected the social activities of millions of people around the world. During this lockdown, people have used social networks, especially Twitter, to express their feelings and thoughts about COVID-19. For instance, people shared their feelings about different topics impacted by COVID-19, e.g., tracking mental health (stress, anxiety) and vaccination impact (Guntuku et al., 2020; Eibensteiner et al., 2021). These tweets result into different trends on global coronavirus (Fernandes et al., 2020). For instance, these trends encompass epidemiological aspects such as variations in case numbers, recoveries, and mortalities over time, as well as public opinion trends encompassing concerns, misinformation, and preventive measures. These trends were helpful for the health officials and other stake-holders by realizing the health crisis and its impact over different regions (WHO, 2020; Organization et al., 2020). Due to the massive flow of tweets regarding the COVID-19 pandemic, it is difficult to analyse the big flow of information. For instance, in a research study (Sheikha, 2020), Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) were proposed as techniques for dimensionality reduction to determine optimal model parameters. However, it is worth noting that this study did not consider the spatial aspect of tweets. Another research (Decoupes et al., 2021) proposed a Hierarchy-based measure for tweet analysis (H-TFIDF) features from COVID-19 tweets by considering spatial and temporal dimensions. H-TFIDF captures important features that reflect local concerns as by taken into account spatio-temporal aspect. These features illustrate various ways of exploring tweets in the health context of the coronavirus COVID-19 pandemic. By using an adaptive interest of these features, it gives a global insight of the evolution of features over space and time. Furthermore, the H-TFIDF features contribute to greater semantic information richness, which could be helpful for sentiment classification of COVID-19 tweets. To evaluate the impact of H-TFIDF features, it is equally essential to assess their performance in comparison to other feature models such as Bag-of-Words (BOW), Term Frequency-Inverse Document Frequency (TF-IDF), and Bidirectional Encoder Representations from Transformers (BERT) (Aizawa, 2003; Devlin et al., 2018).

4.2 State-of-the-art

The state-of-the-art highlights the contributions of previous work associated with spatial opinion mining in the context of EBS systems. The potential of spatial opinion mining is significantly increasing to complement other EBS systems in order to aware researchers and health experts about the impact of events (Prado et al., 2023). Several studies have been carried out to extract and analyse opinions and sentiments associated with various events, contributing to strengthen situational awareness during outbreaks or other events of interest, e.g., including AI epidemic detection, negative sentiments concerning Zika of targeted population, COVID-19 topics and sentiments (Aramaki, Maskawa, and Morita, 2011; Chandrasekaran et al., 2020; Mamidi et al., 2019; Lamsal, Harwood, and Read, 2022). Spatial opinion is performed by scrutinizing geotagged tweets about health related topics to specific geographical locations. The information in the geotagged tweets not only provide the information about an event but also to gain a deeper understanding of the public sentiments, concerns, and reactions within distinct regions (Kumar, 2022). For instance, people have used Twitter during the first lockdown and after to express their feelings and thoughts about COVID-19 evolution over different time (Fernandes et al., 2020). This enriched contextual information strengthens early event detection, the spatial mapping of outbreak dynamics, and also the formulation of precisely targeted response strategies. Therefore, it is an essential complementary method for increasing sensitivity, timeliness and improved situational awareness during disease outbreaks of EBS systems (Meckawy et al., 2022).

Spatial opinion mining is crucial for understanding the spatial distribution, patterns, and trends associated with the events detected by EBS (Kanankege et al., 2020). Spatial opinion mining is promising, however there are several challenges that must be addressed to fully leverage its potential. Due to the massive flow of social media e.g. tweets data, filtering out irrelevant, noisy or misleading information while preserving context is a substantial challenge (Cheng et al., 2022). Moreover, these opinions in the social media change rapidly and tracking these changes over time while considering spatial dimensions can be computationally intensive and resource-demanding (Xavier et al., 2022). However, handling spatial opinion mining near real time or real time can be useful by associating it with EBS detected events. Moreover, the understanding of the context of tweets data that include sarcasm, irony, or cultural differences, can be challenging for automated spatial opinion mining approaches resulting into misinterpretations (Cortis, 2022). Biases presented in social media data resulting into misleading generalization or skewed sentiments of regional situation, ultimately impacting the post consequences (Shabahang et al., 2023). Additionally, it is equally crucial to ethically collect geotagged tweet data, ensuring the use of consented information and the protection of the privacy of social media users (Zhu et al., 2022). Therefore, addressing these challenges in spatial opinion mining is essential to unlock its potential as a valuable tool for EBS, public health, and various other applications where understanding geographic sentiments and opinions is critical.

A research study (Rosa et al., 2020) proposed an event detection system based on the Deep Belief Network (DBN) to find location based social media opinions for the novel coronavirus (COVID-19) disease. The system claims that it was effective in both the accuracy and timeliness of event, but the focused characteristics of events were limited, i.e., keywords and emotions. In another research, (Edry et al., 2021) proposed spatial opinion mining for the localized stressed situation after the first

lockdown of COVID-19 and performed spatial analysis of how the stress level changes over time. Moreover, (Zhang and Qi, 2022) applied spatial opinion mining for situational awareness on region of interest in order to see the disease spread and population affected. In another research (Umair, Masciari, and Ullah, 2023), spatial opinion mining was performed to analyse people reaction to COVID-19 vaccines in the perspective of correct usage and possible advantages. Though, spatial opinion mining into EBS systems can offer a comprehensive perspective on public opinions and sentiments, enabling more informed situational awareness, the existing literature does not commonly regard it as either a pre-event or post-event step within the EBS pipeline (Fritch et al., 2021).

Different researchers have developed various techniques to analyse spatial sentiment data and gain insights into public sentiment, concerns, and reactions related to events (Jabalameli, Xu, and Shetty, 2022). A research study (Hota, Sharma, and Verma, 2021) proposed lexicon-based approach to analyse the sentiments of specific countries extracted from Twitter COVID-19 specific data to see regional opinions. In another research, (Krishnan et al., 2021) proposed machine learning approaches to perform sentiment analysis on real-time tweets as well as on individual input image/video for early evolved COVID situation. Moreover, in another research, deep learning approaches (Malla and Alphonse, 2021; Fattoh et al., 2022) are used for sentiment classification of COVID-19 tweets. Furthermore, another research (Nimmi et al., 2022) proposed a combination of transformer models with ensemble approach to classify COVID-19 tweets. Moreover, another research (Umair et al., 2022) proposed a combination of BERT and Naive Bayes Support Vector Machine (NBSVM) models for sentiment classification of COVID-19 vaccination psychological impact. Another research proposed concept IDs of dedicated lexicon vocabulary for tweet mentions of adverse drug reaction (ADR) to classify whether a tweet contains a personal mention of one health, a more general discussion of the health issue, or is an unrelated mention for effective health monitoring and surveillance (Weissenbacher et al., 2019). In sentiment classification, feature selection is a crucial process in both supervised learning and unsupervised learning. Improper large feature selection may degrade classifier performance and increase the computational cost (Kumar, 2014).

Feature selection techniques can be used to select an optimal subset of features, reducing the computational cost of training a classifier and potentially improving classification performance (Prusa, Khoshgoftaar, and Dittman, 2015). Another study (Madasu and Elango, 2020) proposed the term frequency inverse document frequency (TF-IDF) as a feature extraction technique to obtain results with different subsets of features. (Wang and Lin, 2020) proposed a new method when selecting a suitable number of features by using the chi-square feature selection algorithm to employ feature selection using a preset score threshold. Another study (Ansari, Ahmad, and Doja, 2019) proposed recursive feature elimination to select the optimal feature set and an evolutionary method based on binary particle swarm optimization of the final feature subset. These approaches were validated for sentiment analysis in five different domain datasets, including movies reviews and Amazon product reviews, but in contrast to health context. Further work (Rustam et al., 2021) proposed a comparison of sentiment classification using different features, i.e., Bag-of-words (BOW), TF-IDF, and concatenation of BOW and TF-IDF to boost the performance. The concatenation of BOW and TF-IDF outperformed as compared to individual feature sets for sentiment classification of COVID-19 tweets. However, the issues with features were the computational cost of model learning and over-fitting of the model. To address this research gap, (Decoupes et al., 2021) proposed a set of features that are extracted from

a COVID-19 tweet dataset by considering the spatial and temporal aspects of COVID-19 data. In this work, the main focus was on the hierarchical characteristics of spatial and temporal dimensions for extracting a more relevant set of features in the context. These important features, i.e., hierarchical term frequency inverse document frequency (H-TFIDF) in the tweets for different regions and time, help determine the local situation, crisis management, and opinions of inhabitants. Moreover, these reduced sets of features (H-TFIDF) may be important for sentiment classification of COVID-19 tweets. Although, there are various techniques for opinion mining in the literature that have shown superior performance for the task, there is a noticeable gap in the literature when it comes to utilizing diverse textual features for spatial opinion mining.

4.3 Objectives and Contributions

The main objective of our research is to enhance situational awareness of events over time in EBS focusing on social media (Twitter) through spatial opinion mining. The subsidiary objectives to support the main objective are as follows:

1. How spatial opinion mining can improve EBS systems, focusing on spatial information through H-TFIDF features, as H-TFIDF highlights discriminative features of local concerns at country level (Decoupes et al., 2021).

2. How sentiment classification of COVID-19 tweets performs through different feature sets in order to find the best features among them.

The research outcome of this chapter is a novel approach to spatial sentiment analysis, aimed at gaining insights into the regional situation during a specific time period. By applying the techniques and methodologies discussed in this chapter, the study introduces the use of spatial features, particularly H-TFIDF for sentiment analysis. The primary focus is on comparing the performance of H-TFIDF with other feature sets like BOW, TF-IDF, and BERT. The findings of this research chapter contribute to the understanding of regional sentiments and emotions expressed by people during the specified time period (January 2020). By leveraging H-TFIDF as a spatial feature, the study provides a nuanced understanding of how sentiments vary across different geographical locations. The insight into these regional variations in sentiment can be vital for public health officials and policymakers. It enables them to more effectively gauge public concerns, allowing for the targeted development of communication campaigns. Furthermore, the comparison of H-TFIDF with traditional feature sets (BOW, TF-IDF, BERT) sheds light on the effectiveness of this approach for sentiment analysis. This comparative analysis enables researchers and practitioners to assess the strengths and limitations of each feature set in spatial context and make informed decisions about which technique is more suitable for sentiment analysis in similar context.

4.4 Datasets

In order to perform spatial opinion mining, we utilized two datasets for this task i.e., 1) Training dataset for training model, 2) COVID-19 geocoded tweet dataset for sentiment analysis. The detail of these datasets are as follows:

The **training dataset** is the well-known Kaggle Sentiment140 dataset for sentiment analysis of tweets for English language only. The dataset is available at: Kaggle Sentiment140 dataset [1] (KazAnova, 2016). It has labelled data for supervised learning for the classification of tweets. The dataset contains 1.6 million annotated tweets, equally balanced between two classes. Tweets are being annotated as (0 = negative) and (4 = positive) and later on by trained model will be used as detection for sentiments for COVID-19 tweets data. The training dataset for learning models will be used for binary classification of tweets. Figure 4.1 shows the detail description of the class distribution of training dataset.

Figure 4.1: Training Dataset Description

The **COVID-19 tweet dataset** is extracted from a dataset collected by E. Chen, which can be found on GitHub repository [2] (Chen, Lerman, and Ferrara, 2020). The dataset is curated to capture tweets from the early beginning of the COVID-19 outbreak. The tweets IDs of the COVID-19 were extracted using Twitter Streaming API by using COVID-19 related keywords. The analysis dataset contains 165,537 tweets. Each tweet contains the information ID, UserID, text, location, country and its creation_date respectively. Moreover, H-TFIDF discriminative terms are also extracted for the same dataset (Decoupes et al., 2021). This dataset is used for the analysis in this study.

[1] https://www.kaggle.com/kazanova/sentiment140
[2] https://github.com/echen102/COVID-19-TweetIDs

4.5 Spatial Opinion Mining

We performed spatial opinion mining of COVID-19 tweets through different features, i.e., H-TFIDF, (BERT+ H-TFDIDF) asBH-TFIDF, BOW, and TF-IDF features. The work consists of two main phases, i.e., training phase and prediction phase. The details of each phase are as follows:

4.5.1 Training Phase

In training phase, we considered three machine learning models (linear and non-linear) for performing the task: LR, SVM with linear kernel (Kalcheva, Karova, and Penev, 2020) and RF respectively. These models are trained on the **training dataset** presented in Section 4.4 for sentiment analysis task.

In the first step, we apply data preprocessing to clean the tweets. This step involves removing noise, such as special characters, punctuations, hashtags, and URLs, as well as tokenizing the text into individual words. The cleaning was performed using the Python library *tweet-preprocessor* (Özcan, 2016). Afterwards, text standardization was applied by converting words into lowercase. The benefit of data preprocessing enhances the quality and utility of the data for effective model training and analysis.

