Détection des influenceurs dans des médias sociaux

Kévin Deturck

Institut National des Langues et Civilisations Orientales

École doctorale n°265

Langues, littératures et sociétés du monde

ERTIM

THÈSE

présentée par

Kévin DETURCK

soutenue le 18 Novembre 2021

pour obtenir le grade de **Docteur de l'INALCO**

en Traitement Automatique des Langues

DÉTECTION DES INFLUENCEURS DANS DES MÉDIAS SOCIAUX

Thèse dirigée par :

Madame Frédérique SEGOND	Directrice de recherche, Inria / Inalco

RAPPORTEURS :

Madame Claudine MOÏSE	Professeure des universités, UGA
Monsieur Patrick PAROUBEK	Ingénieur de recherche (HDR), CNRS

MEMBRES DU JURY :

Madame Claudine MOÏSE	Professeure des universités, UGA
Monsieur Patrick PAROUBEK	Ingénieur de recherche, CNRS
Monsieur Pascal AMSILI	Professeur des universités, Université Paris 3
Monsieur Mathieu VALETTE	Professeur des universités, Inalco
Madame Frédérique SEGOND	Directrice de recherche, Inria / Inalco
Monsieur Damien NOUVEL	Maitre de conférences, Inalco

Résumé

Dans cette thèse, nous présentons la conception et l'évaluation d'un système pour détecter automatiquement les personnes influentes dans les médias sociaux, à partir des manifestations de leur action d'influence dans les communications interpersonnelles.

Les approches pour la détection des influenceurs utilisent généralement, soit la structure de la communication entre les individus, soit l'analyse de son contenu. Le cadre théorique retenu dans notre thèse a la particularité de combiner ces deux types d'approches pour leur complémentarité.

Nous caractérisons l'action des influenceurs à l'échelle d'un individu cible, depuis sa mise en œuvre jusqu'à ses effets, par des traits discursifs relevant aussi bien des messages envoyés par les influenceurs que de ceux envoyés par les individus influencés. La détection automatique de ces traits discursifs est faite avec des méthodes en traitement automatique des langues, basées sur des règles linguistiques et des modèles par apprentissage automatique.

À l'échelle d'un groupe, l'action des influenceurs est caractérisée par leur position centrale dans un graphe social qui représente des actions interpersonnelles ayant eu cours à l'intérieur de ce groupe. L'hybridité de notre système consiste en l'utilisation des informations linguistiques sur les traits discursifs d'influence, extraits automatiquement depuis les messages textuels échangés entre individus, afin de construire les graphes sociaux.

Mots clés : influenceurs, influence, traitement automatique des langues, réseaux sociaux, médias sociaux, analyse de graphe, mesures de centralité, apprentissage automatique

Abstract

In this thesis, we present the design and evaluation of a system to automatically detect influencers in social media, based on the manifestations of their influencing action in interpersonal communications.

Approaches to influencer detection generally use either the structure of communication between individuals, or the analysis of its content. The theoretical framework chosen in our thesis has the particularity of combining these two types of approach for their complementarity.

We characterise the action of influencers at the level of a target individual, from its means to its effects, by discursive features of both the messages sent by the influencers and those sent by the influenced individuals. The automatic detection of these discursive features in social media messages is done with methods in natural language processing, based on linguistic rules and machine learning models.

At the group level, the action of influencers is characterised by their central position in a social graph, that represents interpersonal actions within the group. The hybridity of our system consists in the use of linguistic information, automatically extracted in discussions, to construct the social graphs whose structure will be analysed.

Keywords: influencers, influence, social media, social networks, natural language processing, machine learning, graph analysis, centrality measures

Remerciements

D'abord, je remercie beaucoup ma directrice de thèse, Frédérique Segond, instigatrice de cette grande aventure. Tu as cru en moi et m'as donné une ambition que je n'avais pas, celle de faire une thèse de doctorat. Tu m'as beaucoup conseillé, rassuré. Tu as su renforcer ma confiance en moi, et c'est en grande partie grâce à toi que j'en suis là.

Je remercie grandement deux autres membres de mon équipe encadrante, Namrata Patel et Damien Nouvel. Namrata, tu as été très présente auprès de moi pour cette aventure, contrebalançant beaucoup mon solitariat. Tes attentions furent nombreuses et ton entrain communicatif. Damien, malgré la distance géographique, tu as été d'une grande proximité dans le suivi de mes travaux, portant un grand intérêt à mes avancées et participant beaucoup à mes réflexions. Tes suggestions nombreuses ont été salutaires.

J'adresse encore un remerciement, cette fois général, aux membres de mon équipe encadrante : votre patience, votre gentillesse et votre écoute m'ont beaucoup aidé. Vous avez été là pour apporter votre recul et votre expérience, nécessaires pour mener un projet au si long cours.

Les remerciements à mon équipe encadrante ne seraient pas complets sans en adresser aussi à Thibault Celier et Parantapa Goswami. Thibault, en tant que directeur de l'unité recherche et innovation à laquelle nous appartenions à Viseo, tu as œuvré pour valoriser mon travail et m'as donné une liberté d'action bénéfique à ma recherche. Parantapa, même si ce fut court, pour des raisons extérieures à ce travail, tu as été un encadrant disponible, à l'écoute et volontaire ; ce fut agréable et motivant.

Je remercie chaleureusement tous les les étudiants qui se sont investis dans le rôle d'annotateur alors même que la tâche demandée était loin d'être évidente. En particulier, j'ai travaillé de façon prolongée avec Ania Chepaikina, Maëva Leproux, Afala Phaxay et Lili Wu, dans le cadre des stages dédiés à la campagne d'annotation. Vous avez toutes les quatre été patientes, compréhensives, attentives et actives, des qualités qui vous ont consacrées super annotatrices !

Je remercie grandement les rapporteurs de ce travail de thèse, Madame Claudine Moïse et Monsieur Patrick Paroubek, ainsi que les autres personnes extérieures à ce travail, qui ont accepté d'être membres du jury : Monsieur Patrick Amsili, Monsieur Mathieu Valette et Monsieur Nicolas Dugué.

Enfin, bien des personnes m'ont été bénéfiques durant cette longue aventure, en des aspects différents, du plus formel au plus intime. À ces âmes, trop nombreuses pour être énumérées ici, je vous adresse d'aussi nombreux mercis.

Table des matières

Avant-propos

Nous vivons dans une ère où le numérique est la plus grande fenêtre sur le monde. Que ce soit pour observer ou interagir avec le réel, on passe de plus en plus par l'intermédiaire, et littéralement le *biais*, du virtuel.

Internet est la source principale d'information pour beaucoup d'individus du monde entier. Ces derniers utilisent des moteurs de recherche ou des réseaux sociaux afin de répondre instantanément à leurs questions existentielles, comme à leurs problèmes du quotidien.

Les outils du numérique sont des moteurs de la perception que les individus ont du monde. On ne se fait plus un avis sur quelque chose mais on va simplement puiser parmi ce que les autres en disent. L'esprit critique et la capacité de réflexion ne sont plus de mise : toutes les opinions étant représentées, on va vers celle qui nous attire, celle qui nous plait. Ce comportement est un point d'entrée pour influencer les individus : il n'y a qu'à dire ce que le plus grand nombre a envie d'entendre.

Le numérique, avec en particulier Internet, est aussi l'accès principal aux autres. Dans une société mondialisée, la communication entre les individus s'établit instantanément, par l'entremise d'outils numériques, comme les messageries virtuelles et les réseaux sociaux. Ces derniers ont le pouvoir particulier de mettre en contact les individus rapidement et à grande échelle. Il est désormais possible, pour n'importe quelle personne, d'accéder à une large audience afin de se faire connaitre, faire connaitre son œuvre, ou diffuser un message.

Le présent travail de recherche s'inscrit pleinement dans ce contexte où des individus construisent leur vision des choses par Internet, une vision souvent éblouie par la grande quantité d'informations disponible. Ces individus, parfois en perdition, suivent alors d'autres individus qui leur servent de guides.

1. Introduction

1.1. La notion d'*influenceur*

L'activité consistant à influencer autrui n'est pas nouvelle. Elle existe et est étudiée depuis longtemps pour différents domaines où les décisions et le comportement des individus est déterminant, comme en politique ou en marketing. Sa pratique s'organise dans des groupes, comme les lobbies, ou autour d'individus particuliers, par exemple des célébrités. Dans tous les cas, les personnes impliquées ont un rôle d'influenceur. Pour autant, le terme influenceur n'émerge qu'avec les médias sociaux.1.1.

Les médias sociaux, en tant qu'outils de communication accessibles et populaires, ont démocratisé l'activité d'influence, au point d'ériger un statut d'influenceur. Des utilisateurs de médias sociaux se revendiquent influenceurs et en font leur profession. Il existe des répertoires d'influenceurs. Les influenceurs des médias sociaux sont devenus un sujet de société.

Les *nouveaux influenceurs*, qui ont émergé avec les médias sociaux, intéressent à la fois le grand public et les experts. Le grand public s'interroge sur la pratique de cette nouvelle forme d'influence (*Réseaux sociaux, followers et partenariats... Des influenceurs racontent leur quotidien*[1]) tandis que les experts s'intéressent à l'impact des influenceurs sur leurs domaines respectifs (*Les influenceurs et les marques*[2]). C'est cette forme relativement nouvelle d'influence qui nous intéresse dans le cadre de ce travail.

Nous allons élaborer une méthode pour détecter les influenceurs qui agissent dans les médias sociaux et la vérifier sur des données réelles. Avant d'aller plus loin, il faut nous assurer de la définition d'un *influenceur*.

Même si le terme *influenceur* est attesté sur Internet depuis plusieurs années, il fait son entrée dans Le Petit Larousse 2021[3]. Cet évènement reflète l'importance qu'a pris le concept d'influenceur dans la société.

Le dictionnaire en donne la définition suivante : « *Personne qui, par sa position sociale, sa notoriété et/ou son exposition médiatique, a un grand pouvoir d'influence sur l'opinion publique, voire sur les décideurs.* ». Cette définition décrit un influenceur comme un individu particulièrement intégré socialement, ce qui lui permet d'avoir une influence sur

[1] https://france3-regions.francetvinfo.fr/bourgogne-franche-comte/cote-d-or/dijon/reseaux-sociaux-followers-partenariats-influenceurs-racontent-leur-quotidien-1880074.html

[2] https://www.reech.com/fr/influence-etude-reech-2020

[3] https://www.ladepeche.fr/2020/05/21/remontada-influenceur-feminicide-decouvrez-les-nouveaux-mots-du-petit-larousse,8897211.php

les opinions et les décisions d'autres individus. La notion d'*influenceur* recouvre plusieurs concepts qui ont trait à des domaines variés.

Le concept clé dans la définition d'un influenceur est évidemment celui d'*influence*, que nous retrouvons dans son dérivé, *influencer*. Le terme *influence* provient de *influentia*, en latin médiéval, qui désigne l'*écoulement d'un flux provenant des astres et agissant sur les hommes et les choses*[4]. Si cette définition étymologique de l'influence s'applique de façon très générale à des entités du monde, dans le cadre des influenceurs, on parle d'influence entre des personnes.

Étymologiquement, le terme *influence* désigne l'ascendant d'entités sur d'autres entités. Il possède même une connotation divine : l'influence prend sa source d'entités célestes et s'applique, de façon omnipotente, aux entités du monde. On retrouve cela dans le pouvoir et le prestige associés aux influenceurs.

Le terme *influentia* désignera plus tard l'*action lente et continue exercée par une personne, une chose sur une autre personne ou chose*[5]. L'action d'influence est ainsi décrite tel un processus, matérialisé par le flux qui relie l'entité influente à l'entité influencée, et qui reste à caractériser.

Dans le cas des influenceurs, nous traitons l'influence en tant que phénomène social, un phénomène que la sociologie a étudié avant même l'apparition du terme *influenceur*. S'intéresser aux effets de l'influence conduit à un autre domaine, celui de la psychologie.

Nous avons vu que l'influence agit au niveau des opinions et décisions des individus, ce qui relève du domaine de la psychologie sociale, parce qu'il s'agit des effets de contacts sociaux sur la psychologie d'individus. Ces contacts sociaux établissent des canaux servant à la diffusion de l'influence ; il est donc essentiel de considérer ces contacts sociaux afin d'analyser l'action des influenceurs.

Lors d'un processus d'influence, le *flux* qui *s'écoule* entre les individus dénote l'existence d'une communication entre ceux-ci. Pour les influenceurs des médias sociaux, il s'agit d'une communication électronique, avec un échange de messages dont la caractérisation peut aussi bien relever de l'analyse d'image que de l'analyse de texte.

La définition générale d'un influenceur est convenable pour savoir de quoi on parle mais elle reste abstraite : elle ne nous donne pas à voir des influenceurs en action dans un environnement social.

Quand on parle aujourd'hui d'*influenceurs*, on fait principalement référence à des utilisateurs de réseaux sociaux. Nous avons vu qu'un influenceur est particulièrement intégré socialement. Dans les réseaux sociaux, les influenceurs se démarquent avant tout par leur forte communauté.

Le Larousse définit la *communauté* comme un *ensemble de personnes unies par des liens d'intérêts, des habitudes communes, des opinions ou des caractères communs*[6]. Une communauté est donc un groupe dont les membres entretiennent des relations entre eux, par intérêt ou par identification.

[4] https://www.cnrtl.fr/etymologie/influence
[5] https://www.cnrtl.fr/etymologie/influence
[6] https://www.larousse.fr/dictionnaires/francais/communauté/17551

Dans les réseaux sociaux, les influenceurs sont des utilisateurs qui possèdent une communauté, c'est-à-dire qu'ils sont au centre d'un groupe qui les suit et les écoute particulièrement. Les influenceurs génèrent de l'activité dans leur communauté respective. Il reste à savoir pourquoi les influenceurs fédèrent autant.

La popularité des influenceurs se traduit par l'intérêt qu'ils suscitent chez beaucoup d'autres individus. Cela leur assure une audience. Les influenceurs produisent des messages qui ont un impact sur la pensée et le comportement de leur audience. Ces messages peuvent aussi bien être une prescription dans un post du forum Doctissimo[7] qu'une photo postée sur Instagram[8]. La communauté autour d'un influenceur et le contenu qu'il produit sont donc déterminants. Dans le cadre de ce travail, nous analysons ces deux composantes dans l'environnement des médias sociaux.

1.2. Les médiaux sociaux et les influenceurs

Le terme *média social* n'a pas encore sa place dans les dictionnaires classiques. Il est moins populaire que le terme *réseau social*, avec lequel on peut le confondre mais, bien que liés, leurs significations sont distinctes.

Le site *L'internaute*[9] propose une définition d'un *média social* qui inclut ses caractéristiques essentielles : une *plateforme sur Internet qui permet aux gens de créer du contenu, d'organiser ce contenu, de le modifier ou de le commenter, un média social mélange interaction, technologie et création de contenu*[10]. Cette définition inscrit les médias sociaux dans l'ère du Web 2.0, dont un principe est le *contenu généré par les utilisateurs* (*User Generated Content*).

Le média social est le symbole d'une évolution d'Internet, qui met les utilisateurs au centre, en faisant appel à leur participation. Les utilisateurs d'un média social peuvent participer en générant du contenu et en le commentant, que ce soit le leur ou celui d'autres utilisateurs. Il s'agit donc bien d'un média parce qu'il est un moyen pour ses utilisateurs de diffuser du contenu. Ce média est dit *social* parce qu'il met ses utilisateurs en contact autour du contenu diffusé.

Avoir des contacts sociaux et générer du contenu sont des actes essentiels pour les influenceurs (cf. section 1.1). Les médias sociaux sont si hétérogènes à cet égard qu'il convient d'en faire une typologie.

Dans les médias sociaux, nous distinguons deux types de contacts. Il y a d'une part les contacts ponctuels, autour du contenu échangé. Cela concerne principalement les forums, les blogs et les wikis. Il y a d'autre part les contacts durables, qui lient les utilisateurs entre eux ; ils sont caractéristiques des réseaux sociaux. En considérant la nature des contacts,

[7] doctissimo.fr
[8] instagram.com
[9] linternaute.fr
[10] https://www.linternaute.fr/dictionnaire/fr/definition/media-social/

les réseaux sociaux sont plus favorables à la constitution d'une audience pérenne, ce qui peut être crucial pour les influenceurs. Cependant, la génération de contenu est aussi déterminante pour les influenceurs.

Nous distinguons deux formes de génération de contenu dans les médias sociaux. Dans un cas, le contenu est produit sous la forme d'un journal personnel, comme dans les blogs ainsi que les réseaux sociaux Facebook, Instagram, Youtube et Tiktok. Dans l'autre cas, le contenu provient de discussions entre utilisateurs ; c'est ce qu'on trouve dans les forums, les wikis ou le réseau social Twitter. Le format du journal personnel permet aux influenceurs de fédérer une communauté, en étant au centre des échanges.

Ces différents types de médias sociaux forment une palette de modes de communication pour les institutions ou les marques qui peuvent faire appel aux influenceurs comme des leviers de communication.

Pendant la crise sanitaire de la Covid-19, l'État français a collaboré avec des influenceurs pour qu'ils relaient des messages de responsabilité à la population, et en particulier aux jeunes[11]. C'est aussi pour toucher une jeune population que la marque de cosmétiques NYX Cosmetics a travaillé avec des influenceurs Instagram[12]. Dans ce domaine, des marques comme Clarins et Bourjois, visent une communication d'authenticité et de proximité avec des influenceurs relativement peu connus, appelés *micro-influenceurs*[13] par rapport aux beaucoup plus célèbres *macro-influenceurs*[14].

Afin de mettre en relation les influenceurs avec leurs potentiels clients, il existe des catalogues ou des agences d'influenceurs[15] qui nécessitent un processus majoritairement manuel de sélection des influenceurs. C'est justement au niveau du repérage des influenceurs que se situe le présent travail.

Notre objectif est de créer un système de détection automatique des influenceurs dans les médias sociaux, pour identifier plus aisément, d'après des données terrain, les influenceurs qui conviennent à un certain besoin, selon le public ou la thématique visée. Ce serait en particulier un atout pour découvrir des *micro-influenceurs*. Nous allons formaliser la problématique afférente au développement d'un tel système.

1.3. La problématique et ses difficultés

La problématique motivant ce travail est de trouver une solution pour détecter automatiquement les influenceurs à partir de données collectées dans des médias sociaux. Nous modélisons cette problématique en trois tâches successives.

[11] https://www.huffingtonpost.fr/entry/covid-19-comment-les-gouvernements-se-sont-tournes-vers-les-influenceurs-pour-communiquer_fr_5fc0c562c5b6e4b1ea4a5c21

[12] https://www.launchmetrics.com/fr/ressources/blog/instagram-influenceur

[13] https://potion.social/fr/blog/micro-influence-3-exemples-campagnes-efficaces

[14] https://www.journalducm.com/macro-influenceurs

[15] https://junto.fr/blog/top-10-agences-influenceurs/

Nous allons d'abord construire un modèle théorique des influenceurs dans les médias sociaux. Il mettra en exergue des caractéristiques qui permettent de distinguer les influenceurs des autres utilisateurs. Nous développerons ensuite un outil mettant en application ce modèle théorique. Enfin, nous évaluerons l'outil développé pour voir dans quelle mesure il répond à la problématique.

Pour réaliser la première tâche, nous allons traiter les deux composantes des médias sociaux, mises en évidence dans cette introduction, qui sont essentielles pour les influenceurs : la structure sociale et le contenu échangé. C'est un type d'analyse qui est novateur par rapport à l'état de l'art parce que la plupart des approches proposées se concentrent sur l'une ou l'autre des composantes (cf. section 2).

L'hybridité de notre analyse est une difficulté parce qu'elle requiert une solution composite et donc complexe à réaliser. Le traitement de chacune des composantes induit aussi des difficultés spécifiques.

L'analyse des structures sociales dans les médias sociaux soulève des questions de modélisation qui ont trait à des domaines nombreux et variés dont, principalement, la théorie des graphes et la sociologie. La nature des informations présentes dans les médias sociaux est riche et variée, avec notamment des relations et interactions de diverses natures ; il sera d'autant plus complexe de trouver celles qui sont pertinentes à utiliser et la bonne façon de les allier.

Notre spécialité étant le Traitement Automatique des Langues (TAL), nous allons nous intéresser particulièrement à l'analyse du contenu textuel lorsqu'il est le vecteur de processus d'influence. Nous ferons appel à des notions de linguistique pour caractériser les messages ayant trait à l'action des influenceurs, qu'ils soient envoyés par les influenceurs ou les individus influencés.

Pour l'approche textuelle, il y a deux difficultés importantes. La première difficulté, fondamentale, concerne la définition et la modélisation des traits discursifs de l'influence. La seconde difficulté, d'ordre plus pratique, concerne la détection de ces traits discursifs avec des outils de TAL.

Pour les deux axes de notre travail, une difficulté est d'obtenir des données de médias sociaux qui nous permettent d'observer des influenceurs identifiés, que ce soit à l'aune de leur intégration sociale ou de leur participation à des conversations.

Nous allons nous focaliser sur les réseaux sociaux et les forums. Ils ont en commun d'être des *médias conversationnels*, avec des particularités complémentaires pour notre étude : les réseaux sociaux sont, par définition, centrés autour des relations et donc favorables à l'analyse de la structure sociale, tandis que les forums sont basés sur les discussions, générant du contenu textuel.

Afin de répondre à la problématique présentée dans ce premier chapitre, nous proposons une approche focalisée sur la caractérisation des manifestations d'influence dans la structure des contacts sociaux et le contenu des messages textuels (Deturck, 2020). Pour analyser la structure des contacts sociaux, nous adopterons une approche fondée sur la théorie des graphes. Pour analyser le contenu des messages textuels, nous adopterons une approche linguistique. Pour la suite, nous adoptons le plan qui suit.

Dans le chapitre 2, nous faisons un état des lieux sur la caractérisation des influenceurs et leur détection automatique. Ensuite, dans le chapitre 3, nous présentons l'approche que nous proposons pour détecter les influenceurs dans les médias sociaux. Le chapitre 4 est consacré aux jeux de données utilisés pour mettre en pratique notre modélisation. S'en suit la description des expériences réalisées : dans le chapitre 5, sur la détection structurelle des influenceurs, dans le chapitre 6, sur la détection linguistique, et dans le chapitre 7, sur la détection hybride. Nous concluons ce travail dans le chapitre 8.

2. État de l'art

2.1. Détection des influenceurs basée sur la structure des contacts sociaux

2.1.1. La genèse du concept de la centralité

Le concept de la *centralité* prend racine dans l'analyse structurelle des contacts sociaux. Ce type d'analyse est amené par le courant de la *psychologie topologique* qui part du principe que l'environnement d'un individu influence son comportement. L'analyser du comportement d'un individu est réalisée par la topologie de son environnement (des objets aux évènements) en utilisant des concepts mathématiques (Lundberg & Lewin, 1939).

Le concept de la *centralité* émerge quant à lui avec des travaux qui se focalisent sur l'environnement social des individus. Nous allons présenter ces travaux en utilisant le terme de *réseau social* pour désigner un ensemble de liens entre des individus et celui de *graphe social* pour la représentation d'un réseau social suivant la théorie des graphes.

Le psychosociologue Alex Bavelas est le premier à parler de *centralité*. Il émet l'hypothèse que la communication dans un groupe est structurée autour de positions *centrales*, qui permettent d'avoir le contrôle sur la transmission de l'information et, par-là, d'influencer autrui (Bavelas, 1948).

Dans un réseau social, la centralité d'un individu est la grandeur qui indique à quel point celui-ci est proche d'une position centrale définie selon certaines propriétés topologiques. Bavelas associe ces positions centrales au concept d'*intermédiarité*, qui indique à quel point un individu est présent sur les chemins les plus courts entre les autres individus. Il va s'agir ensuite de vérifier cette hypothèse.

C'est ce que proposent Bavelas (Bavelas, 1950) et le psychologue Leavitt (Leavitt, 1951) en étudiant la communication dans des petits groupes d'individus. Tous deux observent que l'émergence de leaders est en lien avec leur centralité dans le réseau. Le concept de la *centralité* est aussi vérifié dans d'autres types de réseaux sociaux, comme les réseaux de connaissance : il y a des individus centraux qui permettent de mettre en rapport d'autres individus (Travers & Milgram, 1969). Ces travaux empiriques sont complétés par des travaux théoriques qui formalisent la centralité et ses mesures.

Le premier travail formalisant l'hypothèse de Bavelas est celui du sociologue Linton Clarke Freeman (Freeman, 1977), qui propose plusieurs mesures de centralité basées sur le concept d'intermédiarité. L'année suivante, il passe en revue l'ensemble des mesures de centralité proposées dans l'état de l'art et en fait la synthèse pour les rendre plus accessibles (Freeman, 1978).

Freeman met en évidence trois grandes conceptions de la centralité : la centralité par adjacence ou degré, la centralité par proximité et la centralité par intermédiarité. Dans ces travaux, les mesures de centralité ne tiennent compte que des chemins les plus courts entre les individus et considèrent les liens de façon binaire : ils existent ou pas. Nous allons voir que d'autres mesures de centralité tiennent compte de l'ensemble des liens du graphe et les pondèrent.

En 1953, le statisticien Leo Katz propose une nouvelle méthode pour mesurer l'importance qu'a un individu dans un groupe social (Katz, 1953). Il s'agit de mesurer le *statut social* d'un individu qui, en plus de prendre en compte le nombre d'individus qui lui sont liés, prend aussi en compte leur statut respectif. Ce principe est aussi utilisé dans les mesures de Bonacich (Bonacich, 1987) et même dans des mesures que nous verrons dans la suite pour le web, comme PageRank (Page et al., 1998).

Après nous être intéressés à la genèse du concept et des mesures de centralité, dont l'application était prévue sur des réseaux d'individus regroupés physiquement, nous allons voir comment la recherche sur la centralité s'est adaptée à la démocratisation du web.

2.1.2. La centralité pour le web

Avec la popularité croissante d'Internet dans les années 1990, de nouvelles mesures de centralité sont apparues pour mesurer l'importance des pages web. Ce type de mesures repose sur l'analyse des graphes de liens hypertextes pour déterminer quelles sont les pages web les plus centrales. PageRank et Hits sont deux mesures célèbres qui reposent sur un principe similaire, celui d'un parcours de graphe itératif pour optimiser le calcul des poids des liens et des scores des pages.

La mesure PageRank (Page et al., 1998) a montré son efficacité en apportant un avantage compétitif indéniable au moteur de recherche Google. La mesure Hits (Kleinberg, 1999) a la particularité de proposer deux types de centralité complémentaires, la centralité *d'autorité* et la centralité *de relais* : le score d'autorité d'une page est proportionnel à son nombre de liens entrants et au score de *relais* des pages sources, le score de *relais* est proportionnel au nombre de liens sortants et au score d'autorité des pages cibles.

Il y a eu un travail (Devi et al., 2014) pour comparer PageRank et Hits, qui a conclu que la mesure PageRank est moins coûteuse en termes de calcul mais moins précise que Hits. PageRank est donc à favoriser pour des grands graphes tandis que Hits est plus intéressante sur des graphes aux dimensions modestes. Outre l'invention de ces mesures de centralité pour le web, un nouvel algorithme a été créé pour adapter la mesure classique de centralité par intermédiarité aux grands graphes du web (Brandes, 2001).

Internet a aussi donné naissance aux *réseaux sociaux en ligne* (ou *réseaux sociaux*), des plateformes numériques qui permettent à leurs utilisateurs de se lier entre eux afin de partager du contenu. Parmi les plus populaires : LinkedIn (2002) est un réseau social professionnel, Facebook (2004) est fondé sur les liens d'amitié et le partage de contenu

multimédia, YouTube (2005) est destiné à la publication de vidéos, Twitter (2006) à celle de courts messages textuels, Instagram (2010) et Snapchat (2011) permettent de créer et de partager des photos et des courtes vidéos, Tiktok (2016), des courtes vidéos.

2.1.3. La centralité pour les réseaux sociaux en ligne

Depuis la fin des années 2000, la popularité et la diversité des réseaux sociaux en ligne dynamisent la recherche sur la détection des influenceurs par centralité. Beaucoup de travaux sont motivés par des applications en marketing, définissant les influenceurs comme les individus qui ont une forte capacité à promouvoir des produits dans les réseaux sociaux car ils ont un impact important sur la communication qui s'y déroule (Kiss & Bichler, 2008), (Heidemann et al., 2010). Il s'agit alors de trouver les mesures de centralité et les modélisations en graphe les plus aptes à identifier ces influenceurs.

Twitter est un réseau social qui est très présent dans la recherche sur la centralité car il est riche en différents types d'actions interpersonnelles et ses données sont plus accessibles que celles d'autres réseaux sociaux populaires, comme Facebook.

Une partie de cette recherche a utilisé des mesures de centralité préexistantes : des mesures généralistes, comme la mesure par intermédiarité (Xu et al., 2014), ou destinées au web, comme la mesure de Katz et PageRank (Rosa et al., 2018). Nous avons aussi relevé un travail d'adaptation de la mesure de centralité PageRank à l'environnement Twitter (Lü et al., 2011).

Facebook a aussi été étudié pour la détection d'influenceurs : dans le cadre d'une évaluation comparative, des graphes représentant différents types d'actions interpersonnelles du réseau social ont été analysés avec différentes mesures de centralité (Khadangi & Bagheri, 2017). La centralité a aussi été utilisée pour identifier les chercheurs qui sont des influenceurs, avec des graphes de co-auteurs (Mariani et al., 2014).

Les mesures de centralité sont fondées sur des hypothèses concernant la dynamique des échanges dans les réseaux sociaux, de façon à définir ce qu'est une position centrale. Dans la section suivante, nous allons voir des approches qui analysent les dynamiques des échanges dans les réseaux sociaux, de façon à optimiser la propagation d'influence à partir d'utilisateurs clés.

2.1.4. Les approches basées sur la propagation de l'influence

Les approches basées sur la propagation de l'influence associent l'intégration sociale des influenceurs et leur pouvoir d'influence à une forte propagation de leurs propos ou actions à travers un réseau social. L'analyse en propagation consiste à analyser la dynamique créée par les actions des individus à l'intérieur d'un réseau social.

La problématique centrale de ce type d'approche est celle de la maximisation d'influence, motivée par le marketing viral (Chen et al., 2010). Elle consiste à identifier, dans un réseau social, un ensemble optimal d'individus à partir desquels une information se propagera maximalement dans le réseau, autrement dit ceux qui rendront une information la plus virale. Des chercheurs travaillent sur des modèle et algorithmes pour trouver une approximation de la solution optimale à ce problème.

La première solution proposée consiste en l'utilisation d'un modèle probabiliste (Domingos & Richardson, 2001) tandis que le travail le plus célèbre se base sur des modèles de diffusion issus des mathématiques pour la sociologie (Kempe et al., 2003).

D'autres travaux vont s'inspirer de ce dernier et chercher à créer des algorithmes plus efficaces, que ce soit par un travail théorique sur leur complexité (Borgs et al., 2014) ou un travail empirique sur les paramètres qui favorisent la diffusion d'une action (Dave et al., 2011) ou d'une opinion (Gionis et al., 2013).

Outre l'intégration sociale, nous avons vu que le contenu échangé lors des contacts sociaux a aussi une importance dans le déclenchement d'un phénomène d'influence.

2.2. Détection des influenceurs basée sur l'analyse du contenu textuel

Un des axes de notre travail consiste à détecter les influenceurs par l'analyse des messages textuels. Le type d'approches le plus commun consiste à caractériser la façon dont s'expriment les influenceurs afin de les détecter en analysant des messages.

Une approche empirique est la comparaison des messages d'influenceurs à ceux d'individus non-influenceurs (Quercia et al., 2011). Une autre approche est l'utilisation de traits comportementaux des influenceurs, définis dans une théorie en psychologie, pour les associer à des traits discursifs qui devront être identifiés dans des messages (Rosenthal & McKeown, 2017).

Un autre type d'approches consiste à identifier les effets des influenceurs dans les échanges, comme la détection de changements de thématique dans les conversations (Nguyen et al., 2014).

Nous avons vu deux grandes catégories de travaux qui portent sur les analyses structurelles et textuelles prises indépendamment. Dans la suite, nous allons voir d'autres travaux qui ont cherché à combiner ces deux axes pour détecter les influenceurs dans les médias sociaux.

2.3. Détection hybride des influenceurs

Certaines approches par hybridation partent du principe que les influenceurs sont au centre de communautés thématiques. Ces approches consistent à intégrer, dans le calcul de la centralité, l'information sur la similarité sémantique des messages entre les utilisateurs (Weng et al., 2010), (Katsimpras et al., 2015). Nous allons voir qu'une autre famille d'approches par hybridation combine la centralité avec des indicateurs d'influence issus de l'analyse textuelle.

L'originalité du contenu est une caractéristique textuelle récurrente pour détecter les influenceurs par hybridation : elle est utilisée seule avec la centralité des auteurs (Song et al., 2007) ou en combinaison avec une caractérisation du niveau d'expertise de leurs messages (Li et al., 2013). La centralité est aussi utilisée en association avec une mesure de la qualité d'écriture des messages et de leur niveau d'engagement, identifié à leur degré de polarisation (Bigonha et al., 2012).

Nous venons de passer en revue les différents types d'approches qui ont été proposés pour détecter les influenceurs dans les médias sociaux. Nous avons vu qu'une majeure partie de ces approches est basée sur le concept de la centralité dans les graphes sociaux.