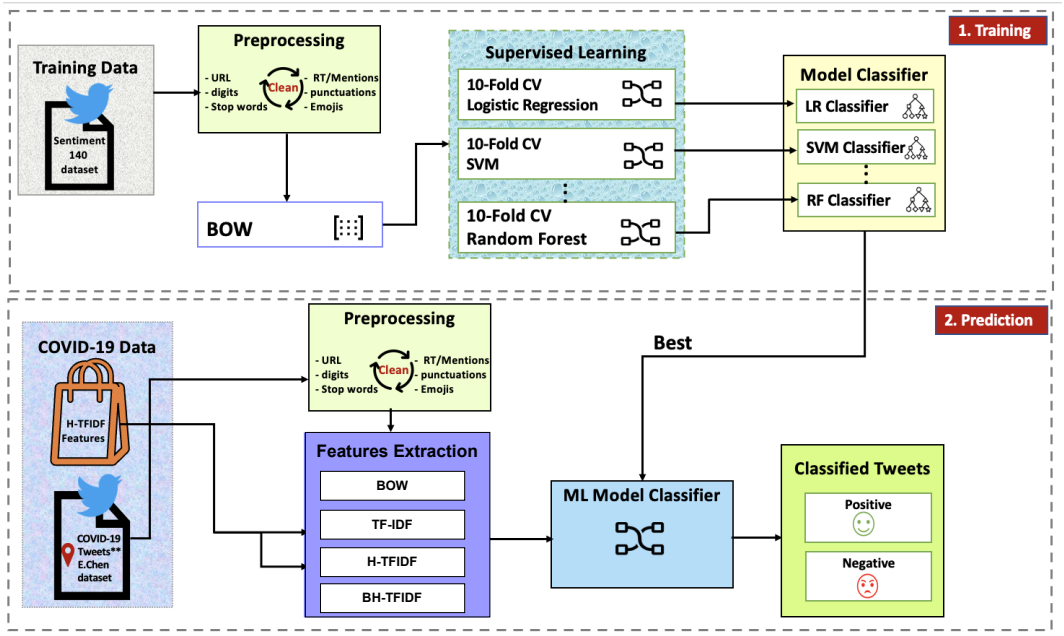


Figure 4.2: Spatial Opinion Mining Pipeline

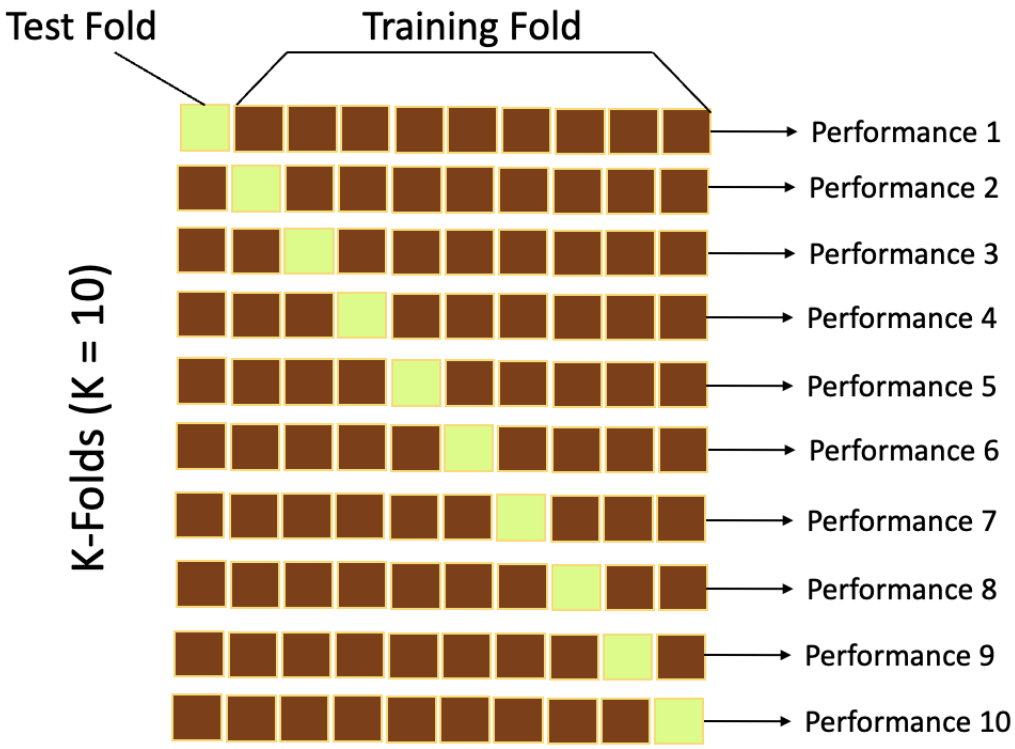


Figure 4.3: Cross Validation

In second step, we used various machine learning models, such as, Logistic Regression (LR), Support Vector Machine (SVM) classifier, Random Forest (RF). These models use the BOW features to make predictions. In order to reduce model overfitting (Berrar, 2019), a cross-validation strategy (Raschka, 2018) is applied to split dataset into multiple subsets for training and evaluating the model performance. For experimental setup, a train-test validation scheme of 90% and 10% is used with 10-fold cross validation as shown in Figure 4.3. Figure 4.2 shows the process pipeline of spatial opinion work. Table 4.1 shows the average performance of models in terms of precision, recall and F-score. We chose the best model (LR) with 0.79 F-score for the sentiment classification task.

	BOW		
	Precision	Recall	F-Score
LR	0.80	0.79	0.79
SVM	0.71	0.70	0.70
RF	0.61	0.63	0.61

Table 4.1: Machine learning models performance with 10-fold cross validation

4.5.2 Prediction Phase

In prediction phase as shown in Figure 4.2, we utilized different feature sets i.e. BOW, TF-IDF, H-TFIDF, and BH-TFIDF for best model for sentiment classification task. COVID-19 tweets dataset (see Section 4.4) is utilized for the sentiment prediction. The tweets are initially preprocessed to

supply to the model. Few examples of preprocess tweets are as follows:

```
<Tweet1>: @pearlylondon Don't worry, if she does contract
a fatal dose of coronavirus at least she will have a
dignified burial \n#Blackadder https://t.co/8KdpMIItki

<Tweet2>:5 Confirmed cases of #coronavirus in Brighton.
In the meantime, local news...
#Brighton https://t.co/KTXkQCOApg

<Preprocessed Tweet1>: do not worry she does contract
fatal dose coronavirus least she will have dignified
burial

<Preprocessed Tweet2>: confirmed cases brighton
meantime local news
```

In next step, we extracted features i.e. BOW, TF-IDF, H-TFIDF and BH-TFIDF for the sentiment classification task. We employed these features to evaluate their performance in the sentiment classification task. The details of these features are discussed as follows:

1. **BOW**: The Bag-of-Words (BOW) model represents text data as a collection of words, disregarding their order and structure but keeping track of their frequency (Qader, Ameen, and Ahmed, 2019). We utilized these features from prediction dataset (COVID-19 unlabelled dataset) to perform sentiment classification.

2. **TF-IDF**: Term Frequency-Inverse Document Frequency (TF-IDF) is commonly used in information retrieval and text mining to evaluate the importance of a word within a document relative to a collection of documents (corpus). (Qaiser and Ali, 2018; Yahav, Shehory, and Schwartz, 2018). We utilized these features from prediction dataset (COVID-19 unlabelled dataset) to perform sentiment classification. TF-IDF is formulated as follows:

$$tf\text{-}idf_{t,d} = tf_{t,d} * idf_t \tag{4.1}$$

$$tf_t = f_t \Big/ f_{tot} \tag{4.2}$$

$$idf_t = \log(N \Big/ df_t) \tag{4.3}$$

3. **H-TFIDF**: H-TFIDF features are the discriminative features extracted by considering spatial and temporal window from the early beginning of the outbreak (Decoupes et al., 2021) in the prediction dataset. We utilized these features from prediction dataset (COVID-19 unlabelled dataset) to perform sentiment classification.

4. **BH-TFIDF**: BERT generates word embeddings for each subword token in a text. These embeddings are contextually rich, meaning they take into account the entire sentence context when representing each word (Devlin et al., 2018). The main purpose of integrating BERT is

to enhance H-TFIDF features in terms of enriching the contextual vocabulary. We combined BERT with H-TFIDF named as (BH-TFIDF) from prediction dataset (COVID-19 unlabelled dataset) to perform sentiment classification.

The workflow of prediction phase is shown in Figure 4.2. Finally, the tweets are classified using these features with the best-performing logistic regression (LR) model. The results and accompanying discussions are presented in the following section.

Data and Software Availability

The experiments were performed to classify spatial tweet groups through various features set in order to know the local situation. The whole workflow is divided into two main components, i.e., 1) Sentiment analysis of COVID-19 Tweets Dataset and 2) Visualization of results through Venn representation. The code, datasets, and results are available at GitHub repository [3].

4.6 Results & Discussion

The results are presented and analysed in three ways, i.e., 1) country-wise comparative analysis, 2) features comparison for sentiment classifications and 3) Venn representation of classified tweets. These results are discussed as follows:

4.6.1 Country-wise comparative analysis

We presented the initial outcomes of sentiment classification using the H-TFIDF feature set to analyze sentiments on a country-by-country basis. The classification results spatial location tweet groups into positive and negative. In particular, we selected the tweets from three countries i.e., Italy, France, and United Kingdom are analysed for the results, as the focus of MOOD [4] project is European countries. The overall sentiments of all tweets are slightly more negative than positive. A same pattern of the sentiment is observed for the United Kingdom, France, and Italy in the early beginning of outbreak. This could be different for the different pandemic period. Unlikely to France and UK, the sentiments in Italy were less negative. Figure 4.4 shows the detail of sentiment analysis for spatial tweet groups.

4.6.2 Features comparison for Sentiment Classification

Binary classification as positive and negative were predicted with 4 different features. The tweets were classified into positive and negative. Overall classified positive tweets and negative tweets using different feature sets are shown in Table 4.2.

[3] https://github.com/mehtab-alam/spatial_opinion_mining
[4] https://mood-h2020.eu/

	Classification	
	Positive	**Negative**
BOW	79000	90538
TF-IDF	96522	73016
H-TFIDF	77452	92086
BH-TFIDF	97536	72002

Table 4.2: Overall sentiment classification count

It is difficult to label all tweets of the prediction dataset. However, we manually analysed and labelled a subset of tweets from the prediction dataset. For evaluation, we labelled 500 tweets as positive or negative, which was considered as gold standard. Furthermore, the state-of-the-art evaluation of the performance of a classification task were measured for each feature results i.e. BOW, TF-IDF, H-TFIDF, and BH-TFIDF with gold standard for class 'positive' and 'negative' respectively. The table 4.3 presents the results of sentiment analysis for positive tweets, showing the precision scores for different feature sets used in the sentiment analysis. The precision measure is used because the classified positive tweet can either be 'True Positive' or 'False positive'. Precision is one of the evaluation metrics used to measure the model performance in correctly predicting positive sentiments (Rustam et al., 2021).

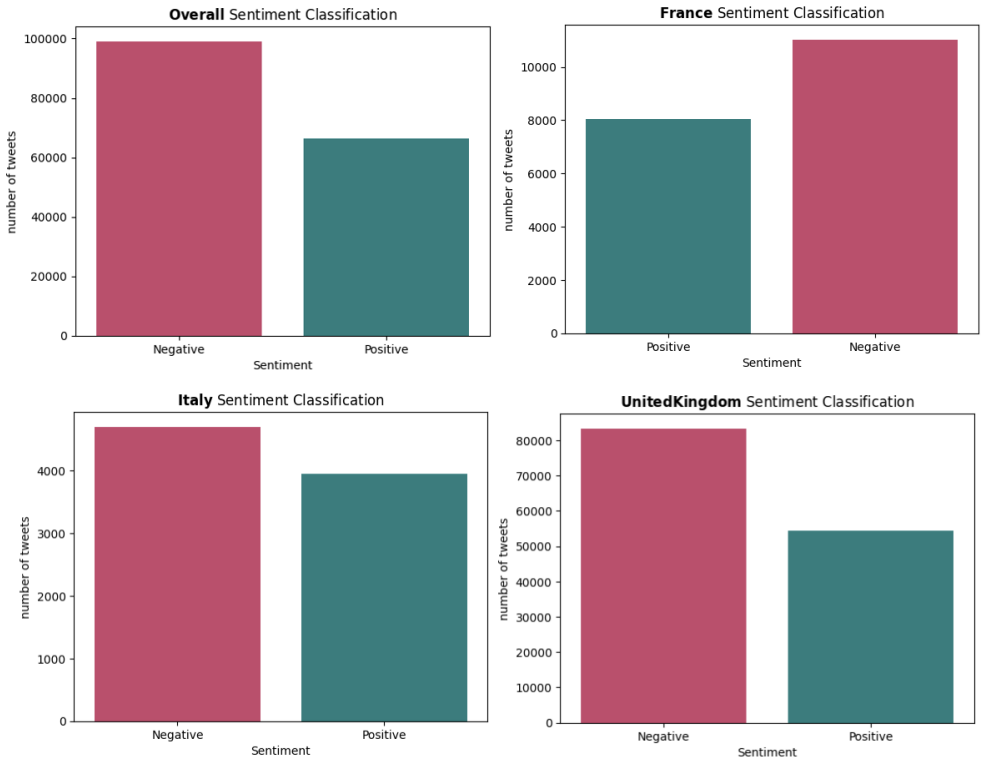


Figure 4.4: Sentiment tweet groups by H-TFIDF features

Features	Precision
BH-TFIDF	0.84
H-TFIDF	0.34
TF-IDF	0.81
BOW	0.42

Table 4.3: Precision of LR with different Features for **Positive** tweets classification

The model using BH-TFIDF features achieved the highest precision score of 0.840. This means that when the model predicted a tweet to be positive, it was correct for 84% of the time. Similarly, the model using H-TFIDF features achieved a precision score of 0.34. It had the lowest precision among the featured sets, indicating that the model's positive predictions were accurate only for about 34% of the time. Moreover, the model using TF-IDF features obtained a precision score of 0.81, performing well in identifying positive tweets, with correct results around 81%. Furthermore, the model using Bag-of-Words (BOW) features achieved a precision score of 0.42, indicating that the model's positive predictions were correct around 42% of the time. In summary, the BH-TFIDF feature set outperformed the other feature sets in terms of precision for positive tweets, while H-TFIDF had the lowest precision. The results provide insights into the effectiveness of different feature sets in identifying positive sentiments in the sentiment analysis of tweets data.

In the Table 4.4, the results of sentiment analysis for negative tweets are presented, and the precision scores for different feature sets used in the analysis are shown.

Features	Precision
BH-TFIDF	0.35
H-TFIDF	0.58
TF-IDF	0.35
BOW	0.80

Table 4.4: Precision of LR with different Features for **Negative** tweets classification

The model using BH-TFIDF features obtained a precision score of 0.35. This means that when the model predicted a tweet to be negative, it was correct about 35% of the time. Subsequently, the model using H-TFIDF features achieved a higher precision score of 0.58. Afterwards, the model using traditional TF-IDF features achieved a precision score of 0.35, which is similar to the BH-TFIDF precision score. Thereafter, the model using Bag-of-Words (BOW) features obtained the highest precision score of 0.80. This suggests that when the model predicted a tweet to be negative, it was correct about 80% of the time. In summary, the BOW feature sets outperformed the other feature sets in terms of precision for negative tweets. The models using these features demonstrated produced significant results in identifying negative sentiments in the sentiment analysis of tweets data.

4.6.3 Venn's representation of classified tweets

We compared classified tweets by different features with a visualization technique called Venn diagram (Ho and Tan, 2021) in order to find common and exclusive classified tweets. The results produced through these features are defined by Equation 4.4 and 4.5.

```
Let B be the BOW set
Let T be the TF-IDF set
Let H be the H-TFIDF set and
Let BH be the BH-TFIDF set
```

$$\begin{cases} Only(B) = B - (B \cap H) - (B \cap T) \\ Only(H) = H - (H \cap B) - (H \cap T) \\ Only(T) = T - (T \cap B) - (T \cap H) \\ Only(B \cap T) = (B \cap T) - (B \cap T \cap H) \\ Only(B \cap H) = (B \cap H) - (B \cap T \cap H) \\ Only(T \cap H) = (T \cap H) - (T \cap H \cap BH) \\ COMMON_{B,H,T} = B \cap H \cap T \end{cases} \tag{4.4}$$

$$\begin{cases} Only(B) = B - (B \cap H) - (B \cap BH) \\ Only(H) = H - (H \cap B) - (H \cap BH) \\ Only(BH) = BH - (BH \cap B) - (BH \cap H) \\ Only(BH \cap B) = (BH \cap B) - ((BH \cap B) \cap T) \\ Only(BH \cap H) = (BH \cap H) - (BH \cap H \cap T) \\ Only(H \cap B) = (H \cap B) - (H \cap B \cap BH) \\ COMMON_{B,H,BH} = B \cap H \cap BH \end{cases} \tag{4.5}$$

For instance, Equation 4.4 defines the classified tweets through BOW, TF-IDF, H-TFIDF features. The first '$Only(B)$' shows the exclusive classified tweets using BOW features, but differently classified by TF-IDF and H-TFIDF features. Similarly, '$Only(H)$' represents the exclusive classified tweets using H-TFIDF features, but differently classified by TF-IDF and BOW features. Moreover, '$Only(B \cap T)$' represents the common classified tweets using BOW features and TF-IDF features, but differently classified by H-TFIDF features. Similarly, Equation 4.5 defines the classified tweets through BOW, H-TFIDF, BH-TFIDF features. The interpretation of the results of the experiments are discussed in the subsequent section.

The Venn diagram represents the relationships and overlaps between the feature extraction method e.g. common and exclusive features of BOW, TF-IDF and H-TFIDF, and the relationship and overlap of classified tweets using these features. For instance, Figure 4.5a shows the positive classified tweets relationship between BOW, TF-IDF and H-TFIDF features. There are 79,000 positive tweets uniquely classified by the Bag of Words (BOW) features, which represent approximately 32.53% of

all positive tweets. Similarly, there are 77,452 positive tweets uniquely classified by the H-TFIDF features, accounting for approximately 31.85% of all positive tweets. Conversely, there are 96,522 positive tweets uniquely classified by the Term Frequency-Inverse Document Frequency (TF-IDF) method, making up approximately 39.62% of all positive tweets. BOW & H-TFIDF in common classified 1,020 positive tweets with 0.42% of total positives tweets, comparatively less than BOW & TF-IDF classified positive tweets 16,302 (6.69%) and H-TFIDF & TF-IDF classified 9,519 (3.91%) of all positive tweets. There are 59,094 positive tweets classified as positive by all three feature extraction method (BOW, H-TFIDF, and TF-IDF). These tweets represent the largest subset and constitute approximately 24.33% of all positive tweets. In Figure 4.5b, there are 97,536 positive tweets uniquely classified by the BH-TFIDF (BERT + H-TFIDF) method, making up approximately 39.15% of all positive tweets. In these results, there is significant overlap between BOW and TF-IDF of 6.44%, and TF-IDF and H-TFIDF of 3.76% and a common overlap of BOW & H-TFIDF & BH-TFIDF of 23.64%. Both sets of results show a similar percentage of positive tweets that are common among all three methods, indicating consistency in their ability to identify common positive sentiments. Figure 4.5 shows that BH-TFIDF and TF-IDF have more capabilities to classify positive tweets.

Similarly, in Figure 4.6, it appears that BOW has the highest number of uniquely classified tweets, while the overlap between all four features contributes to the largest portion of tweet classification. The important features extracted can also be visualized through Venn representations. In Figure 4.7a, the Venn diagram visualizes the common and distinct features extracted from three different feature extraction techniques i.e., BOW, TF-IDF, H-TFIDF for sentiment classification of COVID-19 tweets. The BOW technique captures features like “coronavirus", “china", “people", “virus", “outbreak", “deaths", “economy", “media", “risk", and “work" among others. These features are unique to the BOW method, representing specific frequent words in the tweet dataset. Similarly, the TF-IDF technique identifies features such as “hope", “kill", “italy", “cases", “bbc", “ship", “disease", “trade" and "health" among others. These features are unique to the TF-IDF method and highlight important terms with high importance scores in specific tweets. Moreover, the H-TFIDF technique captures features like “case", “nazi", “positive", “public", “impact", “spread", “hospital", “soon" and “infected" among others. These features are unique to the H-TFIDF method, indicating hierarchical term representations that are significant and uncommon in the context of different spatial region tweets. The Venn diagram also shows the overlap between the different feature sets. For instance, “coronavirus", “china", “virus", “corona" and “outbreak" are common features shared by all three techniques. These terms are likely to be highly relevant and frequent in the tweet dataset, making them significant for sentiment analysis. Some other features appear in combinations of two techniques, showing the commonalities between BOW, TF-IDF, and H-TFIDF. For example, “people", “media", “risk", “kill", “trade", “disease", “stop", “hospital" and “hell" are shared by BOW and TF-IDF, indicating their importance in the sentiment analysis across these two techniques.

Figure 4.7b shows a Venn diagram with three circles, representing the common and distinct features extracted from three different feature extraction techniques i.e., BOW, H-TFIDF, and BH-TFIDF for sentiment classification of COVID-19 tweets. The BOW technique captures features like “coronavirus", “china", “people", “virus", “outbreak", “deaths", “economy", “media", “cost", “japan" and "confirmed," among others. Similarly, the H-TFIDF technique identifies features such as “coronavirus", “china", “people", “virus", “outbreak", “media", “case", “nazi", “positive", “public", “im-

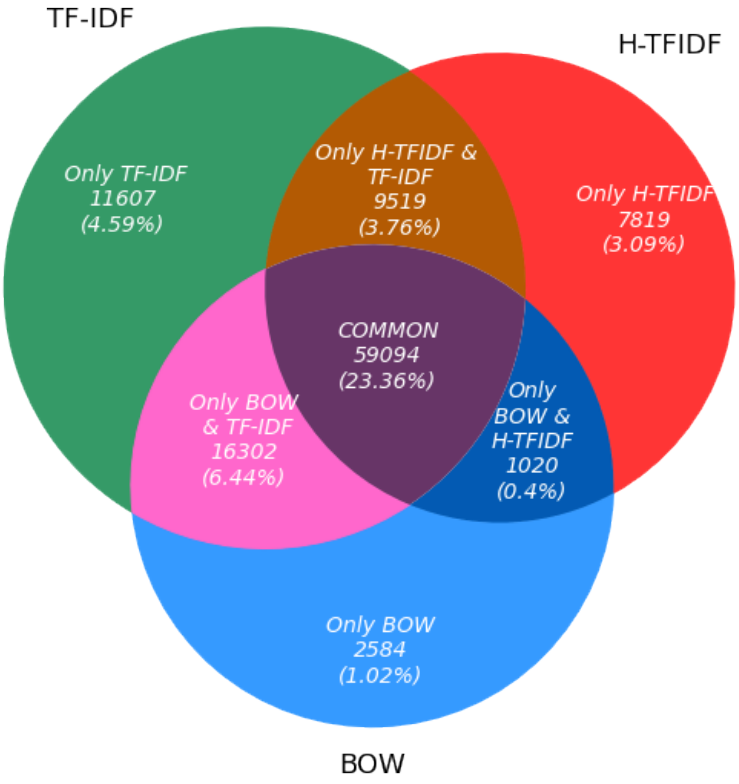
pact", “kill", “deaths", “quarantine", “hospital" and “infected" among others. The BH-TFIDF technique captures features like “coronavirus", “china", “people", “virus" “outbreak", “economy", “risk", “government", “pandemic", “fear", “epidemic", “paedophile", “wuhan", “uk" and “world" among others. These features are unique to the BH-TFIDF method, which combines features from the BERT model and H-TFIDF, representing contextual embeddings and hierarchical term representations for sentiment analysis. The Venn diagram shows the overlap between the different feature sets. For instance, “coronavirus", “china", “people", “virus", “outbreak" and “media" are common features shared by all three techniques. These terms are likely to be highly relevant and frequent in the tweet dataset, making them significant for sentiment analysis.

Some other features appear in combination of two techniques, showing the commonalities between BOW, H-TFIDF, and BH-TFIDF. For example, “cost", “come", “twitter", “kill", “impact", “fault", “confirmed", “year" and “public" are shared by BOW and H-TFIDF or BH-TFIDF, indicating their importance in the sentiment analysis across these techniques.

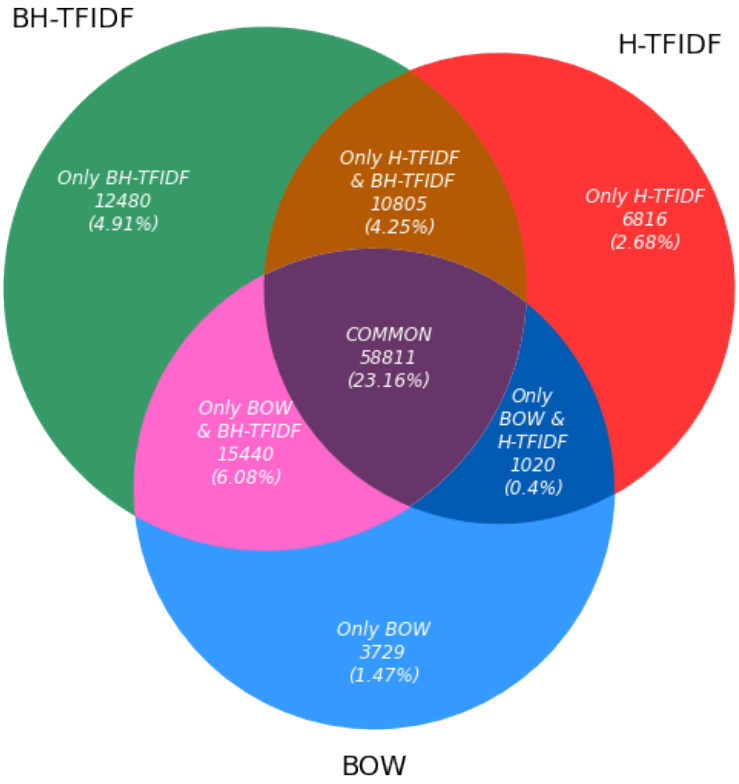
BOW	**TF-IDF**	**H-TFIDF**	**BH-TFIDF**
coronavirus	coronavirus	coronavirus	coronavirus
china	china	china	china
health	death	death	death
spread	health	health	health
cases	news	chinese	*spread*
deaths	pandemic	public	world
travel	*want*	kill	*wuhan*
disease	right	impact	*fault*
trade	travel	fault	kill
economy	hospital	travel	impact

Table 4.5: Top Features of BOW, TF-IDF, H-TFIDF and BH-TFIDF

Table 4.5 shows the top 10 feature terms in the corpus of COVID-19 tweets. The table allows us to compare the common and distinct features extracted by each technique. “Coronavirus" and “china" appear in all four techniques, indicating their strong presence and importance in the dataset across all features methods. Similarly, “health" and “death" are important features identified by BOW, TF-IDF, and H-TFIDF techniques, indicating their relevance in the sentiment analysis task. “spread", “kill", “impact" and “fault" are identified by both H-TFIDF and BH-TFIDF, showcasing their significance in H-TFIDF and BH-TFIDF representations. For instance, BH-TFIDF identified some significant terms that are not highlighted by other methods. Wuhan, China, was the initial epicentre of the COVID-19 outbreak. The term ‘Wuhan’ signifies the geographic origin of the virus and its early impact. Similarly, ‘fault’ investigating the causes and potential areas of improvement in managing the pandemic. In summary, the table provides an insightful comparison of the top features extracted by each technique, offering valuable insights into the relevant and influential terms for sentiment analysis of COVID-19 tweets dataset (see Section 4.4).