Il nous semble que la faiblesse des approches par centralité est de considérer la sélection d'une mesure de centralité comme une fin en soi, mettant de côté la caractérisation des influenceurs et de leur action. Nous pensons qu'il faut bien plus s'intéresser à la nature des graphes sociaux sur lesquels on applique les mesures de centralité, parce qu'elle contribue à déterminer la signification de la centralité mesurée.

La critique est similaire concernant les approches sur la propagation d'influence. Elles se focalisent sur la technique pour optimiser la propagation d'information dans un réseau social, qui n'est vu que comme une structure, laissant de côté la caractérisation du contenu qui est propagé et la façon dont l'audience réagit à ce contenu. Or, nous pensons que ce sont des éléments essentiels dans les processus d'influence et qu'ils contribuent ainsi à détecter les influenceurs.

Notre approche utilise des mesures de centralité parce qu'elles sont pratiques pour analyser globalement les rapports entre les individus représentés dans un graphe social. Cependant, nous travaillons davantage à caractériser l'action d'influence entre les individus de façon à en rendre compte dans la représentation en graphe. Cette caractérisation se fera par une approche textuelle.

Les approches textuelles que nous avons relevées se focalisent sur la caractérisation de l'expression des influenceurs mais elles ne tiennent pas compte des effets sur l'audience. Dans notre approche, nous ajoutons à la caractérisation du discours des influenceurs celle du discours des individus influencés car c'est une preuve potentielle de la présence d'influenceurs. Cela permet d'avoir une vue plus globale sur l'action d'influence.

Ce travail est réalisé sur la base d'une observation de processus d'influence à l'intérieur de conversations. Nous avons cherché en particulier ce qui, dans le discours des influenceurs, produit un phénomène d'influence. Cette approche empirique a l'avantage de s'assurer que les traits discursifs sélectionnés ont trait à l'action des influenceurs.

Les approches structurelles et les approches textuelles constituent deux pans de l'état de l'art qui, nous l'avons vu, sont majoritairement dissociés. Nous avons tout de même relevé quelques approches hybrides mais elles consistent plus en une addition des types d'analyse qu'en une véritable intégration de ces dernières. Dans ce travail, nous allions une description linguistique approfondie du phénomène d'influence à sa représentation structurelle dans un seul modèle hybride.

3. Modélisation de l'influence et définitions

3.1. Introduction

3.1.1. Les influenceurs par leur action d'influence

Le sujet de la détection des influenceurs dans des médias sociaux ainsi que le domaine dans lequel s'inscrit ce travail, le TAL, ont des exigences d'applicabilité auxquelles notre modélisation doit répondre de manière pragmatique. Dans cette optique, nous avons choisi de modéliser les influenceurs par leur action d'influence dans les médias sociaux.

Nous caractérisons l'action des influenceurs dans les médias sociaux par deux composantes : l'*influence individuelle*, qui désigne l'impact sur une prise de décision à l'échelle d'un individu, et l'*influence globale*, qui représente la multiplication de l'influence individuelle à travers un réseau d'individus en interaction. Nous allons voir en quoi les médias sociaux favorisent l'influence.

3.1.2. L'influence dans les médias sociaux

Le principe d'un média social est de mettre en contact ses utilisateurs. Les contacts entre les utilisateurs des médias sociaux sont les vecteurs de l'influence qui s'y exerce. Nous distinguons trois types de contacts entre les utilisateurs d'un média social.

Le premier type de contact est l'*action interpersonnelle*. Nous définissons une *action interpersonnelle* dans un média social comme la production par un utilisateur d'un effet à destination d'au moins un autre utilisateur. Par exemple, l'envoi d'un tweet, même sans la mention d'un destinataire, est une action interpersonnelle parce qu'il produit un message pour l'ensemble des utilisateurs de Twitter. Une action interpersonnelle peut être spontanée ou correspondre à une réaction à une autre action interpersonnelle. Elle peut entrainer le type de contact suivant, l'interaction sociale.

Le Larousse définit une interaction sociale comme une *relation interpersonnelle entre deux individus au moins, par laquelle les comportements de ces individus s'influencent mutuellement*[16]. Une interaction sociale implique donc des actions et des réactions interpersonnelles mutuelles ; c'est le cas, par exemple, dans une conversation. Remarquons que l'influence, telle que nous l'avons définie (cf. section 3.1.1), ne requiert

[16] https://www.larousse.fr/dictionnaires/francais/interaction/43595/locution?q=interaction#180273

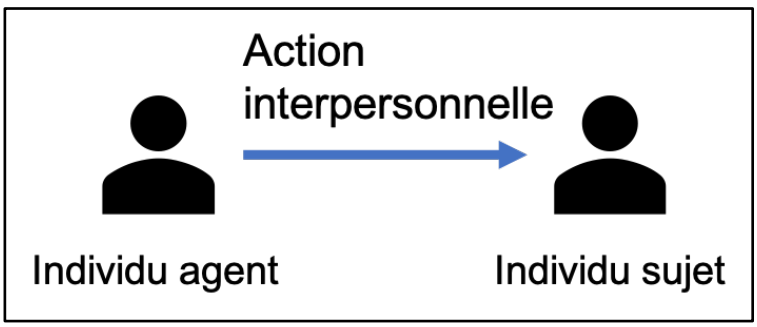
pas d'interaction sociale parce qu'elle peut s'exercer sans qu'il y ait des réactions mutuelles entre l'influenceur et l'influencé. L'action interpersonnelle comme l'interaction sociale peuvent conduire au troisième et dernier type de contact social, la *relation sociale*.

Dans un média social, nous définissons la *relation sociale* comme un contact qui s'établit durablement entre deux individus. Cette relation peut être réciproque, comme la relation d'amitié dans Facebook, ou unilatérale, comme la relation consistant à suivre l'activité d'un utilisateur sur Twitter. Dans le premier cas, la relation résulte d'une interaction sociale entre les deux individus amis, dans le second cas, elle résulte d'une action interpersonnelle de l'utilisateur suiveur vers l'utilisateur suivi.

Nous constatons que l'action interpersonnelle est le type de contact fondamental, à la base des interactions et des relations sociales. Ainsi, nous allons nous focaliser sur la description des actions interpersonnelles en vue de modéliser l'influence dans les médias sociaux.

Une action interpersonnelle possède une structure que nous caractérisons par un lien orienté entre les individus parties prenantes : ce lien part des individus qui produisent l'action, que nous appelons *individus agents*, et va jusqu'aux individus sur lesquels porte cette action ou qui ont produit le contenu ayant induit l'action, que nous appelons *individus sujets* (cf. Figure 1).

Figure 1 : Structure d'une action interpersonnelle



Une action interpersonnelle possède aussi une sémantique qui découle de son contenu. Les actions interpersonnelles dans les médias sociaux ont des types spécifiques de structure et de contenu (cf. Tableau 1) que nous allons expliciter.

Tableau 1 : Typologie des actions interpersonnelles dans les médias sociaux

Sémantique / Portée	Portée directe	Portée indirecte
Sémantique fermée	Abonnement à un individu	Partage d'un message
Sémantique ouverte	Message à un individu	Réponse à un message

Nous caractérisons la structure d'une action interpersonnelle d'après la nature de sa *portée*, qui désigne l'objet de cette action, à travers deux modalités de contact entre les individus agents et les individus sujets.

La portée est *directe* lorsque l'action interpersonnelle porte directement sur l'individu sujet. C'est le cas, par exemple, de l'envoi d'un message à un utilisateur, ou des actions *S'abonner* sur Instagram et *Suivre* sur Twitter, qui consistent toutes les deux à entrer directement en relation avec un utilisateur.

La portée d'une action interpersonnelle est *indirecte* lorsque cette dernière porte sur le contenu qu'un individu a produit ; nous considérons alors que cet individu est indirectement l'individu sujet de l'action. Par exemple, dans Twitter, c'est le cas de l'action *Retweeter*, qui consiste à partager le tweet d'un autre utilisateur. Pour compléter cette typologie, il faut nous intéresser à la sémantique des actions interpersonnelles.

Du point de vue sémantique, nous distinguons deux types d'actions interpersonnelles dans les médias sociaux, celles dont la sémantique est *fermée* et celles dont la sémantique est *ouverte*.

La sémantique d'une action interpersonnelle est *fermée* lorsqu'elle est prédéfinie par le média social : c'est par exemple le cas de l'action *J'aime*[17] présente dans plusieurs réseaux sociaux comme Facebook, Twitter et Instagram. La sémantique d'une action interpersonnelle est *ouverte* lorsque ce sont les utilisateurs qui la construisent, par exemple lors de l'envoi d'un message.

Nous avons défini l'action des influenceurs dans les médias sociaux en deux aspects complémentaires : l'*influence individuelle* et l'*influence globale*, la première étant intrinsèque à la seconde (cf. section 3.1.1). Nous formulons l'hypothèse que, à l'intérieur d'un média social, l'influence individuelle apparait dans la sémantique des actions interpersonnelles et l'influence globale dans leur structure. Pour analyser l'influence individuelle et l'influence globale, nous allons focaliser notre analyse sur certains types d'actions interpersonnelles.

Pour l'influence individuelle, nous allons nous intéresser tout particulièrement aux actions interpersonnelles dont la sémantique est ouverte, parce qu'elles laissent libre cours à l'expression des individus, avec en particulier du texte, que nous analyserons par des approches en TAL.

Afin de voir si l'influence globale apparait davantage au niveau des utilisateurs eux-mêmes ou du contenu qu'ils produisent, nous allons comparer l'utilisation d'actions interpersonnelles à la portée directe par rapport à celles dont la portée est indirecte.

En section 3.2, nous proposons la modélisation de l'influence globale avec une approche par graphe, puis, en section 3.3, la modélisation de l'influence individuelle par une approche conversationnelle. Enfin, la section 3.4 portera sur l'hybridation des deux types d'approches précédents pour affiner la modélisation de l'influence globale.

[17] La sémantique du *J'aime* a même été enrichie par Facebook qui propose désormais une gamme de réactions comme « J'adore » ou «Ha ha».

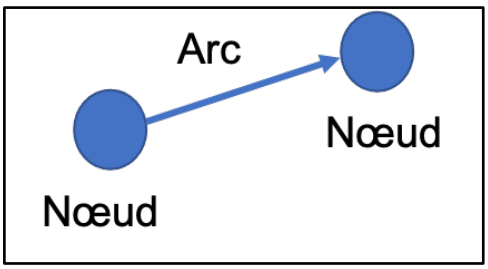
3.2. L'influence globale par le réseau

3.2.1. Approche par l'analyse de graphe

Nous avons précédemment formulé l'hypothèse que l'influence globale apparaissait à travers la structure des actions interpersonnelles (cf. section 3.1.2). Pour analyser la structure des actions interpersonnelles, nous devons la formaliser. Pour y parvenir, nous utilisons la théorie des graphes.

Par définition, un graphe est un outil permettant de représenter un ensemble d'éléments pouvant être reliés les uns aux autres. Ces éléments sont appelés des *nœuds* et les liens qui les relient des *arcs*.

Figure 2 : Éléments de base d'un graphe



Ainsi, on définit formellement un graphe G tel que $G = < N, A >$, où N est un ensemble de nœuds et A un ensemble d'arcs. Dans le cadre de notre travail, nous construisons des graphes sociaux.

Dans un graphe social, les nœuds représentent des individus et les arcs indiquent des liens entre ces individus. Dans notre cas, les liens correspondent à des actions interpersonnelles (cf. section 3.1.2). Par exemple, à partir de Twitter, nous pouvons créer un graphe de retweets dans lequel les arcs représenteraient les actions de retweet entre des utilisateurs. L'analyse de graphes fait appel à des propriétés que nous définissons ci-dessous.

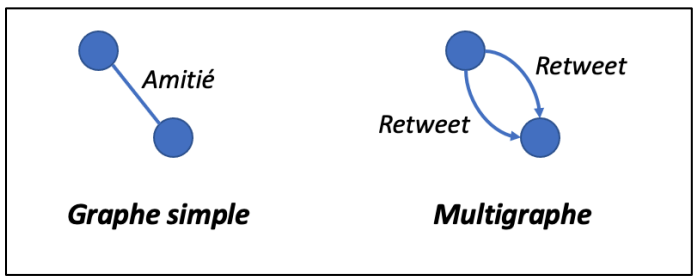
- Complexité du graphe

La *complexité* du graphe dépend de la possibilité de trouver plusieurs arcs (des arcs multiples) entre deux nœuds *voisins*, c'est-à-dire deux nœuds directement reliés par un arc. Un graphe est *simple* lorsque deux nœuds sont reliés par au maximum un seul arc et qu'il n'y a pas de boucle (un arc qui relie un nœud à lui-même). Un graphe est un *multigraphe* lorsque deux nœuds peuvent être reliés par plus d'un arc ou qu'il y a une

boucle. Cette distinction permet d'étudier l'importance de la multiplicité des interactions sociales dans l'action des influenceurs.

Un exemple de *graphe simple* est un graphe d'amitié issu de Facebook : la relation d'amitié entre deux individus est unique, alors, il n'est pas possible de trouver plus d'un arc entre deux nœuds. Au contraire, un graphe de retweet issu de Twitter peut être un multigraphe dans le cas où un utilisateur a retweeté plusieurs tweets d'un autre utilisateur. Nous représentons en figure 3 un exemple pour chacun des deux types de graphe.

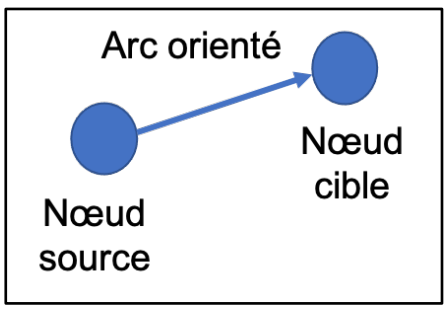
Figure 3 : Illustration des concepts de *graphe simple* et de *multigraphe*



- ***Orientation du graphe***

Un graphe est dit orienté si chacun de ses arcs est orienté, c'est-à-dire avec un nœud source et un nœud cible (cf. Figure 4). Dans le cadre de l'étude des influenceurs, cette propriété du graphe est déterminante puisqu'elle permet en particulier d'identifier la source d'une action d'influence. Nous analyserons donc essentiellement des graphes orientés.

Figure 4 : Illustration du graphe orienté



- ***Degré d'un nœud***

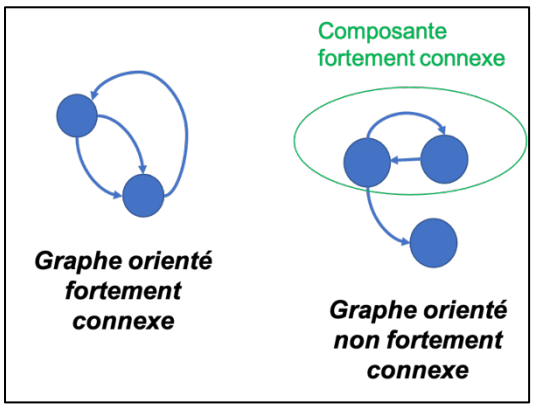
Le *degré d'un nœud* correspond au nombre d'arcs qui lui sont incidents. Dans un graphe orienté, le *degré entrant* ne tient compte que des arcs orientés vers le nœud considéré tandis que le *degré sortant* ne tient compte que de ceux orientés vers d'autres nœuds. Les

mesures de centralité se basent principalement sur cette propriété parce qu'elle rend compte de l'intégration d'un nœud dans un graphe.

- ***Connexité d'un graphe orienté***

Dans un graphe orienté, la *connexité* a trait à l'existence d'un chemin entre les nœuds du graphe. Un *chemin* entre deux nœuds *n1* et *n2* d'un graphe orienté *G* est un ensemble d'arcs appartenant à *G* qui relient *n1* à *n2*. Un graphe est *fortement connexe* si cette propriété est vraie pour tous les nœuds du graphe (à gauche dans la Figure 5). Dans un graphe non fortement connexe, il est possible d'identifier des *composantes fortement connexes* (à droite dans la Figure 5). La connexité est un indicateur de la cohésion des individus, favorable à la propagation de l'influence.

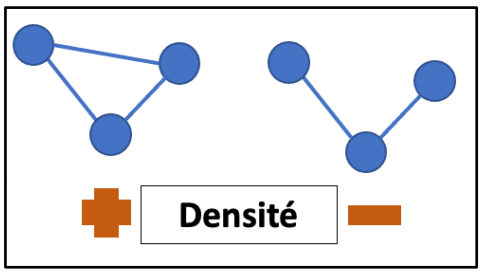
Figure 5 : Illustration de la notion de *connexité d'un graphe orienté*



- ***Densité du graphe***

La *densité* d'un graphe correspond à la proportion d'arcs uniques (les arcs reliant des paires de nœuds différentes) existant dans le graphe par rapport à la quantité d'arcs uniques qu'il est possible d'y établir (cf. Figure 6).

Figure 6 : Illustration de la notion de densité du graphe



Dans le cas d'un graphe social, la densité indique à quel point les individus considérés ont des liens entre eux ; une densité importante peut contribuer à une plus grande influence globale. La densité δ d'un graphe orienté contenant M nœuds et N arcs peut être calculée avec la formule ci-dessous.

$$\delta = \frac{N}{M\,(M-1)}$$

Dans la section suivante, nous allons voir comment nous utilisons la représentation en graphe pour modéliser l'influence globale.

3.2.2. Modélisation de l'influence globale par la centralité

Le concept de la centralité provient d'une intuition mathématique du psychosociologue Bavelas (Bavelas, 1948), formalisée par le sociologue structuraliste Freeman (Freeman, 1978), qui se base sur une vision des réseaux d'interactions sociales comme ayant une topologie en étoile, c'est-à-dire avec des interactions qui s'articulent autour d'une position centrale.

Bavelas associe la capacité d'influencer un groupe social à une position centrale dans la communication de ce même groupe, une idée que nous allons reprendre pour modéliser l'influence globale.

En nous inspirant de l'idée de Bavelas présentée auparavant, nous formulons l'hypothèse suivante : dans un réseau social, l'influence globale d'un individu se manifeste par son positionnement central dans la structure des actions interpersonnelles, caractéristiques de son action d'influence individuelle. Ainsi, pour évaluer l'influence globale des individus dans un réseau social, nous allons utiliser la *centralité*.

Appliquée à la représentation d'un réseau en graphe, la centralité est une grandeur qui indique à quel point les nœuds de ce graphe possèdent les propriétés topologiques spécifiques à une position centrale. Différentes spécifications d'une telle position sont possibles (Freeman, 1978), ce qui donne lieu à différentes mesures de centralité, c'est-à-dire différentes modélisations de la centralité.

Les différentes mesures de centralité se distinguent par leurs définitions respectives de ce qu'est une position centrale dans un graphe. Toutes calculent un score pour chacun des nœuds d'un graphe avec une formule spécifique qui tient nécessairement compte des arcs auxquels chaque nœud est relié. Plus le score d'un nœud est élevé, plus il est considéré comme ayant une position centrale dans le graphe.

Nous venons de nous intéresser à la façon de représenter et d'analyser la structure des contacts interpersonnels pour mesurer l'influence globale des individus dans un réseau

social. Nous avons défini l'influence globale comme une multiplicité d'influences individuelles (cf. section 3.1.1) et nous avons formulé l'hypothèse que l'identification de l'influence individuelle passe par une analyse du contenu des contacts interpersonnels (cf. section 3.1.2). Nous allons proposer une modélisation de l'influence individuelle centrée sur la description linguistique des messages échangés dans les médias sociaux.

3.3. L'influence individuelle par le message

3.3.1. Approche linguistique

Poursuivant notre objectif d'une modélisation pragmatique, nous avons choisi de modéliser l'influence individuelle à l'aune de ses manifestations langagières, dans des messages de médias sociaux (Nouvel et al., 2019). Afin de concevoir une telle modélisation, notre approche, empirique, part de discussions incluant des influenceurs pour analyser à la fois leurs messages et ceux des autres participants.

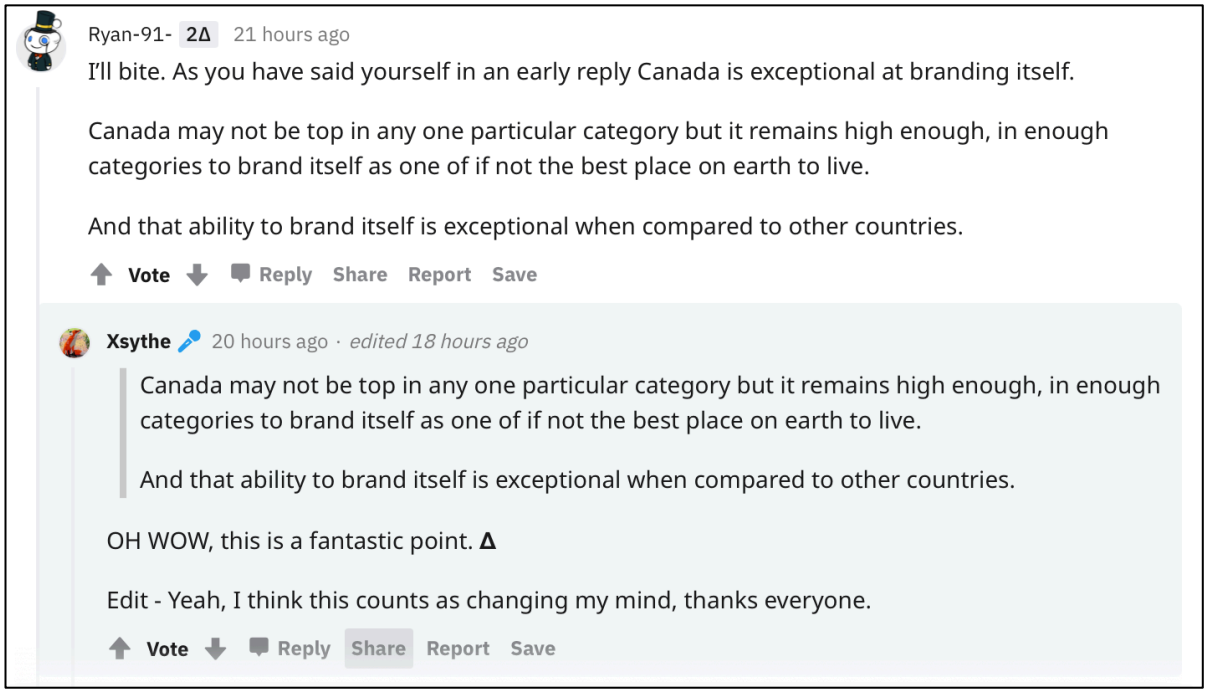
Notre objectif est d'identifier des régularités qui nous permettent de modéliser l'influence individuelle par le langage. Une première étape a été de constituer un jeu de données qui contienne des influenceurs identifiés et des conversations auxquelles ils ont participé.

Nous avons choisi un jeu de données issu du forum *Change my view*[18] (Tan et al., 2016). Dans ce forum, l'auteur d'un fil de discussion expose, dans un message initial, son opinion sur un sujet quelconque. Ensuite, les autres participants discutent cette opinion pour faire changer d'avis l'auteur initial, ce qui revient à essayer d'exercer une influence individuelle sur ce dernier (cf. section 3.1.1). Si ce forum convient bien à notre étude de l'influence individuelle par son principe, nous allons voir qu'il est aussi un atout par son système de récompense des participants influenceurs.

Lorsqu'un message fait changer d'avis un participant (même s'il n'est pas l'auteur initial), ce dernier doit citer le post correspondant (incluant le texte du message et son auteur) et expliquer son changement d'avis avec un token *delta*. Si l'explication est jugée suffisante (avec au moins 50 caractères), le robot modérateur valide l'attribution du delta. La Figure 7 montre un extrait d'une discussion issue du forum *Change my view*. On peut y voir une réponse à un message avec un changement d'avis signalé par un symbole delta.

[18] https ://www.reddit.com/r/changemyview/

Figure 7 : Extrait d'une discussion dans le forum *Change my view*



Le système *delta* nous offre une annotation en influence *ad hoc*. En outre, nous avons vu que les processus d'influence apparaissent dans des messages, ce qui convient bien à une analyse avec du TAL. La prochaine étape consiste à caractériser les discours de l'influence individuelle à partir de ce corpus.

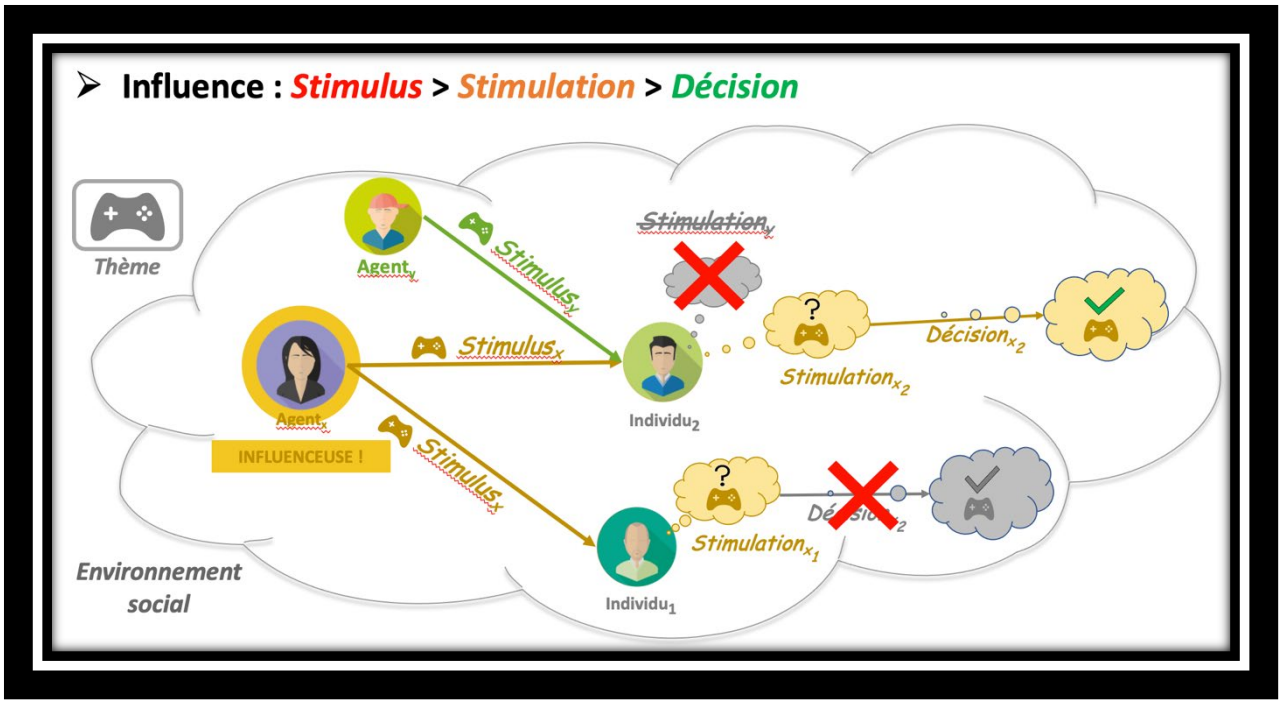
Nous avons analysé manuellement une dizaine de fils de discussion contenant des changements d'avis (et donc des actions d'influence individuelle) pour identifier des régularités en termes de (1) structure et (2) discours.

Chaque action d'influence individuelle se manifeste par un message avec un changement d'avis validé par la modération du forum. Un tel message cite le message source du changement d'avis (la source d'influence) et donne l'explication de ce qui l'a stimulé.

Nous avons généralisé ces observations pour modéliser l'action d'influence individuelle comme un processus en trois étapes, présenté en section 3.3.2. Pour chacune de ces étapes, nous avons identifié des types de discours réguliers que nous définissons dans les sections 3.3.3 et 3.3.4.

3.3.2. Modèle de l'influence individuelle

Figure 8: Modèle de l'influence individuelle



L'influence individuelle, telle que nous l'avons définie en section 3.1.1, pose un regard à l'échelle microscopique sur le phénomène social qu'est l'influence (cf. section 1). Elle concerne la façon dont l'influence s'exerce entre un individu et un autre.

Le modèle que nous présentons dans cette section s'inscrit dans la lignée de travaux en psychologie sociale, comme (Mason et al., 2007), et en sciences de la communication, comme (Dillard & Wilson, 2014). Ces travaux décrivent l'influence en tant que processus avec des individus sources qui impactent le mental d'individus cibles par l'échange de messages (respectivement les agents et les individus dans la Figure 8).

Notre modèle de l'influence individuelle contient trois composantes ou étapes (cf. Figure 8) : le *stimulus*, la *stimulation* et la *décision*. Les composantes de *stimulus* et de *stimulation* correspondent à un cadre théorique en psychologie sociale, décrit dans (Turner & Oakes, 1986), qui donne l'environnement social d'un individu comme porteur de *stimuli* pouvant modifier (ou *stimuler*) l'état psychologique des individus s'y trouvant.

La composante de *décision* fait référence au processus de prise de décision, particulièrement étudié en psychologie sociale, comme dans (Ajzen, 1996). Dans la suite, nous définissons chacune de ces trois composantes.

- **Stimulus**

Le *stimulus* est le principe actif d'un influenceur. Dans un environnement social, un agent influenceur produit un *stimulus*, qui consiste en une action de sa part, parvenant à d'autres individus dans l'environnement. Cette action porte sur un thème en adéquation avec l'environnement. Dans l'illustration de la Figure 8, les agents *x* et *y* produisent le $Stimulus_x$ et le $Stimulus_y$, respectivement à destination des individus 1 et 2, sur le thème des jeux-vidéo. Ces *stimuli*, parvenus aux individus cibles, peuvent générer la deuxième composante que nous définissons ci-après, la *stimulation*.

- **Stimulation**

L'arbitrage (ou la prise de décision) d'un individu utilise des informations cognitives ou affectives (Finucane et al., 2003), ce sont donc les deux points de stimulation pour un influenceur. Nous appelons *stimulation* la modification de l'état affectif ou cognitif chez un individu, en réaction à un *stimulus*. Cette modification n'entraine pas nécessairement une décision de l'individu cible et donc une influence sur ce dernier.

Dans la Figure 8, nous observons les *stimulations* x_1 et x_2, respectivement engendrées chez les individus 1 et 2 par les *stimuli* de l'agent *x*. Nous remarquons que le *stimulus* transmis par l'agent *y* à l'individu 2 n'a pas produit de *stimulation*. Le processus d'influence est donc un échec pour l'agent *y*. Pour l'agent *x*, il se poursuit pour peut-être aboutir à la génération de la troisième et dernière composante, définie ci-après : la décision.

- **Décision**

Un individu a une action d'influence individuelle dans un environnement social s'il parvient à impacter l'arbitrage d'un autre individu présent dans cet environnement. Un tel phénomène se manifeste par l'expression d'une nouvelle décision chez l'individu influencé, en réaction à une ou plusieurs stimulation(s). Dans la Figure 8, nous observons que la stimulation *x2* n'a pas produit de décision chez l'individu 2, tandis que la stimulation *x1* en a généré une chez l'individu 1. Nous pouvons donc conclure, pour cet exemple, que l'agent *x* a exercé une action d'influence individuelle sur l'individu 1 mais pas sur l'individu 2.

Les composantes de *stimulation* et de *décision* ont trait aux réactions psychologiques des individus face aux actions des influenceurs (représentées dans notre modèle par la composante de *stimulus*). Ces réactions, par leur nature psychologique, peuvent rester latentes chez les individus parce qu'elles ne sont pas nécessairement verbalisées comme dans notre environnement d'étude, le forum *Change my view* ; elles seraient alors indétectables par notre système.

Toutefois, il est possible de composer avec des données qui ne contiendraient qu'une représentation partielle de l'influence individuelle. Chacune des composantes de notre

modèle se détecte indépendamment des autres. Ainsi, notre système est capable d'identifier une partie des composantes d'un phénomène d'influence individuelle pour quand même détecter l'influenceur correspondant.

Au cours de notre exploration des messages de *Change my view*, nous avons relevé diverses manifestations langagières de chacune des composantes. Nous les avons catégorisées en différents traits discursifs que nous définissons et modélisons en section 3.3.3 pour les *stimuli* et 3.3.4 pour la *stimulation* et la *décision*.

Nos modélisations sont sous la forme de patrons linguistiques, représentés par des arbres décrivant les structures des différents traits discursifs, avec des couleurs qui représentent le type d'analyse linguistique utilisé pour les identifier automatiquement dans des messages.

3.3.3. Les traits discursifs *stimuli*

Dans un fil de discussion *Change my view*, les *stimuli* sont principalement contenus dans les messages produits par des participants autres que l'auteur initial, parce qu'il s'agit de générer un changement d'avis chez l'auteur initial.

Les messages sont en anglais et n'ont pas de taille imposée. Ils peuvent être écrits en réponse à l'auteur initial ou bien à un autre participant à la discussion. Parmi ces messages, nous avons analysé les discours de ceux qui ont été désignés par un auteur initial comme ayant conduit à son changement d'avis (les influenceurs). Cela nous a permis de mettre en évidence trois types de *stimuli* que nous décrivons dans la suite.

- ***Claim* (*Assertion*)**

Un *claim* est un type d'expression par laquelle un individu délivre une description comme étant factuelle, c'est-à-dire une affirmation de ce qui est prétendument un fait dans le monde (Saurí & Pustejovsky, 2012). Un *claim* peut n'être factuel qu'en apparence, c'est-à-dire qu'il peut faire une description concrète avec certitude sans que celle-ci soit vraie.

Ce trait discursif a la particularité d'être désigné par un terme en anglais (*claim*) car c'est le premier terme que nous avons trouvé comme désignation satisfaisante. Nous avons pris l'habitude de l'utiliser mais nous pensons que *assertion*[19] est une traduction correcte en français, ayant la dénotation de tenir quelque chose pour vrai sans apporter de preuve.

Nous présentons les deux patrons linguistiques que nous avons identifiés pour les claims en Figure 9 et Figure 10. Ils correspondent à deux structures syntaxiques comportant des caractéristiques linguistiques qui, dans différentes combinaisons, couvrent l'ensemble des cas de claims que nous avons observés. Un claim a la forme d'une

[19] https://www.larousse.fr/dictionnaires/francais/assertion/5806

expression d'observation ou de connaissance, il contient donc un lexique concret et neutre.

- Exemple : « *Le crash d'un avion en Iran* » ; il s'agit de la description d'un événement avec seulement un syntagme nominal, ce qui correspond au motif de claim SN décrit dans la Figure 9.

Figure 9 : Modélisation linguistique du claim SN

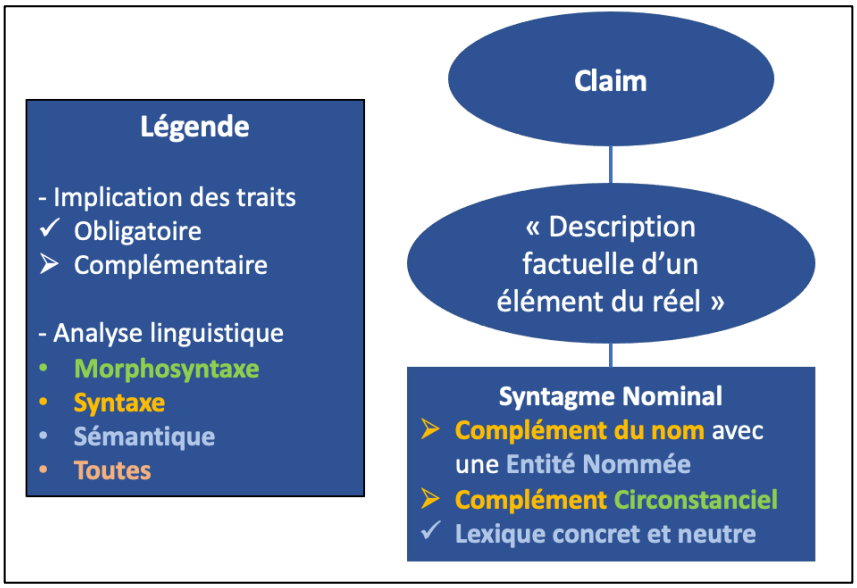
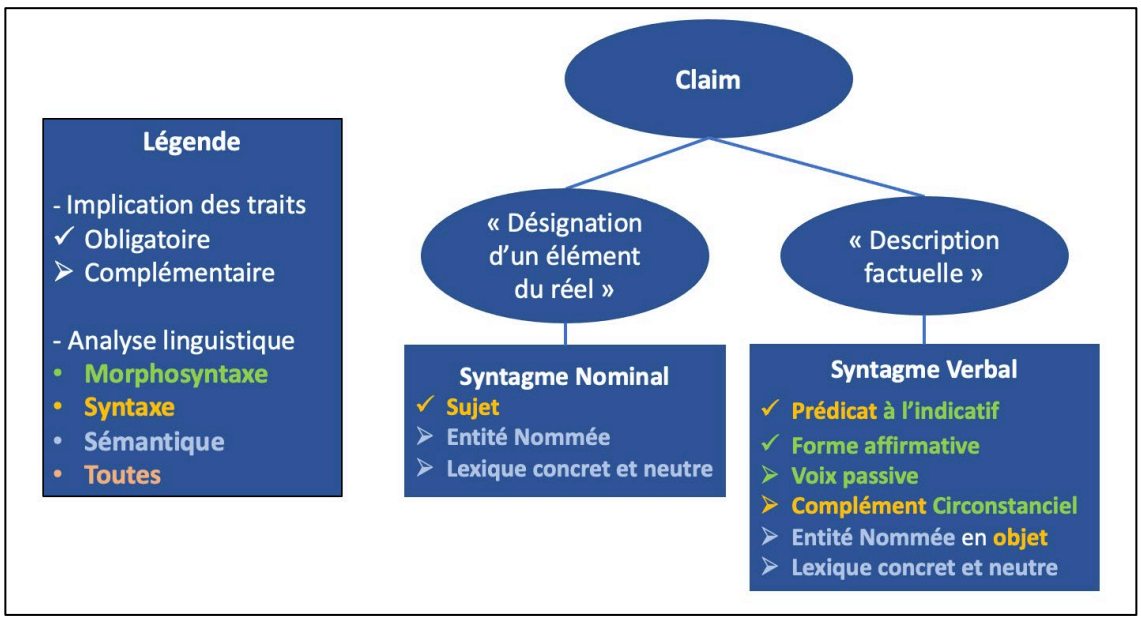


Figure 10 : Modélisation linguistique du claim SN-SV



L'auteur d'un claim affirme ce qu'il prétend être une réalité, ainsi, le modèle SN-SV (cf. Figure 10) se caractérise par une forme affirmative et l'utilisation du mode de la certitude, l'indicatif. Dans ce modèle, la voix passive est souvent utilisée pour mettre en avant l'objet d'une action, comme dans *Deux hommes ont été repérés hier soir dans cette rue*.

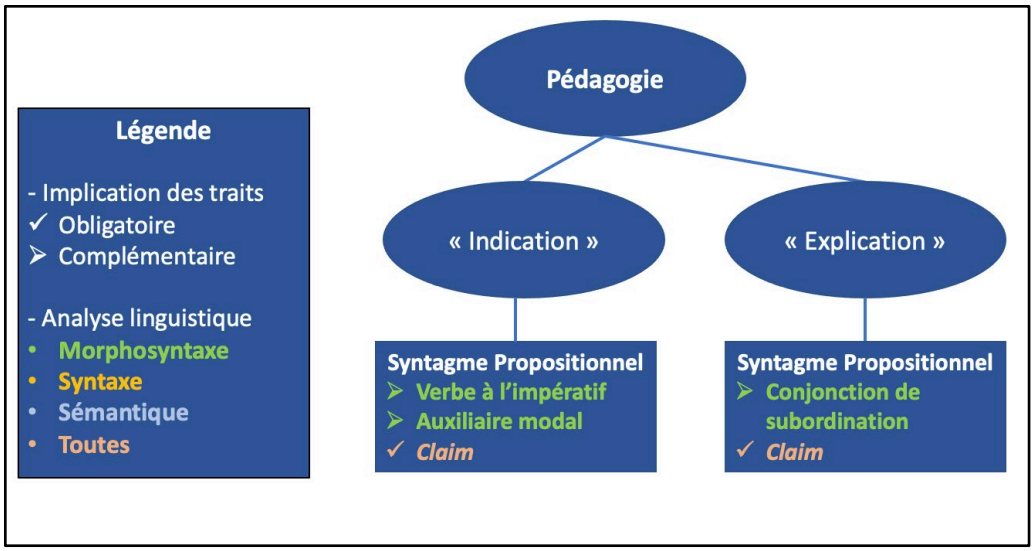
- Exemple : « *Salman hold talks with* Putin » ; il s'agit de la description d'un événement, ayant donc bien l'apparence d'un fait. Cet exemple correspond au motif SN-SV décrit dans la Figure 10, avec la désignation d'une personne par le syntagme nominal « *Salman* » et d'une action associée avec le syntagme verbal « *hold talks with Putin* ».

- *Pédagogie*

La *pédagogie* est le discours d'un individu qui guide autrui dans sa compréhension du monde ou son comportement dans le monde. La pédagogie avait déjà été identifiée par (Dillard & Wilson, 2014) comme une motivation d'influence.

Ce type de discours prend la forme d'une leçon ou d'un conseil, modélisés dans les Figure 11 et Figure 12. La leçon comporte une indication, sous la forme d'un claim ou d'un conseil, accompagnée de son explication, avec possiblement des exemples et une rhétorique logique. Le conseil contient des suggestions ou des recommandations, avec des verbes au conditionnel ou à l'impératif. Un discours pédagogique inclut bien souvent un destinataire (celui qu'il veut guider), avec notamment l'emploi de la deuxième personne.

Figure 11 : Modélisation linguistique de la pédagogie *leçon*



- Exemple : « *Turn it off so they can stay in the darkness of their misguidance* » ; il s'agit de l'expression d'un conseil directement adressé à son destinataire, avec

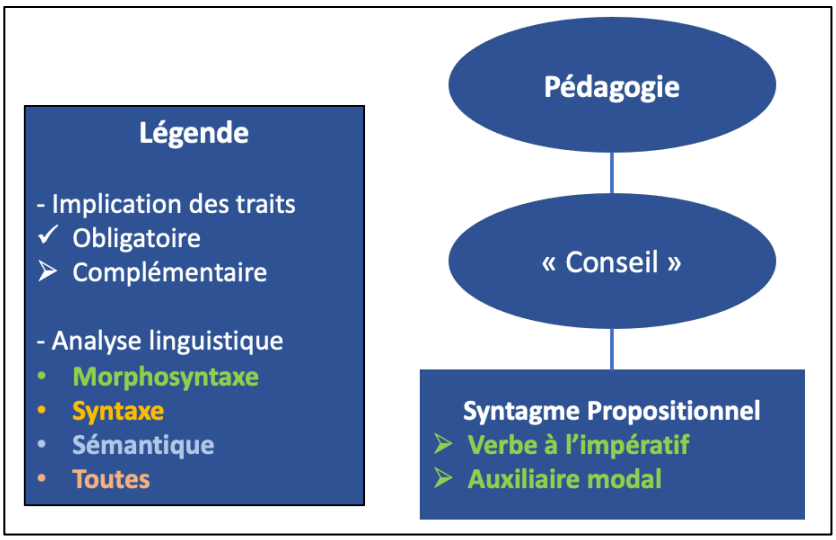
une explication, pour le guider dans son comportement et sa pensée. Cet exemple correspond au modèle de pédagogie leçon décrit dans la

- Figure 11.

- Exemple : « *Turn it off » ;* par comparaison avec l'exemple précédent, celui-ci est une pédagogie avec un conseil seul, exprimé avec un verbe à l'impératif, ce qui correspond au modèle de pédagogie *conseil* décrit dans la Figure 12.

- Exemple : « *You should do it.* » ; dans cet exemple, il s'agit aussi d'une pédagogie *conseil* (cf. Figure 12) parce qu'il y a un conseil seul, cette fois-ci avec un auxiliaire modal.

Figure 12 : Modélisation linguistique de la pédagogie *conseil*

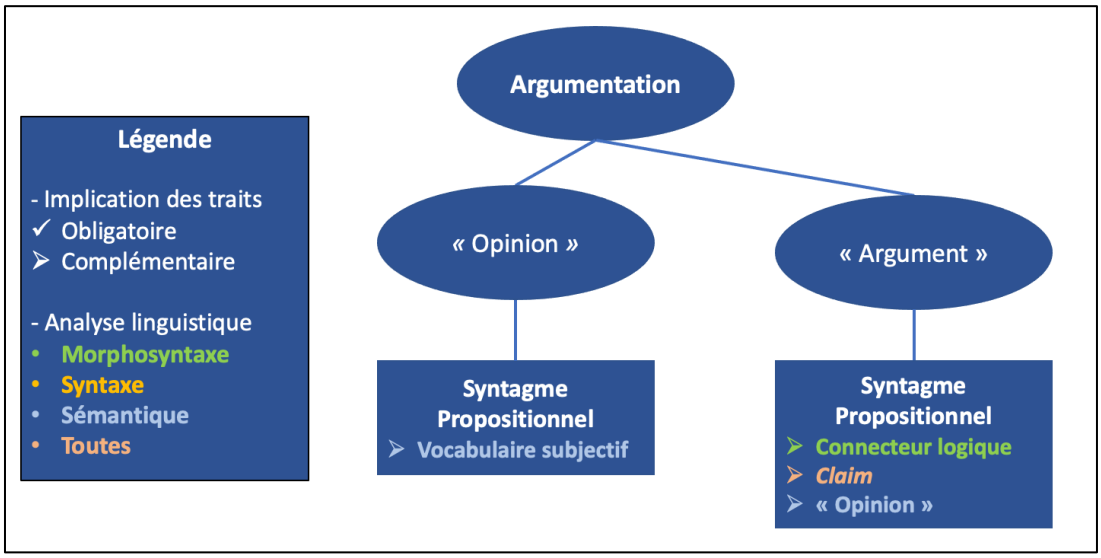


- *Argumentation*

L'argumentation est un type de discours qui consiste à soutenir la véracité d'un propos avec un ou plusieurs arguments articulés de façon logique (Eckle-Kohler et al., 2015). Nous parlons d'*opinion* pour le propos soutenu par l'argumentation car il n'est pas vrai a priori, justifiant la démarche de l'argumenter. Il exprime un point de vue de l'énonciateur, avec un vocabulaire caractéristique de la subjectivité. Les arguments peuvent être des *claims* ou des opinions, parfois reliés au propos défendu avec un connecteur logique (cf. Figure 13).

- Exemple : « *It appears that ISIS are the best diplomats on Earth since they work for Iran, America, Turkey, Saudi and Israel.* » ; il y a ici la présence d'une opinion (« *the best diplomats on Earth* »), liée à son explication (« *they work for…* ») par une conjonction de subordination (« *since* »).

Figure 13 : Modélisation linguistique de l'argumentation



Nous venons de décrire les traits discursifs que nous avons mis en exergue dans les messages des influenceurs. Ils donnent lieu à des réactions de l'audience. Nous avons aussi décrit ces réactions par des traits discursifs que nous présentons dans la section suivante.

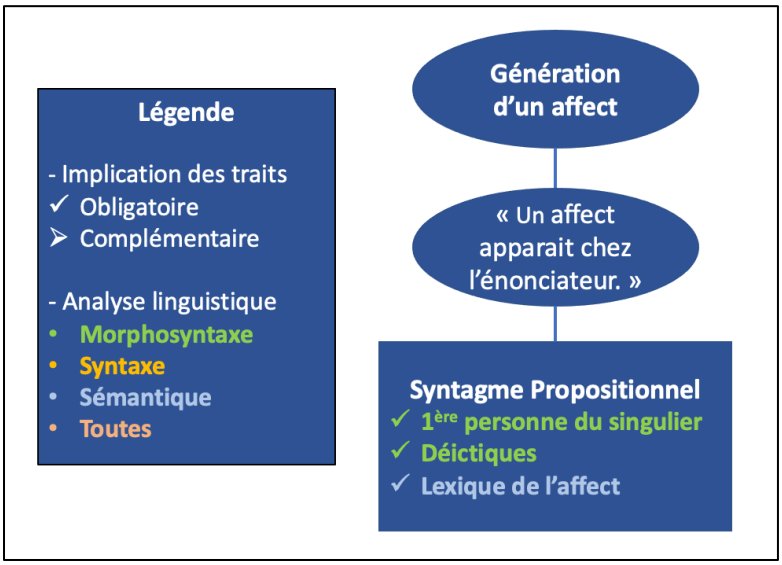
3.3.4. Les traits discursifs de la *stimulation* et de la *décision*

Dans une discussion *Change my view*, les participants désignent les messages qui les conduisent à changer d'avis avec un message contenant un symbole *delta* et l'explicitation de leur nouvelle position et ce qui l'a engendrée. Un tel message contient à la fois l'expression d'une *décision* et de sa *stimulation*. Nous avons analysé ces messages spécifiques et mis en exergue les traits discursifs présentés ci-dessous.

- ***Génération d'un affect***

Nous parlons de *génération d'un affect* pour toute réaction relatant l'expérience d'un sentiment ou d'une émotion par l'énonciateur, en lien avec la situation d'énonciation. Le discours correspondant est modélisé dans la Figure 14. Il se caractérise par un lexique désignant des émotions ou des sentiments et leur apparition. L'énonciateur emploie la première personne du singulier pour parler de lui-même, et possiblement un ou plusieurs déictiques pour désigner ce qui est à l'origine de sa réaction. L'influence de l'affect sur la prise de décision constitue un sujet de recherche (Binali et al., 2010).

Figure 14 : Modélisation linguistique de la *génération d'un affect*

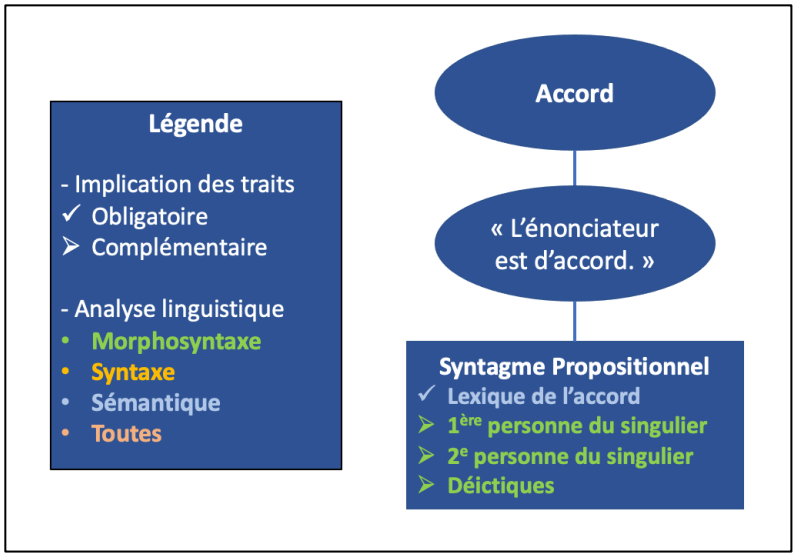


- Exemple : « *You gave me some hope for the oils* » ; l'énonciateur exprime bien l'apparition d'un nouveau sentiment chez lui (« *gave me some hope* »), en lien avec le destinataire de son message (« *You* »).

- *Accord*

L'expression d'un *accord* apparait dans un énoncé où l'énonciateur pose une équivalence entre son point de vue ou ses actes et le point de vue ou les actes d'autrui, à quelque degré que ce soit. L'accord est étudié, de façon similaire à l'affect, en lien avec la prise de décision des individus (Germesin & Wilson, 2009). Nous en présentons une caractérisation linguistique dans la Figure 15.

Figure 15 : Modélisation linguistique de l'accord



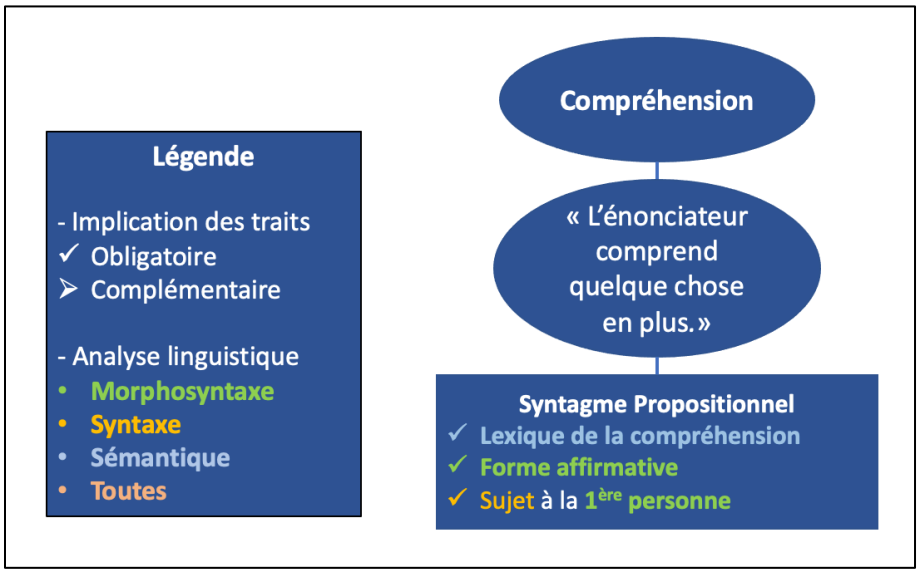
Le trait discursif de l'accord se caractérise par un lexique ayant trait à l'opinion et la similarité, par un emploi de la première personne du singulier pour l'énonciateur et de la deuxième personne du singulier pour le destinataire de la réaction. Il peut y avoir l'utilisation de déictiques afin de préciser ce qui, dans la situation d'énonciation, engendre l'accord.

- Exemple : « *I do agree that the left has similar issues* » ; le verbe « *agree* », appartenant évidemment au lexique de l'accord et utilisé à la première personne du singulier, exprime l'accord de l'énonciateur, conformément au modèle de la Figure 15.

- *Compréhension*

La *compréhension* se manifeste dans le discours d'un individu faisant état du raisonnement qu'il a réussi à produire grâce à un message. Ce type d'expression fait le lien avec la recherche en psychologie sociale qui considère le processus de compréhension d'un message comme un facteur important pour l'impact de la communication (Wyer & Shrum, 2015). Le discours de la compréhension se caractérise par un lexique spécifique au raisonnement, dans une forme affirmative, avec un sujet à la première personne du singulier (cf.
Figure 16).

Figure 16 : Modélisation linguistique de la compréhension



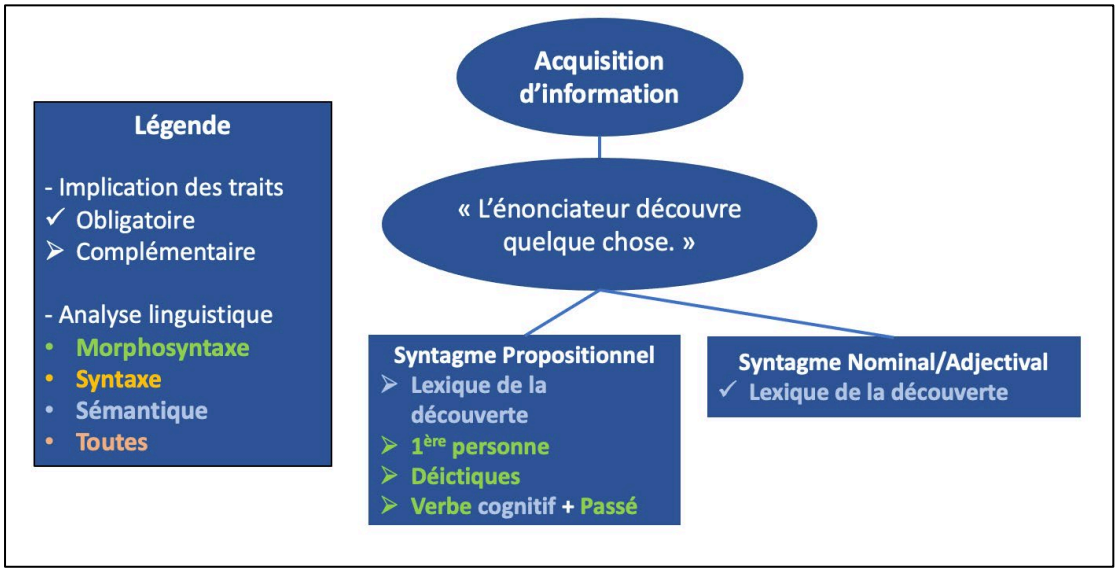
➢ Exemple : « *Yours was the first comment to make me understand how changing the definition would render the word useless* » ; il y a bien ici l'expression de l'acquisition intellectuelle d'un sujet par l'énonciateur (« *make me understand* »), en lien avec un message du destinataire (« *Yours was the first comment to* (…) »).

- Acquisition d'information

L'*acquisition d'information* apparait dans tout énoncé où l'énonciateur indique recevoir une information comme étant nouvelle, sans mettre en doute. Elle apparait aussi implicitement lorsqu'un énonciateur remet en cause ses propres connaissances ou des informations reçues auparavant, ce que nous considérons comme un effet de l'acquisition d'information. L'acquisition d'information correspond à une stimulation de l'intellect (Hidi & Baird, 1986).

Notre modélisation linguistique de ce type de discours (cf. Figure 17) tient pour beaucoup dans l'expression de la découverte, avec un lexique caractéristique ayant trait à la nouveauté. Dans le cas d'une proposition, il peut aussi y avoir un verbe ayant trait à la pensée (que nous appelons *verbe cognitif*), à la première personne du singulier et conjugué au passé, qui peut exprimer de la part de l'énonciateur une révision de ses connaissances. Des déictiques peuvent s'ajouter pour faire le lien entre l'expression de la découverte et les éléments de la situation d'énonciation ayant engendré cette découverte.

Figure 17 : Modélisation linguistique de l'*acquisition d'information*



- Exemple : « *I didn't know that NHS and Medicare are so different.* » ; il y a ici l'expression par l'énonciateur de la découverte d'une information.

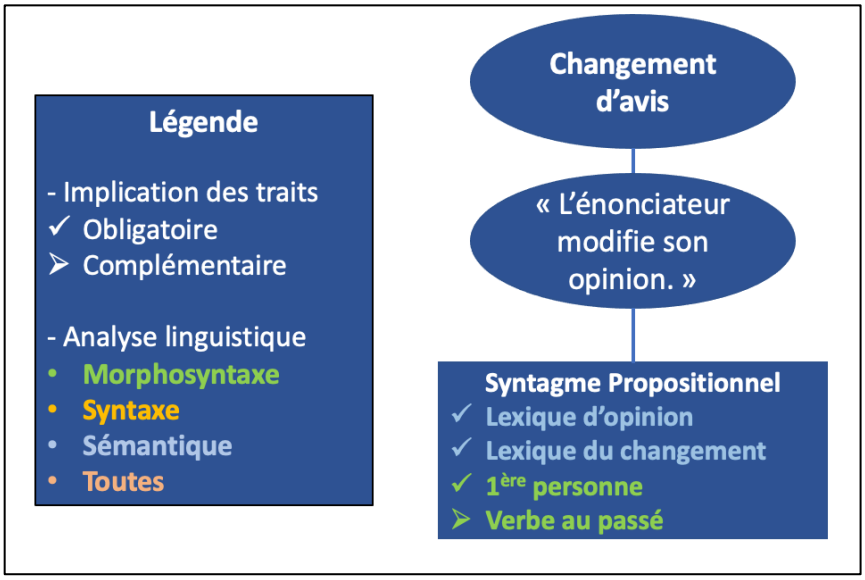
- *Décision : le changement d'avis*

Il n'y a qu'un type de discours pour la *décision* parce que, dans *Change my view*, la finalité d'une action d'influence est de faire changer d'avis. Nous identifions l'expression d'un changement d'avis à tout énoncé dans lequel l'énonciateur indique une remise en cause ou une évolution de son avis, quel que soit le degré de cette remise en cause ou de cette évolution. C'est aussi tout énoncé où l'énonciateur fait référence à son avis passé car cela dénote indirectement qu'il a changé d'avis.

Ce type de discours est important pour notre travail parce que le processus de formation (et de changement) de l'opinion est au cœur de la recherche sur l'influence sociale (Friedkin & Johnsen, 1990).

Nous caractérisons linguistiquement ce trait discursif (cf. Figure 18) par des lexiques spécifiques aux expressions d'opinion et de changement, par l'utilisation du passé, qui peut servir à revenir sur un avis antérieur, et par l'emploi de la première personne, spécifique à l'expression d'une opinion de l'énonciateur.

Figure 18 : Modélisation linguistique du *changement d'avis*



- Exemple : « *I suppose I could be worrying about something else* » ; cet énoncé est un exemple de *changement d'avis* parce qu'il exprime bien un changement de position avec un verbe d'opinion, « *suppose* », introduisant l'expression au passé « *I could be* ».

Nous avons présenté notre modélisation de l'action d'influence par l'analyse de la structure des liens interpersonnels, dans la section 3.2, et par l'analyse discursive entre des interlocuteurs, dans la section 3.3. Nous allons voir, dans la section suivante, comment nous avons combiné ces deux types d'analyse pour caractériser l'action des influenceurs.

3.4. L'influence globale par hybridation

3.4.1. Complémentarité de l'approche linguistique et de l'approche basée sur les graphes

Pour détecter l'influence globale des individus par leur intégration dans la structure des contacts sociaux, nous avons utilisé le concept de la centralité (cf. section 3.2). Cependant, les mesures de centralité ne disent rien sur la sémantique des liens auxquels

on les applique. C'est pourtant une caractéristique cruciale pour donner tout son sens à ce qu'on mesure et ainsi mieux déterminer ce que la centralité modélise réellement.

Les liens interpersonnels sont les unités d'analyse de l'influence globale dans les graphes sociaux. Or, dans notre modèle de l'influence (cf. section 3.1), nous définissons l'influence globale comme une composition d'influences individuelles. Nous choisissons alors de créer les liens de nos graphes d'influence à partir de l'analyse linguistique des messages, suivant notre modèle de l'influence individuelle (cf. section 3.3).

Notre hypothèse avec l'approche hybride est qu'utiliser les informations linguistiques, issues de notre modélisation de l'influence individuelle, à l'intérieur de la représentation en graphe des contacts sociaux, permettra d'obtenir une meilleure représentation de l'influence globale entre les individus et donc de mieux détecter les influenceurs. Il reste à modéliser la combinaison des informations linguistiques et des informations structurelles pour représenter l'influence globale dans les graphes sociaux.

3.4.2. Modèle hybride de l'influence globale

Nous enrichissons par hybridation notre modèle initial de l'influence globale, fondé sur la centralité des individus dans un réseau social (cf. section 3.2). Nous y associons une caractérisation sémantique des arcs du graphe qui correspond aux trois composantes de l'influence individuelle (cf. section 3.3) : le *stimulus*, la *stimulation* et la *décision*. Nous avons défini trois composantes pour le modèle hybride, elles correspondent aux trois phases de l'hybridation.

- Le graphe de *stimuli*

L'influence individuelle passe nécessairement par des stimuli circulant entre les individus (cf. section 3.3.2). Le graphe de *stimuli* représente cette circulation et, ainsi, la structure le long de laquelle des processus d'influence peuvent avoir lieu. Les arcs du graphe représentent la circulation des *stimuli* et les nœuds représentent les individus. Nous allons voir les deux propriétés essentielles du graphe de *stimuli*.

Le graphe de stimuli est un graphe orienté (cf. section 3.2.1), c'est-à-dire que ses arcs possèdent une direction d'un nœud vers un autre. Cela permet de distinguer, dans la transmission des *stimuli*, les individus sources des individus cibles : cette distinction sera utile pour identifier les individus qui sont à la source des processus d'influence (les influenceurs).

En plus d'être orienté, le graphe de *stimuli* est un graphe simple (cf. section 3.2.1), c'est-à-dire qu'il contient au maximum un arc orienté d'un nœud vers un autre. Chaque arc orienté d'un individu *Orig* vers un individu *Dest* signifie qu'il y a eu au moins un *stimulus* transmis de *Orig* à *Dest*. Il y a un arc orienté depuis un nœud représentant un individu

Orig vers un nœud représentant un individu *Dest* si un *stimulus* émis par l'individu *Orig* a été transmis à l'individu *Dest*. Nous considérons que cela se produit dans l'un des trois cas suivants :

- l'individu *Orig* a envoyé un message contenant un *stimulus* à l'individu *Dest*,
- l'individu *Dest* a répondu à un message de l'individu *Orig* qui contient un *stimulus*,
- l'individu *Dest* a répondu à l'individu *Orig* par un message contenant un trait discursif de *stimulation* ou de *décision* (cf. section 3.3.4).

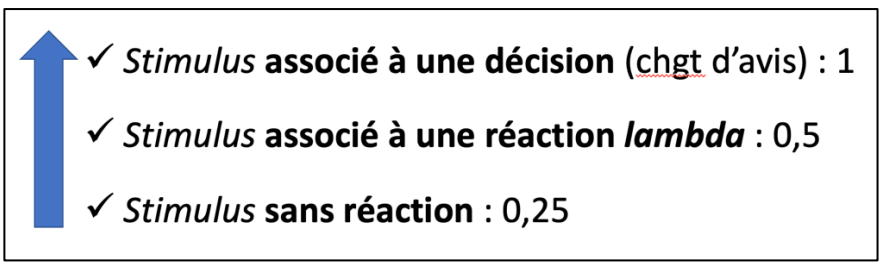
- La pondération du graphe de *stimuli*

Pour modéliser le flot d'influence individuelle dans le graphe de *stimuli*, nous pondérons chaque arc orienté du graphe par la probabilité qu'y circule de l'influence depuis l'individu représenté par le nœud source vers l'individu représenté par le nœud cible. Nous calculons la probabilité d'influence le long d'un arc *stimulus* en faisant une moyenne des poids des *stimuli* transmis le long de cet arc, comme décrit dans l'équation ci-dessous.

$$Poids\ (Arc(Orig, Dest)) = \frac{\sum Poids(stimuli(Orig, Dest))}{|Stimuli(Orig, Dest)|}$$

Le poids de chaque *stimulus* rend compte de la probabilité que ce dernier déclenche un phénomène d'influence individuelle depuis l'individu source vers l'individu cible. Nous considérons que cette probabilité est fonction de la réaction de l'individu cible, avec une échelle de valeurs ajustable, allant croissant entre 0 et 1, depuis l'absence de réaction jusqu'à la prise d'une décision, en passant par toute autre réaction (cf. Figure 19). L'hypothèse sous-jacente est l'identification des *stimuli* et *décisions* dans les messages.

Figure 19 : Échelle indicative de poids d'un *stimulus*



Par exemple, pour pondérer un arc orienté depuis un nœud *Orig* vers un nœud *Dest*, il faut recenser les *stimuli* transmis depuis l'individu représenté par le nœud *Orig* vers l'individu représenté par le nœud *Dest* et les pondérer selon les réactions qu'ils ont

engendrées chez ce dernier. Nous proposons un exemple ci-dessous, en utilisant notre échelle indicative de poids (cf. Figure 19).