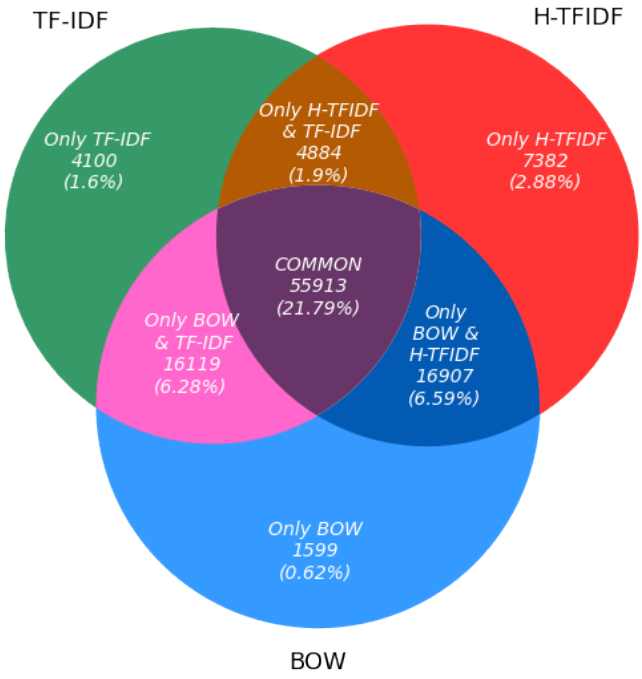


(a) Features: BOW, TF-IDF and H-TFIDF

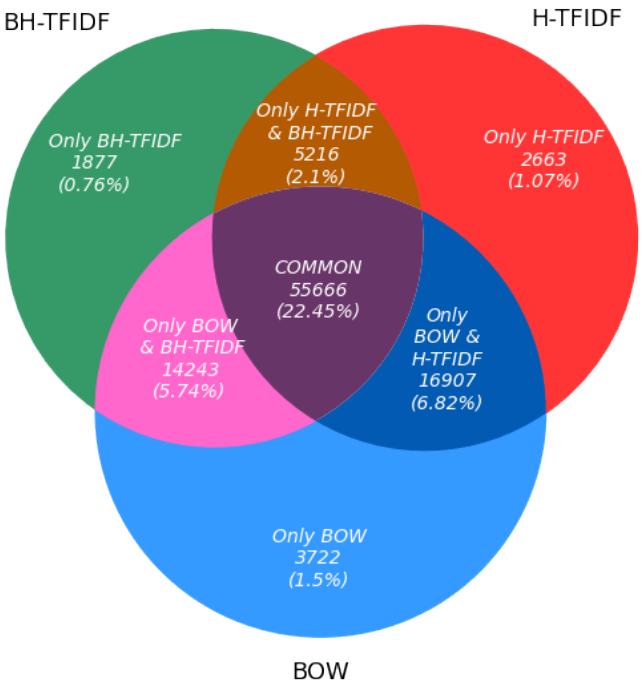


(b) Features: BOW, H-TFIDF and BH-TFIDF

Figure 4.5: Positive tweets comparison by features

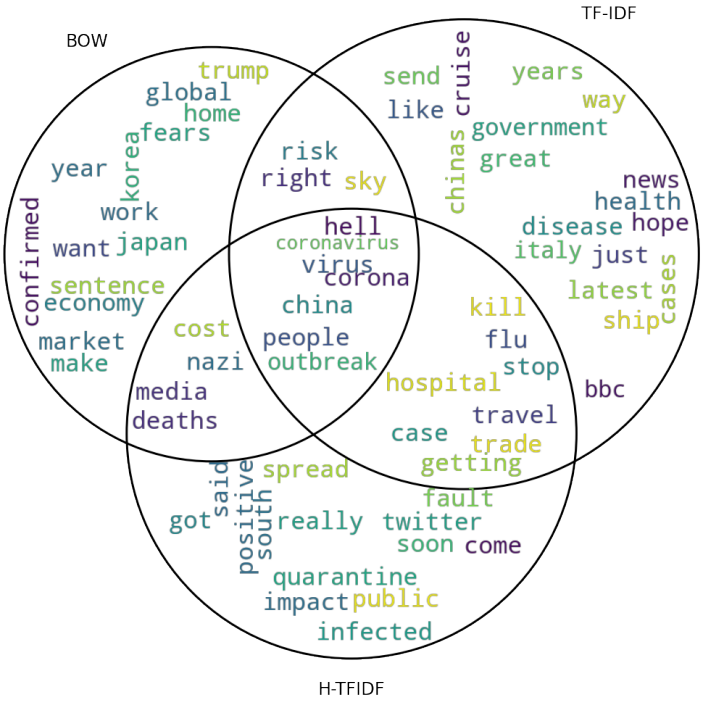


(a) Features: BOW, TF-IDF and H-TFIDF

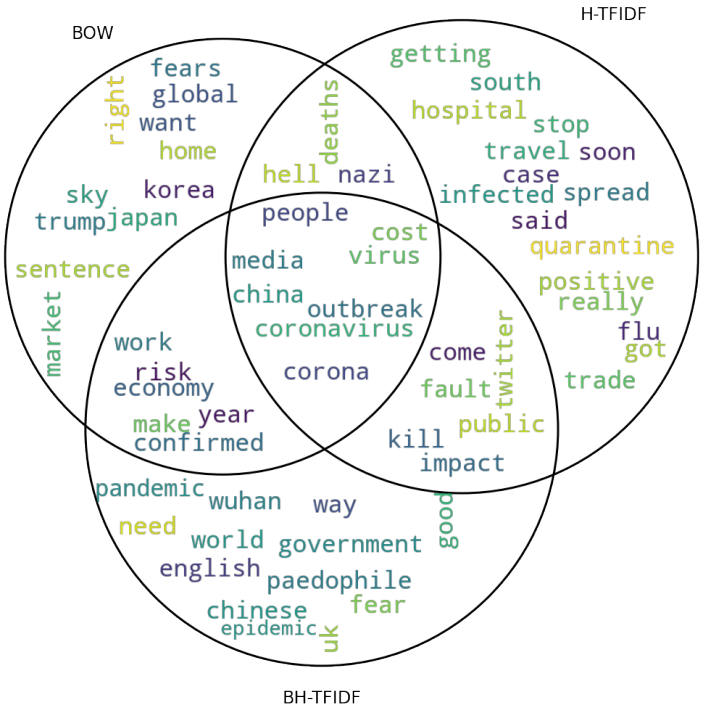


(b) Features: BOW, H-TFIDF and BH-TFIDF

Figure 4.6: Negative tweets comparison by features



(a) Top Features: BOW, TF-IDF and H-TFIDF



(b) Top Features: BOW, H-TFIDF and BH-TFIDF

Figure 4.7: Top features extracted from COVID-19 unlabelled dataset

4.7 Conclusion and Perspectives

In this work, we presented spatial opinion mining in order to gain insights into the regional situation during a specific time period. Additionally, we performed sentiment classification through different feature sets, including Bag-of-words (BOW), TF-IDF, H-TFIDF, and BH-TFIDF in order to find the best features for sentiment classification task. This work focused on spatial opinion mining and the performance of various features within this context. This work can be enhanced using advanced terminology extraction methods. These methods offer advantages like weak supervision and scalability, emphasizing both single-word and multi-word term extraction for a deeper understanding of sentiment expressions. The goal is to create comprehensive, accurate disease-specific terminology dictionaries, improving sentiment classification without heavy reliance on manual labelling and enabling scalability to larger datasets.

The current work was focused on generalized sentiment analysis for specific locations and targeted populations. For deeper understanding of events, we could explore dynamic topic identification, diving deeper into sentiments related to specific topics within these events. This approach would provide a more fine-grained analysis, highlights how sentiments evolve around different aspects of events, ultimately contributing to a more comprehensive understanding. In the context of MOOD project, the real-time or near real-time sentiment analysis of tweets with spatial information enables a deeper understanding of public opinions and emotions within specific geographic locations. This spatial context is a significant enhancement for EBS, offering insights into how events are perceived and experienced in diverse regions.

Part V

Conclusion and Perspectives

CHAPTER 5

CONCLUSION AND PERSPECTIVES

This chapter summarises all the work carried out during our studies about different prospects of event extraction in the context of EBS systems. These are essentially the scientific or methodological contributions and the perspectives that we propose for each chapter. This chapter is subdivided into two sections. First, we discuss the contributions in Section 5.1, and the prospects in Section 5.2.

5.1 Contributions

In this thesis, the three main contributions are, i.e., 1) Ensure data quality 2) Geographical accuracy of event and 3) Situational awareness of event to improve EBS pipeline. The overview of these contributions are as follows:

Data Quality Assurance for EBS System

In chapter 2 (Data Quality Assurance), we delved into the importance of ensuring data quality in online news sources within the context of the EBS pipeline. This focus on data quality is crucial for obtaining relevant and credible information from these news sources. In the data driven score based approach, we classified news articles through data quality score (DQS). The DQS is computed from the state-of-the-art metadata score (MS) based on metadata attributes in addition to content score (CS) calculated using attributes extracted from the article content. Furthermore, an extension called epidemiological entity extraction score (E3S) is introduced into the framework. E3S is computed from the epidemic-related attributes along with spatio-temporal attributes extracted from the content of the news article. The resulting DQS classifies news articles into 'Relevant' and 'Irrelevant' categories, demonstrating its effectiveness using avian-influenza-related article dataset. This approach is valuable in terms of influence of attributes, thereby contributing to enhancing the aspect of explainability unlike machine learning approaches.

Secondly, we proposed metadata-based news article classification through machine learning approach. In this approach, we obtained important metadata from news articles using an automated method. Afterwards, we classified relevant news articles through machine learning model using extracted important metadata features. This approach allows the systematic analysis of various metadata features associated with news articles to determine their relevance or irrelevance. As a result, we found 'title' and 'keywords' of the news articles the most important for classification task.

Lastly, we introduced metadata-based news source classification through machine learning approach. This source metadata includes source category, geographical coverage, topic coverage, and world press freedom index. It is extracted through automated approach and evaluated further to classify the relevant news sources. We proposed a clustering algorithm to categorized sources into specialized and generalized. Moreover, a dedicated algorithm categorized the geographical coverage of source e.g. local, national and international. Finally, classify the news sources into relevant and irrelevant. This approach allows ranking/labelling the potential news sources in the context of EBS. The publications associated to this chapter are cited as (Syed et al., 2022a) and (Syed et al., 2023a).

Geographical Accuracy of Events: The role of Spatial Relations

In chapter 3 (Geographical Accuracy of Events: The role of Spatial Relations), the research focused on extracting spatial relation associated with locations from textual data and determining their geographical coordinates for visualization on GIS, benefiting geographical accuracy of reported such

locations of events. We proposed a rule based NLP approach to extract geospatial relations locations from text and evaluated with disease dataset having spatial relations locations. Consequently, we proposed an algorithm to compute the geographical coordinates of the spatial relations location and evaluated the polygons qualitatively by groups of end users. As a result, it improved the geographical accuracy of events to identify more precise affected region. The publication associated to this work is cited as (Syed et al., 2022c).

Spatial Opinion Mining for Situational Awareness of Events

In chapter 4 (Spatial Opinion Mining for Situational Awareness of Events), the research highlights the potential of spatial opinion mining, especially during health events, for obtaining real-time insights into public sentiments and perceptions for effective awareness of public health situations. We performed sentiment analysis applied to geotagged tweets during the COVID-19 pandemic that provides insights into local sentiments and situations related to the ongoing epidemic across different regions. Afterwards, more comprehensive spatial opinion mining analysis, employing various features such as BERT, H-TFIDF, TF-IDF, and BOW. The aim is to assess the significance of these features in sentiment classification within the context of EBS. The publications associated to this chapter are cited as (Syed et al., 2022d) and (Syed et al., 2022b).

5.2 Perspectives

The primary future perspective revolves around integrating the following aspects into the EBS system: 1) ensuring the data quality to classify relevant news article, 2) enhancing the geographical accuracy of events, and 3) elevating situational awareness regarding events. The primary focus of this research is to integrate these aspects into the EBS pipeline, particularly within PADI-web. A promising perspective within this research is the integration of these diverse contributions. For instance, the integration of spatial opinion mining and precise spatial information extraction could serve as quality criteria, enhancing the overall research framework. The future perspectives of each aspect are as follows:

Data Quality Assurance for EBS System

In future work, we will focus on additional source metadata attributes. This includes temporal relevance (timeliness of events), sentiment analysis (public opinions about reporting of media outlets) and automated topic coverage of news sources. Furthermore, we will utilize this metadata to assign the quality labels or quality indices to news sources, thereby further enhancing the credibility and reliability of the information within the EBS context.

Furthermore, we expand our work through a two-level classification approach for news articles. In case of relevant articles, we will dig in deeper into their contextual relevance by classifying them

into specific categories. These categories could encompass outbreak declarations, risk assessments, disease transmission, preventive measures, control strategies, and more. This fine-grained classification will offer a more comprehensive understanding of the diverse aspects of health events, enabling more targeted and effective responses within the EBS framework. Apart from that, we will focus on building a fine-tuned language model for epidemiological news report detection for different European languages (Mutuvi et al., 2020). After the improvements and generalization of this work, the goal is to integrate the quality labels of news sources and fine-grained articles classification into EBS pipeline in the context of One Health.

Geographical Accuracy of Events: The role of Spatial Relations

There is significant room for improvement of geocoding algorithm. Our first priority is to enhance the algorithm for compound spatial relation polygons by doing some minor adjustments. Furthermore, we will expand our geocoding dataset by including additional cities and countries polygons. The objective of this expansion is to enable a comprehensive qualitative analysis of these polygons, with the aim of ensuring the reliability of our geocoding method. In future work, we will investigate the geographical aspects between official events and those detected by EBS for assessing system performance and alignment with real-world geographical dynamics. This analysis provides valuable insights and equips decision-makers with accurate information for more effective interventions. For instance, in the case of avian influenza, various organizations report official events, such as WAHIS by the WOAH and EMPRES-i by the FAO, each reporting events at different geographical granularity levels, including Admin level-1, Admin level-2, and Admin level-3. We will check the role of spatial relations in comparison to official events. If EBS can provide finer spatial information than official reports, it can significantly improve the accuracy and effectiveness of event surveillance and response efforts. The ultimate goal is to seamlessly integrate the geoparsing and geocoding methods, incorporating spatial relation locations, to enhance the geographical accuracy of events within the EBS system.

Spatial Opinion Mining for Situational Awareness of Events

Our future work aims to enhance sentiment analysis of COVID-19 tweets using advanced terminology extraction methods. These techniques, including weakly supervised and unsupervised approaches, will enable us to effectively leverage extensive unlabelled data. We will focus on term extraction, covering both single-word and multi-word terms. Moreover, we will focus on aspect-based sentiment analysis for health events, focusing on specific topics like severity, economic impact, preventive measures and risks, offers valuable perspectives for understanding public sentiment and enhancing crisis response. For instance, **severity assessment** sentiments may indicate trust in containment measures or could signal concerns about the outbreak seriousness. Moreover, **economic impact** sentiments concerning economic consequences provides information about how the public perceives the financial risks and implications of the outbreak. In future work, we will incorporate the real-time or near real-time sentiment analysis of tweets in the spatial context of events to get a deeper

understanding of public opinions and emotions within specific geographic locations. This spatial context is a significant enhancement for EBS, offering insights into how events are perceived and experienced in diverse regions.

Part VI

Résumé de la thèse en français

CHAPTER 6

RÉSUMÉ DE LA THÈSE EN FRANÇAIS

6.1 Introduction

Les épidémies de maladies infectieuses constituent de graves menaces pour la santé humaine, animale et végétale (One Health) (Kim et al., 2020). Les épidémies de maladies infectieuses affectent non seulement la santé individuelle, mais aussi l'économie et le commerce national et international (Rees et al., 2019). Il est donc important de mettre en œuvre des méthodes de surveillance sanitaire pour reconnaître les épidémies potentielles de maladies infectieuses et minimiser leurs effets dévastateurs sur la population touchée et, indirectement, sur la société. Dans la littérature existante (WHO, 2008; Rees et al., 2019), il existe deux principaux types de surveillance : 1) la surveillance fondée sur les événements (EBS) et 2) la surveillance fondée sur les indicateurs (IBS). La surveillance fondée sur des indicateurs utilise des sources officielles pour détecter les épidémies importantes (Runge-Ranzinger et al., 2008). Ces sources produisent des données structurées et fiables, offrant un large éventail d'informations sur l'agent pathogène, la source du foyer, les espèces, les signes cliniques, etc. En raison des procédures officielles, la déclaration des foyers accuse un retard considérable. En revanche, la surveillance fondée sur les événements (EBS) fait référence à la collecte d'informations concernant des événements susceptibles de présenter des risques pour la santé et signalés dans des sources de données non structurées telles que des données textuelles, c'est-à-dire des articles d'actualité, des mises à jour sur les médias sociaux, etc. Selon l'Organisation mondiale de la santé (OMS), environ 60% des épidémies sont identifiées par des sources informelles (Abbood et al., 2020). Ces deux stratégies de surveillance se complètent en termes d'avantages grâce à leurs processus uniques de collecte, de vérification, d'évaluation et d'interprétation des données (WHO, 2008) et sont considérées comme fondamentales dans l'élaboration d'un système de surveillance complet (Balajee et al., 2021). Notre recherche s'est principalement concentrée sur les EBS.

Les systèmes de type EBS permettent de détecter des informations sous forme d'événement provenant de différentes sources de données non structurées (Balajee et al., 2021). Ils permettent aux autorités concernées d'être mieux préparées à l'apparition de maladies endémiques et pandémiques en fonctionnant comme un élément clé d'un système d'alerte précoce efficace (WHO, 2008; Balajee et al., 2021). Pour l'acquisition de ces événements, les sources d'information en ligne telles que les articles de presse, les blogs, les médias sociaux (tels que Twitter, etc.) et d'autres rapports ad hoc ont fait l'objet d'une grande attention dans la mise en œuvre de systèmes EBS (Valentin, 2020).

6.2 Motivation

Un événement fait référence à une occurrence spécifique, un modèle ou un groupe d'incidents liés à la santé dans une région géographique ou une population communautaire qui sont surveillés et étudiés afin d'évaluer leur importance en termes de santé humaine (PH)/santé animale (AH) (Shakeri Hossein Abad et al., 2021). Ces événements comprennent les épidémies de maladies infectieuses, l'augmentation inhabituelle du nombre de cas de maladies, l'émergence de nouvelles maladies ou tout autre incident lié à la santé qui nécessite une attention particulière, une enquête et une réponse de la part des autorités de santé publique (Welby-Everard, Quantick, and Green, 2020). Il y a cinq

aspects pour définir un événement à savoir: *quand*, *où*, *laquelle*, *qui* et *pourquoi* (Ibrahim, 2020; Sims, Park, and Bamman, 2019; Yu et al., 2020). Ces aspects sont importants car, en épidémiologie, il s'agit des caractéristiques potentielles d'une épidémie. Par exemple, *quand* est la date de début de l'événement, *où* représente la région touchée, *laquelle* maladie, *qui* est touché hôte (par exemple l'homme, l'animal), l'agent pathogène et les parties prenantes et *pourquoi* représente la cause de la maladie. Il existe plusieurs approches pour extraire les événements à partir de sources de données textuelles, notamment les approches basées sur des règles, l'apprentissage automatique supervisé, l'apprentissage semi-supervisé, l'apprentissage non supervisé, les modèles d'apprentissage profond et les approches basées sur les ontologies, etc. Il est important d'améliorer ces différentes approches afin d'obtenir des informations précises sur les événements, afin de refléter fidèlement la situation réelle.

La qualité des données est le fondement d'une surveillance efficace des maladies, car elle garantit que les informations recueillies sont pertinentes, fiables, dignes de confiance et à jour (Kington et al., 2021). Des informations inexactes ou peu fiables peuvent conduire à des réponses erronées en matière de santé publique et à une évaluation a posteriori par les épidémiologistes. Par conséquent, l'assurance de la qualité des données est essentielle pour éliminer le bruit, filtrer les sources non fiables et réduire les biais afin de prendre les bonnes décisions en matière de contrôle et de surveillance.

La précision de la localisation des foyers, qui implique l'identification de la région géographique exacte d'un événement, influe considérablement sur les méthodes d'extraction automatisée des événements dans les EBS. Atteindre cette précision peut s'avérer difficile, car il faut discerner les détails spécifiques de la localisation mentionnée dans le texte, y compris les relations spatiales complexes telles que "au nord de Paris" ou "près de la frontière française". Par la suite, après avoir extrait le lieu, il est tout aussi crucial d'identifier la zone géographique précise (point ou polygone) ou la représentation géospatiale de l'événement. Sans ce niveau de précision dans les informations géographiques, les efforts de surveillance peuvent manquer du contexte spatial nécessaire pour cibler efficacement les interventions (Friis and Sellers, 2020) et peuvent également affecter les études d'épidémiologie spatiale (Kirby, Delmelle, and Eberth, 2017). Cette lacune pourrait potentiellement conduire à une propagation rapide des maladies à l'intérieur des régions et d'une région à l'autre en raison du manque d'informations ou d'informations moins précises sur les épidémies potentielles.

La fouille d'opinion spatiale à partir de médias sociaux tels que Twitter constitue une source d'information dynamique et en temps réel pour la connaissance de la situation (Vernier, Farinosi, and Foresti, 2019). Elle consiste à analyser le contenu des tweets géolocalisés pour en extraire des opinions, des sentiments et des informations liés aux événements sanitaires (Steiger, De Albuquerque, and Zipf, 2015). Des techniques de traitement automatique du langage naturel (TALN) sont appliquées pour comprendre le contexte des tweets. Dans le contexte d'un événement, elles peuvent révéler le sentiment du public concernant la gravité de l'épidémie, les symptômes, les mesures de prévention, les expériences personnelles, etc.

Cette thèse est principalement financée par le projet '**MOnitoring Outbreaks for Disease surveillance in a data science context (MOOD[1])**'. Le projet MOOD vise à exploiter les techniques les plus récentes de fouille de texte et d'analyse de données volumineuses provenant de sources multiples afin

[1] https://mood-h2020.eu/

d'améliorer la veille sanitaire des maladies (ré)-émergentes en Europe.

6.3 Objectifs

Notre principale question de recherche est la suivante : comment l'extraction d'événements peut-elle être améliorée dans le cadre des systèmes de surveillance basés sur les événements (EBS) entièrement automatisés ? Les questions de recherche subsidiaires pour compléter la question principale sont les suivantes:

1. Quelles stratégies peuvent être employées pour améliorer la qualité des données des sources d'information afin de garantir des informations fiables pour l'extraction d'événements? En outre, dans quelle mesure l'amélioration de la qualité des données des sources d'information contribue-t-elle à une extraction plus précise des événements?

2. Comment l'extracton automatique d'information spatiale précise peut améliorer l'extraction d'évenements? Quelles techniques peuvent être utilisées pour identifier avec précision la localisation des événements, par exemple les relations spatiales associées aux lieux?

3. De quelle manière la fouille d'opinion spatiale des plateformes de médias sociaux comme Twitter peut-elle améliorer la connaissance de la situation des événements?

6.4 Contributions

Pour atteindre ces objectifs, nous avons trois contributions principales permettant d'améliorer 1) l'évaluation de la qualité des données 2) la précision géographique des événements extraits et 3) l'analyse de sentiments liées aux événements. Un résumé de ces contributions est donné ci-dessous:

La qualité des données pour les systèmes de surveillance EBS

Dans le chapitre 2 (Guarantir la qualité des données), nous avons étudié l'importance d'assurer la qualité des données dans le contexte du pipeline des systèmes EBS. Cette attention portée à la qualité des données est cruciale pour obtenir des informations pertinentes et crédibles à partir de ces sources de données. Dans l'approche basée sur les données, nous avons utilisé le Score de Qualité des Données (DQS) pour classer les articles d'actualité. Le DQS est calculé en combinant le score des métadonnées (MS), qui se base sur les attributs des métadonnées, et le score de contenu (CS), qui est calculé à partir des attributs extraits du contenu de l'article. De plus, nous avons introduit une extension appelée score d'extraction des entités épidémiologiques (E3S), qui est calculé à partir des attributs liés à l'épidémie et des attributs spatiotemporels extraits du contenu de l'article. Le score DQS résultant permet de classifier les articles de presse comme "pertinents" ou "non pertinents", démontrant ainsi son efficacité à l'aide d'un ensemble de données d'articles liés à la grippe aviaire.

Cette approche est précieuse pour considérer et discuter l'influence des attributs, contribuant ainsi à améliorer la dimension explicative contrairement aux approches d'apprentissage automatique.

En second lieu, nous avons introduit une classification des articles de presse basée sur leurs métadonnées grâce à une méthode d'apprentissage automatique. Dans cette démarche, nous avons extrait les métadonnées essentielles des articles. Ensuite, nous avons utilisé un modèle d'apprentissage automatique pour classifier les articles pertinents en se basant sur les caractéristiques des métadonnées extraites. Cette approche permet une analyse systématique des diverses métadonnées associées aux articles d'actualité afin d'évaluer leur pertinence. Nous avons ainsi constaté que le "titre" et les "mots-clés" des articles étaient les plus déterminants pour la tâche de classification.

En dernière étape, nous avons instauré une classification des sources d'information en nous appuyant sur les métadonnées grâce à une méthode d'apprentissage automatique. Ces métadonnées, englobant la catégorie de la source, la portée géographique, la thématique traitée et l'indice mondial de liberté de la presse, sont extraites puis évaluées afin de classer les sources d'information pertinentes. Nous avons élaboré un algorithme de regroupement pour diviser les sources en deux catégories (spécialisées et généralistes). De plus, nous avons proposé un algorithme spécifique pour caractériser la portée géographique des sources (locale, nationale et internationale). Nous avons classé ainsi les sources d'information en deux catégories : celles qui sont pertinentes et celles qui ne le sont pas. Cette approche permet de classer/étiqueter les sources d'information potentielles. Cette contribution scientifique est associée aux publications (Syed et al., 2022a) et (Syed et al., 2023a).