- 1 *stimulus* sans réaction : poids de 0,25
- 1 *stimulus* ayant engendré un changement d'avis : poids de 1
- 2 *stimuli* ayant engendré des réactions sans changement d'avis : poids de 0,5

Le calcul du poids de l'arc se fera alors en appliquant la formule donnée précédemment, consistant en une moyenne des poids des *stimuli*, comme ci-dessous.

$$Poids(Arc(Orig, Dest) = \frac{(1 * 0{,}25 \ + 2 \ * 0{,}5 + 1 \ * 1 \ )}{4} = 0{,}56$$

- La centralité des individus représentés dans le graphe

Nous associons l'influence globale des individus à leur centralité dans le graphe de *stimuli* pondéré en influence individuelle. Cette centralité doit indiquer dans quelle mesure chaque individu propage de l'influence à travers le graphe. La propagation d'information à partir d'un nœud dans un graphe orienté se faisant par ses liens sortants (dirigés vers un autre nœud), la mesure de centralité utilisée doit donc se focaliser sur ce type de liens ; c'est le cas de la mesure par degré sortant (Shaw, 1954) et de *Hits relais* (Kleinberg, 1999).

Conformément à notre modèle de l'action des influenceurs dans un réseau social (cf. section 3.1), notre approche hybride considère l'influence globale de chaque individu d'après la place qu'il occupe dans un système d'influences individuelles. Dans la section suivante, nous allons présenter les jeux de données que nous avons utilisés pour vérifier chacune des approches de notre travail.

4. Jeux de données pour les expériences

Nous présentons dans ce chapitre les différents jeux de données que nous avons utilisés pour nos expériences. Leurs types variés ont permis de vérifier nos hypothèses et d'évaluer les composantes de notre modèle d'influence.

Le jeu d'utilisateurs Twitter (cf. section 4.1) nous a servi à évaluer la composante de l'influence globale modélisée par le concept de la centralité appliqué à des graphes sociaux (cf. section 3.2). Nous l'utilisons donc pour les expériences sur la détection structurelle des influenceurs (cf. section 5.1).

Le corpus issu du forum *Change my view* (cf. section 4.2) est celui que nous avons le plus utilisé. C'est en se basant sur une partie de ce corpus que nous avons modélisé l'influence individuelle (cf. section 3.3). Nous l'avons aussi utilisé pour la campagne d'annotation (cf. section 6.1) sur les traits discursifs d'influence (cf. sections 3.3.3 et 3.3.4) et pour réaliser les expériences sur la détection automatique de ces traits (cf. sections 6.2 et 6.3). Enfin, ce corpus nous a servi à évaluer le modèle hybride (cf. section 7).

Le corpus de tweets commentant des festivals culturels (cf. section 4.3) nous a uniquement servi à travailler sur la détection automatique du trait discursif de l'argumentation (cf. section 3.3.3), dans le cadre du Lab MC2 de la campagne d'évaluation CLEF 2018 (Deturck et al., 2018). Ce travail est en lien avec notre caractérisation de l'argumentation (cf. section 3.3.3) et nos expériences pour détecter automatiquement ce trait discursif des influenceurs (cf. section 6.2).

Nous avons travaillé avec le corpus de tweets d'individus étiquetés comme radicalisés (cf. section 4.4) avec la perspective d'une collaboration dans le cadre du projet de recherche européen Trivalent[20] (cf. section 8.3). Nous l'avons utilisé pour la campagne d'annotation (cf. section 6.1) sur les traits discursifs d'influence (cf. sections 3.3.3 et 3.3.4) ainsi que pour les expériences sur la détection automatique des traits *stimuli* (cf. section 6.2) et celles sur le modèle hybride (cf. section 7).

4.1. Utilisateurs Twitter

Twitter est à la fois un média conversationnel et un réseau social, c'est donc une ressource importante pour de nombreux travaux et campagnes d'évaluation portant sur l'analyse de l'influence. Nous présentons ici un jeu d'utilisateurs Twitter constitué dans le cadre de la campagne d'évaluation *RepLab 2014*[21] (Amigó et al., 2014).

[20] https://trivalent-project.eu
[21] http://nlp.uned.es/replab2014/

Une tâche de *RepLab 2014* consistait à classer les utilisateurs selon leur degré d'influence (Amigó et al., 2014). Les organisateurs ont défini l'influence comme la capacité à impacter les opinions, ce qui correspond bien à notre définition de l'influence individuelle (cf. section 3.1.1). La constitution de ce jeu de données est présentée en chiffres dans le Tableau 2. Nous allons la commenter.

Tableau 2 : Statistiques sur les utilisateurs

Catégories	*Influenceur*			*Non-influenceur*			Toutes		
Domaines / Partitions	*Entr.*	*Test*	*Toutes*	*Entr.*	*Test*	*Toutes*	*Entr.*	*Test*	*Toutes*
Automobile	379	765	1 144	808	1 589	2 397	1 186	2 353	3 539
Banque	419	713	1 132	896	1 795	2 691	1 314	2 507	3 821
Divers	N/A	88	88	N/A	44	44	N/A	131	131
Tous	796	1 563	2 359	1 704	3 428	5 132	2 500	4 991	7 491

Les utilisateurs Twitter ont été sélectionnés dans deux domaines que sont la banque et l'automobile, pour avoir au moins 1000 suiveurs et pour s'exprimer en anglais ou en castillan. Ils ont été manuellement annotés en tant qu'influenceur ou non-influenceur par des experts de la réputation en ligne, avec des critères comme les caractéristiques personnelles ou le nombre de suiveurs.

L'ensemble est divisé en jeux d'entrainement et de test. Le jeu de test contient en supplément une catégorie mélangeant des profils de divers domaines pour éprouver la robustesse des systèmes.

Le principal intérêt de ce jeu de données est de contenir des annotations en influenceurs qui correspondent à notre définition du concept. Un autre intérêt est qu'il a été utilisé dans le cadre d'une campagne d'évaluation dont les différents systèmes participants constituent autant de points de comparaison pour notre propre système. Enfin, il contient les identifiants Twitter des comptes, ce qui nous a permis d'extraire des informations supplémentaires avec l'API Twitter lors de nos expériences (cf. section 5.1).

Si nous avons retenu ce jeu de données pour l'analyse structurelle de l'influence, le jeu de données que nous présentons dans la section suivante est, quant à lui, destiné principalement à l'analyse linguistique.

4.2. Discussions issues du forum *Change my view*

Pour l'analyse linguistique de l'influence individuelle, nous utilisons un corpus extrait du forum *Change my view* (Tan et al., 2016). Rappelons le principe de ce forum, que nous avons déjà décrit en détails dans la section 3.3.1 : les participants à un fil de discussion doivent faire changer d'avis son auteur (l'auteur initial) et ce dernier doit désigner les participants et les messages qui l'ont fait changer d'avis. Les données contiennent ainsi une annotation *ad-hoc* correspondant à de l'influence individuelle.

Nous présentons dans le Tableau 3 les statistiques sur la constitution du jeu de données. Les fils de discussion ont été publiés de 2013 à 2015. Ils représentent un total de plus de 1 000 000 de messages pour environ 80 000 participants. Il y a une partition de développement (90% des données) et une partition d'évaluation (10% des données). Quelques filtrages ont été appliqués successivement au jeu de données initial pour arriver à celui que nous avons finalement utilisé.

Il y a d'abord un prétraitement effectué par les auteurs du jeu de données initial : il s'agissait d'appliquer une contrainte de participation minimale pour chaque fil de discussion (au moins dix participants et une réponse de l'auteur initial) pour s'assurer de leur pertinence (le corpus *préfiltré* dans le Tableau 3). Nous partons du corpus *préfiltré* et appliquons un nouveau filtrage.

Pour focaliser les données sur la caractérisation de l'expression du changement d'avis, nous avons opéré un nouveau filtrage du corpus pour ne conserver que les fils de discussion dans lesquels l'auteur initial change d'avis au moins une fois. Ces fils de discussion représentent 30% du corpus *préfiltré* (le corpus *Changement d'avis* dans le Tableau 3). Nous obtenons un ratio moyen de près d'une quinzaine de messages par fil de discussion.

Tableau 3 : Statistiques sur le corpus *Change my view*

Corpus / Caractéristiques	#fils de discussion pour l'entrainement	#messages pour l'entrainement	#fils de discussion pour l'évaluation	#messages pour l'évaluation
Initial	18 363	1 114 533	2 263	145 733
Préfiltré	10 743	128 901	1 529	20 883
Changement d'avis	3 191	42 776	672	9 463

4.3. Les festivals culturels en tweets

Ce corpus est issu du projet ANR GAFES[22] qui vise à valoriser le contenu publié par les participants à des festivals culturels. Nous l'avons utilisé dans le cadre de l'atelier CLEF MC2 2018 qui proposait notamment une tâche sur la détection d'argumentation dans les tweets de festivaliers (Cossu et al., 2018).

La détection d'argumentation dans des tweets est une tâche intéressante pour notre travail puisque l'argumentation fait partie des traits discursifs que nous avons identifiés comme *stimuli* d'influence (cf. section 3.3.3). Les participants devaient classer les tweets selon leur *argumentativité*, soit leur probabilité d'être des tweets argumentatifs. Le cas d'usage était d'aider les organisateurs de festivals à obtenir des avis constructifs.

Tableau 4 : Statistiques sur les tweets GAFES

Langue	Multilingue	Anglais	Anglais	Français	Français
Jeu de données	Jeu de données initial	Filtré sur l'anglais	Filtré sur l'anglais + les noms de festival	Filtré sur le français	Filtré sur le français + les noms de festival
Nombre de tweets	63M	34M	2M	3M	200k
Nombre d'auteurs différents	45M	9M	1M	1M	100k
Nombre de tokens[23]	960M	532M	25M	41M	3M
Nombre de lemmes différents[24]	N/A[25]	7M	61k	252k	15k
Valeur de subjectivité moyenne[26]	N/A[25]	0,28	0,28	0,26	0,15
Valeur de polarité moyenne[26]	N/A[25]	0.18	0,14	0,13	0,07

Dans le Tableau 4, nous présentons des statistiques sur l'ensemble du corpus selon des propriétés qui permettent de rendre compte de sa diversité. Ce jeu de données est une

22 http://anr-gafes.univ-avignon.fr
23 Obtained using Unix 'wc' command
24 Lemmatized using http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
25 Not designed to support all languages of the original dataset
26 Subjectivity and polarity magnitudes are obtained using https://github.com/sloria/textblob

collection de plus de 50 millions de tweets qui doivent tous contenir le mot-clé *festival* sans restriction de langue. Ils ont été extraits avec une API Twitter privée qui a permis de parcourir les archives de la période allant de mai 2015 à mai 2016.

Les tweets sont fournis avec des métadonnées : le nom de l'auteur, l'identifiant du tweet, le code ISO de la langue, le client utilisé pour l'envoi ainsi que la date de l'envoi. L'ensemble des données est enregistré dans des formats de fichier qui doivent faciliter leur indexation et les requêtes : CSV et XML. Les tweets originaux sont distingués des retweets.

Un premier atout de ce jeu de données est son volume important qui contient beaucoup d'exemples (positifs ou négatifs) pouvant servir de référence pour développer notre module de détection d'argumentation, et favorisant l'utilisation de méthodes par apprentissage.

Un deuxième atout de ce jeu de données est la faible spécification de son extraction qui, combinée au large volume extrait, offre un très large champ à l'analyse, que ce soit du point de vue de la langue ou des propos contenus dans les tweets.

Un troisième atout est la popularité de la thématique abordée (les festivals culturels). Elle promet une grande diversité de contenu et un engagement des auteurs qui induit une pertinence du contenu à analyser, en particulier concernant l'argumentation.

Le très grand volume initial des données nous a conduits à opérer deux phases de filtrage (cf. Tableau 4) afin de travailler avec un volume optimisé pour la tâche de MC2 et ainsi alléger les traitements informatiques.

La première phase de filtrage consiste à ne conserver que les messages dans les deux langues requises par la tâche : l'anglais et le français. La seconde phase de filtrage est destinée à ne garder que les messages comportant des noms de festivals spécifiques à chaque langue et imposés par les organisateurs de la tâche.

4.4. Tweets d'individus pro-État islamique

Dans le but de développer un contre-discours face à l'État islamique (EI), l'agence en stratégie du numérique *FifthTribe*[27] a extrait un ensemble de tweets identifiés comme provenant d'utilisateurs qui soutiennent l'organisation terroriste.

Nous supposons que cet ensemble de tweets exprime un engagement pour l'EI, avec notamment des tentatives d'influence en sa faveur. Pour déterminer la généricité des procédés discursifs que nous avons identifiés comme des *stimuli* d'influence dans le forum *Change my view,* il est intéressant de voir si nous arrivons à les retrouver dans un contexte d'application différent, comme celui du discours radical. C'est aussi l'occasion d'aborder un autre genre textuel que celui du forum.

[27] https://www.fifthtribe.com/

Le corpus contient plus de 17 000 tweets (cf. Tableau 5) extraits dans une période de trois mois après les attentats de Paris en janvier 2015. L'ensemble des tweets est accompagné de métadonnées : le nom de l'auteur, la date de publication et l'identifiant du tweet. Le jeu de données est mis à disposition publiquement sur *Kaggle*[28].

Tableau 5 : Statistiques sur les volumes du jeu de tweets de radicalisés

Nombre de tweets	*17 350*
Nombre d'auteurs	*111*

Nous allons voir quels critères ont été utilisés pour identifier des utilisateurs Twitter soutiens de l'EI et distinguer, en particulier, les d'utilisateurs qui soutiennent l'EI de ceux qui sont contre l'occident mais ne soutiennent pas l'EI.

Parmi les critères utilisés pour repérer des utilisateurs Twitter soutiens de l'EI, il y a la présence de certains mots-clés dans leurs noms, leur description de profil ou leurs tweets : par exemple *Dawla* (marque une révérence à l'EI) et *Baqiyyah* (dénote un militantisme pour l'expansion de l'EI). Les images publiées par les utilisateurs sont aussi utilisées (pour la présence du drapeau de l'EI ou de ses leaders) ainsi que les réseaux de suivi des utilisateurs.

Nous allons maintenant passer à la description des expériences réalisées à partir de ces jeux de données, en commençant par celles qui concernent la détection structurelle des influenceurs.

[28] https://www.kaggle.com/fifthtribe/how-isis-uses-twitter

5. Détection structurelle des influenceurs

5.1. Évaluation de mesures de centralité

5.1.1. Objectif

L'objectif de cette partie du travail est de déterminer si une analyse structurelle des actions interpersonnelles, selon différentes mesures de centralité, est pertinente pour détecter les influenceurs. Plus précisément, il s'agira d'évaluer comparativement une sélection de mesures de centralité dans leur capacité à détecter les influenceurs à partir des données d'un réseau social (Deturck et al., 2018).

L'analyse par centralité est une analyse de l'influence globale, définie comme une composition d'influences individuelles (cf. section 3.1.1), que nous avons supposé apparaitre dans la sémantique des actions interpersonnelles (cf. section 3.1.2).

Nous voulons étudier l'impact de la sémantique des actions interpersonnelles sur les performances de l'analyse structurelle pour la détection des influenceurs. Afin de simplifier ces expériences focalisées sur l'analyse de la structure des contacts sociaux, en évitant d'intégrer de l'analyse linguistique, il ne s'agira pour cette étude que d'actions interpersonnelles dont la sémantique est fermée (cf. section 3.1.2), c'est-à-dire préétablie par le média social considéré. Dans la section suivante, nous décrivons la méthodologie mise en œuvre afin de répondre à cet objectif.

5.1.2. Méthodologie

Nous avons choisi six mesures de centralité, la plupart parce qu'elles sont des classiques dans l'état de l'art sur l'analyse de graphes par centralité (cf. section 2.1) et l'une, *LeaderRank*, parce qu'elle est une adaptation aux réseaux sociaux de l'algorithme classique *PageRank*. Nous y ajoutons un classement aléatoire des utilisateurs pour mieux évaluer l'apport des mesures de centralité. Chacune des mesures possède sa conception particulière de la centralité, que nous présentons ci-dessous.

- ***Degré entrant*** (Freeman, 1978) : c'est une mesure de centralité *locale* parce qu'elle attribue un score aux nœuds d'un graphe orienté selon leur nombre respectif de liens entrants

- ***Intermédiarité*** (Freeman, 1978) : le score d'un nœud est relatif au nombre de fois où il est présent le long du chemin le plus court entre les autres nœuds du graphe

- ***Proximité*** (Freeman, 1978) : un nœud est d'autant plus central que sa distance par rapport aux autres nœuds du graphe est petite

- ***PageRank*** (Page et al., 1998) : calcule itérativement la centralité d'un nœud dans un graphe, selon la centralité de chaque nœud pointant vers lui, avec un paramètre intégrant une probabilité d'arrivée sur un nœud sans suivre aucun arc du graphe

- ***LeaderRank*** (Lü et al., 2011) : adapte l'algorithme *PageRank* à la diversité des réseaux sociaux en proposant un nouvel algorithme qui, au lieu de nécessiter le réglage a priori d'un paramètre concernant la probabilité d'arriver sur un nœud sans suivre aucun arc du graphe, calcule cette dernière *ad hoc* pour qu'elle soit inversement proportionnelle au nombre de nœuds ayant des arcs entrants

- ***Hits*** (Kleinberg, 1999) : mesure simultanément une centralité d'*autorité* et une centralité de *relais* ; la centralité de relais d'un nœud est d'autant plus grande qu'il possède des liens sortants vers des nœuds ayant une grande centralité d'autorité, la centralité d'autorité d'un nœud est d'autant plus grande qu'il possède des liens entrants provenant de nœuds ayant une grande centralité de relais

Ces mesures de centralité attribuent des scores de centralité potentiellement différents à chaque nœud d'un graphe, produisant des classements d'utilisateurs aussi différents, permettant d'évaluer comparativement les mesures. Afin d'évaluer les mesures de centralité, nous avons choisi des données de référence.

Pour cette étude, nous avons choisi comme référence le jeu d'utilisateurs Twitter que nous avons décrit dans la section 4.1. Avec ce jeu de données comme référence, nous allons évaluer la capacité des mesures de centralité à détecter l'influence d'individus en se basant sur la structure d'actions interpersonnelles dans Twitter.

Le média social Twitter est un atout pour une telle évaluation. Il propose une variété d'actions interpersonnelles à la sémantique fermée, par exemple *Aimer*, *Retweeter* et *Suivre*. Nous allons ainsi pouvoir faire varier, durant nos expériences, le type des actions interpersonnelles représentées, ce qui permettra d'évaluer l'impact de leur sémantique sur les performances de la détection structurelle des influenceurs.

À partir d'un sous-ensemble du jeu d'utilisateurs Twitter, nous avons extrait des actions interpersonnelles de différents types. Nous avons choisi un sous-ensemble des

utilisateurs pour limiter le temps d'extraction et la complexité des analyses. Les actions interpersonnelles extraites impliquent les utilisateurs initiaux et d'autres utilisateurs. Nous allons expliquer comment nous avons constitué le sous-ensemble d'utilisateurs.

Nous avons sélectionné un sous-ensemble de cinquante utilisateurs appartenant tous à un même domaine pris au hasard, la banque. Nous supposons que l'unicité du domaine permettra d'obtenir des graphes plus denses (cf. section 3.2.1).

Nous avons fait en sorte que notre échantillon d'individus contienne la même proportion d'influenceurs que le jeu de données original, soit 1/3 d'influenceurs, afin de conserver une certaine comparabilité avec les systèmes ayant participé à la compétition *RepLab*, pour laquelle ce jeu de données a été constitué (cf. section 4.1). À partir de ce sous-ensemble d'utilisateurs, nous avons extrait certains types d'actions interpersonnelles Twitter dont nous allons décrire la sélection.

Pour les expériences, un premier objectif est de sélectionner les types d'actions interpersonnelles qui soient les plus révélateurs d'une influence entre les utilisateurs qui ont produit l'action et ceux qui sont l'objet de l'action ou qui ont produit le contenu sur lequel porte l'action. Un second objectif est d'évaluer comparativement la pertinence des actions portant sur le contenu et des actions portant sur les utilisateurs pour la détection des influenceurs.

Nous évaluons le niveau d'engagement induit par les différents types d'actions interpersonnelles de façon à en sélectionner deux, l'un à la portée directe, dont l'objet est directement un utilisateur, et l'autre à la portée indirecte, dont l'objet est un tweet d'un utilisateur que nous considérons comme l'*utilisateur sujet* de l'action (cf. section 3.1.2).

Dans le Tableau 6, nous décrivons les différents types d'actions interpersonnelles Twitter qui sont publiques (accessibles à notre analyse), avec en gras ceux que nous avons sélectionnés pour les expériences.

Tableau 6 : Description des types d'actions interpersonnelles et publiques sur Twitter

Nom	*Nature de l'objet*	*Portée*	*Effet principal*
Retweeter	Tweet	Indirecte	Envoi du tweet aux suiveurs
Aimer	Tweet	Indirecte	Distinction publique d'un tweet
Répondre	Tweet	Indirecte	Tweet en réponse à un tweet
Mentionner	Utilisateur	Directe	Tweet adressé à un utilisateur
Suivre	Utilisateur	Directe	Abonnement à un utilisateur

Suivre est le type d'action portant sur un utilisateur qui marque l'engagement le plus important parce que son effet est durable (une relation avec l'utilisateur suivi), contrairement à *Mentionner* qui a un effet ponctuel (une notification de l'utilisateur mentionné) ; nous choisissons donc *suivre* pour le type d'action à la portée directe. Nous voulons comparer ce type d'action avec un autre type d'action qui porte sur un contenu, donc à la portée indirecte.

Répondre nécessite une analyse linguistique pour accéder à sa sémantique, ce qui est hors du champ de l'étude, ainsi, nous l'éliminons des types candidats. *Retweeter* et *Aimer* indiquent tous les deux l'intérêt d'un utilisateur pour le contenu cible avec un engagement que nous considérons plus fort pour *Retweeter* parce que cela revient à s'approprier le contenu cible quand *Aimer* consiste seulement à le distinguer publiquement. Nous retenons donc *Retweeter* comme second type d'action interpersonnelle.

Nous avons extrait toutes les relations de suivi associées à notre échantillon d'utilisateurs. Pour les retweets, afin de limiter le temps consacré à l'extraction des données, nous avons choisi de ne considérer que les dix derniers tweets de chaque utilisateur et d'extraire les retweets associés, dans la limite imposée par l'API de 100 par tweet. Il nous reste à choisir comment représenter dans des graphes les actions interpersonnelles extraites.

Nous construisons un graphe pour chacun des deux types d'action interpersonnelle sélectionnés (*Suivre* et *Retweeter*). Dans chaque graphe, les nœuds représentent les individus et les arcs sont orientés de façon à représenter l'existence d'au moins une action interpersonnelle du type considéré depuis un individu vers un autre.

Pour évaluer la capacité des mesures de centralité à détecter les influenceurs à partir des graphes résultants, il nous a fallu choisir une mesure d'évaluation qui compare les classements d'individus produits par les mesures de centralité aux annotations de référence.

Dans l'optique de comparer les performances de notre approche à celles des approches de la compétition *RepLab* 2014, nous avons choisi d'utiliser la même mesure d'évaluation : *Mean Average Precision* (MAP). Cette mesure permet de comparer les résultats des systèmes sous la forme d'un classement à l'annotation de référence qui est binaire.

La mesure MAP produit des valeurs comprises entre 0 et 1, 0 étant la pire valeur et 1 la meilleure. Nous en présentons la formule ci-dessous, avec n le nombre d'influenceurs, N le nombre total d'entités considérées, $p(i)$ la précision au rang i (en considérant les i premières entités) et $R(i) = 1$ si l'entité au rang i est pertinente et $R(i) = 0$ si l'entité au rang i n'est pas pertinente.

$$MAP = \frac{1}{n}\sum_{i=1}^{N} p(i)\, R(i)$$

Cette mesure d'évaluation classique du domaine de la recherche d'information, considère qu'un classement est d'une qualité d'autant plus grande qu'il place en premier les résultats pertinents. Pour notre étude, nous allons évaluer les mesures de centralité à travers la qualité des classements d'individus qu'elles produisent, en considérant que les plus influents doivent être les mieux classés.

Dans la section suivante, nous allons décrire les graphes que nous avons construits à partir des actions interpersonnelles extraites, puis, nous présenterons les résultats de l'évaluation des mesures de centralité sur ces graphes.

5.1.3. Résultats

Nous présentons des statistiques sur les graphes de suivi et de retweets respectivement dans le Tableau 7 et le Tableau 8.

Tableau 7 : Caractéristiques du graphe de suivi

Nombre de nœuds	5,067,480
Nombre d'arcs	5,149,491
Densité	10^{-7}

Tableau 8 : Caractéristiques du graphe de retweets

Nombre de nœuds	2099
Nombre d'arcs	2051
Densité	10^{-4}

Le nombre de nœuds dans chacun des deux graphes est largement supérieur au nombre des cinquante utilisateurs initiaux. Cela indique que nous avons extrait beaucoup d'actions interpersonnelles ayant eu cours avec des nouveaux utilisateurs plutôt qu'entre les utilisateurs de l'ensemble initial. Ce résultat est confirmé par un autre point commun aux deux graphes : leur densité.

Nous observons que les graphes obtenus ont une faible densité. Cela signifie que les utilisateurs du corpus initial (à partir desquels nous avons extrait les actions interpersonnelles) se suivent et se retweetent peu. Cette information relativise les résultats des mesures de centralité sur ces graphes puisque celles-ci se basent précisément sur la présence de liens entre les nœuds pour faire leurs calculs.

Le nombre de nœuds et d'arcs est beaucoup plus important dans le graphe de suivi que dans le graphe de retweets ; rappelons que nous avons limité l'extraction des retweets et pas celle des suiveurs. Cette différence de dimensions est à prendre en compte pour la comparaison des résultats entre les deux types de graphe.

Nous présentons l'évaluation d'une partie des mesures de centralité sélectionnées pour la détection des influenceurs sur le graphe de suivi dans le Tableau 9 et sur le graphe de retweets dans le Tableau 10.

Tableau 9 : Les résultats d'évaluation sur le graphe de suivi

Centralité	*Aléatoire*	*Degré entrant*	*PageRank*	*LeaderRank*	*Hits*
MAP (%)	**38,67**	**43,49**	**44,25**	**44,53**	**51,68**

Tableau 10 : Les résultats d'évaluation sur le graphe de retweets

Centralité	*Aléatoire*	*Degré entrant*	*PageRank*	*LeaderRank*	*Hits*
MAP (%)	**38,67**	**40,91**	**40,91**	**40,91**	**40,91**

Nous ne présentons pas de résultats pour les mesures de centralité par proximité et par intermédiarité car elles calculent les chemins les plus courts entre tous les nœuds du graphe, ce qui requiert un graphe fortement connexe (cf. section 3.2.1), une propriété que n'ont pas les graphes obtenus.

L'ensemble des mesures de centralité produisent des meilleurs résultats que le classement aléatoire des individus. Cela signifie que les graphes de suivi et de retweets sont porteurs d'informations que l'approche par centralité parvient à exploiter pour détecter les influenceurs.

La comparaison avec les systèmes de la compétition *RepLab* 2014 (Amigó et al., 2014) sur le domaine de la banque est plutôt favorable à notre approche, avec des résultats MAP généralement un peu meilleurs. Il faut tout de même nuancer ce constat en rappelant que nos expériences ne se font que sur un échantillon des données de référence qui ont été utilisées par les systèmes de la compétition.

Les résultats de la détection des influenceurs avec les mesures de centralité sont cependant moyens ou mauvais à travers l'ensemble des configurations expérimentales. Une seule configuration (*Hits* sur le graphe de suivi) dépasse une valeur MAP de 50%.

Le faible score de la mesure par degré entrant, qui ne prend en compte que le nombre de liens entrants pour calculer la centralité d'un nœud, indique que le seul nombre de suiveurs et de retweets n'est pas suffisant pour détecter les influenceurs.

Il y a une similitude entre les résultats de la mesure de centralité par degré entrant et ceux de la plupart des autres mesures de centralité, qui valorisent les nœuds ayant des liens provenant de nœuds plus centraux. Cela signifie que les influenceurs de notre référence ne sont pas particulièrement suivis ou retweetés par des individus qui le sont particulièrement eux-mêmes.

Hits obtient le meilleur résultat sur le graphe de suivi mais il ne se distingue pas des autres mesures de centralité sur le graphe de retweets (cf. Tableau 10). Cela montre peut-être que, dans notre cas, l'information sur les suiveurs est plus significative que l'information sur les retweets pour détecter les influenceurs.

5.1.4. Conclusion

Les résultats des mesures de centralité pour la détection d'influenceurs sont prometteurs. Leurs résultats sont globalement meilleurs qu'une approche aléatoire et la comparaison avec les systèmes de la compétition *RepLab* 2014, à relativiser à cause de

notre échantillonnage des données, leur est favorable. La mesure *Hits*, qui a la particularité de distinguer deux types de centralité, s'est montrée bien plus performante que les autres mesures sur le graphe de suivi uniquement. Le suivi est peut-être plus significatif que le retweet.

Il reste que les résultats sont généralement faibles : une seule configuration a une valeur MAP au-dessus de 50%. Les faibles résultats de la mesure par degré entrant sur les graphes de suivi et de retweet peuvent signifier que les données sur la quantité de suiveurs ou de retweets ne sont pas suffisantes. Il serait alors intéressant d'intégrer des informations sur le contenu des messages dans la représentation en graphe.

5.2. Étude d'impact de la dimension des graphes sur les mesures de centralité

5.2.1. Objectifs et méthodologie

Nous allons analyser les performances d'une sélection de mesures de centralité suivant la taille du graphe utilisé et donc la quantité d'actions interpersonnelles qu'il représente. Par rapport à l'expérimentation précédente et forts de ses résultats (cf. section 5.1), nous nous focalisons sur l'action interpersonnelle *Suivre* qui nous est apparue comme la plus pertinente pour détecter les influenceurs et sur les mesures de centralité *PageRank*, *LeaderRank* et *Hits*, qui n'ont pas besoin d'un graphe connexe pour calculer un résultat.

Nous utilisons, comme dans l'expérimentation précédente, le jeu d'utilisateurs Twitter (cf. section 4.1) en créant cette fois-ci des échantillons de 50 utilisateurs correspondant à différents intervalles de nombre de suiveurs (cf. Tableau 11). Nous faisons en sorte que chaque échantillon comprenne la même proportion d'influenceurs, celle du jeu de données d'origine : 1/3 d'influenceurs. Nous construisons un graphe de suivi pour chacun des cinq segments et nous évaluons les mesures de centralité avec la mesure MAP.

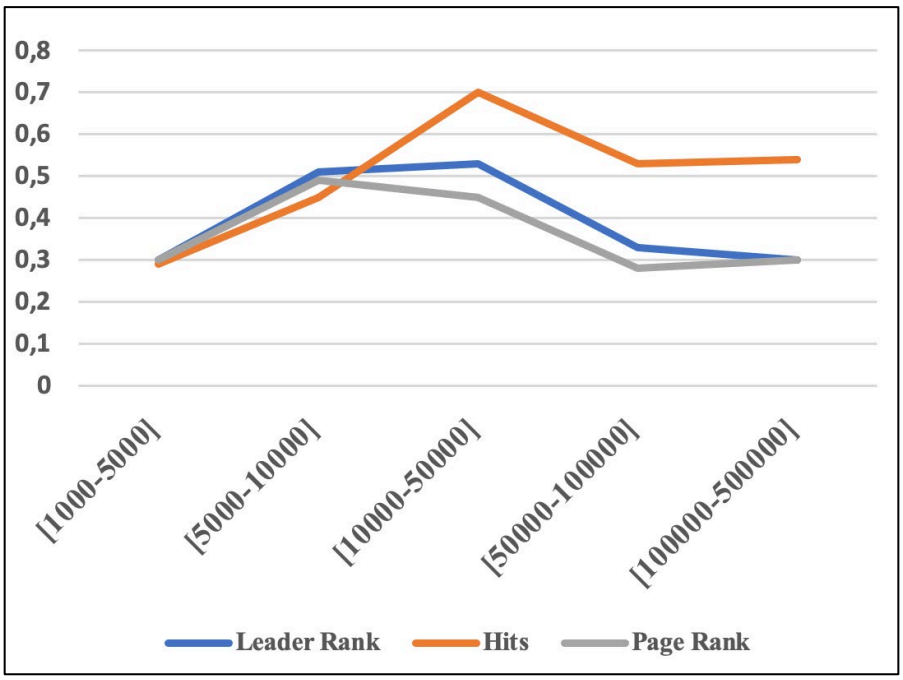
Tableau 11 : Statistiques sur les graphes de suivi construits

Segments/Caractéristiques	**[1000-5000]**	**[5000-10,000]**	**[10,000-50,000]**	**[50,000-100,000]**	**[100,000-500,000]**
Nombre de nœuds	122,060	361,422	1,000,180	3,156,671	9,708,482
Nombre d'arcs	124,747	372,939	1,034,110	3,322,007	10,521,92
Densité	10^{-6}	10^{-6}	10^{-6}	10^{-7}	10^{-7}

5.2.2. Résultats

Avec la Figure 20, nous représentons les résultats de l'évaluation avec la mesure MAP pour chacune des trois mesures de centralité sur les différents graphes de suivi que nous avons construits. Chaque graphe correspond à un intervalle de nombre de suiveurs des utilisateurs représentés.

Figure 20 : Résultats de MAP par segment



La performance des trois mesures augmente (jusqu'à un certain point) avec la taille des segments. Or, plus l'intervalle du nombre de suiveurs est grand, plus l'écart possible entre les utilisateurs augmente, ce qui accroît le pouvoir discriminant de cette information pour distinguer les influenceurs. Cela peut signifier que l'information sur le nombre de suiveurs aide effectivement à distinguer les influenceurs.

Il y a une baisse globale des performances entre le segment [10,000-50,000] et le segment [50,000-100,000]. Notre hypothèse est que le taille du graphe, beaucoup plus importante avec le second segment (des millions de nœuds et d'arcs en plus), a un impact sur les performances de toutes les mesures. Hormis ces tendances globales, nous remarquons quelques différences entre les trois mesures de centralité.

Nous constatons que deux des mesures ont des performances plus stables : *PageRank* et *LeaderRank*. Cette similitude peut s'expliquer par le fait que la seconde s'inspire de la première. Ces deux mesures ont la particularité de se détacher quelque peu de la structure

du graphe en intégrant de l'aléatoire dans leur calcul respectif. Cela peut expliquer pourquoi leurs résultats sont moins sensibles aux variations de la taille du graphe.

Pour *Hits*, les performances sont plus variables (50% de progression entre [1000-5000] et [10,000-50,000] et 15% de baisse entre [10,000-50,000] et [100,000-500,000]). Contrairement aux deux algorithmes précédents, *Hits* suit strictement la structure du graphe : il n'intègre pas de *saut* d'un nœud à l'autre ni de dilution du score d'un nœud à travers ses liens sortants. Il est donc plus sensible à la modification de la taille du graphe.

Le fait que l'augmentation de la marge du nombre de suiveurs entre les utilisateurs aide l'algorithme qui se base le plus sur la structure indique aussi la significativité de l'information de suivi.

La Figure 20 montre que, lorsque la taille des segments augmente, *LeaderRank* fait mieux que *PageRank*. Cela peut s'expliquer par le fait que *LeaderRank* utilise la structure du graphe pour réguler la part d'aléatoire dans son parcours du graphe, tandis que *PageRank* utilise uniquement un paramètre numérique en entrée.

6. Détection linguistique des influenceurs

Un système de détection des influenceurs, fondé sur notre modèle de l'influence individuelle (cf. section 3.3.2), doit être capable d'en repérer automatiquement les traits discursifs présentés dans les sections 3.3.3 et 3.3.4.

Dans cette partie de la thèse, nous présentons l'implémentation et l'évaluation des modules destinés à la détection automatique des traits *stimuli* (cf. section 6.2) et du trait de *décision* (cf. section 6.3). Nous n'avons pas eu le temps de faire des expériences sur les traits de *stimulation*. Ce dernier point ne nous a toutefois pas empêché d'évaluer la détection d'influenceurs par hybridation (cf. section 7).

Les difficultés principales sont de formaliser les caractéristiques linguistiques de chaque trait discursif et de les détecter automatiquement avec des outils de TAL. Dans la section suivante, nous présentons la campagne d'annotation des traits d'influence.

6.1. Campagne d'annotation

6.1.1. Principes généraux

Objectif

Notre modèle de l'influence individuelle repose sur trois composantes (cf. section 3.3.2) que nous avons caractérisées linguistiquement par différents traits discursifs (cf. sections 3.3.3 et 3.3.4). Manquant de données de référence sur ces traits pour le développement et l'évaluation d'un système basé sur ce modèle, nous avons organisé une campagne d'annotation afin de construire un corpus de référence.

Tâche d'annotation

Il existe différents types de tâche d'annotation. Notre tâche correspond au type *unitizing* (Krippendorff, 1995). Une annotation de type *unitizing* consiste à extraire, dans le *continuum* d'un texte, des unités ou segments et à les catégoriser. Dans notre cas, il s'agit d'identifier, dans des messages de médias sociaux, les segments de texte qui correspondent à l'un des traits discursifs de l'influence individuelle (cf. sections 3.3.3 et 3.3.4). C'est particulièrement difficile parce que les annotateurs doivent identifier à la fois les frontières de texte pertinentes et la catégorie correspondante.

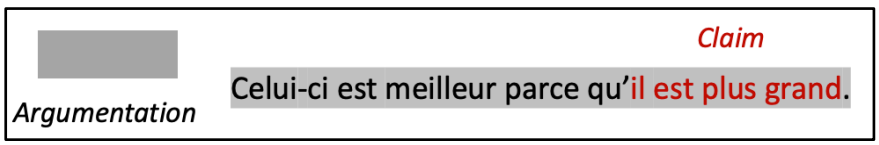
Évaluation de la qualité des annotations

Évaluer la qualité des annotations produites consiste à s'assurer qu'elles correspondent aux données attendues afin de les exploiter dans un corpus de référence. L'indice de qualité habituellement utilisé est l'accord inter-annotateurs (Artstein, 2017). Il vise à évaluer la fiabilité du travail d'annotation, c'est-à-dire dans quelle mesure les annotations qui en découlent correspondent a priori au schéma d'annotation prévu, sur la base du critère théorique de la reproductibilité (Krippendorff, 2004) : une annotation est d'autant plus fiable qu'elle est reproduite par plusieurs annotateurs.

Choix d'une mesure d'accord inter-annotateurs

Il existe différentes mesures d'accord inter-annotateurs selon le type de tâche d'annotation à évaluer. Pour une annotation du type *unitizing* , la littérature propose, à notre connaissance, deux références : une variante du coefficient Alpha de Krippendorff pour l'*unitizing* (Krippendorff, 1995) et la famille de mesures Gamma (Mathet et al., 2015). La mesure de Krippendorff ne permet pas de mesurer un accord quand des annotations de catégories différentes se chevauchent (cf. Figure 21), c'est pour cela que nous avons choisi d'utiliser les mesures Gamma.

Figure 21 : Illustration du chevauchement entre catégories



Mesures d'accord inter-annotateurs Gamma

Nous utiliserons les trois mesures de la famille Gamma : *Gamma*, *GammaCat* et *GammaK*. *Gamma* (Mathet et al., 2015) compare les annotations selon leurs étiquettes et leurs bornes tandis que *GammaCat* (Mathet, 2017) se focalise sur les étiquettes et donc sur la distinction des catégories. *GammaK* détaille *GammaCat* pour chacune des catégories.

Le guide d'annotation

Pour cadrer le travail d'annotation, nous avons rédigé un guide (Deturck, 2021). Il sert à contextualiser et expliquer la tâche d'annotation. Il définit les différentes catégories à annoter, avec des exemples pour aider leur compréhension par les annotateurs. Ce travail de rédaction est particulièrement difficile parce que les définitions doivent être à la fois précises et accessibles. Le guide a été amélioré tout au long des sessions d'annotation, et ce, en collaboration avec les annotateurs.

Typologie des erreurs d'annotation

Pour améliorer le guide d'annotation, nous avons réalisé une typologie des erreurs parmi les divergences entre les annotateurs ; il s'agit de mettre en exergue ce qui ne correspond pas au schéma d'annotation. Nous présentons ci-dessous les types d'erreurs que nous avons relevés, en les désignant tels qu'ils le seront dans les tableaux statistiques de chaque séance.

- ***Troncature*** : l'annotation ne contient qu'une partie du segment textuel attendu
  - Exemple : [[*He came yesterday evening*]Annotation *and talked to the president.*]Segment attendu

- ***Bruit*** : l'annotation contient le segment de texte attendu mais aussi du texte superflu dont le contenu ne fait pas partie des exclusions dans la définition du trait discursif annoté
  - Exemple : [*Let's go to the next step.* [*He came yesterday evening and talked to the president.*]Segment attendu]Annotation

- ***Nom d'une catégorie*** : confusion avec la catégorie désignée
  - Exemple : [*He came yesterday evening.*]Pédagogie annotée (*Claim* attendu)

- ***Invalidité*** : annotation ne correspondant à aucune catégorie ou comportant du contenu qui correspond à une catégorie mais aussi du contenu exclu par la définition de cette dernière
  - Exemple : [[*He came yesterday evening.*]Contenu attendu [*It was fun!*]Contenu exclu]*Claim* annoté

Nous distinguons aussi deux types d'annotations divergentes : celles qui se chevauchent, que nous appelons ***similaires*** (cf. Figure 22) et celles qui sont disjointes, que nous appelons ***différentes*** (cf. Figure 23). L'hypothèse sous-jacente est qu'il y a des erreurs spécifiques à chacun de ces deux types.

Figure 22 : Illustration des annotations *similaires*

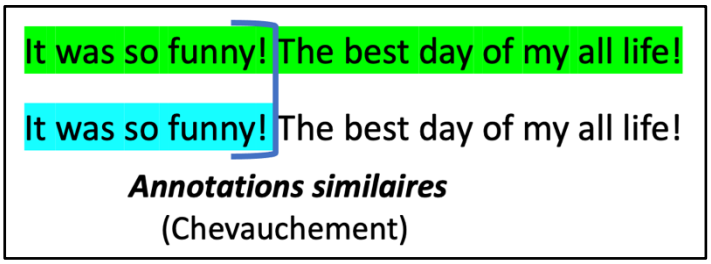
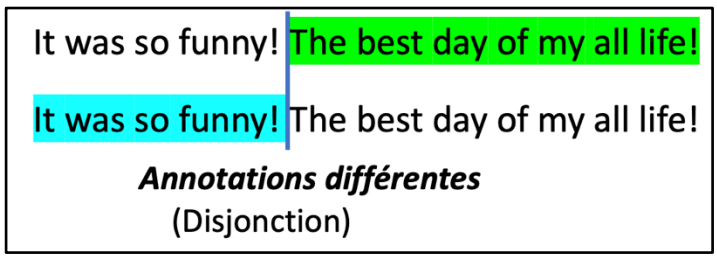


Figure 23 : Illustration des annotations *différentes*



Choix des données

Pour favoriser la variété de notre corpus de référence, nous l'avons constitué à partir de deux jeux de données qui correspondent à deux genres textuels complémentaires : le jeu de discussions du forum *Change my view* (cf. section 4.2), dans lequel les auteurs doivent développer leurs propos, et le corpus de tweets d'individus pro-État islamique (cf. section 4.4), dont les messages sont contraints à un nombre limité de caractères.

Partition des données

Nous avons partitionné les données à annoter pour les répartir entre annotateurs et ainsi maximiser la quantité de messages annotés lors d'une session. Afin de simplifier l'annotation et ainsi favoriser sa qualité, nous avons fait en sorte que chaque partition ne contienne qu'un seul genre de messages (*Change my view* ou Twitter). Dans la description des sessions d'annotation, nous désignerons chaque jeu de données par son genre textuel et un chiffre identifiant, par exemple *Tweet 3*.

Taille des jeux de données

Nous avons dimensionné chaque jeu de données de façon à ce qu'il puisse être traité par un seul annotateur en deux heures maximum, ce qui correspond à la durée minimale des sessions d'annotation que nous avons organisées. Nous avons estimé empiriquement le temps d'annotation d'un seul message d'après son genre textuel (un tweet ou un message de forum) : 45 secondes pour un tweet et 80 secondes pour un message de forum, les tweets tendant à être plus succincts que les messages de forum. Cela nous a conduits à créer des jeux de tweets contenant 100 messages et des jeux de messages de forum contenant 80 messages.

Logiciel d'annotation

Nous avons utilisé le logiciel *Gate* (Cunningham et al., 2013) comme outil d'annotation. Il propose un ensemble de modules pour l'analyse de texte, et notamment pour l'annotation, avec une interface graphique pour faciliter le travail des annotateurs.

Constitution de groupes d'annotateurs

Mesurer la qualité des annotations requiert de pouvoir calculer un accord inter-annotateurs. Il nous fallait donc prévoir des données identiques pour différents annotateurs. Nous avons constitué des groupes d'annotateurs de sorte que tous les membres d'un groupe traitent individuellement un même jeu de données. Afin de maximiser le nombre de groupes et donc la quantité de jeux de données traités, nous avons constitué les plus petits groupes possibles, avec deux voire trois annotateurs lorsque leur nombre était impair.

Protocole d'une séance d'annotation

Nous avons intégré, au début de chaque session d'annotation, une phase d'échanges avec les annotateurs sur le guide afin de régler les incompréhensions éventuelles. À la fin de chaque session d'annotation, chaque groupe discute en interne de ses divergences lors d'une phase dite de *réconciliation*. Les solutions trouvées ou les conflits résiduels sont mis en commun pour créer une dynamique de progression entre les annotateurs et améliorer le guide d'annotation. Une séance d'annotation classique adopte la structure présentée ci-dessous.

1. Mise au point sur le guide d'annotation
2. Annotation en binômes voire trinômes
3. Phase de réconciliation interne
4. Discussion en commun des cas problématiques
5. Mise à jour du guide d'annotation

6.1.2. Séance d'annotation n°1 « Montpellier »

Contexte

Nous présentons la configuration de la première session d'annotation dans le Tableau 12. Pour cette première séance, nous n'avons fait annoter que le *claim*. Les annotateurs sont quatorze étudiants en Master 1 *Intelligence artificielle* à l'université Montpellier 3. Nous avons organisé cette séance dans le cadre d'un TD de 2h dédié à l'annotation de données pour l'apprentissage automatique. Les annotateurs n'ont pas tous eu le temps de prendre connaissance du guide d'annotation avant la séance, ce qui nous a conduits à présenter oralement, en introduction, le contexte et la tâche.

Tableau 12 : Configuration de la session d'annotation n°1

Catégories à annoter	*Claim*
Nombre d'annotateurs	14
Groupes	7 binômes
Jeux de données	5 *Twitter*, 2 *Forum*

Annotations réalisées

Le Tableau 13 montre que les jeux de messages du forum *Change my view* comportent globalement plus d'annotations que les jeux de tweets, ce que nous expliquons par le fait que ce forum favorise les messages développés et donc relativement longs. Avec le Tableau 14, on peut constater que l'écart de quantité d'annotations produites entre les annotateurs d'un même groupe est globalement inférieur à 50% des annotations ; ce résultat est positif quant à l'intersubjectivité présente dans les annotations.

Tableau 13 : Nombre d'entités annotées

Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Tweet 5*	*Forum 1*	*Forum 2*
Claim	158	86	76	107	194	158	212

Tableau 14 : Écart inter-annotateurs de nombre d'annotations par groupe (% du total)

Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Tweet 5*	*Forum 1*	*Forum 2*
Écart (%)	4	28	16	16	36	38	0

Accord inter-annotateurs

Il n'y a pas de scores *GammaK* pour cette session parce qu'elle n'a qu'une seule catégorie d'annotation. Nous mesurons des accords inter-annotateurs inégaux et majoritairement faibles à travers les différents groupes (cf. Tableau 15). Les annotateurs n'ont été que très peu formés par manque de temps, ce qui peut expliquer les faibles accords mesurés. Le Tableau 16 confirme le désaccord important, avec une majorité d'annotations différentes (ou disjointes). C'est indicateur d'une session d'annotation très peu fiable.

Tableau 15 : Scores d'accord inter-annotateurs

Mesure / Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Tweet 5*	*Forum 1*	*Forum 2*
Gamma	0,59	0,27	0,70	0,14	0,49	0,26	0,24

Tableau 16 : Proportions (%) des types d'appariements

Type/Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Tweet 5*	*Forum 1*	*Forum 2*
Identiques	*3*	*3*	*20*	*1*	*11*	*6*	*15*
Similaires	*65*	*31*	*46*	*28*	*47*	*21*	*11*
Différents	*32*	*66*	*44*	*71*	*42*	*73*	*74*

Typologie des erreurs dans les annotations similaires

Le Tableau 17 montre que, sur l'ensemble des annotations similaires, la majorité des erreurs a trait aux frontières textuelles (*bruit* et *troncature*). Ce ne sont pas des erreurs critiques parce qu'elles concernent des annotations qui contiennent tout de même du contenu conforme au guide d'annotation. Pour le bruit, l'annotation contient en plus du contenu superflu qui n'est toutefois pas exclu par la définition de la catégorie visée (cf. Figure 24). Pour la troncature, c'est que l'annotation n'inclut pas le segment textuel maximum attendu (cf. Figure 25).

Figure 24 : Illustration du bruit avec des annotations similaires

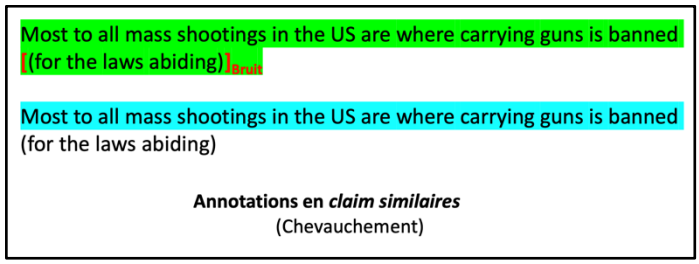
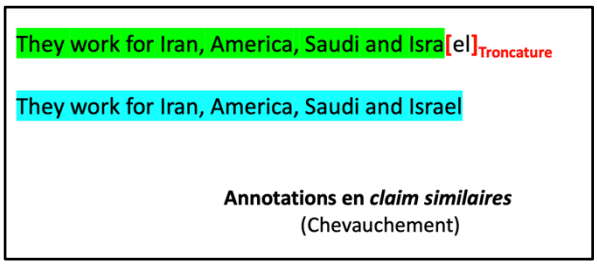


Figure 25 : Illustration de la troncature avec des annotations similaires



Ces types d'erreur sont davantage présents dans les tweets à cause de spécificités, comme les mots-dièses, les marques de mention d'autres utilisateurs et la présence plus fréquente d'URL (cf. Figure 26 et Figure 27) ; le choix de les inclure ou pas dans l'annotation n'était pas évident.

Figure 26 : Illustration de la troncature avec les tweets

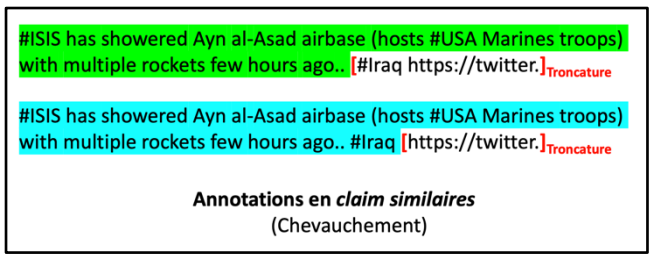
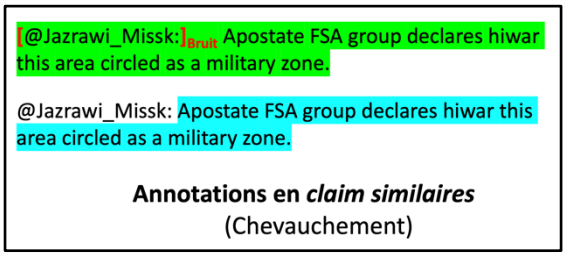


Figure 27 : Illustration du bruit avec les tweets

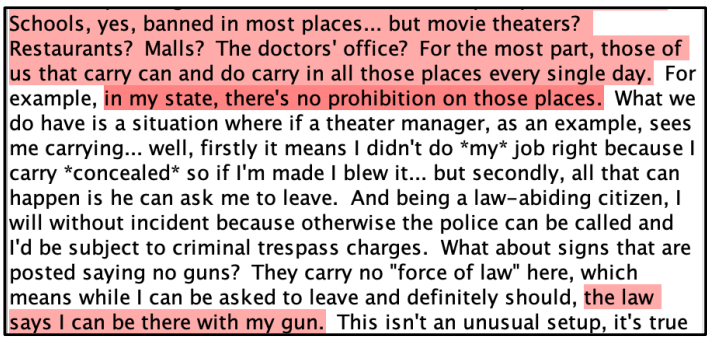


Il y a une proportion d'annotations invalides globalement plus grande dans les messages du forum *Change my view* (cf. Tableau 17), ce que nous expliquons par un contenu qui peut être plus développé et ainsi plus complexe que celui des tweets (cf. Figure 28).

Tableau 17 : Proportions (%) des types d'erreurs relevés dans les annotations similaires

Type/Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Tweet 5*	*Forum 1*	*Forum 2*
Bruit	*9*	*47*	*79*	*21*	*15*	*26*	*50*
Invalidité	*4*	*45*	*0*	*0*	*12*	*36*	*32*
Troncature	*83*	*8*	*21*	*64*	*56*	*15*	*18*
Argumentation	*0*	*0*	*0*	*0*	*0*	*15*	*0*
Pédagogie	*4*	*0*	*0*	*15*	*17*	*8*	*0*

Figure 28 : Extrait d'un message de forum



Typologie des erreurs dans les annotations différentes

Dans le Tableau 18, on peut observer que les annotations différentes comportent plus d'erreurs critiques (des invalidités et des confusions) que les annotations similaires. Une divergence plus forte entre les annotateurs semble donc être indicatrice de problèmes plus importants avec le schéma d'annotation.

Tableau 18 : Proportions (%) des types d'erreurs relevés dans les annotations différentes

Type/Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Tweet 5*	*Forum 1*	*Forum 2*
Invalidité	*34*	*92*	*25*	*27*	*32*	*78*	*82*
Troncature	*41*	*0*	*25*	*72*	*24*	*8*	*0*
Argumentation	*0*	*0*	*0*	*0*	*0*	*1*	*13*
Pédagogie	*25*	*8*	*50*	*1*	*44*	*13*	*5*

Ces types d'erreur sont plus graves parce qu'ils concernent des annotations ne comportant aucun contenu en rapport avec l'une des catégories visées ou alors du contenu exclu par le guide. Par exemple, l'annotation en *claim* de *Perhaps, then, we should consider (…)* constitue une invalidité parce que, même si ce segment de texte est déclaratif, il contient une expression de possibilité, exclue dans la définition de la catégorie *Claim*.

Une difficulté majeure tient dans la distinction du contenu subjectif, qui s'oppose au caractère factuel du claim. Par exemple, le segment de texte *In USA I don't think many people own and carry guns with them* contient une description concrète, comme un claim, mais pour exprimer un point de vue, ce qui est exclu de la définition d'un claim.

Il y a aussi une confusion importante du claim avec la pédagogie parce que cette dernière peut inclure des claims. Par exemple, le segment de texte *Plz read the press release published by the family of Sheikh Abu Hamza regarding his situation* n'est pas un claim en tant que tel mais une suggestion qui contient un claim, dans le segment de texte *press release published by the family of Sheikh Abu Hamza regarding his situation*.

Conclusion partielle

Les annotations résultantes ne sont a priori pas de bonne qualité (avec un faible accord inter-annotateurs). Toutefois, beaucoup d'erreurs ne relèvent que de frontières incorrectes, avec du contenu tout de même acceptable. Les erreurs sérieuses, c'est-à-dire les contradictions avec le guide d'annotation, sont peu variées, facilitant leur appréhension et l'amélioration du guide d'annotation.

6.1.3. Séance d'annotation n°2 « Paris »

Contexte

Dans le Tableau 19, nous présentons la configuration de la deuxième session d'annotation. Les annotateurs sont treize étudiants du M1 TAL de l'INaLCO. Elle a lieu dans le cadre d'un TD de 2h portant sur la détection de l'argumentation. Nous avons traité l'ensemble des catégories de *stimuli*, à savoir *Claim*, *Pédagogie* et *Argumentation*. Avant la séance, les annotateurs ont pris connaissance du guide et ont bénéficié d'une explication à l'oral du contexte et de la tâche. Il y a un trinôme d'annotateurs, désigné par *Forum 1*.

Tableau 19 : Configuration de la session d'annotation n°2

Catégories à annoter	*Claim*, *Pédagogie*, *Argumentation*
Nombre d'annotateurs	13
Groupes	5 binômes, 1 trinôme
Jeux de données	4 *Twitter*, 2 *Forum*

Annotations réalisées

Le Tableau 20 montre que les messages de forum contiennent globalement moins d'annotations. Les annotateurs nous ont indiqué que la complexité des messages avait ralenti leur travail. Il y a logiquement une prépondérance de l'argumentation dans le forum de débat et du claim, le type de discours le moins complexe, dans les tweets, qui sont plutôt brefs (cf. Tableau 21). L'écart de volume annoté est particulièrement important à l'intérieur du groupe *Tweet 1* (cf. Tableau 22). Cet écart est peut-être lié à une acquisition déséquilibrée de la tâche entre les annotateurs du groupe.

Tableau 20 : Nombre d'entités annotées

Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Forum 1*	*Forum 2*
Claim	70	91	46	46	27	27
Pédagogie	43	10	35	18	31	24
Argumentation	16	6	46	9	31	36
Total	129	107	127	73	89	87

Tableau 21 : Proportions (%) des types d'entités annotées

Type / Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Forum 1*	*Forum 2*
Claim	54	85	36	63	30	31
Pédagogie	33	9	28	25	35	28
Argumentation	13	6	36	12	35	41

Tableau 22 : Écart inter-annotateurs de nombre d'annotations par groupe (% du total)

Type / Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Forum 1*	*Forum 2*
Claim	66	28	0	18	27	12
Pédagogie	76	0	0	0	16	24
Argumentation	0	34	18	0	19	32

Accord inter-annotateurs

Dans le Tableau 23, nous relevons un accord inter-annotateurs inégal entre les groupes, y compris sur un même genre textuel. La proportion majoritaire d'annotations différentes à travers les trois catégories (cf. Tableau 24, Tableau 25 et Tableau 26) semble indiquer une difficulté à définir les catégories. La catégorie Claim affiche globalement un meilleur accord (cf. Tableau 23), sûrement parce que c'est la catégorie dont l'expression est la plus simple (cf. section 3.3.3). Le Tableau 23 affiche un accord particulièrement important concernant la catégorie Pédagogie dans les groupes Tweet 2 et Tweet 3.

Tableau 23 : Scores d'accord inter-annotateurs

Mesure / Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Forum 1*	*Forum 2*
Gamma	0,17	0,51	0,40	0,35	0,14	0,36
GammaCat	0,54	0,54	0,67	0,49	0,33	0,60
GammaK - Claim	0,75	0,61	0,64	0,73	0,39	0,82
GammaK - Pédagogie	0,43	0,63	0,82	0,31	0,24	0,46
GammaK - Argumentation	0,49	0,39	0,36	0,45	0,42	0,50

Tableau 24 : Proportions des types d'appariements pour le claim

Type/Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Forum 1*	*Forum 2*
Identiques	12	8	14	12	3	15
Similaires	30	50	43	27	13	20
Différents	68	42	43	61	84	65

Tableau 25 : Proportions des types d'appariements pour la pédagogie

Type/Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Forum 1*	*Forum 2*
Identiques	8	0	20	13	1	0
Similaires	8	25	20	7	26	26
Différents	74	75	60	80	73	74

Tableau 26 : Proportions des types d'appariements pour l'argumentation

Type/Groupe	*Tweet 1*	*Tweet 2*	*Tweet 3*	*Tweet 4*	*Forum 1*	*Forum 2*
Identiques	0	0	11	14	2	3
Similaires	23	20	17	14	27	30
Différents	77	80	72	72	71	67

Nouveau mode de typologie des erreurs

Nous modifions quelque peu notre présentation de la typologie à partir de cette session parce qu'il y a désormais plusieurs catégories à annoter. La distinction des types d'erreur ne se fait plus par groupe mais par trait. Nous avons fait ce choix pour la lisibilité et une meilleure compréhension des difficultés à travers les catégories d'annotation. Le principe de la distinction des annotations similaires et différentes est conservé.

Typologie des erreurs dans les annotations similaires

Dans le Tableau 27, nous présentons la proportion des types d'erreurs relevés dans les annotations similaires. Pour le claim, la majorité des erreurs ont trait aux frontières (*troncature*). Ce n'est pas le cas pour la pédagogie, dont la majorité des erreurs sont plus graves (*invalidité*), dues surtout à la présence d'expressions subjectives, exclues dans la définition de cette catégorie.

Les annotateurs ont eu tendance à annoter la pédagogie comme de l'argumentation. La difficulté afférente à la distinction de ces deux catégories est qu'elles ont trait à deux formes d'explication : l'argumentation explique un point de vue tandis que la pédagogie explique un fait.

Tableau 27 : Proportions (%) des types d'erreurs relevés dans les annotations similaires

Type/Trait	*Claim*	*Pédagogie*	*Argumentation*
Bruit	17	11	5
Invalidité	8	61	15
Troncature	71	23	12
Claim	N/A	0	18
Pédagogie	3	N/A	50
Argumentation	1	5	N/A

Typologie des erreurs dans les annotations différentes

On retrouve une plus grande proportion d'erreurs graves (*invalidité* et *confusion*) dans les annotations différentes (cf. Tableau 28). Une grande difficulté concerne l'expression subjective, exclue dans les définitions du claim et de la pédagogie, mais difficile à distinguer. Pour l'argumentation, l'erreur est de considérer une expression d'opinion seule alors que le principe de l'argumentation est de soutenir une opinion avec une

justification. On retrouve des confusions déjà vues entre certaines catégories : un claim est souvent pris pour de la pédagogie et la pédagogie pour de l'argumentation.

Tableau 28 : Proportions (%) des types d'erreurs relevés dans les annotations différentes

Type/Trait	*Claim*	*Pédagogie*	*Argumentation*
Bruit	1	0	0
Invalidité	43	46	40
Troncature	39	2	3
Claim	N/A	39	21
Pédagogie	15	N/A	37
Argumentation	2	13	N/A

Conclusion partielle

Cette première session avec plusieurs catégories montre une difficulté afférente à la subjectivité, exclue du claim et de la pédagogie mais essentielle dans l'argumentation. Nous avons aussi observé une confusion importante entre les catégories, à cause des points communs que sont la possibilité d'une expression à l'aspect factuel dans les trois catégories et la fonction explicative que peuvent prendre la pédagogie et l'argumentation.

6.1.4. Séance d'annotation n°3 « Stages-1 »

Contexte

La troisième session d'annotation, dont la configuration est décrite dans le Tableau 29, a eu lieu dans le cadre d'un stage d'un mois, avec quatre étudiants du M1 TAL de l'INaLCO. Il s'agit de la première des trois sessions d'annotation qui ont eu lieu lors de ce stage. Elle concerne, comme la session précédente, les catégories *stimuli*, à savoir *Claim*, *Pédagogie* et *Argumentation*. Le binôme ayant reçu un jeu de données *forum* a mis deux fois plus de temps (une journée) que ceux ayant reçu un jeu de tweets, c'est pourquoi ces derniers ont annoté un jeu de données en plus.

Tableau 29 : Configuration de la séance d'annotation n°3

Catégories à annoter	*Claim*, *Pédagogie*, *Argumentation*
Nombre d'annotateurs	4
Groupes	2 binômes
Jeux de données	2 *Twitter*, 1 *Forum*

Annotations réalisées

Sur les tweets, le nombre d'annotations est quelque peu différent entre les groupes et par catégorie (cf. Tableau 30). Nous avons deux hypothèses pour expliquer cela : des compositions différentes entre les jeux de données ou une acquisition inégale du schéma d'annotation entre les annotateurs. Nous privilégions la seconde hypothèse parce que les jeux de données proviennent d'une même ressource. On retrouve une présence particulièrement importante de l'argumentation dans les messages du forum de débat, et une prépondérance du claim dans les tweets (cf. Tableau 31).

Tableau 30 : Nombre d'entités annotées

Groupe	*Tweet 1*	*Tweet 2*	*Forum 1*
Claim	80	105	81
Pédagogie	6	38	26
Argumentation	21	10	48
Total	107	153	155

Tableau 31 : Proportions (%) des types d'entités

Groupe	*Tweet 1*	*Tweet 2*	*Forum 1*
Claim	75	69	52
Pédagogie	6	25	17
Argumentation	19	6	31

Le Tableau 32 affiche des écarts de volume annoté qui sont majoritairement faibles, ce qui peut être le signe d'un équilibre dans l'acquisition du schéma d'annotation entre les annotateurs. Il faut toutefois nuancer cette constatation en remarquant les deux écarts forts (supérieurs à 50%) pour des catégories dont les annotations sont nombreuses (*Claim* et *Argumentation*).

Tableau 32 : Écart inter-annotateurs du nombre d'annotations par groupe (% du total)

Groupe	*Tweet 1*	*Tweet 2*	*Forum 1*
Claim	10	10	56
Pédagogie	34	6	22
Argumentation	58	0	4

Accord inter-annotateurs

Nous mesurons un accord inter-annotateurs globalement meilleur pour les tweets que pour les messages de forum (cf. Tableau 33). Les annotateurs nous ont indiqué que ces derniers avaient un contenu plus complexe que les tweets, rendant leur annotation plus difficile.

Tableau 33 : Scores d'accord inter-annotateurs

Mesure / Groupe	*Tweet 1*	*Tweet 2*	*Forum 1*
Gamma	0,55	0,69	0,20
GammaCat	0,74	0,87	0,46
GammaK - Claim	0,76	0,84	0,64
GammaK - Pédagogie	0,66	1	0,42
GammaK - Argumentation	0,61	0,69	0,23

La majorité des divergences tient dans des annotations différentes plutôt que similaires pour les trois catégories (cf. Tableau 34 à Tableau 36). Cela peut indiquer des disparités profondes entre les annotateurs concernant l'acquisition du schéma d'annotation.