La précision géographique des événements: Le rôle des relations spatiales

Dans le chapitre 3 (Précision géographique des événements: le rôle des relations spatiales), notre recherche s'est concentrée sur l'extraction de relations spatiales associées à des lieux à partir de données textuelles et sur la détermination de leurs coordonnées géographiques. Ceci qui permet d'améliorer la précision géographique des lieux des événements signalés. Nous avons proposé une approche TALN basée sur des règles pour extraire les lieux des relations géospatiales à partir de textes. Ainsi, nous avons proposé un algorithme pour calculer les coordonnées géographiques de l'emplacement des relations spatiales et évalué qualitativement les polygones produits par des groupes d'utilisateurs finaux. Cet algorithme a permis d'améliorer la précision géographique des événements et d'identifier plus précisément les régions touchées. Cette contribution scientifique est associée à la publication (Syed et al., 2022d).

Fouille d'opinions spatiales pour la connaissance de la situation des événements

Dans le chapitre 4 (Fouille d'opinions spatiales pour la connaissance de la situation des événements), notre recherche met en évidence le potentiel de la fouille spatiale d'opinions, en particulier lors d'événements sanitaires, pour obtenir des informations en temps réel sur les sentiments et les per-

ceptions du public en vue d'une sensibilisation efficace. Nous avons effectué une analyse des sentiments appliquée aux tweets géolocalisés pendant la pandémie de COVID-19, qui donne un aperçu des sentiments locaux et des situations liées à l'épidémie en cours dans différentes régions. Ensuite, nous avons effectué une analyse plus complète de l'extraction d'opinions spatiales, en utilisant diverses représentations qui s'appuient sur BERT, H-TFIDF, TF-IDF et BOW. L'objectif est d'évaluer l'importance de ces caractéristiques pour la classification des sentiments dans le contexte des systèmes de surveillance. Cette contribution scientifique est associée aux publications (Syed et al., 2022d) et (Syed et al., 2022b).

L'objectif de cette thèse était d'étudier différentes stratégies afin d'améliorer la détection des événements pour une surveillance efficace des maladies. L'objectif principal de nos travaux est d'intégrer ces aspects dans la pipeline d'un système EBS, en particulier dans PADI-web. Une perspective prometteuse de cette thèse de recherche réside dans l'intégration de ces diverses contributions. Par exemple, l'intégration de la fouille d'opinions spatiales et de l'extraction d'informations spatiales précises pourrait servir de critères de qualité, améliorant ainsi le cadre général de la surveillance.

BIBLIOGRAPHY

[1] Heidi Abbas et al. “Usage of social media in epidemic intelligence activities in the WHO, Regional Office for the Eastern Mediterranean”. In: *BMJ Global Health* 7.Suppl 4 (2022), e008759.

[2] Auss Abbood et al. “EventEpi—A natural language processing framework for event-based surveillance”. In: *PLoS computational biology* 16.11 (2020), e1008277.

[3] Philip Abdelmalik et al. “The Epidemic Intelligence from Open Sources initiative: a collaboration to harmonize and standardize early detection and epidemic intelligence among public health organizations/L’initiative« Epidemic Intelligence from Open Sources»: une collaboration visant a harmoniser et a standardiser les procedures de detection precoce et de renseignement epidemiologique entre les organisations de sante publique”. In: *Weekly Epidemiological Record* 93.20 (2018), pp. 267–270.

[4] Akiko Aizawa. “An information-theoretic perspective of tf–idf measures”. In: *Information Processing & Management* 39.1 (2003), pp. 45–65.

[5] Hussein S. Al-Olimat et al. “Towards Geocoding Spatial Expressions (Vision Paper)”. In: SIGSPATIAL ’19. Chicago, IL, USA: Association for Computing Machinery, 2019, 75–78. ISBN: 9781450369091. DOI: 10.1145/3347146.3359356. URL: https://doi.org/10.1145/3347146.3359356.

[6] Oscar Alomar et al. “Development and testing of the media monitoring tool Med IS ys for the monitoring, early identification and reporting of existing and emerging plant health threats”. In: *EFSA Supporting Publications* 13.12 (2016), 1118E.

[7] Álvaro Alonso Casero. “Named entity recognition and normalization in biomedical literature: a practical case in SARS-CoV-2 literature”. PhD thesis. ETSI_Informatica, 2021. URL: https://oa.upm.es/67933/.

[8] Gunjan Ansari, Tanvir Ahmad, and Mohammad Najmud Doja. “Hybrid Filter–Wrapper feature selection method for sentiment classification”. In: *Arabian Journal for Science and Engineering* 44.11 (2019), pp. 9191–9208.

[9] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. “Twitter catches the flu: detecting influenza epidemics using Twitter”. In: *Proceedings of the 2011 Conference on empirical methods in natural language processing*. 2011, pp. 1568–1576.

[10] Elena Arsevska et al. “Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System”. In: *PLoS One* 13.8 (2018), e0199960.

[11] Philipp Bachmann, Mark Eisenegger, and Diana Ingenhoff. “Defining and Measuring News Media Quality: Comparing the Content Perspective and the Audience Perspective”. In: *The International Journal of Press/Politics* (2021), p. 1940161221999666.

[12] S Arunmozhi Balajee et al. “The practice of event-based surveillance: concept and methods”. In: *Global Security: Health, Science and Policy* 6.1 (2021), pp. 1–9.

[13] Zach Bastick. "Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation". In: *Computers in human behavior* 116 (2021), p. 106633.

[14] Carlo Batini, Monica Scannapieco, et al. "Data and information quality". In: *Cham, Switzerland: Springer International Publishing. Google Scholar* 43 (2016).

[15] Franz-Josef Behr. "Geocoding: Fundamentals, Techniques, Commercial and Open Services". In: *AGSE 2010* (2010), p. 111.

[16] Edina Berlinger et al. "Press freedom and operational losses: The monitoring role of the media". In: *Journal of International Financial Markets, Institutions and Money* 77 (2022), p. 101496.

[17] Daniel Berrar. "Cross-Validation". In: *Encyclopedia of Bioinformatics and Computational Biology - Volume 1*. Ed. by Shoba Ranganathan et al. Elsevier, 2019, pp. 542–545. DOI: 10.1016/b978-0-12-809633-8.20349-x. URL: https://doi.org/10.1016/b978-0-12-809633-8.20349-x.

[18] Md Momen Bhuiyan et al. "Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020), pp. 1–26.

[19] Élise Bigeard and Natalia Grabar. "Detection and analysis of medical misbehavior in online forums". In: *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE. 2019, pp. 7–12.

[20] Michael Blench. "Global Public Health Intelligence Network (GPHIN)". In: *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Government and Commercial Uses of MT*. Waikiki, USA: Association for Machine Translation in the Americas, 2008, pp. 299–303. URL: https://aclanthology.org/2008.amta-govandcom.2.

[21] Tessel Bogaard et al. "Metadata categorization for identifying search patterns in a digital library". In: *Journal of Documentation* 75.2 (2019), pp. 270–286.

[22] Verena Brinks and Oliver Ibert. "From corona virus to corona crisis: The value of an analytical and geographical understanding of crisis". In: *Tijdschrift voor economische en sociale geografie* 111.3 (2020), pp. 275–287.

[23] Michael A. Cacciatore. "Misinformation and public opinion of science and health: Approaches, findings, and future directions". In: *Proceedings of the National Academy of Sciences* 118.15 (2021), e1912437117. DOI: 10.1073/pnas.1912437117. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1912437117. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1912437117.

[24] Malwina Carrion and Lawrence C. Madoff. "ProMED-mail: 22 years of digital surveillance of emerging infectious diseases". In: *International Health* 9.3 (June 2017), pp. 177–183. ISSN: 1876-3413. DOI: 10.1093/inthealth/ihx014. eprint: https://academic.oup.com/inthealth/article-pdf/9/3/177/24324986/ihx014.pdf. URL: https://doi.org/10.1093/inthealth/ihx014.

[25] Livio Cascone, Pietro Ducange, and Francesco Marcelloni. "Exploiting online newspaper articles metadata for profiling city areas". In: *International Conference on Intelligent Data Engineering and Automated Learning*. Springer. 2019, pp. 203–215.

[26] Kenrick D Cato, Bevin Cohen, and Elaine Larson. "Data elements and validation methods used for electronic surveillance of health care-associated infections: A systematic review". In: *American journal of infection control* 43.6 (2015), pp. 600–605.

[27] Lois Mai Chan et al. "A faceted approach to subject data in the Dublin Core metadata record". In: *Journal of Internet Cataloging* 4.1-2 (2001), pp. 35–47.

[28] Ranganathan Chandrasekaran et al. "Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study". In: *Journal of medical Internet research* 22.10 (2020), e22624.

[29] Angel X Chang and Christopher D Manning. "Sutime: A library for recognizing and normalizing time expressions." In: *LREC*. Vol. 3735. 2012, p. 3740.

[30] Hutchatai Chanlekha and Nigel Collier. "A methodology to enhance spatial understanding of disease outbreak events reported in news articles". In: *International Journal of Medical Informatics* 79.4 (2010). Human Factors Engineering for Healthcare Applications Special Issue, pp. 284–296. ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2010.01.014. URL: https://www.sciencedirect.com/science/article/pii/S1386505610000274.

[31] Emily Chen, Kristina Lerman, and Emilio Ferrara. "Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set". In: *JMIR Public Health and Surveillance* 6.2 (2020), e19273.

[32] Hao Chen, Maria Vasardani, and Stephan Winter. "Geo-referencing place from everyday natural language descriptions". In: *arXiv preprint arXiv:1710.03346* (2017). DOI: 10.48550/arXiv.1710.03346.

[33] Wai Khuen Cheng et al. "A review of sentiment, semantic and event-extraction-based approaches in stock forecasting". In: *Mathematics* 10.14 (2022), p. 2437.

[34] Konstantin Clemens. "Geocoding with OpenStreetMap Data". In: 2015. URL: https://api.semanticscholar.org/CorpusID:38962500.

[35] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. "Automatic deception detection: Methods for finding fake news". In: *Proceedings of the association for information science and technology* 52.1 (2015), pp. 1–4.

[36] Keith Cortis. "Multidimensional opinion mining from social data". PhD thesis. Dublin City University, 2022.

[37] Adam T Craig et al. “Early warning epidemic surveillance in the Pacific island nations: an evaluation of the Pacific syndromic surveillance system”. In: *Tropical Medicine & International Health* 21.7 (2016), pp. 917–927.

[38] Rémy Decoupes et al. “H-TFIDF: What makes areas specific over time in the massive flow of tweets related to the covid pandemic?” In: *AGILE: GIScience Series* 2 (2021), pp. 1–8.

[39] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[40] Giandomenico Di Domenico et al. “Fake news, social media and marketing: A systematic review”. In: *Journal of Business Research* 124 (2021), pp. 329–341.

[41] Akash Dutt Dubey. “Twitter Sentiment Analysis during COVID-19 Outbreak”. In: *Available at SSRN 3572023* (2020).

[42] Tamar Edry et al. “Real-time geospatial surveillance of localized emotional stress responses to COVID-19: a proof of concept analysis”. In: *Health & Place* 70 (2021), p. 102598.

[43] Fabian Eibensteiner et al. “People’s willingness to vaccinate against COVID-19 despite their safety concerns: Twitter poll analysis”. In: *Journal of Medical Internet Research* 23.4 (2021), e28973.

[44] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. “A novel approach for selecting hybrid features from online news textual metadata for fake news detection”. In: *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*. Springer. 2019, pp. 914–925.

[45] Mohamed Ben Ellefi and Pierre Drap. “A Bit on the Right, A Bit on the Left: Towards Logical Bundle Adjustment”. In: *International Journal of Innovative Technology and Exploring Engineering* (2020).

[46] Marwa Essam and Tamer Elsayed. “Why is That a Background Article: A Qualitative Analysis of Relevance for News Background Linking”. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 2009–2012.

[47] Ibrahim Eldesouky Fattoh et al. “Semantic sentiment classification for covid-19 tweets using universal sentence encoder”. In: *Computational intelligence and neuroscience* 2022 (2022).

[48] Blossom Fernandes et al. “The impact of COVID-19 lockdown on internet use and escapism in adolescents”. In: *Revista de psicología clínica con niños y adolescentes* 7.3 (2020), pp. 59–65.

[49] Jacques Fize, Ludovic Moncla, and Bruno Martins. “Deep learning for toponym resolution: Geocoding based on pairs of toponyms”. In: *ISPRS International Journal of Geo-Information* 10.12 (2021), p. 818.

[50] Clark C. Freifeld et al. “HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports”. In: *Journal of the American Medical Informatics Association* 15.2 (2008), pp. 150–157. ISSN: 1067-5027. DOI: https://doi.org/10.1197/jamia.M2544. URL: https://www.sciencedirect.com/science/article/pii/S1067502707003441.

[51] Robert H Friis and Thomas Sellers. *Epidemiology for public health practice*. Jones & Bartlett Learning, 2020.

[52] William M Fritch et al. “Application of the Haddon matrix to COVID-19 prevention and containment in nursing homes”. In: *Journal of the American Geriatrics Society* 69.10 (2021), pp. 2708–2715.