Tableau 34 : Proportions (%) des types d'appariements pour le claim

Type/Groupe	*Tweet 1*	*Tweet 2*	*Forum 1*
Identiques	50	67	10
Similaires	17	0	9
Différents	33	33	81

Tableau 35 : Proportions (%) des types d'appariements pour la pédagogie

Type/Groupe	*Tweet 1*	*Tweet 2*	*Forum 1*
Identiques	0	63	5
Similaires	20	9	14
Différents	80	26	81

Tableau 36 : Proportions (%) des types d'appariements pour l'argumentation

Type/Groupe	*Tweet 1*	*Tweet 2*	*Forum 1*
Identiques	6	20	3
Similaires	11	0	20
Différents	83	80	77

Typologie des erreurs dans les annotations similaires

Le Tableau 37 montre que, parmi les annotations similaires, la majorité des erreurs ne sont que superficielles (*troncature* et *bruit*). Il y a toutefois, en minorité (12%), des annotations en claim invalides, encore à cause de la présence d'expressions subjectives.

Tableau 37 : Proportions (%) des types d'erreurs relevés dans les annotations similaires

Type/Trait	*Claim*	*Pédagogie*	*Argumentation*
Bruit	32	14	30
Invalidité	12	0	0
Troncature	56	86	70
Claim	N/A	0	0
Pédagogie	0	N/A	0
Argumentation	0	0	N/A

Typologie des erreurs dans les annotations différentes

On retrouve une prépondérance des erreurs graves (*invalidités* et *confusion*) dans les annotations différentes (cf. Tableau 38). Les invalidités sont plus présentes avec les catégories *Claim* et *Pédagogie*, une fois encore à cause de la difficulté à distinguer la subjectivité, exclue pour ces catégories. Comme précédemment, il y a une confusion de la pédagogie et de l'argumentation avec le claim.

Tableau 38 : Proportions (%) des types d'erreurs pour les annotations différentes

Type/Trait	*Claim*	*Pédagogie*	*Argumentation*
Bruit	0	0	0
Invalidité	61	50	25
Troncature	15	0	0
Claim	N/A	30	50
Pédagogie	24	N/A	25
Argumentation	0	20	N/A

Conclusion partielle

Les annotations résultantes sont relativement bonnes pour les tweets et mauvaises pour les messages de forum. Le contenu plus développé de ces derniers rend, d'après les annotateurs eux-mêmes, leur annotation plus difficile. Nous avons remarqué que la majorité des erreurs est superficielle. Si les cas d'erreur plus sérieux semblent difficiles à résoudre parce que persistants d'une séance à l'autre, nous supposons qu'ils sont moins prépondérants du fait de l'amélioration de l'accord inter-annotateurs.

6.1.5. Séance d'annotation n°4 « Stages-2 »

Contexte

Cette séance a lieu après une phase de discussion concernant les difficultés éprouvées lors de la session précédente, donnant lieu à l'amélioration du guide d'annotation. C'est l'occasion d'une nouvelle configuration en binômes (cf. Tableau 39) afin d'éviter des biais de groupe. Nous restons sur les catégories de *stimuli* pour voir si l'expérience des annotateurs et les discussions améliorent l'annotation.

Tableau 39 : Configuration de la séance n°4

Catégories à annoter	*Claim*, *Pédagogie*, *Argumentation*
Nombre d'annotateurs	4
Groupes	2 binômes
Jeux de données	1 *Twitter*, 1 *Forum*

Annotations réalisées

Dans le Tableau 40, nous constatons que le nombre d'entités annotées est beaucoup plus important pour le jeu de données issu du forum *Change my view*, ce qui peut s'expliquer par la taille des messages, généralement plus longue dans ce média social. Le Tableau 41 montre encore une présence particulièrement importante de l'argumentation dans les messages du forum et de claims dans les tweets. Le Tableau 42 affiche des écarts de volume annoté relativement faibles, ce qui est positif quant à l'intersubjectivité présente dans les annotations.

Tableau 40 : Nombre d'entités annotées

Groupe	*Tweet 1*	*Forum 1*
Claim	131	84
Pédagogie	40	91
Argumentation	18	105
Total	189	280

Tableau 41 : Proportions (%) des types d'entités

Groupe	*Tweet 1*	*Forum 1*
Claim	69	30
Pédagogie	22	33
Argumentation	9	37

Tableau 42 : Écart inter-annotateurs de nombre d'annotations par groupe (% du total)

Groupe	*Tweet 1*	*Forum 1*
Claim	2	16
Pédagogie	6	0
Argumentation	12	26

Accord inter-annotateurs

Nous mesurons un accord inter-annotateurs particulièrement élevé à travers les groupes et les catégories (cf. Tableau 43), corroboré par une large majorité d'annotations identiques dans les tweets (cf. Tableau 44 à Tableau 46). Il y a toutefois une majorité d'annotations différentes dans les messages de forum (cf. Tableau 44 à Tableau 46) : la comparaison entre la faible valeur de *Gamma* et les valeurs hautes de *GammaCat* et *GammaK* (cf. Tableau 43), qui ne prennent pas en compte les frontières, indique que ces annotations différentes sont plutôt dûes à des désaccords sur les frontières que sur la distinction des catégories.

Tableau 43 : Scores d'accord inter-annotateurs

Mesure / Groupe	*Tweet 1*	*Forum 1*
Gamma	0,91	0,34
GammaCat	0,94	0,82
GammaK - Claim	0,92	0,95
GammaK - Pédagogie	0,97	0,86
GammaK - Argumentation	0,95	0,68

Tableau 44 : Proportions des types d'appariements pour le claim

Type/Groupe	*Tweet 1*	*Forum 1*
Identiques	90	13
Similaires	3	12
Différents	7	75

Tableau 45 : Proportions des types d'appariements pour la pédagogie

Type/Groupe	*Tweet 1*	*Forum 1*
Identiques	81	18
Similaires	0	34
Différents	19	48

Tableau 46 : Proportions des types d'appariements pour l'argumentation

Type/Groupe	*Tweet 1*	*Forum 1*
Identiques	80	7
Similaires	0	33
Différents	20	60

Typologie des erreurs dans les annotations similaires

Nous présentons les proportions des types d'erreurs pour les annotations similaires dans le Tableau 47. Au regard de cette session aussi, la majorité des erreurs dans les annotations similaires ne sont que formelles, puisqu'il s'agit principalement de troncature. On relève à nouveau une confusion du claim avec la pédagogie.

Tableau 47 : Proportions (%) des types d'erreurs pour les annotations similaires

Type/Trait	*Claim*	*Pédagogie*	*Argumentation*
Bruit	27	15	21
Invalidité	9	9	18
Troncature	37	67	50
Claim	N/A	0	0
Pédagogie	27	N/A	11
Argumentation	0	9	N/A

Typologie des erreurs dans les annotations différentes

Nous présentons une typologie des erreurs dans les annotations différentes avec le Tableau 48. On retrouve une proportion importante d'invalidités dans les annotations différentes, pour les catégories *Claim* et *Argumentation*. Pour la catégorie *Claim*, cela provient de la difficulté à distinguer le contenu subjectif, exclu de sa définition. Quant à l'argumentation, les invalidités correspondent à des expressions d'opinion sans justification. On retrouve aussi une confusion entre certaines des catégories, probablement liée à la difficulté que nous avons rencontrée pour les définir à la fois assez largement et précisément.

Tableau 48 : Proportions (%) des types d'erreurs pour les annotations différentes

Type/Trait	*Claim*	*Pédagogie*	*Argumentation*
Bruit	0	0	0
Invalidité	43	0	58
Troncature	0	0	0
Claim	N/A	50	13
Pédagogie	52	N/A	29
Argumentation	5	50	N/A

Conclusion partielle

Les annotateurs de cette séance ont un accord globalement très bon, aussi bien pour les tweets que pour les messages de forum. Nous retrouvons toutefois des difficultés déjà observées sur les séances précédentes, que nous ne sommes pas arrivés à totalement résorber. Il faudrait améliorer le guide d'annotation en supprimant ou en précisant la notion de subjectivité, qui pose beaucoup de problèmes. Une autre piste d'amélioration est d'expliciter les critères de distinction entre les catégories.

6.1.6. Séance d'annotation n°5 « Stages-3 »

Contexte

La séance d'annotation n°5 est la dernière. Nous présentons sa configuration dans le Tableau 49. Des nouveaux binômes sont constitués afin d'éviter des biais de groupe. Cette séance est la seule à traiter les catégories de *stimulation* et de *décision* : la *stimulation* concerne les réactions à des *stimuli,* comprenant l'accord, le changement d'affect, la compréhension et l'acquisition d'information. La *décision* a trait à l'expression d'un changement d'avis, suite à une stimulation (cf. section 3.3.4). Nous n'utilisons que des messages du forum *Change my view* parce qu'eux seuls contiennent ces réactions.

Tableau 49 : Configuration de la séance n°5

Catégories à annoter	*Accord*, *Affect*, *Compréhension*, *Information*, *Chgt d'avis*
Nombre d'annotateurs	4
Groupes	2 binômes
Jeux de données	2 *Forum*

Annotations réalisées

Dans le Tableau 50 et le Tableau 51, nous constatons un net déséquilibre dans les quantités d'annotations à travers les catégories, avec une forte tendance en faveur de la catégorie *Accord*. Les annotateurs nous ont rapporté qu'ils ont parfois eu du mal à distinguer certaines des autres catégories. Le Tableau 52 affiche un écart important des quantités annotées à l'intérieur des binômes, ce qui révèle un manque d'homogénéité dans la compréhension des catégories entre les annotateurs.

Tableau 50 : Nombre d'entités annotées

Groupe	*Forum 1*	*Forum 2*
Accord	55	24
Affect	1	5
Compréhension	3	3
Information	8	1
Chgt d'avis	13	3
Total	80	36

Tableau 51 : Proportions (%) des types d'entités

Groupe	*Forum 1*	*Forum 2*
Accord	69	67
Affect	1	14
Compréhension	4	8
Information	10	3
Chgt d'avis	16	8

Tableau 52 : Écart inter-annotateurs de nombre d'annotations par groupe (% du total)

Groupe	*Forum 1*	*Forum 2*
Accord	10	0
Affect	100	20
Compréhension	34	34
Information	0	100
Chgt d'avis	8	100

Accords inter-annotateurs

Les scores d'accord inter-annotateurs sont présentés dans le Tableau 53. Nous observons un taux d'accord globalement élevé pour la catégorie *Accor*d, laissant présager une fiabilité des annotations. Concernant les autres catégories, les scores sont trop peu significatifs, compte tenu des très faibles quantités annotées, parfois par un seul des annotateurs. Les Tableaux 54 à 58 montrent des proportions de types d'appariements disparates entre les groupes pour une même catégorie, ce qui évoque une instabilité dans l'annotation et donc des lacunes dans la compréhension de ces catégories.

Tableau 53 : Scores d'accord inter-annotateurs

Mesure / Groupe	*Forum 1*	*Forum 2*
Gamma	0,87	0,55
GammaCat	1	0,82
GammaK - Accord	1	0,87
GammaK - Affect	N/A	1
GammaK - Compréhension	1	-0,05
GammaK - Information	1	N/A
GammaK - Chgt d'avis	1	N/A

Tableau 54 : Proportions (%) des types d'appariements pour l'accord

Type/Groupe	*Forum 1*	*Forum 2*
Identiques	90	40
Similaires	0	20
Différents	10	40

Tableau 55 : Proportions (%) des types d'appariements pour l'affect

Type/Groupe	*Forum 1*	*Forum 2*
Identiques	N/A	67
Similaires	N/A	0
Différents	N/A	33

Tableau 56 : Proportions (%) des types d'appariements pour la compréhension

Type/Groupe	*Forum 1*	*Forum 2*
Identiques	50	0
Similaires	0	0
Différents	50	100

Tableau 57 : Proportions (%) des types d'appariements pour l'information

Type/Groupe	*Forum 1*	*Forum 2*
Identiques	60	N/A
Similaires	0	N/A
Différents	40	N/A

Tableau 58 : Proportions (%) des types d'appariements pour le changement d'avis

Type/Groupe	*Forum 1*	*Forum 2*
Identiques	57	N/A
Similaires	29	N/A
Différents	14	N/A

Typologie des erreurs dans les annotations similaires

Le Tableau 59 montre que les annotations similaires ne contiennent que des erreurs superficielles (*troncature* et *bruit*), ce qui est un indice positif concernant la compréhension des catégories.

Tableau 59 : Proportions (%) des types d'erreurs pour les annotations similaires

Erreur/Trait	*Accord*	*Affect*	*Compréhension*	*Information*	*Chgt d'avis*
Bruit	N/A	50	N/A	N/A	50
Invalidité	N/A	0	N/A	N/A	0
Troncature	N/A	50	N/A	N/A	50
Accord	N/A	0	N/A	N/A	0
Affect	N/A	N/A	N/A	N/A	0
Compréhension	N/A	0	N/A	N/A	0
Information	N/A	0	N/A	N/A	0
Chgt d'avis	N/A	0	N/A	N/A	N/A

Typologie des erreurs dans les annotations différentes

Le Tableau 60 montre que les annotations différentes relèvent d'erreurs d'annotation profondes, avec une majorité d'annotations qui ne correspondent à aucune des définitions du guide. Ces résultats sont à relativiser par le très faible nombre d'annotations et par le fait qu'il s'agit de la première (et unique) session concernant les catégories visées.

Tableau 60 : Proportions (%) des types d'erreurs pour les annotations différentes

Type/Trait	*Accord*	*Affect*	*Compréhension*	*Information*	*Chgt d'avis*
Bruit	0	0	0	0	0
Invalidité	100	100	0	0	0
Troncature	0	0	0	0	0
Claim	0	0	0	0	0
Pédagogie	0	0	0	0	0
Accord	N/A	0	0	0	0
Affect	0	N/A	100	0	0
Compréhension	0	0	N/A	100	0
Information	0	0	0	N/A	100
Chgt d'avis	0	0	0	0	N/A

Conclusion partielle

Nous retenons de cette session la très faible quantité d'annotations. Les annotateurs nous ont indiqué avoir peu annoté à la fois parce qu'ils n'étaient pas assurés d'avoir bien compris les catégories et parce qu'elles leur ont semblé moins présentes que les catégories *stimuli*. C'était la première séance d'annotation sur ces catégories, avec donc peu d'expérience pour les annotateurs comme pour nous-mêmes. Il est aussi possible que ces catégories soient moins présentes que les *stimuli* parce qu'on peut supposer qu'il y a plus de tentatives d'influence que de réactions à ces tentatives.

6.1.7. Bilan global des séances d'annotation

Dans le Tableau 61, nous présentons le nombre d'annotations par trait discursif et genre textuel face au nombre de messages qu'il y avait à annoter. Les différences du nombre d'annotations entre les traits discursifs sont cohérentes avec les différences du nombre de messages à annoter. Ce n'est cependant pas le cas entre les genres textuels : nous relevons en effet des tendances contraires pour la pédagogie et l'argumentation. Cela reflète probablement des caractéristiques discursives différentes entre les tweets et les messages du forum *Change my view*.

Tableau 61 : Statistiques sur les quantités annotées par trait discursif et genre textuel

Statistique	*Nombre de messages à annoter*		*Nombre d'annotations*	
Genre textuel	*Tweet*	*Forum*	*Tweet*	*Forum*
Claim	1200	480	1190	589
Pédagogie	700	320	137	172
Argumentation	700	320	126	220
Accord	0	160	0	79
Affect	0	160	0	6
Compréhension	0	160	0	6
Information	0	160	0	9
Changement d'avis	0	160	0	16

Le Tableau 62 présente la proportion moyenne des types d'erreurs à travers les traits discursifs. Nous avons regroupé les types d'erreurs *troncature* et *bruit* parce qu'ils relèvent tous les deux d'erreurs de frontière non critiques. Il faut relativiser les résultats concernant les traits de *stimulation* et de *décision*, au regard de leur faible nombre d'annotations (cf. Tableau 61).

Tableau 62 : Proportion moyenne des types d'erreur par trait discursif

Type d'erreur	*Troncature ou bruit*	*Confusion*	*Invalidité*
Claim	27	39	34
Pédagogie	28	60	12
Argumentation	24,5	21	38,5
Accord	0	0	100
Affect	50	0	50
Compréhension	0	100	0
Information	0	100	0
Changement d'avis	50	50	0

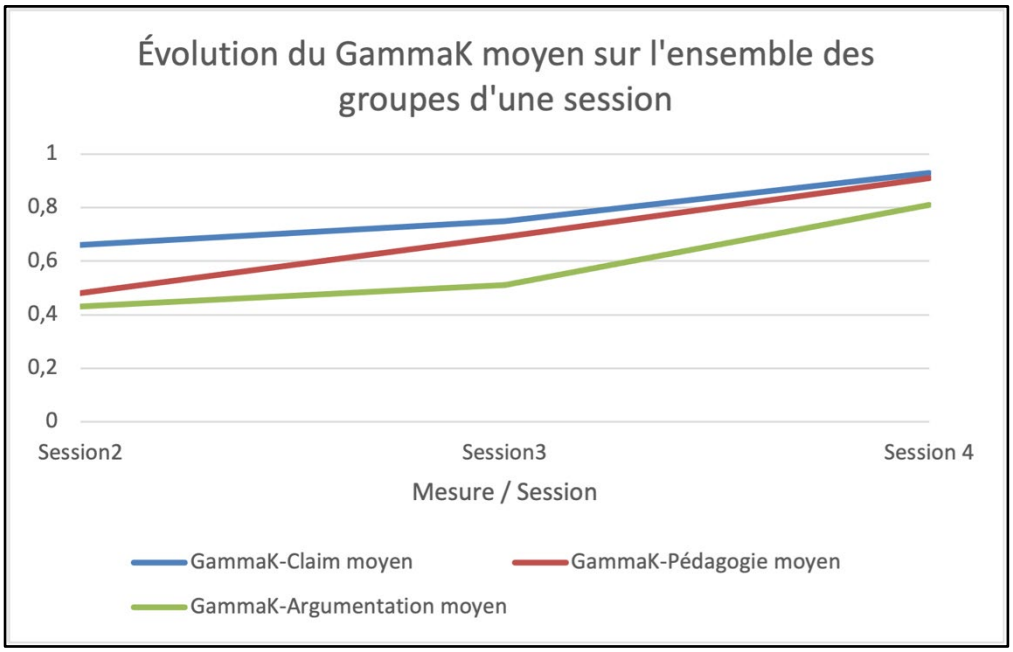
Il y a une part importante de confusion entre les catégories d'annotation. À cet égard, nos pistes pour améliorer le guide d'annotation sont soit d'expliciter les critères de distinction entre les catégories, soit d'unifier les catégories pour lesquelles les critères de distinction sont trop peu pertinents.

La majorité des erreurs sont des invalidités pour les catégories *Argumentation* et *Accord*. Pour l'argumentation, les invalidités correspondent à l'annotation d'expressions d'opinions seules. Il faudrait souligner dans le guide que l'argumentation a pour principe d'exprimer une opinion en la justifiant. Concernant la catégorie *Accord*, il s'agit d'une seule invalidité, qui correspond à l'expression d'une opinion positive concernant le propos de quelqu'un ; il faudrait ajouter dans le guide qu'une opinion positive n'est pas synonyme d'accord.

- Exemple : « What you're saying is interesting » ; ceci est une expression d'opinion positive concernant le propos de quelqu'un mais elle n'indique pas pour autant que l'énonciateur est d'accord avec ce qui est dit

La Figure 29 présente l'évolution du score *GammaK* moyen par trait discursif, à travers des sessions d'annotation portant sur les mêmes traits. Cela permet de constater la progression de l'accord inter-annotateurs au cours de la campagne d'annotation. C'est sûrement indicateur d'une amélioration de la compréhension de la tâche par les annotateurs et ainsi d'une amélioration de la qualité des annotations.

Figure 29 : Évolution des scores d'accord inter-annotateurs



Le bilan global de cette campagne d'annotation montre que la tâche était difficile mais que son organisation et l'implication des annotateurs ont permis d'aller vers des annotations de meilleure qualité. La difficulté de la tâche provient principalement de la complexité des traits discursifs à annoter, qui a nécessité un important travail définitoire et une remise en question permanente au cours de la campagne d'annotation, en collaboration avec les annotateurs.

Dans la section suivante, nous décrivons comment nous avons constitué un corpus de référence à partir des annotations produites.

6.1.8. Constitution du corpus de référence

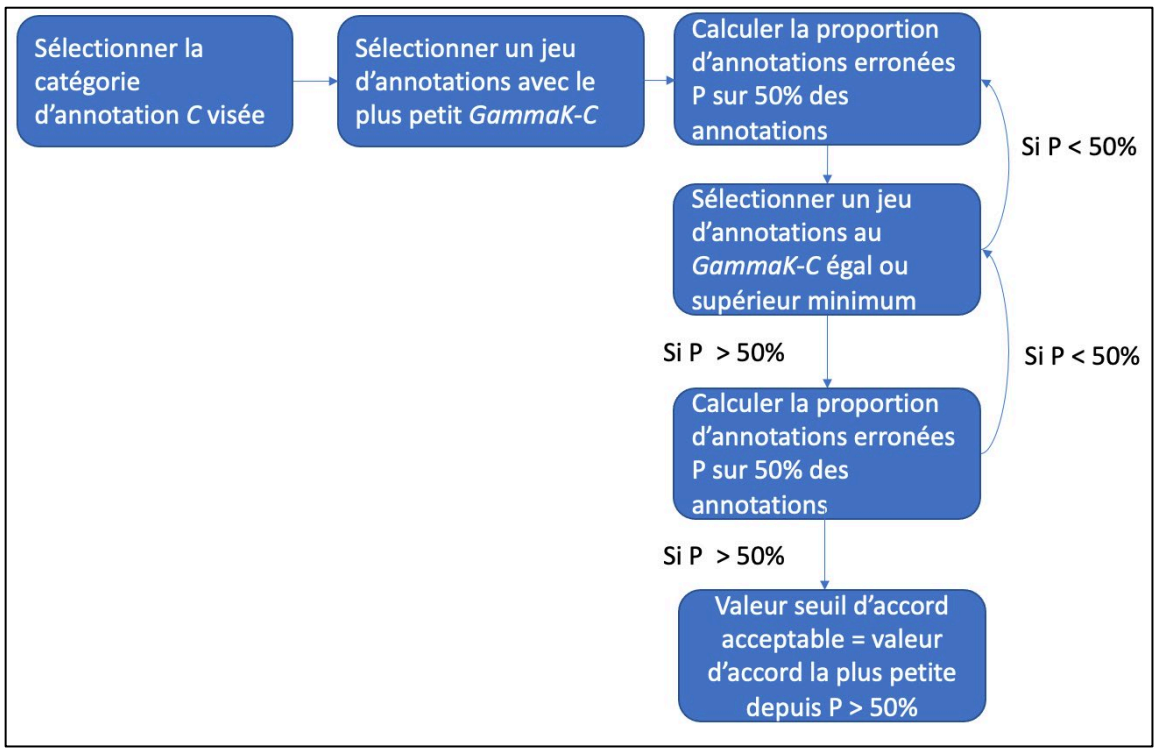
Introduction

L'objectif de cette campagne d'annotation était de constituer un corpus de référence pour les différents traits discursifs de notre modèle d'influence individuelle (cf. sections 3.3.3 et 3.3.4). Pour ce faire, il était nécessaire de déterminer, parmi les annotations produites, lesquelles étaient d'assez bonne qualité pour faire partie de notre corpus de référence.

Nous avons cherché dans la littérature quels critères devaient être utilisés pour déterminer si un jeu d'annotations était d'une qualité suffisante pour servir de référence. Si l'accord inter-annotateurs est communément utilisé pour évaluer la qualité d'un jeu d'annotations, nous avons constaté qu'il n'y a pas un seuil d'accord acceptable qui soit universel pour toutes les tâches d'annotation (Artstein & Poesio, 2008). Nous avons alors créé une méthode pour déterminer un seuil *ad-hoc* d'accord acceptable.

Méthodologie

Figure 30 : Protocole de sélection des annotations de référence



La méthode que nous avons mise en œuvre (cf. Figure 30) repose sur une analyse manuelle de la qualité des annotations. Pour une catégorie d'annotation donnée, le principe est de comparer la valeur d'accord de quelques jeux d'annotations à leurs proportions respectives de messages annotés correctement. Cela permet d'estimer une valeur seuil d'accord acceptable pour chacune des catégories d'annotation. Cela permet aussi d'avoir une vue sur les catégories les mieux annotées, en comparant, par catégorie, les scores d'accord à la valeur seuil identifiée.

Une valeur seuil d'accord acceptable est une valeur d'accord à partir de laquelle les jeux d'annotations ont une proportion *acceptable* de messages annotés correctement. Nous considérons que, pour une catégorie d'annotation donnée, un jeu d'annotations contient une proportion *acceptable* de messages annotés correctement lorsqu'une majorité (plus de 50%) de ses messages contient l'ensemble des annotations attendues.

Nous privilégions l'utilisation des valeurs d'accord *GammaK*, disponibles lorsqu'il y a plusieurs catégories annotées (cf. section 6.1.1), parce que cette mesure se concentre sur la reconnaissance des catégories, sans tenir compte des désaccords sur les frontières textuelles, qui ne donnent pas forcément lieu à des erreurs critiques. La session d'annotation n°1 (cf. section 6.1.2) n'a pas de scores *GammaK* parce qu'elle ne comporte qu'une seule catégorie d'annotation ; nous utiliserons alors les valeurs *Gamma* pour les jeux d'annotation de cette session.

Nous avons procédé en analysant d'abord, pour chaque catégorie, un jeu d'annotations ayant la plus petite valeur d'accord. Nous parcourons ensuite les jeux d'annotations par ordre croissant de valeur d'accord, tant que nous ne trouvons pas une proportion d'annotations correctes acceptable. Nous privilégions l'alternance des genres textuels (les tweets et les messages de forum). Nous avons traité 50% de chaque jeu d'annotations sélectionné, avec l'hypothèse que ce serait assez représentatif de l'ensemble.

Pour une catégorie d'annotation, lorsqu'un jeu de données contient une proportion acceptable de messages annotés correctement, nous estimons si sa valeur d'accord est une valeur seuil d'accord acceptable en analysant le jeu de données, si possible de genre textuel différent, qui possède une valeur d'accord égale ou supérieure la plus proche. Si ce nouveau jeu de données a aussi une proportion acceptable de messages annotés correctement, nous validons la valeur d'accord du jeu d'annotations précédent comme valeur seuil d'accord acceptable. Sinon, nous reprenons le parcours des jeux d'annotation.

Une fois que nous avons estimé les valeurs seuils d'accord acceptable par catégorie d'annotation, nous résolvons les divergences d'annotation pour les jeux de d'annotations dont la valeur d'accord est supérieure ou égale à la valeur seuil. Les divergences d'annotation pour un même message pourraient constituer du bruit dans la référence.

Différentes méthodes peuvent être utilisées pour résoudre les divergences dans un jeu d'annotations (Mathet & Widlöcher, 2016). Nous en avons choisi deux, avec l'objectif de maximiser la qualité des annotations produites et de minimiser les biais.

L'une des méthodes consiste à utiliser la réconciliation entre les annotateurs. Avec la méthode par réconciliation, ce sont les annotateurs qui discutent pour décider des annotations à conserver parmi celles qui sont divergentes.

L'autre méthode que nous avons appliquée lorsqu'il n'y avait pas eu de réconciliation entre les annotateurs est celle de la *super-annotation* : une seule personne sélectionne, parmi les annotations divergentes, celles qui lui semblent correctes et qui seront conservées. Ayant conçu la tâche d'annotation, nous avons pris le rôle de *super-annotateur*.

La méthode de la réconciliation permet de limiter les biais éventuels relatifs à la décision d'un seul annotateur, comme dans le cas de la super-annotation. Cependant, la méthode de la super-annotation peut accroitre la fiabilité des annotations en faisant appel à un annotateur qui a plus d'expérience que les annotateurs initiaux.

Résultats

Dans le Tableau 63, nous donnons les valeurs minimales de *GammaK* pour chacune des catégories d'annotation, ce qui permet de les comparer aux valeurs seuils qui ont été validées dans chacune des catégories.

Nous constatons que, pour chaque catégorie d'annotation, ce sont les valeurs *GammaK* minimales qui ont été validées comme des valeurs seuils d'accord acceptables. Ces valeurs d'accord sont basses mais correspondent pourtant à deux jeux d'annotation qui contiennent une majorité de messages exploitables après la résolution des divergences. En plus, ces proportions (les taux de validité) sont généralement élevées.

Tableau 63 : Valeurs minimales de *GammaK* par catégorie d'annotation

Catégorie	**Claim**	**Pédagogie**	**Argumentation**
***GammaK* minimal**	0,14	0,24	0,23

Dans les Tableaux 64 à 66, nous présentons les résultats de notre analyse des jeux d'annotation pour, respectivement, les catégories *Claim*, *Pédagogie* et *Argumentation*. Les taux de validité sont la proportion de messages que nous avons identifiés comme contenant les annotations correctes. Ils sont donc soumis à la qualité de notre analyse manuelle des annotations.

Tableau 64 : Résultats de la validation d'une valeur seuil d'accord acceptable pour la catégorie *Claim*

Jeu d'annotations/Évaluation	***GammaK-Claim***	**Taux de validité (%)**
Session 1, *tweet 5*	0,14	98%
Session 1, *forum 1*	0,24	57%

Tableau 65 : Résultats de la validation d'une valeur seuil d'accord acceptable pour la catégorie *Pédagogie*

Jeu d'annotations/Évaluation	***GammaK-Pédagogie***	**Taux de validité (%)**
Session 2, *forum 1*	0,24	78%
Session 2, *tweet 4*	0,31	96%

Tableau 66 : Résultats de la validation d'une valeur seuil d'accord acceptable pour la catégorie *Argumentation*

Jeu d'annotations/Évaluation	***GammaK-Argumentation***	**Taux de validité (%)**
Session 3, *forum 6*	0,23	92%
Session 2, *tweet 10*	0,36	92%

Les valeurs seuils d'accord acceptable que nous avons validées correspondent toutes aux plus petites valeurs d'accord pour chaque catégorie. Elles nous conduisent donc à exploiter l'ensemble des données produites lors de la campagne d'annotation. Nous avons alors appliqué nos méthodes de résolution des divergences à tous les jeux d'annotations.

Dans le Tableau 67, nous présentons les taux d'utilisation des deux méthodes de résolution des divergences d'annotation que nous avons utilisées, à savoir la réconciliation des annotateurs et la super-annotation. La majeure partie des divergences sont résolues par la méthode de la super-annotation parce qu'une grande partie des groupes d'annotateurs n'avaient pas fait de réconciliation.

Tableau 67 : Taux d'utilisation des méthodes de résolution des divergences d'annotation

Méthode de résolution des divergences d'annotation	**Réconciliation**	**Expert**
Taux d'application (%)	28%	72%

Conclusion

Nous avons constaté que des jeux d'annotations avec des valeurs d'accord basses semblent tout de même exploitables, après une phase de résolution des divergences. En effet, même si les annotateurs ont été en désaccord, l'ensemble de leurs annotations peut quand même contenir les annotations attendues.

À la vue de ces résultats, nous pensons que l'accord inter-annotateurs ne doit pas être pris comme une référence absolue de la qualité des annotations, d'autant qu'il n'existe pas de valeur de référence pour déterminer l'acceptabilité d'un jeu d'annotations.

La méthode que nous avons développée, fondée sur une analyse directe de la qualité des annotations, nous a permis d'utiliser l'ensemble des jeux de données produits

lors de la campagne d'annotation. Cette méthode pourrait servir à la constitution de données de référence à partir d'autres jeux d'annotation.

Après avoir constitué notre corpus de référence, nous allons voir comment nous l'avons utilisé pour travailler à la détection automatique des traits discursifs avec des techniques de TAL.

6.2. Détection automatique des *stimuli*

6.2.1. Introduction

Nous allons présenter notre approche pour la détection automatique des traits discursifs *stimuli*, à savoir le *claim*, la *pédagogie* et l'*argumentation*, qui correspond à une tâche d'annotation automatique de textes, que nous avons traitée en deux modalités.

Au début de nos expériences, nous avons opté pour une segmentation, dans les messages, des expressions correspondant à nos traits discursifs. Ce type d'annotation est plus difficile qu'une classification des messages parce qu'elle nécessite d'extraire les segments de textes pertinents. De plus, elle n'est pas nécessaire parce que nous avons besoin d'identifier les messages qui contiennent les traits discursifs, ce à quoi répond une approche par classification. C'est pour cela que nous avons mis en œuvre une approche par segmentation uniquement pour le trait *Claim*.

La majorité de notre travail sur l'annotation automatique des traits *stimuli* se fait avec une approche par classification automatique, consistant à distinguer des messages de médias sociaux selon qu'ils contiennent ou pas des expressions correspondant aux traits discursifs.

Nous avons mis en œuvre les deux grands types d'approches pour faire de l'annotation automatique de textes : le développement de règles linguistiques et le développement de modèles par apprentissage automatique. Il sera intéressant de voir lequel des deux produira les meilleurs résultats seul, et si leur combinaison est pertinente.

6.2.2. Données de référence

Dans un premier temps, nous avons constitué un petit jeu de référence en claims (une *micro-référence*) à partir de 363 messages sur les 560 annotés lors de la session n°2 (cf. section 6.1.3). Cela nous a permis de débuter nos expériences sur la détection de claims avant d'avoir fini la campagne d'annotation.

Pour nous assurer de la qualité des annotations utilisées comme référence, nous avons opéré un filtrage manuel consistant à évincer les erreurs d'annotation au sens de la typologie décrite dans la section 6.1.1.

Dans le tableau 65, nous présentons des statistiques sur le jeu de données obtenu à l'issue de ce filtrage. Les tweets sont beaucoup plus présents que les messages de forum ; cette statistique est cohérente avec la constitution originale des jeux de données (cf. section 6.1) : les jeux de tweets contiennent plus de messages parce que nous les avons supposés plus rapides à annoter du fait de leur limitation en nombre de caractères.

Tableau 65 : Statistiques sur la micro-référence pour les claims

Nombre de messages	363
% de tweets	85
% de messages de forum	15
Nombre de claims	142

Nous avons partitionné ce jeu de référence de façon à obtenir un jeu de développement et trois jeux d'évaluation. Cette partition suit des proportions classiques, à savoir 70% des données réservées au jeu de développement et 10% pour chacun des trois jeux d'évaluation. Dans le Tableau 66, nous présentons des statistiques sur chacune des partitions obtenues.

Tableau 66 : Statistiques sur les partitions de la micro-référence pour les claims

Caractéristique / Partition	*Développement*	*Éval1*	*Éval2*	*Éval3*
Nombre de messages	263	37	33	30
Nombre de claims	103	12	13	14

Nous recensons un total de 103 claims dans la partition de développement. Cet ensemble de 103 claims a servi de référence pour, d'une part, concevoir les règles linguistiques et, d'autre part, entrainer nos modèles par apprentissage automatique.

Dans un second temps, après avoir estimé les valeurs seuils d'accord inter-annotateurs acceptable concernant les annotations des traits *stimuli* (cf. section 6.1.8), nous avons utilisé la totalité des annotations produites lors de la campagne d'annotation comme référence pour la détection de ces traits. Dans le Tableau 68 et le Tableau 69, nous présentons des statistiques sur les partitions de ces données de référence. Les données ont été distribuées suivant le schéma suivant : 75% des données pour l'entrainement des algorithmes d'apprentissage et 15% pour chacune des deux partitions d'évaluation.

Les données de référence ont une meilleure composition pour le trait *Claim* que pour *Pédagogie* et *Argumentation* : les messages sont plus nombreux et en proportion plus

grande dans la classe positive (les messages contenant le trait ciblé). C'est un avantage pour l'apprentissage du trait *Claim*.

Tableau 68 : Statistiques sur les partitions des données de référence totales pour *Claim*

Partition	*Tout*	*Entrainement*	*Éval1*	*Éval2*
Nombre de messages	1126	789	169	168
% de positifs *Claim*	45%	47%	43%	39%
Nombre de *claims*	627	452	93	82

Tableau 69 : Statistiques sur les données de référence pour *Pédagogie* et *Argumentation*

Partition	*Tout*	*Entrainement*	*Éval1*	*Éval2*
Nombre de messages	716	500	108	108
% de positifs *Pédagogie*	14%	14%	14%	14%
% de positifs *Argumentation*	7%	6%	9%	7%
Nombre de *pédagogies*	117	85	15	17
Nombre d'*argumentations*	57	39	13	9

6.2.3. Méthodologie

Nous commençons avec un prétraitement des messages en vue de leur analyse par nos différents algorithmes (les règles linguistiques et les modèles par apprentissage automatique). Ce prétraitement consiste à annoter les tokens avec différentes informations linguistiques qui correspondent à notre caractérisation de chacun des traits discursifs *stimuli* présentée en section 3.3.3 (cf. Figure 31, Figure 32 et Figure 33).

Toutes les informations linguistiques ont été obtenues avec la librairie *Stanford CoreNLP*[29], sauf les scores calculés à partir de lexiques. Nous détaillons dans la suite l'ensemble des informations linguistiques extraites.

Figure 31 : Chaine d'analyse des messages pour le trait discursif *Claim*

[29] https://stanfordnlp.github.io/CoreNLP/

Figure 32 : Chaine d'analyse des messages pour le trait discursif *Pédagogie*

Figure 33 : Chaine d'analyse des messages pour le trait discursif *Argumentation*

- ***Token*** : le token tel qu'il apparait dans le texte
- ***POS*** : la catégorie morphosyntaxique du token
- ***Dep*** : la relation de dépendance syntaxique du token
- ***NER*** : la catégorisation du token comme appartenant ou pas à une entité nommée ; c'est une information utilisée spécifiquement pour le claim, ayant observé une utilisation particulièrement importante des entités nommées dans les expressions ayant un caractère factuel

- ***PUNCT*** : indique pour chaque token s'il s'agit d'une ponctuation finale affirmative ou interrogative ; ce trait vise à distinguer le caractère affirmatif des claims

- ***Lex_conc*** : le *score lexical du concret* est basé sur un lexique (Brysbaert et al., 2014) qui attribue un score à près de 40 000 lemmes anglais selon que leur sémantique a trait à quelque chose de concret, 1 étant le degré minimal et 5, le degré maximal. Nous reprenons les scores tels quels, sans contextualisation, sur les tokens lemmatisés des messages, en attribuant *-1* aux lemmes non présents dans le lexique. Ce score est en lien avec l'aspect factuel des claims (cf. section 3.3.3), que nous caractérisons notamment par une évocation ce qui est concret.

- ***Lex_senti*** : le *score lexical de l'émotion* est basé sur un lexique (Warriner et al., 2013) qui attribue trois scores à un ensemble de près de 14 000 lemmes anglais selon leur degré d'évocation de la joie, de l'excitation et de la domination, sur une échelle allant de 1 à 9. Nous utilisons, pour chaque lemme, la moyenne des trois scores, sans contextualisation, avec -1 pour les lemmes absents du lexique. Les

hypothèses sous-jacentes sont que les claims, ayant une forme factuelle, expriment peu d'émotions et inversement pour l'argumentation, qui défend un point de vue.

Pour nos expériences avec de l'apprentissage automatique, nous avons beaucoup utilisé la librairie Scikit-learn[30] (Pedregosa et al., 2011). Elle contient les implémentations d'algorithmes par apprentissage automatique, des méthodes pour l'évaluation et l'exploitation des résultats, ainsi qu'une documentation visant à rendre cet ensemble d'outils accessibles et à illustrer leurs applications.

Pour guider l'apprentissage automatique, nous utilisons différents descripteurs des messages, produits à partir de certaines des informations linguistiques issues du prétraitement.

Beaucoup de nos descripteurs sont basés sur une représentation des messages avec la méthode de calcul TF-IDF[31] (*Term Frequency-Inverse Document Frequency*). Il s'agit d'une pondération des séquences de caractères dans un texte suivant la spécificité de leur présence par rapport à l'ensemble d'un corpus. Nous utilisons la classe *TfidfVectorize*r de Scikit-learn sur les séquences d'unigrammes et bigrammes de tokens, de bigrammes et trigrammes de caractères, ainsi que les étiquettes morphosyntaxiques et de dépendances syntaxiques des séquences de tokens dans chaque message.

Une autre partie de nos descripteurs linguistiques est basée sur une représentation des messages avec des plongements lexicaux[32] (*word embeddings*). Il s'agit d'une représentation vectorielle des mots d'un texte selon leur contexte d'apparition. Nous avons choisi deux modèles de plongements lexicaux en langue anglaise pour leur caractère générique et leurs sources différentes : le modèle *base-uncased*[33], produit par l'algorithme Bert (Devlin et al., 2019), et le modèle *en_vectors_web_lg*[34], produit par l'algorithme Glove (Pennington et al., 2014).

Il y a en plus un descripteur particulier qui correspond à la sortie binaire des règles linguistiques, indiquant, pour chaque message, si le trait discursif visé est présent ou non. C'est une façon d'hybrider les approches par règles et par apprentissage automatique afin de renforcer notre système (Deturck et al., 2019).

Pour la détection automatique des traits discursifs *stimuli* par l'approche symbolique, nous avons défini des règles linguistiques. Ces règles sont fondées sur notre modélisation linguistique de chacun des traits (cf. section 3.3.3). Elles font appel à des informations linguistiques extraites lors du prétraitement des messages. Nous décrivons ci-dessous l'ensemble des règles que nous avons définies, avec des exemples.

[30] https://scikit-learn.org/stable/
[31] https://monkeylearn.com/blog/what-is-tf-idf/
[32] https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa
[33] https://huggingface.co/bert-base-uncased
[34] https://github.com/explosion/spacy-models/blob/master/meta/en_vectors_web_lg-2.3.0.json

- **Règle *Claim 1*** : **une proposition déclarative avec une entité nommée comme argument sujet ou objet du prédicat est un claim**

    - Exemple : « *#ISIS reportedly shot down an #Assad army helicopter.* »

- **Règle *Claim 2*** : **un nom modifié par un complément du nom contenant une entité nommée est un claim** ; la distinction entre les règles 1 et 2 se justifie par la différence de syntaxe.

    - Exemple : « *Clashes between SDF and FSA in Malikia village* »

- **Règle *Claim 3* : une proposition affirmative à la voix passive est un claim**

    - Exemple : « *1 Filipina, Canadians and Norwegian hostages held by #IS Soldiers* »

- **Règle *Claim 4* : une proposition affirmative contenant un verbe modifié par un complément circonstanciel ou un modifieur adverbial est un claim** ; cette règle s'explique par notre définition d'un claim, donné comme une description adoptant une forme factuelle (cf. section 3.3.3) et qui est caractérisée notamment par l'expression de circonstances.

    - Exemple : « *Public react to hearing Qur'an for the first time!* »

- **Règle *Pédagogie 1* : un message contenant un impératif relève de la pédagogie**

    - Exemple : « *Look at the definition* »

- **Règle *Pédagogie 2* : un message contenant un auxiliaire modal relève de la pédagogie** ; nous avons observé que les auxiliaires modaux sont souvent utilisés pour prodiguer des conseils

    - Exemple : « *You may want to check* »

- **Règle *Pédagogie 3* : un message contenant une question relève de la pédagogie** ; poser une question est une façon d'orienter la pensée du destinataire

    - Exemple : « *Did u realize?* »

- **Règle *Pédagogie 4* : un message contenant la construction *need to* relève de la pédagogie** ; c'est une construction que nous avons assez souvent relevée, similaire à la précédente qui utilise des auxiliaires modaux

    - Exemple : « *You need to watch it.* »

- **Règle *Argumentation 1* : une proposition contenant un auxiliaire modal est une *expression d'opinion*** ; l'utilisation d'un auxiliaire modal permet ici d'accentuer le fait qu'il s'agit d'une opinion et non d'un fait certain

    - Exemple : « *It may be good for you.* »

- **Règle *Argumentation 2* : une proposition contenant un attribut du sujet, un adverbe ou un verbe présent dans un lexique de la subjectivité** (Wilson et al., 2005) **est une *expression d'opinion***

    - Exemple : « *That's terrible!* »

- **Règle *Argumentation 3* : une phrase adjacente d'une *expression d'opinion* est une *justification***

    - Exemple : « *That's terrible! He should be here since Saturday!* »

- **Règle Argumentation *4* : une proposition contenant un mot parmi "*so*", "*therefore*", "*because*", "*hence*", "*consequently*", "*since*", contient une *justification***

    - Exemple : « *Because it helps you!* »

- **Règle *Argumentation 5* : une phrase contenant une *expression d'opinion* et une *justification* relève de l'argumentation**

    - Exemple : « *I think it's a good thing because it helps you.* »

- **Règle *Argumentation 6* : une phrase contenant une *expression d'opinion* et une phrase adjacente forment une l'argumentation**

    - Exemple : « *I think it's a good thing. It helps you.* »

Pour implémenter ces règles linguistiques, nous avons utilisé deux modules de la librairie *Stanford CoreNLP*[35]. L'un de ces modules est *Stanford TokensRegex*[36]. Il permet de capturer des séquences de tokens suivant des caractéristiques linguistiques spécifiques, mais il exclut l'information sur les dépendances syntaxiques, pourtant nécessaire à nos règles. C'est pour cela que nous avons aussi utilisé le module *Stanford Semgrex*[37] qui, lui, permet de définir des motifs basés sur les dépendances syntaxiques.

À chaque message, les règles sont appliquées de la numéro 1 à la numéro 4, au niveau de la phrase. Nous avons choisi ce niveau d'analyse parce qu'un claim est nécessairement situé à l'intérieur d'une phrase. Dès qu'une règle détecte un claim dans une phrase, l'analyse s'arrête, même si toutes les règles n'ont pas été appliquées : c'est suffisant pour catégoriser le message comme contenant le trait claim et, pour la segmentation, nous extrayons la phrase entière comme expression de claim, par simplification.

Pour la détection automatique de claims par apprentissage automatique, nous utilisons deux types d'algorithmes selon que nous opérons par segmentation ou pas classification. Pour la segmentation des claims dans les messages, nous avons choisi un algorithme standard pour ce type de tâche, le CRF[38] (*Conditional Ramdom Fields*).

Le CRF est un algorithme d'apprentissage supervisé, il a donc besoin qu'on le guide en enrichissant les données avec des descripteurs. Un modèle de type *CRF* analyse et annote un texte par token, en séquence. Les descripteurs utilisés pour guider l'apprentissage doivent donc être au niveau du token.

Nous utilisons les descripteurs obtenus lors du prétraitement des messages (cf. Figure 31). Nous y ajoutons le trait *RULE*, qui est l'annotation des tokens au format BIO (*Begin*, *Inside*, *Outside*), indiquant s'ils sont au début, à l'intérieur ou à l'extérieur d'une expression de claim, d'après les règles linguistiques précédemment définies. Ce trait permet d'hybrider l'approche symbolique et l'approche par apprentissage automatique, afin de

[35] https://stanfordnlp.github.io/CoreNLP/
[36] https://nlp.stanford.edu/software/tokensregex.html
[37] https://nlp.stanford.edu/software/tregex.shtml
[38] https://towardsdatascience.com/conditional-random-fields-explained-e5b8256da776

voir si leur alliance est bénéfique à la détection des claims. Outre le choix des descripteurs, l'apprentissage dépend aussi du choix des paramètres du CRF.

Pour créer des modèles *CRF*, nous avons utilisé la librairie Wapiti[39]. Nous avons utilisé les valeurs par défaut dans Wapiti, à l'exception du paramètre de régularisation, que nous avons fait varier progressivement dans l'ensemble *{0,00001 ; 0,25 ; 0,5 ; 0,75 ; 1}.*

Pour la classification automatique des messages selon qu'ils contiennent ou pas des claims, nous avons sélectionné quatre algorithmes classiques pour une tâche de classification automatique de textes avec un apprentissage supervisé : un algorithme de la famille SVM, l'algorithme des forêts aléatoires, un autre fondé sur la régression logistique et un perceptron multicouche.

Nous avons utilisé les implémentations de ces quatre types d'algorithmes dans la librairie Scikit-learn. Concernant SVM, qui est une famille d'algorithmes, nous avons choisi l'implémentation *SVC*[40] parce qu'elle est destinée à faire de la classification et possède relativement peu de paramètres. Nous avons aussi utilisé Scikit-learn pour l'optimisation des données et du paramétrage des algorithmes.

L'optimisation des données consiste, d'une part, en la normalisation des valeurs numériques des descripteurs, avec l'algorithme *StandardScaler*[41], de façon à en faciliter l'apprentissage par les algorithmes.

L'optimisation des données consiste d'autre part en une sélection des descripteurs les plus significatifs pour la classification des messages. Pour ce faire, nous utilisons l'algorithme *SelectKBest*[42] qui conserve les *k* descripteurs les plus significatifs pour la classification des messages, d'après l'analyse de leur variance avec des tests statistiques lors de l'apprentissage. La valeur de *k* entre dans le cadre de l'optimisation des paramètres algorithmiques.

L'optimisation du paramétrage des algorithmes consiste à trouver la combinaison des valeurs de paramètres qui produit les meilleurs résultats. Nous réalisons cette optimisation indépendamment pour chacun des quatre algorithmes sélectionnés, en incluant celle du paramétrage de la valeur *k* pour l'algorithme *SelectKBest*.

Pour la recherche des paramètres optimaux, nous avons utilisé l'algorithme *GridSearchCV*[43] avec une validation croisée (une méthode d'évaluation combinant différentes itérations) sur les données d'entrainement. L'algorithme consiste à identifier le meilleur paramétrage pour l'apprentissage en parcourant une grille de valeurs associées à un ensemble de paramètres, ces paramètres et ces valeurs étant sélectionnés par l'utilisateur.

[39] https://wapiti.limsi.fr

[40] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[41]https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html#sklearn.preprocessing.StandardScaler

[42] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

[43] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Dans le Tableau 70, nous présentons l'ensemble des valeurs parcourues pour l'optimisation du paramètre *k* de *SelectKBest*. Nous avons choisi cet ensemble de valeurs pour qu'il soit varié sans être trop grand, pour limiter la durée de l'optimisation.

Tableau 70 : Espace de recherche pour l'optimisation du paramètre de *SelectKBest*

Paramètre	Grille des valeurs
k	[5, 10, 20, 30]

Le Tableau 71 indique les paramètres et valeurs que nous avons sélectionnés pour l'optimisation du paramétrage de chacun des algorithmes de classification. Nous avons choisi les paramètres les plus généraux et sélectionné des ensembles de valeurs avec les mêmes critères d'efficience que pour l'optimisation de *SelectKBest*. Les paramètres non mentionnés sont réglés sur leur valeur par défaut dans Scikit-learn.

Tableau 71 : Espace de recherche pour l'optimisation des paramètres algorithmiques

	Paramètre	Grille des valeurs
SVM	*C*	[0,00001 ; 1 ; 1000]
	kernel	Toutes les valeurs permises
	gamma	Toutes les valeurs permises
Forêt aléatoire	*n_estimators*	[10, 100, 500]
	criterion	Toutes les valeurs permises
	max_features	Toutes les valeurs permises
Régression logistique	*C*	[0,00001 ; 1 ; 1000]
	penalty	[l1, l2]
	solver	[liblinear, saga]
Perceptron multicouche	*max_iter*	[1000, 5000, 10 000]
	activation	Toutes les valeurs permises
	solver	Toutes les valeurs permises

Pour évaluer les règles et modèles par apprentissage, nous avons utilisé trois mesures d'évaluation classiques en TAL : la *précision*, le *rappel* et la *F-Mesure*. Ces trois mesures permettent d'évaluer, selon des modalités respectives que nous allons détailler, la qualité de l'annotation automatique par rapport aux annotations de référence.

La précision indique le bruit présent dans les résultats de l'annotation automatique en calculant la proportion d'annotations correctes parmi les annotations produites par le système. Le rappel indique le silence de l'annotation automatique en calculant la proportion des annotations attendues qui ont été trouvées par le système. Enfin, la F-mesure est une combinaison des deux mesures précédentes.

$$Précision\ (P) = \frac{|\ corrects \wedge trouvés|}{|trouvés|}$$

$$Rappel\ (R) = \frac{|\ corrects\ \wedge trouvés|}{|corrects|}$$

$$\text{F-mesure}\ = \frac{2 * P *\ R}{P + R}$$

Dans la section suivante, nous présentons les résultats de l'évaluation des règles linguistiques et des différentes configurations de modèles par apprentissage automatique.

6.2.4. Résultats

Dans le Tableau 72, nous donnons les résultats de l'évaluation des règles linguistiques pour la segmentation des claims sur la partition *Éval1* de la micro-référence et sur la partition *Éval1* de la référence complète (cf. section 6.2.2). Nous faisons remarquer que, pour le calcul de ces valeurs, nous incluons dans les annotations correctes les segments textuels qui ne correspondent que partiellement à un segment de référence : notre objectif est plus de détecter la présence de claims dans les messages que leurs frontières textuelles exactes.

Tableau 72 : Résultats de l'évaluation des règles linguistiques pour la segmentation des claims

Données / Mesure d'évaluation	Précision	Rappel	F-mesure
Micro-Éval1	0,26	0,71	0,38
Total-Éval1	0,50	0,84	0,63

Les résultats montrent que les règles linguistiques ont globalement un bon rappel mais une faible précision. Elles ont donc tendance à engendrer du bruit, c'est-à-dire que beaucoup des annotations qu'elles produisent ne correspondent pas à des claims dans la référence.

Ce bruit s'explique en grande partie par une définition trop extensive de nos règles. La règle 4 (cf. section 6.2.3) apporte beaucoup de bruit parce qu'elle inclut tous les modifieurs adverbiaux. La présence de subjectivité, exclue dans les claims, déjà difficile à

distinguer pour des annotateurs humains (cf. section 6), l'est d'autant plus pour un système d'annotation automatique.

Le bruit peut être atténué en restreignant ou en enrichissant la définition des règles à partir des types d'erreurs relevés. Le fait que les performances s'améliorent en augmentant la taille des données d'évaluation est positif quant à leur robustesse.

Nous allons maintenant voir les résultats du CRF pour cette même tâche de segmentation des claims. Les modèles ont été entrainés sur la partition *Développement* de la micro-référence et évalués sur les mêmes données que les règles, à savoir la partition Éval*1* de la micro-référence.

Pour le CRF, nous utilisons l'outil d'évaluation de Wapiti, qui calcule la précision et le rappel en considérant comme annotations correctes seulement les segments strictement identiques à la référence, contrairement à l'évaluation des règles qui inclut les segments partiellement identiques.

Tableau 73 : Évaluation du CRF pour la segmentation des *claims*

Descripteurs	*Régularisation L2*		***Précision***		***Rappel***		***F-mesure***	
	Micro	*Total*	*Micro*	*Total*	*Micro*	*Total*	*Micro*	*Total*
Token – POS – Dep - NER – Lex_conc	1	0,00001	0,56	0,50	**0,27**	**0,32**	**0,36**	**0,39**
Token – POS – NER – Lex_conc	0,5	0,25	0,54	0,52	0,26	0,30	0,35	0,38
Token – POS - NER	0,5	0,00001	0,55	0,51	0,26	0,29	0,35	0,37
Token – POS	0,00001		0,72	**0,55**	0,22	0,26	0,34	0,35
Token – POS – Dep – Lex_conc	1	0,00001	0,59	0,49	0,24	**0,32**	0,34	**0,39**
Token	0,00001		0,41	0,47	0,25	0,25	0,31	0,33
Token – POS – Dep - Lex_conc – Lex_senti	0,5	N/A	0,45	N/A	0,23	N/A	0,31	N/A
Token – POS – Dep	1	0,25	**0,75**	0,52	0,19	0,26	0,30	0,35
Token – POS – Dep - NER	0,00001		0,49	0,52	0,21	0,29	0,29	0,37
Token – POS – Dep – NER – Rules	0,00001	N/A	0,36	N/A	0,23	N/A	0,28	N/A

Dans le Tableau 73, nous présentons les meilleurs résultats avec chaque descripteur, en termes de F-mesure, dans l'ordre décroissant ; seul le descripteur *PUNCT* n'est pas présent car les combinaisons qui l'incluaient ont toutes produit 0 de F-mesure.

Nous présentons aussi les résultats de combinaisons alternatives, afin de voir l'apport de chaque descripteur. Nous n'avons pas évalué les descripteurs *Lex_senti* et *Rule* sur le corpus *Total* parce que nous avons observé des résultats moins bons lorsqu'ils sont ajoutés lors de l'évaluation sur la micro-référence.

Le meilleur résultat global (0,36 et 0,39 de F-mesure) est obtenu par la combinaison de descripteurs classiques en TAL (*Token*, *POS*, *Dep* et *NER*) et du descripteur spécifique au claim : *Lex_conc*. L'ajout de ce descripteur à la combinaison *Token – POS – Dep - NER* fait gagner 7% de *F-Mesure* et 4% par rapport à la combinaison *Token – POS – Dep*. C'est un résultat qui semble valider notre hypothèse selon laquelle les claims ont un vocabulaire ayant particulièrement trait aux choses concrètes.

Globalement, le CRF est meilleur en précision qu'en rappel. Nous obtenons jusqu'à 75% de précision avec la combinaison *Token – POS – Dep* mais le rappel est faible (30%). Un tel système peut être acceptable. Par exemple, pour lutter contre la radicalisation, détecter une petite partie des influenceurs peut être intéressant, en particulier dans un environnement restreint, où l'impact est probablement plus important. Il en est de même pour le cas du marketing, où il peut être suffisant d'accéder à quelques influenceurs.

L'inclusion du descripteur *Rule* ne fait pas gagner en performance globale. En regardant la différence entre les combinaisons *Token – POS – Dep - NER* et *Token – POS – Dep – NER – Rule*, nous constatons une baisse de précision (13%) avec l'ajout du descripteur *Rule*, ce qui s'explique par la mauvaise précision des règles linguistiques, que nous avions observée précédemment. Nous relevons tout de même une légère amélioration du rappel, ce qui est en accord avec le rappel particulièrement élevé des règles.

Il est remarquable que le descripteur *NER* produise si peu de gain de performance alors que ce descripteur a trait aux entités nommées, pourtant prédominantes dans notre caractérisation linguistique du claim. En comparant les résultats, il semble que le CRF fonctionne relativement bien avec des données sur la forme des tokens et que l'information sémantique agit en complément pour améliorer la détection, comme entre les combinaisons *Token-POS* et *Token-POS-NER*.

Le passage de la micro-référence vers des données beaucoup plus grandes fait globalement gagner en rappel mais il fait perdre en précision. L'apprentissage a appris sur plus de données, il parvient donc à détecter plus de cas, mais la plus grande généralisation des modèles conduit aussi à détecter des faux positifs.

Il est logique de constater que, pour éviter le surapprentissage, le facteur de régularisation est globalement plus élevé avec le jeu de données le plus petit car il est plus facilement *appris* par l'algorithme.

Nous allons maintenant voir les résultats d'évaluation des règles linguistiques et de l'apprentissage automatique, sur les données de référence complètes, pour la tâche de classification des messages, selon qu'ils contiennent ou pas un trait discursif *stimulus*.

Nous commençons, comme précédemment, avec les résultats d'évaluation des règles linguistiques pour la détection des claims (cf. Tableau 74).

Tableau 74 : Résultats d'évaluation des règles linguistiques *Claim* pour la classification

Précision	Rappel	F-mesure
0,44	0,85	0,48

Comme pour l'évaluation des règles linguistiques destinées à la segmentation des claims, nous observons une balance précision-rappel grandement favorable au rappel, la précision étant encore faible. Nous pensons encore que c'est une définition trop extensive des règles qui engendre ce fort déséquilibre. L'analyse des erreurs permet cependant de nuancer ce résultat : le système par règles détecte des claims qui n'étaient pas présents dans la référence, ce qui fait faussement baisser sa précision.

Le défaut d'exhaustivité de la référence ne signifie pas pour autant qu'elle est majoritairement mal annotée et qu'elle n'est donc pas fiable pour l'évaluation. Cela ne remet pas non plus en cause la méthode utilisée pour sa constitution : la réconciliation inter-annotateurs comme la super-annotation visaient à choisir les annotations correctes parmi celles déjà présentes et non à en trouver des nouvelles.

Dans le Tableau 75, nous présentons les résultats d'évaluation de l'apprentissage automatique pour la classification des messages des deux partitions d'évaluation, selon qu'ils contiennent au moins un claim (la classe positive) ou aucun (la classe négative).

Les scores sont supérieurs à ceux obtenus avec les règles linguistiques et tous sont au-dessus de la valeur médiane 0,5. La balance précision-rappel est plus équilibrée. Ces résultats sont obtenus avec uniquement des descripteurs issus des plongements lexicaux Glove, sélectionnés par *SelectKBest* pour leur significativité. Ces observations induisent que l'apprentissage automatique parvient à détecter relativement bien les messages comportant des claims, tels qu'annotés dans la référence.

Tableau 75 : Résultats d'évaluation de l'apprentissage pour la classification en claim

Algorithme / Mesure	Précision		Rappel		F-mesure	
	Éval1	*Éval2*	*Éval1*	*Éval2*	*Éval1*	*Éval2*
SVM	0,63	0,69	0,64	0,68	0,63	0,69
Régression logistique	0,58	0,65	0,64	0,68	0,61	0,67
Perceptron multicouche	**0,64**	0,68	**0,65**	0,68	**0,65**	0,68
Forêt aléatoire	0,62	**0,71**	0,61	**0,77**	0,61	**0,74**

En comparant les résultats entre les deux partitions d'évaluation, nous constatons qu'ils sont globalement meilleurs sur *Éval2* que sur *Éval1* et les algorithmes produisant les meilleurs résultats sont différents d'une partition à l'autre. Cela peut signifier que les deux partitions contiennent des représentations différentes du trait *Claim* et que les claims présents dans *Éval2* sont plus ressemblants à ceux du corpus d'apprentissage.

Dans le Tableau 76, nous donnons les résultats d'évaluation des règles linguistiques pour la classification des messages en pédagogie. Comme pour le trait *Claim*, les règles produisent une précision faible et un fort rappel. C'est encore une fois dû à une trop grande extensivité de leur définition.

Tableau 76 : Résultats d'évaluation des règles linguistiques *Pédagogie* pour la classification

Précision	Rappel	F-mesure
0,44	0,85	0,48

Dans le Tableau 77, nous présentons les résultats d'évaluation de l'apprentissage automatique pour la classification en pédagogie. Les résultats sont globalement moins bons que pour le trait *Claim*. Nous avions remarqué précédemment que les données de référence pour le claim étaient de meilleure composition que pour les deux autres traits *stimuli*.

Tableau 77 : Résultats d'évaluation de l'apprentissage pour la classification en pédagogie

Algorithme / Mesure	Précision		Rappel		F-mesure	
	Éval1	*Éval2*	*Éval1*	*Éval2*	*Éval1*	*Éval2*
SVM	0,28	0,20	0,73	0,60	0,41	0,30
Régression logistique	0,32	**0,26**	**0,80**	**0,73**	**0,45**	**0,39**
Perceptron multicouche	0,30	0,22	0,20	0,33	0,24	0,26
Forêt aléatoire	**0,67**	0,20	0,27	0,13	0,38	0,16

Contrairement à l'évaluation sur la détection des claims, les performances globales baissent en passant de la partition *Éval1* à la partition *Éval2*, ce qui signifie que les deux partitions d'évaluation ont des constitutions différentes et que la seconde a une représentation de la pédagogie qui correspond moins aux données d'apprentissage.

La précision globale des algorithmes est particulièrement faible (en-dessous de 50%). Ainsi, le système n'est pas assez fiable parce qu'il a tendance à produire trop de bruit. Cela

peut être dû à une difficulté pour formaliser les expressions de pédagogie à partir des descripteurs que nous avons fournis.

Les descripteurs qui ont été conservés par l'algorithme *SelectKBest* comme étant les plus significatifs sont cette fois-ci de différents types. Nous présentons ci-dessous un classement des types de descripteurs retenus suivant l'ordre décroissant de leurs poids.

- *TF-IDF_POS* : 54
- Plongements lexicaux BERT : 41
- *TF-IDF_Dep* : 37
- Plongements lexicaux Spacy : 36

Tableau 78 : Résultats d'évaluation des règles linguistiques *Argumentation* pour la classification

Précision	Rappel	F-mesure
0,19	0,90	0,31

Dans le Tableau 78, nous présentons les résultats des règles linguistiques pour la classification des messages en argumentation. Comme avec l'évaluation des règles concernant les traits discursifs précédents, nous avons un rappel fort et une faible précision, ici de façon plus accentuée. Pour la détection du trait *Argumentation* aussi, les règles linguistiques sont trop extensives.

Tableau 79 : Résultats d'évaluation de l'apprentissage pour la classification en argumentation

Algorithme / Mesure	Précision		Rappel		F-mesure	
	Éval1	*Éval2*	*Éval1*	*Éval2*	*Éval1*	*Éval2*
SVM	0,54	0,30	**0,70**	**0,75**	0,61	0,43
Régression logistique	0,50	0,23	**0,70**	0,62	0,58	0,33
Perceptron multicouche	0,62	0,42	0,50	**0,75**	0,56	**0,64**
Forêt aléatoire	**0,83**	**0,50**	0,50	0,25	**0,62**	0,33

Dans le Tableau 79, nous présentons les résultats d'évaluation de l'apprentissage automatique pour la classification des messages en argumentation. Ces résultats sont, comme ceux sur la pédagogie, moins bons que les résultats sur le trait *Claim*. Nous l'expliquons là aussi par une moins bonne qualité des annotations de référence. Les

résultats sont tout de même globalement meilleurs que ceux pour la pédagogie, en particulier au regard de la précision. Peut-être que les annotations sur l'argumentation sont plus cohérentes que celles concernant la pédagogie.

Les performances des algorithmes sont de différents ordres : la forêt aléatoire se distingue par sa précision tandis que les trois autres algorithmes parviennent à détecter plus de cas (ayant un rappel globalement plus élevé).

Comme précédemment pour la détection de la pédagogie, nous remarquons des performances globalement à la baisse en passant de *Éval1* à *Éval2*. Cela semble confirmer une différence de constitution entre les deux partitions, avec, dans ce cas aussi, une représentation du trait visé, en l'occurrence l'argumentation, qui est moins fidèle dans *Éval2* à la représentation établie lors de l'apprentissage. Le perceptron multicouche se distingue tout de même avec un gain de rappel important (25%) de *Éval1* à *Éval2*, peut-être par une plus grande capacité de généralisation.

6.2.5. Conclusion

Nous avons expérimenté les deux principales approches pour faire de l'annotation automatique des traits discursifs *stimuli*, à savoir le développement de règles linguistiques et celui de modèles par apprentissage automatique.

Nous avons constaté que les règles linguistiques sont globalement définies de façon trop extensive car elles engendrent trop de bruit. Il faudrait améliorer leur précision d'après une analyse des erreurs et réévaluer l'utilisation de leurs sorties en tant que descripteurs des messages pour l'approche par apprentissage automatique. Cette dernière permet d'obtenir des résultats plus acceptables, en particulier concernant le trait *Claim*, dont les données de référence ont une meilleure constitution au regard des volumes annotés et des accords inter-annotateurs.