[53] Sean Gillies et al. *Shapely: manipulation and analysis of geometric objects*. toblerity.org, 2007. URL: https://github.com/Toblerity/Shapely.

[54] Michael A Gisondi et al. *A deadly infodemic: social media and the power of COVID-19 misinformation*. 2022.

[55] Shilpa Gite et al. “Textual feature extraction using ant colony optimization for hate speech classification”. In: *Big data and cognitive computing* 7.1 (2023), p. 45.

[56] Cyril Goutte and Eric Gaussier. “A probabilistic interpretation of precision, recall and F-score, with implication for evaluation”. In: *European conference on information retrieval*. Springer. 2005, pp. 345–359. DOI: 10.1007/978-3-540-31865-1_25.

[57] Milan Gritta et al. “What’s missing in geographical parsing?” In: *Language Resources and Evaluation* 52 (2018), pp. 603–623.

[58] Sharath Chandra Guntuku et al. “Tracking mental health and symptom mentions on Twitter during COVID-19”. In: *Journal of general internal medicine* 35 (2020), pp. 2798–2800.

[59] Aakansha Gupta and Rahul Katarya. “Social media based surveillance systems for healthcare using machine learning: a systematic review”. In: *Journal of biomedical informatics* 108 (2020), p. 103500.

[60] Kai Hakala and Sampo Pyysalo. “Biomedical named entity recognition with multilingual BERT”. In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. 2019, pp. 56–61. DOI: 10.18653/v1/D19-5709.

[61] Andrew Halterman. “Mordecai: Full Text Geoparsing and Event Geocoding.” In: *J. Open Source Softw.* 2.9 (2017), p. 91.

[62] Linda L Hill. *Georeferencing: The geographic associations of information*. Mit Press, 2009.

[63] Sung Yang Ho and Tan. “What can Venn diagrams teach us about doing data science better?” In: *International Journal of Data Science and Analytics* 11.1 (2021), pp. 1–10.

[64] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". 2017. DOI: 10.5281/zenodo.1212303.

[65] HS Hota, Dinesh K Sharma, and Nilesh Verma. "Lexicon-based sentiment analysis using Twitter data: a case of COVID-19 outbreak in India and abroad". In: *Data Science for COVID-19*. Elsevier, 2021, pp. 275–295.

[66] Xuke Hu et al. "GazPNE2: A general place name extractor for microblogs fusing gazetteers and pretrained transformer models". In: *IEEE Internet of Things Journal* 9.17 (2022), pp. 16259–16271.

[67] Yang Hu et al. "Discovering authoritative news sources and top news stories". In: *Asia Information Retrieval Symposium*. Springer. 2006, pp. 230–243.

[68] Nahla Khamis Ibrahim. "Epidemiologic surveillance for controlling Covid-19 pandemic: types, challenges and implications". In: *Journal of infection and public health* 13.11 (2020), pp. 1630–1638.

[69] Md Rafiqul Islam et al. "Deep learning for misinformation detection on online social networks: a survey and new perspectives". In: *Social Network Analysis and Mining* 10.1 (2020), pp. 1–20.

[70] Shaghayegh Jabalameli, Yanqing Xu, and Sujata Shetty. "Spatial and sentiment analysis of public opinion toward COVID-19 pandemic using twitter data: At the early stage of vaccination". In: *International Journal of Disaster Risk Reduction* 80 (2022), p. 103204.

[71] Karwan Jacksi et al. "Clustering documents based on semantic similarity using HAC and K-mean algorithms". In: *2020 International Conference on Advanced Science and Engineering (ICOASE)*. IEEE. 2020, pp. 205–210.

[72] Nastaran Jafarpour et al. "Quantifying the determinants of outbreak detection performance through simulation and machine learning". In: *Journal of biomedical informatics* 53 (2015), pp. 180–187.

[73] Kelsey Jordahl et al. *geopandas/geopandas: v0.8.1*. Version v0.8.1. July 2020. DOI: 10.5281/zenodo.3946761. URL: https://doi.org/10.5281/zenodo.3946761.

[74] Neli Kalcheva, Milena Karova, and Ivaylo Penev. "Comparison of the accuracy of SVM kemel functions in text classification". In: *2020 International Conference on Biomedical Innovations and Applications (BIA)*. IEEE. 2020, pp. 141–145.

[75] Kaushi ST Kanankege et al. "An introductory framework for choosing spatiotemporal analytical tools in population-level eco-epidemiological research". In: *Frontiers in Veterinary Science* 7 (2020), p. 339.

[76] KazAnova. *Sentiment140 dataset*. https://www.kaggle.com/kazanova/sentiment140. 2016. URL: https://www.kaggle.com/kazanova/sentiment140.

[77] Muzammil Khan et al. “Normalizing digital news-stories for preservation”. In: *2016 Eleventh International Conference on Digital Information Management (ICDIM)*. IEEE. 2016, pp. 85–90.

[78] Mira Kim et al. “Automated Classification of Online Sources for Infectious Disease Occurrences Using Machine-Learning-Based Natural Language Processing Approaches”. In: *International Journal of Environmental Research and Public Health* 17.24 (2020), p. 9467.

[79] Raynard S Kington et al. “Identifying credible sources of health information in social media: Principles and attributes”. In: *NAM perspectives* 2021 (2021).

[80] Russell S Kirby, Eric Delmelle, and Jan M Eberth. “Advances in spatial epidemiology and geographic information systems”. In: *Annals of epidemiology* 27.1 (2017), pp. 1–9.

[81] Ken P Kleinman and Allyson M Abrams. “Assessing surveillance using sensitivity, specificity and timeliness”. In: *Statistical methods in medical research* 15.5 (2006), pp. 445–464.

[82] Margarita Kokla and Eric Guilbert. “A review of geospatial semantic information modeling and elicitation approaches”. In: *ISPRS International Journal of Geo-Information* 9.3 (2020), p. 146. DOI: 10.3390/ijgi9030146.

[83] Hema Krishnan et al. “Machine learning based sentiment analysis of coronavirus disease related twitter data”. In: *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*. IEEE. 2021, pp. 459–464.

[84] S Vanaja K Ramesh Kumar. “Analysis of feature selection algorithms on classification: a survey”. In: (2014).

[85] Vaibhav Kumar. “Spatiotemporal sentiment variation analysis of geotagged COVID-19 tweets from India using a hybrid deep learning model”. In: *Scientific Reports* 12.1 (2022), p. 1849.

[86] Haewoon Kwak et al. “What is Twitter, a social network or a news media?” In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 591–600.

[87] Rabindra Lamsal, Aaron Harwood, and Maria Rodriguez Read. “Socially enhanced situation awareness from microblogs using artificial intelligence: A survey”. In: *ACM Computing Surveys* 55.4 (2022), pp. 1–38.

[88] Jochen L Leidner and Michael D Lieberman. “Detecting geographical references in the form of place names and associated spatial natural language”. In: *Sigspatial Special* 3.2 (2011), pp. 5–11.

[89] Gaël Lejeune et al. “Multilingual event extraction for epidemic detection”. In: *Artificial intelligence in medicine* 65.2 (2015), pp. 131–143.

[90] Julien Lesbegueries, Christian Sallaberry, and Mauro Gaio. “Associating spatial patterns to text-units for summarizing geographic information”. In: *ACM SIGIR 2006. GIR, Geographic Information Retrieval, Workshop*. 2006, pp. 40–43.

[91] Linna Li et al. “1.22-spatial data uncertainty”. In: *Comprehensive geographic information systems. Amsterdam: Elsevier* (2018), pp. 313–340.

[92] Michael Y Lin et al. “Quality of traditional surveillance for public reporting of nosocomial bloodstream infection rates”. In: *JAMA* 304.18 (2010), pp. 2035–2041.

[93] Shu-Yu Lin et al. “Analysing WAHIS Animal Health Immediate Notifications to Understand Global Reporting Trends and Measure Early Warning Capacities (2005–2021)”. In: *Transboundary and Emerging Diseases* 2023 (2023).

[94] Jens P Linge et al. “Internet surveillance systems for early alerting of health threats”. In: *Eurosurveillance* 14.13 (2009), p. 19162.

[95] Jens P Linge et al. “MedISys: medical information system”. In: *Advanced ICTs for disaster management and threat detection: Collaborative and distributed frameworks*. IGI Global, 2010, pp. 131–142.

[96] Steffen Lohmann et al. “Concentri cloud: Word cloud visualization for multiple text documents”. In: *2015 19th International Conference on Information Visualisation*. IEEE. 2015, pp. 114–120.

[97] Paul A Longley and James A Cheshire. “Geographical information systems”. In: *The Routledge Handbook of Mapping and Cartography*. Routledge, 2017, pp. 251–258.

[98] Avinash Madasu and Sivasankar Elango. “Efficient feature selection techniques for sentiment analysis”. In: *Multimedia Tools and Applications* 79.9 (2020), pp. 6313–6335.

[99] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. “A review: Data preprocessing and data augmentation techniques”. In: *Global Transitions Proceedings* (2022).

[100] SreeJagadeesh Malla and PJA Alphonse. “COVID-19 outbreak: An ensemble pretrained deep learning model for detecting informative tweets”. In: *Applied Soft Computing* 107 (2021), p. 107495.

[101] Ravali Mamidi et al. “Identifying key topics bearing negative sentiment on Twitter: insights concerning the 2015-2016 Zika epidemic”. In: *JMIR public health and surveillance* 5.2 (2019), e11036.

[102] Jane Mandalios. “RADAR: An approach for helping students evaluate Internet sources”. In: *Journal of information science* 39.4 (2013), pp. 470–478.

[103] Katherine McDonough, Ludovic Moncla, and Matje van de Camp. “Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora”. In: *International Journal of Geographical Information Science* 33.12 (2019), pp. 2498–2522. DOI: 10.1080/13658816.2019.1620235.

[104] Rehab Meckawy et al. “Effectiveness of early warning systems in the detection of infectious diseases outbreaks: a systematic review”. In: *BMC public health* 22.1 (2022), pp. 1–62.

[105] Amine Medad et al. “Comparing supervised learning algorithms for spatial nominal entity recognition”. In: *AGILE: GIScience Series* 1 (2020), pp. 1–18. DOI: 10.5194/agile-giss-1-15-2020,2020.

[106] Stuart E Middleton et al. “Location extraction from social media: Geoparsing, location disambiguation, and geotagging”. In: *ACM Transactions on Information Systems (TOIS)* 36.4 (2018), pp. 1–27.

[107] Ludovic Moncla. “Automatic reconstruction of itineraries from descriptive texts”. PhD thesis. Pau, 2015.

[108] Ludovic Moncla et al. “Geocoding for texts with fine-grain toponyms: an experiment on a geoparsed hiking descriptions corpus”. In: *Proceedings of the 22nd acm sigspatial international conference on advances in geographic information systems*. 2014, pp. 183–192.

[109] Ghulam Mujtaba et al. “Clinical text classification research trends: systematic literature review and open issues”. In: *Expert systems with applications* 116 (2019), pp. 494–520.

[110] Stephen Mutuvi et al. “Multilingual epidemiological text classification: a comparative study”. In: *COLING, International Conference on Computational Linguistics*. 2020, pp. 6172–6183.

[111] K Nimmi et al. “Pre-trained ensemble model for identification of emotion during COVID-19 based on emergency response support system dataset”. In: *Applied Soft Computing* 122 (2022), p. 108842.

[112] Yoshiko Nozato. “Credibility of online newspapers”. In: *Convención Anual de la Association for Education in Journalism and Mass Communication. Washington, DC Disponible en: http://citeseerx. ist. psu. edu/viewdoc/summary* (2002).

[113] OpenStreetMap contributors. *Planet dump retrieved from https://planet.osm.org*. https://www.openstreetmap.org. 2017. URL: https://www.openstreetmap.org.

[114] Eyal Oren and Heidi E Brown. *Infectious Disease Epidemiology: An Introduction*. Springer Publishing Company, 2022.

[115] World Health Organization et al. “ApartTogether survey: preliminary overview of refugees and migrants self-reported impact of Covid-19”. In: *Geneva, World Health Organization* (2020).

[116] World Health Organization et al. *Early detection, assessment and response to acute public health events: implementation of early warning and response with a focus on event-based surveillance: interim version*. Tech. rep. World Health Organization, 2014.

[117] Jesse O’Shea. “Digital disease detection: A systematic review of event-based internet biosurveillance systems”. In: *International journal of medical informatics* 101 (2017), pp. 15–22.

[118] Malte Ostendorff et al. “Enriching bert with knowledge graph embeddings for document classification”. In: *arXiv preprint arXiv:1909.08402* (2019).

[119] Tatiana Prado et al. “Wastewater-based epidemiology for preventing outbreaks and epidemics in Latin America–Lessons from the past and a look to the future”. In: *Science of The Total Environment* 865 (2023), p. 161210.

[120] Agung B Prasetijo et al. “Hoax detection system on Indonesian news sites based on text classification using SVM and SGD”. In: *2017 4th international conference on information technology, computer, and electrical engineering (ICITACEE)*. IEEE. 2017, pp. 45–49.

[121] Joseph D Prusa, Taghi M Khoshgoftaar, and David J Dittman. “Impact of feature selection techniques for tweet sentiment classification”. In: *The Twenty-eighth international flairs conference*. 2015.

[122] James Pustejovsky et al. “TimeML: Robust specification of event and temporal expressions in text.” In: *New directions in question answering* 3 (2003), pp. 28–34.

[123] Wisam A Qader, Musa M Ameen, and Bilal I Ahmed. “An overview of bag of words; importance, implementation, applications, and challenges”. In: *2019 international engineering conference (IEC)*. IEEE. 2019, pp. 200–204.

[124] Shahzad Qaiser and Ramsha Ali. “Text mining: use of TF-IDF to examine the relevance of words to documents”. In: *International Journal of Computer Applications* 181.1 (2018), pp. 25–29.

[125] Sebastian Raschka. “Model evaluation, model selection, and algorithm selection in machine learning”. In: *arXiv preprint arXiv:1811.12808* (2018).

[126] EE Rees et al. “Early detection and prediction of infectious disease outbreaks”. In: *CCDR* 45 (2019), p. 5.

[127] *Reporters Without Borders RSF*. 2012. URL: https://www.loc.gov/item/lcwaN0017598/.

[128] Philip Resnik and Jimmy Lin. “11 evaluation of NLP systems”. In: *The handbook of computational linguistics and natural language processing* 57 (2010). DOI: 10.1002/9781444324044.ch11.

[129] Jenn Riley. “Understanding metadata”. In: *Washington DC, United States: National Information Standards Organization (http://www. niso. org/publications/press/UnderstandingMetadata. pdf)* 23 (2017), pp. 7–10.

[130] Renata Lopes Rosa et al. “Event detection system based on user behavior changes in online social networks: Case of the covid-19 pandemic”. In: *IEEE Access* 8 (2020), pp. 158806–158825.

[131] Charlotte Rudnik et al. “Searching News Articles Using an Event Knowledge Graph Leveraged by Wikidata”. In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW ’19. San Francisco, USA: Association for Computing Machinery, 2019, 1232–1239. ISBN: 9781450366755. DOI: 10.1145/3308560.3316761. URL: https://doi.org/10.1145/3308560.3316761.

[132] Silvia Runge-Ranzinger et al. “What does dengue disease surveillance contribute to predicting and detecting outbreaks and describing trends?” In: *Tropical Medicine & International Health* 13.8 (2008), pp. 1022–1041.

[133] Furqan Rustam et al. “A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis”. In: *Plos one* 16.2 (2021), e0245909.

[134] Naseer Ahmed Sajid et al. “A Novel Metadata Based Multi-Label Document Classification Technique.” In: *Computer Systems Science & Engineering* 46.2 (2023).

[135] João Santos, Ivo Anastácio, and Bruno Martins. “Using machine learning methods for disambiguating place references in textual documents”. In: *GeoJournal* 80 (2015), pp. 375–392.

[136] Reza Shabahang et al. ““Give your thumb a break” from surfing tragic posts: Potential corrosive consequences of social media users’ doomscrolling”. In: *Media Psychology* 26.4 (2023), pp. 460–479.

[137] Zahra Shakeri Hossein Abad et al. “Digital public health surveillance: a systematic scoping review”. In: *NPJ digital medicine* 4.1 (2021), p. 41.

[138] Hassan Sheikha. “Text mining Twitter social media for Covid-19: Comparing latent semantic analysis and latent Dirichlet allocation”. In: *University of Gävle, Faculty of Engineering and Sustainable Development, Department of Computer and Geospatial Sciences. diva-portal.org* (2020).

[139] Matthew Sims, Jong Ho Park, and David Bamman. “Literary event detection”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 3623–3634.

[140] Laura Slaughter et al. “A framework for capturing the interactions between laypersons’ understanding of disease, information gathering behaviors, and actions taken during an epidemic”. In: *Journal of biomedical informatics* 38.4 (2005), pp. 298–313.

[141] Enrico Steiger, Joao Porto De Albuquerque, and Alexander Zipf. “An advanced systematic literature review on spatiotemporal analyses of t witter data”. In: *Transactions in GIS* 19.6 (2015), pp. 809–834.

[142] Mehtab Alam Syed et al. “A Data-Driven Score Model to Assess Online News Articles in Event-Based Surveillance System”. In: *Information Management and Big Data: 8th Annual International Conference, SIMBig 2021, Virtual Event, December 1–3, 2021, Proceedings*. Springer. 2022, pp. 264–280.

[143] Mehtab Alam Syed et al. “A metadata approach to classify domain-specific documents for Event-based Surveillance Systems”. In: *2023 International Conference on Communication, Computing and Digital Systems (C-CODE)*. IEEE. 2023, pp. 1–5.

[144] Mehtab Alam Syed et al. “Feature Selection for Sentiment Classification of COVID-19 Tweets: H-TFIDF Featuring BERT.” In: *HEALTHINF*. 2022, pp. 648–656.

[145] Mehtab Alam Syed et al. *GeospaCy*. Version 1.0.0. 2023. DOI: 10.5281/zenodo.8415401. URL: https://github.com/mehtab-alam/GeospaCy/.

[146] Mehtab Alam Syed et al. "GeoXTag: Relative spatial information extraction and tagging of unstructured text". In: *AGILE: GIScience Series* 3 (2022), pp. 1–10. DOI: 10.5194/agile-giss-3-16-2022,2022.

[147] Mehtab Alam Syed et al. "Spatial opinion mining from COVID-19 twitter data". In: *International Journal of Infectious Diseases* 116 (2022), S27.

[148] Sayali Sunil Tandel, Abhishek Jamadar, and Siddharth Dudugu. "A survey on text mining techniques". In: *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*. IEEE. 2019, pp. 1022–1026.

[149] Areeba Umair, Elio Masciari, and Muhammad Habib Ullah. "Vaccine sentiment analysis using BERT+ NBSVM and geo-spatial approaches". In: *The Journal of Supercomputing* (2023), pp. 1–31.

[150] Areeba Umair et al. "Sentimental Analysis of COVID-19 Vaccine Tweets Using BERT+ NBSVM". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2022, pp. 238–247.

[151] Sarah Valentin. "Extraction and combination of epidemiological information from informal sources for animal infectious diseases surveillance". PhD thesis. Université Montpellier, 2020.

[152] Sarah Valentin, Renaud Lancelot, and Mathieu Roche. "Identifying associations between epidemiological entities in news data for animal disease surveillance". In: *Artificial Intelligence in Agriculture* 5 (2021), pp. 163–174.

[153] Sarah Valentin et al. "PADI-web 3.0: A new framework for extracting and disseminating fine-grained information from the news for animal disease surveillance". In: *One Health* 13 (2021), p. 100357.

[154] Sarah Valentin et al. "PADI-web: an event-based surveillance system for detecting, classifying and processing online news". In: *Human Language Technology. Challenges for Computer Science and Linguistics: 8th Language and Technology Conference, LTC 2017, Poznań, Poland, November 17–19, 2017, Revised Selected Papers 8*. Springer. 2020, pp. 87–101.

[155] Maria Vasardani, Stephan Winter, and Kai-Florian Richter. "Locating place names from place descriptions". In: *International Journal of Geographical Information Science* 27.12 (2013), pp. 2509–2532.

[156] Yuli Vasiliev. *Natural Language Processing with Python and SpaCy: A Practical Introduction*. No Starch Press, 2020.

[157] Reza Vaziri and Mehran Mohsenzadeh. "A questionnaire-based data quality methodology". In: *International Journal of Database Management Systems* 4.2 (2012), p. 55.

[158] Sherry L Vellucci. "Metadata." In: *Annual Review of Information Science and Technology (ARIST)* 33 (1998), pp. 187–222.

[159] Marco Vernier, Manuela Farinosi, and Gian Luca Foresti. "Twitter data mining for situational awareness". In: *Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics*. IGI Global, 2019, pp. 684–695.

[160] Richard Y Wang and Diane M Strong. "Beyond accuracy: What data quality means to data consumers". In: *Journal of management information systems* 12.4 (1996), pp. 5–33.

[161] Zhaoxia Wang and Zhiping Lin. "Optimal feature selection for learning-based algorithms for sentiment classification". In: *Cognitive Computation* 12.1 (2020), pp. 238–248.

[162] Jamie K Wardman. "Recalibrating pandemic risk leadership: Thirteen crisis ready strategies for COVID-19". In: *COVID-19*. Routledge, 2022, pp. 260–288.

[163] Davy Weissenbacher et al. "Overview of the Fourth Social Media Mining for Health (SMM4H) Shared Tasks at ACL 2019". In: *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 21–30. DOI: 10.18653/v1/W19-3203. URL: https://aclanthology.org/W19-3203.

[164] P Welby-Everard, O Quantick, and A Green. "Emergency preparedness, resilience and response to a biological outbreak". In: *BMJ Mil Health* 166.1 (2020), pp. 37–41.

[165] David Westerman, Patric R Spence, and Brandon Van Der Heide. "Social media as information source: Recency of updates and credibility of information". In: *Journal of computer-mediated communication* 19.2 (2014), pp. 171–183.

[166] WHO. "A guide to establishing event-based surveillance". In: *World Health Organization* (2008).

[167] WHO. *WHO announces COVID-19 outbreak a pandemic*. 2020. URL: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic.

[168] Kumanan Wilson and John S Brownstein. "Early detection of disease outbreaks using the Internet". In: *Cmaj* 180.8 (2009), pp. 829–831.

[169] Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. "ensemble named entity recognition (ner): evaluating ner Tools in the identification of Place names in historical corpora". In: *Frontiers in Digital Humanities* 5 (2018), p. 2.

[170] Kehan Wu et al. "Deep learning models for spatial relation extraction in text". In: *Geo-spatial Information Science* 0.0 (2022), pp. 1–13. DOI: 10.1080/10095020.2022.2076619. eprint: https://doi.org/10.1080/10095020.2022.2076619. URL: https://doi.org/10.1080/10095020.2022.2076619.

[171] Diego Ricardo Xavier et al. "Involvement of political and socio-economic factors in the spatial and temporal dynamics of COVID-19 outcomes in Brazil: A population-based study". In: *The Lancet Regional Health–Americas* 10 (2022).

[172] Zhan Xu. "Examining US Newspapers' Partisan Bias in COVID-19 News Using Computational Methods". In: *Communication Studies* 74.1 (2023), pp. 78–96.

[173] Inbal Yahav, Onn Shehory, and David Schwartz. "Comments mining with TF-IDF: the inherent bias and its removal". In: *IEEE Transactions on Knowledge and Data Engineering* 31.3 (2018), pp. 437–450.

[174] Chenchen Ye et al. "Beyond text: Incorporating metadata and label structure for multi-label document classification using heterogeneous graphs". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 3162–3171.

[175] Junting Ye and Steven Skiena. "Mediarank: computational ranking of online news sources". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2469–2477.

[176] Manzhu Yu et al. "Spatiotemporal event detection: A review". In: *International Journal of Digital Earth* 13.12 (2020), pp. 1339–1365.

[177] Victor L Yu and Lawrence C Madoff. "ProMED-mail: an early warning system for emerging diseases". In: *Clinical infectious diseases* 39.2 (2004), pp. 227–232.

[178] Sarah Zenasni et al. "Spatial information extraction from short messages". In: *Expert Systems with Applications* 95 (2018), pp. 351–367.

[179] Daniel Zeng, Zhidong Cao, and Daniel B Neill. "Artificial intelligence–enabled public health surveillance—from local detection to global epidemic monitoring and control". In: *Artificial Intelligence in Medicine*. Elsevier, 2021, pp. 437–453. DOI: 10.1016/B978-0-12-821259-2.00022-3.

[180] Chunju Zhang et al. "Rule-based extraction of spatial relations in natural language text". In: *2009 international conference on computational intelligence and software engineering*. IEEE. 2009, pp. 1–4.

[181] Chunju Zhang et al. "Rule-Based Extraction of Spatial Relations in Natural Language Text". In: *2009 International Conference on Computational Intelligence and Software Engineering*. 2009, pp. 1–4. DOI: 10.1109/CISE.2009.5363900.

[182] Jiawei Zhang and Hua Qi. "Data Mining and Spatial Analysis of Social Media Text Based on the BERT-CNN Model to Achieve Situational Awareness: a Case Study of COVID-19". In: *Journal of Geodesy and Geoinformation Science* 5.2 (2022), p. 38.

[183] Shuai Zhang et al. "Identifying features of health misinformation on social media sites: an exploratory analysis". In: *Library Hi Tech* 40.5 (2022), pp. 1384–1401.

[184] Wei Zhang and Judith Gelernter. "Geocoding location expressions in Twitter messages: A preference learning method". In: *Journal of Spatial Information Science* 9 (2014), pp. 37–70.

[185] Kun Zheng et al. “A knowledge representation model based on the geographic spatiotemporal process”. In: *International Journal of Geographical Information Science* 36.4 (2022), pp. 674–691. DOI: 10.1080/13658816.2021.1962527. eprint: https://doi.org/10.1080/13658816.2021.1962527. URL: https://doi.org/10.1080/13658816.2021.1962527.

[186] Cheng Zhou et al. “Characterizing the dissemination of misinformation on social media in health emergencies: An empirical study based on COVID-19”. In: *Information Processing & Management* 58.4 (2021), p. 102554.

[187] Wenxuan Zhou et al. “UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition”. In: *arXiv preprint arXiv:2308.03279* (2023).

[188] Xiao Xiang Zhu et al. “Geo-information harvesting from social media data”. In: *arXiv preprint arXiv:2211.00543* (2022).

[189] Xiaolan Zhu and Susan Gauch. “Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web”. In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. 2000, pp. 288–295.

[190] Said Özcan. *tweet-preprocessor: Elegant tweet preprocessing*. 2016. URL: https://github.com/s/preprocessor.