La détection de la pédagogie est encore trop peu fiable en termes de précision, quelle que soit l'approche. À cet égard, l'amélioration des règles linguistiques sur ce trait discursif pourrait constituer un levier de progression important.

6.3. Détection du *changement d'avis* (*Décision*)

6.3.1. Introduction

Le *changement d'avis* correspond au trait discursif de la composante de *décision* dans notre modèle de l'influence individuelle (cf. section 3.3.3). L'objectif de cette partie de

notre travail est de développer un module qui détecte automatiquement la présence d'une expression de changement d'avis (cf. section 3.2.4) dans des messages de médias sociaux.

Nous formalisons ce travail comme une tâche de classification binaire consistant à distinguer les messages qui contiennent un changement d'avis de ceux qui n'en contiennent pas. Nous n'avons pas eu le temps de développer des règles linguistiques sur l'expression de changement d'avis, nous avons donc procédé uniquement avec une approche par apprentissage automatique.

Nous avons choisi un apprentissage supervisé afin d'avoir une prise sur les caractéristiques linguistiques utilisées. Ce type d'apprentissage requiert un corpus contenant des exemples de référence pour la classification visée.

Nous avons choisi le forum *Change my view* comme ressource pour ce travail parce que son principe repose justement sur le changement d'avis. De plus, le règlement de ce forum impose un format particulier aux messages exprimant un changement d'avis, avec notamment la présence d'un marqueur delta ; cela constitue une annotation de référence *ad hoc*, nécessaire au développement et à l'évaluation de notre module. Dans la section suivante, nous allons voir la méthode suivie pour mettre en œuvre les expériences.

6.3.2. Méthodologie

Nous avons utilisé le jeu de données *Changement d'avis* du corpus *Change my view* (cf. section 4.2) : il résulte d'un premier filtrage imposant un minimum de participation par discussion (au moins dix participants et une réponse de l'auteur initial) et d'un second filtrage qui ne garde que les discussions contenant au moins un changement d'avis.

Nous utilisons comme référence l'annotation *ad hoc* des messages contenant un changement d'avis suffisamment explicité, faite par le robot modérateur du forum. Les annotations de la campagne d'annotation n'avaient pas encore été réalisées au moment de cette partie du travail.

10% des messages du corpus contiennent un changement d'avis (la classe positive). Ce déséquilibre dans la représentation des deux classes constitue un biais : une bonne classification pourrait être obtenue par chance en faveur de la classe majoritaire. Nous verrons que nous avons fait en sorte de contrebalancer ce biais par le paramétrage de l'apprentissage. Étant donné que le changement d'avis concerne principalement les auteurs initiaux, nous avons choisi d'analyser seulement leurs messages.

Pour créer un modèle par apprentissage capable de détecter automatiquement les messages qui expriment un changement d'avis, nous avons sélectionné les traits linguistiques que nous présentons ci-dessous.

- ***Nombre de tokens*** : c'est le trait le plus simple (notre *baseline*), il caractérise chaque message par le nombre de tokens qu'il contient ; cela permet de voir si la

taille des messages est un trait distinctif des messages contenant un changement d'avis

- ***Sacs de mots*** : représente chaque message d'après le nombre d'occurrences qu'il contient de chaque token recensé dans le corpus ; ce descripteur sert à distinguer le lexique spécifique à l'expression d'un changement d'avis

- ***Catégories morphosyntaxiques*** : il s'agit de calculer le nombre d'occurrences des catégories morphosyntaxiques présentes avec les tokens de chacun des messages ; c'est une information sur le contexte d'utilisation des tokens

- ***Usage du passé*** : c'est un trait binaire qui indique pour chaque message s'il contient au moins un verbe au passé, partant de l'hypothèse que lorsqu'un individu exprime un changement d'avis, il fait notamment un retour sur son précédent point de vue

- ***Connotation lexicale*** : un ensemble de descripteurs d'un message qui évaluent (1) sa subjectivité, en lien avec le caractère personnel de l'expression d'un changement d'avis, et (2) son inclusion d'éléments concrets, visant la factualité, avec pour (1), la proportion de pronoms personnels de la première personne ainsi que la moyenne, pour tous les lemmes, de leurs scores d'évocation de la joie, de l'excitation et de la domination (Warriner et al., 2013), et pour (2), la moyenne de leurs scores concernant l'évocation de choses perceptibles (Brysbaert et al., 2014)

Afin d'implémenter un classifieur en changement d'avis, fondé sur ces traits linguistiques, nous avons choisi l'algorithme d'apprentissage par régression logistique car il convient bien à notre tâche de classification et il est assez facile à mettre en place. Nous avons utilisé l'implémentation *LogisticRegressionCV*[44] (Scikit-learn), qui inclut l'optimisation des paramètres de régularisation avec une validation croisée. Nous avons laissé les valeurs des paramètres par défaut, hormis celles listées ci-dessous :

- *class_weight* : "*balanced*" ; pour contrebalancer le biais dans nos données lié au déséquilibre dans la présence des deux classes, donne un poids d'autant plus grand aux erreurs dans une classe que celle-ci est minoritaire,
- *scoring* : "*roc_auc*" ; mesure d'évaluation concernant la validation croisée pour l'optimisation des paramètres de l'algorithme, identique à celle utilisée pour l'évaluation finale du classifieur et qui sera décrite dans la suite,
- *solver* : "*sag*" ; pour une convergence de l'apprentissage plus rapide,
- *tol* : 0.001 ; idem que ci-dessus,
- *max_iter* : 500 ; pour que l'apprentissage parvienne à converger,

[44]https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegressionCV.html

- *random_state* : 0 ; valeur fixée pour la reproductibilité des résultats.

Pour évaluer notre modèle, nous avons choisi d'utiliser la mesure *Area Under the Receiver Operating Characteristic Curve* (*ROC AUC*), implémentée dans la librairie Scikit-learn[45]. Elle convient bien à l'évaluation d'un classifieur binaire parce qu'elle se focalise sur la distinction entre la classe positive et la classe négative, en comparant le taux de vrais positifs par rapport au nombre de messages dans la classe positive et le taux de faux positifs par rapport au nombre de messages dans la classe négative (Majnik & Bosnić, 2013).

Afin d'éviter un biais d'apprentissage lié au formalisme de *Change my view*, qui impose la présence du symbole *delta* dans les messages exprimant un changement d'avis, nous avons supprimé ce symbole de l'ensemble des messages, sous les formes *delta*, ***Δ*** et *∆*. Nous avons aussi supprimé les tokens *award* et *awarded* qui peuvent leur être associés.

Nous associons à chaque message un attribut *True* ou *False* qui indique respectivement si le message contient ou pas un changement d'avis. Afin d'évaluer la pertinence de chaque descripteur, nous créons des modèles qui les utilisent séparément. Nous présentons les résultats de l'évaluation dans la section suivante.

6.3.3. Résultats

Tableau 80 : Résultats de la classification par descripteur linguistique

Descripteur	*Nombre de tokens*	*Sacs de mots*	*POS*	*Connotation lexicale*	*Passé*
Score AUC (%)	**51,38**	**82,11**	**60,97**	**64,70**	**57,15**

Dans le tableau 80, nous présentons l'ensemble des scores *AUC* obtenus pour chacun des descripteurs sélectionnés. En plus de ces scores, nous avons relevé le poids des différentes informations extraites dans les messages pour chaque descripteur afin de distinguer la classe positive (les messages désignés comme contenant une expression de changement d'avis) et la classe négative (les messages désignés comme ne contenant pas une expression de changement d'avis).

Nous obtenons le meilleur résultat avec la représentation des messages en *sacs de mots* (82,11%). Parmi les tokens les plus discriminants pour la classe positive, il y a « *convinced* » qui dénote bien la réaction à un argument jugé convaincant et qui peut

[45] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

conduire à un changement d'avis. Il y a des termes de concession qui marquent un fléchissement dans la pensée de l'auteur (« *concede* », « *still* »). Le token « *hadn* » dénote quant à lui une évocation du passé. Cela rejoint notre hypothèse de l'importance de l'emploi du passé dans l'expression d'un changement d'avis, corroborée par le résultat du descripteur dédié (57,15%).

Le descripteur *Connotation lexicale* se classe en deuxième position avec son score de 64,70%. Parmi cet ensemble de traits, la proportion des pronoms personnels de première personne arrive en tête pour distinguer les messages exprimant un changement d'avis, ce qui valide notre hypothèse d'une subjectivité importante dans ce type d'expression. La classe négative est quant à elle distinguée par des tokens évoquant l'excitation et des choses concrètes, ce qui semble bien correspondre à des messages de débat plutôt qu'à une résolution, se manifestant par l'expression d'un changement d'avis.

Le descripteur *POS*, fondé sur la fréquence des catégories morphosyntaxiques dans les messages, se positionne en troisième position, avec un score de 60,97%. Les traits discriminants pour la classe positive sont majoritairement les pronoms de la première personne, ce qui corrobore l'hypothèse de la subjectivité dans l'expression d'un changement d'avis. La classe négative est quant à elle caractérisée par l'emploi de coordonnants et d'interjections, ce qui semble correspondre à des messages prenant part à un débat plutôt qu'à une résolution par l'expression d'un changement d'avis.

Le trait *POS* dépasse, comme attendu, le trait de base sur le nombre de tokens qui obtient un score de 51,38%. Cela montre que le nombre de tokens dans les messages n'est pas pertinent pour distinguer les expressions de changement d'avis.

En conclusion, les résultats montrent qu'une représentation des messages fondée sur des traits génériques tels que les *sacs de mots* sont davantage pertinents pour distinguer la présence d'une expression de changement d'avis avec une approche par apprentissage automatique. Ils semblent aussi valider nos hypothèses sur la présence particulière de la subjectivité et de l'emploi du passé dans ce type d'expression. Une perspective à cette étude consiste à combiner les descripteurs, ici utilisés séparément.

7. Détection hybride des influenceurs

7.1. Introduction

Dans cette partie de la thèse, nous décrivons l'approche que nous avons mise en œuvre pour implémenter et évaluer notre modèle hybride destiné à la détection d'influenceurs dans les médias sociaux (cf. section 3.4). Le principe de ce modèle est de mesurer, avec la centralité, l'influence globale d'individus représentés dans un graphe social qui intègre l'information sur les traits discursifs d'influence (cf. section 3.3.3 et 3.3.4) éventuellement contenus dans leurs échanges.

Le but est d'affiner la mesure d'influence globale par centralité (cf. section 3.2) avec les informations linguistiques correspondant à notre modèle d'influence individuelle (cf. section 3.3).

Dans la section 7.2, nous présentons les données de médias sociaux que nous avons sélectionnées pour l'hybridation, la représentation que nous avons conçue pour intégrer l'information sur les traits discursifs d'influence à l'intérieur d'un graphe, les mesures de centralité que nous avons appliquées aux graphes résultants ainsi que le protocole suivi pour évaluer le système hybride. Nous analysons les résultats de l'évaluation en section 7.3 et nous concluons sur l'apport de l'hybridation pour détecter les influenceurs en section 7.4.

7.2. Méthodologie

7.2.1. Données d'évaluation

Pour implémenter et évaluer notre modèle hybride de détection des influenceurs, nous avons choisi d'utiliser le jeu de données *Change my view* (Tan et al., 2016) décrit en section 4.2. Il contient des discussions dont le principe pour un participant est d'envoyer des messages pour faire changer d'avis autrui. Cela correspond bien à l'influence individuelle (cf. section 3.3) que nous voulons détecter avec nos traits discursifs. L'information sur l'échange des messages entre les utilisateurs est adéquate pour une représentation en graphe et l'annotation *ad-hoc* des influenceurs est pratique pour évaluer notre système.

Nous avons utilisé 50 fils de discussion d'un sous-ensemble filtré de ce jeu de données (le corpus *Changement d'avis* dans la section 4.2) : nous rappelons que ce filtrage ajoute

des critères de pertinence sur les fils de discussion conservés (cf. section 4.2) et leur quantité restreinte encadre le temps du prétraitement des messages en vue de leur analyse avec les descripteurs linguistiques. Aussi, les données utilisées pour l'hybridation sont différentes de celles destinées à l'apprentissage des classifieurs sur les traits discursifs afin d'éviter des biais.

Dans le Tableau 81, nous présentons des statistiques générales concernant le jeu de données finalement utilisé pour les expériences sur l'hybridation. Nous remarquons en particulier le faible taux d'influenceurs, qui exige un système particulièrement précis. En complément, le Tableau 82 donne des statistiques sur la participation par discussion. Elles indiquent une activité globalement soutenue, ce qui est un signe positif pour la diversité des observations possibles dans le jeu de données.

Tableau 81 : Statistiques générales sur le jeu de données pour l'hybridation

Nombre de discussions	Nombre de messages	Nombre de participants	Pourcentage d'influenceurs
50	4664	1435	5%

Tableau 82 : Statistiques sur la participation dans les discussions pour l'hybridation

Nombre médian de messages par discussion	Nombre médian de participants par discussion	Pourcentage d'influenceurs par discussion
61	27	5%

Nous allons décrire la méthode utilisée pour implémenter et évaluer notre modèle hybride. La première étape a consisté à appliquer des classifieurs selon nos traits discursifs d'influence à l'ensemble des messages du jeu de données. Nous n'avons pas eu le temps d'implémenter et d'évaluer un système de détection des traits de *stimulation* (cf. section 3.3.4), ces derniers sont donc absents des expériences sur l'hybridation.

7.2.2. Développement des classifieurs sur les traits discursifs

Pour implémenter les classifieurs en traits discursifs, nous avons sélectionné, parmi les combinaisons d'algorithmes et de descripteurs linguistiques que nous avions précédemment expérimentées (cf. sections 6.2 et 6.3), celles qui ont produit les meilleurs résultats pour chaque trait, à condition que ces résultats nous aient semblé suffisants.

Parmi les traits discursifs dont nous avons évalué la détection automatique, nous avons évincé la pédagogie parce que c'est le seul trait pour lequel la meilleure précision est en

dessous de 50%, ce que nous avons jugé insuffisant pour une utilisation dans la chaine de traitement que constitue l'hybridation. Nous rappelons ci-dessous les configurations qui avaient produit les meilleurs résultats pour les autres traits discursifs.

- ***Claim*** : Forêt aléatoire avec des plongements lexicaux Glove
- **Argumentation** : Perceptron multicouches avec des plongements lexicaux Bert et TF-IDF sur les bigrammes et trigrammes de caractères ainsi que les unigrammes et bigrammes de tokens
- **Changement d'avis** : Régression logistique avec la représentation *Sacs de mots*

Dans le Tableau 83, nous rappelons les meilleurs résultats d'évaluation que nous avons obtenus sur les classifieurs concernant chaque trait discursif, avec les configurations décrites précédemment (cf. sections 6.2.4 et 6.3.3).

Tableau 83 : Évaluation des classifieurs sur les traits discursifs retenus pour l'hybridation

Mesure	Précision		Rappel		F-mesure		Score AUC
Évaluation	Éval. 1	Éval. 2	Éval. 1	Éval. 2	Éval. 1	Éval. 2	Éval. unique
Claim	0,62	0,71	0,61	0,77	0,61	0,74	N/A
Argumentation	0,62	0,42	0,50	0,75	0,56	0,54	N/A
Changement d'avis	N/A						0,82

Nous avons implémenté un classifieur avec la meilleure configuration évaluée pour chaque trait discursif, en utilisant un paramétrage des algorithmes d'apprentissage similaire à celui utilisé lors des expériences sur la détection des traits discursifs correspondants (cf. sections 6.2 et 6.3). La seule différence est le nombre maximum d'itérations *max_iter* de la régression logistique pour la détection du changement d'avis, augmenté à 1000 afin de permettre la convergence de l'apprentissage sur un nombre de messages plus grand.

Le choix des données d'apprentissage (cf. Tableau 84) s'est fait de façon à intégrer un maximum d'exemples de chaque classe pour un meilleur apprentissage. Nous avons aussi fait en sorte d'évincer les 50 fils de discussion destinés aux expériences sur l'hybridation, de façon à éviter des biais.

Tableau 84 : Données d'apprentissages des classifieurs en traits discursifs

Trait / Statistique	Nombre de messages	Pourcentage de messages positifs
Claim	1126	45%
Argumentation	716	7%
Changement d'avis	83 555	2%

Pour les traits *Claim* et *Argumentation*, nous avons utilisé les données produites lors de la campagne d'annotation (cf. section 6.1). Pour le changement d'avis, nous avons

utilisé une partie des messages du jeu de données *Changement d'avis* provenant du forum *Change my view* (cf. section 4.2), qui inclut l'annotation *ad hoc* des messages porteurs d'un changement d'avis suffisamment explicité, faite par le robot modérateur du forum. Nous n'avons utilisé qu'une partie des messages pour faciliter la convergence de l'apprentissage. Nous n'avons pas exploité les annotations manuelles parce qu'elles sont peu nombreuses et bruitées (cf. section 6.1.6).

La deuxième étape du protocole d'hybridation est celle de la construction d'un graphe social qui intègre les informations sur les traits discursifs d'influence, conformément à notre modèle hybride (cf. section 3.4).

7.2.3. Construction du graphe social

Dans notre système hybride, le graphe social représente la transmission de *stimuli* entre les individus, pondéré d'après les réactions engendrées par chacun des *stimuli* chez les individus cibles. Cette pondération est croissante depuis les *stimuli* n'ayant engendré aucune réaction jusqu'à ceux ayant produit un changement d'avis, en passant par les *stimuli* ayant entrainé une réaction autre qu'un changement d'avis.

Nous avons construit ce graphe à partir de la classification des messages selon les traits discursifs sélectionnés précédemment. Pour mieux évaluer l'apport des différentes composantes de l'hybridation, nous avons choisi de créer aussi une version intermédiaire de ce graphe, avec seulement les informations sur les *stimuli*, sans la pondération d'après les réactions.

En plus du graphe correspondant à notre modèle hybride, pour une évaluation comparative, nous avons aussi construit un graphe *baseline* : c'est un graphe orienté et simple (au maximum un arc orienté d'un nœud à un autre) qui représente la transmission de messages, quels qu'ils soient, entre les individus. Il y a un arc orienté depuis un nœud représentant un individu *Orig* vers un nœud représentant un individu *Dest* si l'une des deux conditions suivantes est vérifiée :

- l'individu *Orig* a envoyé un message à l'individu *Dest*,
- l'individu *Dest* a répondu à un message de l'individu *Orig*.

7.2.4. Sélection des mesures de centralité

Les graphes obtenus sont utilisés dans la troisième étape, qui est celle de la mesure de centralité des individus représentés. Nous utilisons la mesure de centralité comme un indicateur de l'influence globale des individus, pour détecter les influenceurs. Nous évinçons l'individu *DeltaBot*, qui est le robot modérateur du forum *Change my view*. Deux mesures de centralité seront utilisées :

- le ***degré sortant*** (Shaw, 1954) : la centralité d'un nœud correspond à son nombre d'arcs sortants (ceux qui, depuis le nœud, sont orientés vers un nœud différent),

- ***Hits relais*** (Kleinberg, 1999) : la centralité d'un nœud correspond à son nombre d'arcs sortants, pondérés d'autant que les nœuds vers lesquels ils sont orientés ont beaucoup de chemins entrants (des combinaisons d'arcs orientés menant à eux).

Conformément à la description de notre modèle hybride (cf. section 3.4), nous avons choisi des mesures de centralité qui se focalisent sur les arcs sortants (ceux qui, depuis un nœud, sont orientés vers un nœud différent) pour que la centralité mesurée rende compte de l'importance de chaque individu dans la propagation d'influence.

Les deux mesures de centralité choisies sont complémentaires : le *degré sortant* se focalise sur le voisinage direct d'un nœud tandis que *Hits relais* prend en compte la globalité du graphe en favorisant les arcs sortants vers des nœuds populaires. Notre hypothèse est que *Hits relais* produira des meilleurs résultats parce que, dans *Change my view*, les tentatives d'influence sont focalisées sur les auteurs initiaux des fils de discussion, qui se distinguent ainsi par leur popularité.

7.2.5. Évaluation des résultats

Pour évaluer la capacité des mesures de centralité à indiquer l'influence des individus représentés dans les graphes, nous classerons ces derniers selon leurs scores de centralité respectifs ; les influenceurs devront apparaitre en haut du classement.

Nous allons comparer ce classement avec une référence binaire en influenceurs. Dans cette référence, un individu est considéré comme un influenceur dès lors qu'il a reçu une récompense *Delta* d'un autre individu et que cette récompense a été validée par le robot modérateur du forum *Change my view*.

Pour effectuer la comparaison entre le classement en centralité et la référence binaire, nous allons utiliser la mesure *Mean Average Precision* (MAP) que nous avons déjà présentée dans la section sur l'évaluation des mesures de centralité (cf. section 5.1.2). Le principe de cette mesure est qu'un classement est d'une qualité d'autant plus grande, avec un score allant de 0 à 1, qu'il positionne dans les premières places les objets recherchés, dans notre cas, les influenceurs.

7.2.6. Organisation des expériences

Nous avons mis en œuvre différentes configurations expérimentales de façon à comparer les résultats selon que les graphes de *stimuli* et *baseline* sont pondérés en fonction des réactions ou pas et, pour le graphe de *stimuli*, selon que la pondération distingue ou pas

le changement d'avis parmi les réactions. Notre hypothèse est que plus la pondération est précise et conforme à notre modèle hybride (cf. section 3.4), meilleurs seront les résultats. Nous détaillons les différentes configurations ci-dessous.

- **configuration 1** : pas de pondération des graphes,
- **configuration 2** : pondération des graphes de *stimuli* et *baseline* d'après les réactions aux *stimuli*, sans distinguer le changement d'avis ; chaque arc est pondéré par la moyenne des poids attribués aux *stimuli* (pour le graphe de *stimuli*) ou messages quelconques (pour la *baseline*) dont il représente la transmission ; le poids est de 0,5 lorsqu'il y a une réponse de l'individu cible de l'arc et de 0,25 en l'absence de réponse,
- **configuration 3** : pondération des graphes de *stimuli* avec la distinction du changement d'avis parmi les réactions, conformément à notre modèle hybride (cf. section 3.4) ; en plus des poids précédemment donnés, un poids de 1 est attribué aux *stimuli* ayant engendré une réponse catégorisée comme porteuse d'un changement d'avis.

Les données *Change my view* comportent un biais pour les mesures de centralité. En effet, le principe de ce forum est de faire changer d'avis les auteurs initiaux des fils de discussion ; ce sont donc des individus qui seront centraux dans les interactions, sans pour autant être nécessairement des influenceurs. Pour limiter ce biais, nous avons adopté deux stratégies, une stratégie *locale* et une stratégie *globale*.

La stratégie *locale* consiste à mesurer la centralité des participants sur des graphes construits par discussion, des graphe locaux. Un classement des individus est ensuite établi pour chaque discussion, en évinçant son auteur initial. L'évaluation est faite sur l'ensemble des classements, chaque classement étant comparé à une référence binaire sur les influenceurs dans la discussion correspondante.

La stratégie *globale* consiste quant à elle à mesurer la centralité des participants sur un graphe représentant l'ensemble des discussions, un graphe global. Comme pour la stratégie précédente, on établit un classement des participants par discussion, en évinçant son auteur initial, mais cette fois-ci, les scores de centralité sont mesurés sur le graphe global. L'évaluation se fait comme précédemment.

Même si les graphes locaux contiennent beaucoup moins d'informations que les graphes globaux, notre hypothèse est qu'ils sont plus pertinents pour détecter les influenceurs car les classements des participants sont établis par discussion.

Nous allons analyser les résultats d'évaluation du graphe avec l'hybridation face au graphe *baseline*, dans chacune des configurations mises en œuvre, avec des graphes globaux ou locaux.

7.3. Résultats

Tableau 85 : Résultats d'évaluation du modèle hybride pour la détection des influenceurs

Évaluation	MAP avec l'hybridation				MAP avec la *baseline*			
Type de graphe	Graphe global		Graphes locaux		Graphe global		Graphes locaux	
Mesure de centralité	Degré sortant	Hits	Degré sortant	Hits	Degré sortant	Hits	Degré sortant	Hits
Configuration 1	0,26	0,24	0,32	0,33	0,205	0,204	0,22	0,25
Configuration 2	0,27	0,25	0,349	0,348	0,21	0,20	0,30	0,22
Configuration 3	0,31	0,28	0,48	0,55	Seulement par hybridation			

Tableau 86 : Évaluation avec les traits *stimuli* isolés

Évaluation	MAP avec l'hybridation			
Type de graphe	Graphe global		Graphes locaux	
Mesure de centralité	Degré sortant	Hits	Degré sortant	Hits
Config. 1, *claim* seul	0,22	0,20	0,32	0,33
Config. 1, *arg.* seul	0,22	0,23	0,32	0,33

Dans le Tableau 85, nous présentons les résultats de l'évaluation du modèle hybride pour la détection des influenceurs, face à la *baseline*, dans nos différentes configurations expérimentales.

Nous observons que l'approche par hybridation supplante la *baseline* à travers l'ensemble des configurations et des types de graphes. Ceci montre l'apport de notre approche par hybridation et ainsi la pertinence de notre modèle hybride pour la détection d'influenceurs.

Nous devons tout de même nuancer notre constat sur les résultats de l'approche par hybridation car les meilleurs scores obtenus sont moyens. Cependant, nous n'avons pas utilisé la totalité des composantes de notre modèle, en particulier les traits discursifs de *stimulation* ainsi que le trait discursif *stimulus* de la pédagogie. Il est aussi possible d'améliorer la détection des traits discursifs déjà utilisés afin que notre modèle hybride puisse être pleinement exploité.

Pour déterminer l'apport de chacun des traits *stimuli* représentés dans les expériences (*Claim* et *Argumentation*), nous avons évalué leur utilisation séparément (cf. Tableau 86). Nous ne présentons que les résultats de la configuration 1 parce que ceux des deux autres configurations suivent les mêmes tendances.

Nous constatons que les performances sont un peu moins bonnes lorsqu'on isole les traits discursifs avec les graphes globaux tandis qu'elles sont inchangées avec les graphes locaux. Nous pensons qu'il y a plus de bruit dans les graphes globaux, ce qui rend leur analyse plus sensible à la soustraction d'informations.

Le résultat est meilleur avec le trait *Argumentation* et la mesure Hits. La mesure Hits favorise les arcs sortants vers des nœuds avec une plus grande centralité entrante ; ces nœuds représentent, dans notre cas, les individus auxquels on a le plus envoyé de messages classés comme contenant de l'argumentation. Cela indique peut-être que l'argumentation influence davantage que le claim.

Notre hypothèse sur une amélioration des résultats à mesure que la pondération s'affine est vérifiée : les résultats progressent en complexifiant les configurations pour tous les types de graphes, avec l'hybridation comme avec la *baseline*. Cela montre la pertinence de notre système de pondération des arcs du graphe d'après les réactions pour détecter les influenceurs.

En comparant les résultats entre les graphes globaux et les graphes locaux, notre hypothèse sur la plus grande pertinence des graphes locaux se vérifie parce les résultats avec ce dernier type de graphe sont globalement meilleurs. Il semble donc bien que les graphes par discussion, même s'ils contiennent moins d'informations que les graphes globaux, sont tout de même plus pertinents pour détecter les influenceurs dans chacune des discussions.

La comparaison des résultats entre les deux mesures de centralité est moins en adéquation avec notre hypothèse disant que Hits serait meilleure du fait de sa valorisation plus importante des nœuds populaires et de la focalisation des processus d'influence sur les auteurs initiaux des discussions. L'hypothèse est vérifiée avec les graphes locaux uniquement, lorsqu'il n'y a pas la pondération par les réactions (la configuration 1) ou lorsqu'il y a la pondération par les réactions en distinguant le changement d'avis (la configuration 3).

Le fait que l'hypothèse est vérifiée uniquement sur les graphes locaux est peut-être dû à la présence, dans les graphes globaux, d'un bruit lié au biais des auteurs initiaux. Ce biais réside dans ce type de graphes puisque, représentant plusieurs discussions, on y trouve plusieurs auteurs initiaux, que nous n'avons pas supprimés parce que l'auteur initial d'une discussion peut être influenceur dans une autre.

Lorsque la pondération par les réactions ne distingue pas le changement d'avis (la configuration 2), Hits est moins bon que le degré sortant pour tous les types de graphes, que ce soit avec la *baseline* ou l'hybridation, Cela peut s'expliquer par le fait que les réactions, prises indistinctement, peuvent tout à fait avoir trait à un désaccord plutôt qu'à un processus d'influence. Ceci peut engendrer un bruit auquel serait plus sensible Hits qui valorise les nœuds recevant beaucoup de messages.

7.4. Conclusion

Nous avons évalué notre approche hybride pour la détection d'influenceurs dans le média social *Change my view*. Notre système supplante globalement une *baseline* n'intégrant pas d'analyse linguistique. Cela montre l'apport de l'hybridation et semble valider notre modélisation des traits discursifs d'influence. Les différentes configurations mises en œuvre ont aussi montré la pertinence de notre système de pondération des arcs du graphe par les réactions.

Nous avons expérimenté deux types de graphes, les graphes globaux et les graphes locaux. Même si les graphes globaux contiennent plus d'informations que les graphes locaux, ils ont produit des résultats moins bons, probablement parce qu'ils contiennent du bruit, provenant notamment du biais lié aux auteurs initiaux des discussions, spécifique au forum *Change my view*. Il serait intéressant de refaire cette comparaison sur un autre média social.

Parmi les deux mesures de centralité utilisées, nous avons montré que Hits, qui est la mesure la plus complexe, produit, comme nous l'avions supposé, des meilleurs résultats que la mesure par degré sortant, à condition d'être appliquée à des graphes sans trop de bruit.

Certes, les meilleurs résultats obtenus avec notre approche sont moyens dans l'absolu mais ils sont prometteurs. Nous n'avons pas exploité la totalité de notre modèle et son implémentation, comprenant différents modules pour détecter les traits discursifs d'influence, est perfectible.

8. Conclusion générale

8.1. Contributions

8.1.1. Modèle d'influence

Ce travail de recherche apporte une contribution théorique et originale, fondée sur une hybridation de deux types d'approches préexistants dans la littérature : une approche structurelle et une approche linguistique. Nous avons montré la complémentarité de ces deux types d'approches, à travers les deux composantes de notre modèle, portant sur l'action d'un influenceur : l'*influence globale*, qui désigne l'influence à travers un ensemble d'individus, et l'*influence individuelle*, qui a trait à l'action d'influence sur un seul individu.

Nous avons formulé l'hypothèse que l'influence globale d'un individu, à l'intérieur d'un groupe social, pouvait être mesurée par sa centralité, lorsque cet individu est représenté dans un graphe d'actions entre les membres du groupe ; ces actions interpersonnelles doivent refléter l'éventuelle influence d'un membre du groupe sur un autre.

Pour identifier l'influence individuelle dans des conversations, nous avons élaboré une approche linguistique. Elle est basée sur l'observation manuelle de processus d'influence dans le forum *Change my view*. Cette approche empirique donne à notre modèle un caractère pragmatique, qui favorise son applicabilité.

Notre modèle est pragmatique aussi par son contenu : il décrit un influenceur par son action d'influence, avec des traits discursifs qui représentent les moyens de cette action ainsi que ses effets ; pour identifier un influenceur, nous favorisons l'utilisation des preuves de son influence. Les traits discursifs mis en exergue ont donné lieu à une campagne d'annotation, afin de constituer des données de référence les concernant, nécessaire à l'implémentation et à l'évaluation de notre système.

8.1.2. Corpus annoté en traits discursifs

Nous avons décrit la préparation, la mise en œuvre ainsi que l'évaluation d'une campagne d'annotation de nos traits discursifs d'influence. Nous avons aussi présenté un retour d'expérience sur les difficultés afférentes à la campagne d'annotation et nos solutions pour y répondre.

La préparation de la campagne d'annotation a donné lieu à la rédaction d'un guide d'annotation, nécessitant d'expliciter les définitions de nos différents traits discursifs. Ce guide a été revu et corrigé de façon itérative, en collaboration avec les annotateurs. Nous avons constaté les bénéfices de cette approche pour la qualité des annotations. C'est une méthode *agile* (Voormann & Gut, 2008) parce qu'elle consiste en une analyse fréquente du travail d'annotation. Elle avait déjà été utilisée pour améliorer l'accord inter-annotateurs d'une tâche d'annotation manuelle (Alex et al., 2010).

Le travail des annotateurs a permis de vérifier la pertinence de nos traits discursifs, aussi bien du point de vue de leurs définitions que de leur présence dans des messages issus de médias sociaux, dont les genres textuels étaient différents : le forum *Change my view*, que nous avions utilisé pour la définition des traits discursifs, et Twitter, qui était un environnement inédit pour leur identification.

Nous avons construit un corpus de référence sur les traits discursifs *stimuli*, après avoir évalué la qualité des annotations correspondantes. N'ayant pas trouvé, dans l'état de l'art, de référence suffisante quant à une valeur seuil d'accord inter-annotateurs acceptable pour constituer un corpus de référence, nous avons conçu et mis en œuvre une méthode empirique pour déterminer une telle valeur *ad hoc*. Cette méthode a montré qu'un jeu d'annotations avec un faible score d'accord pouvait tout de même contenir les annotations attendues. Nous avons utilisé le corpus résultant pour travailler sur l'hybridation.

8.1.3. Validation

Nous avons décrit l'implémentation et l'évaluation d'un système fondé sur les deux composantes de notre modèle : l'influence globale, que nous avons associée à la centralité dans un graphe social, et l'influence individuelle, que nous avons reliée à la présence de traits discursifs dans les messages. Ce système hybride repose donc à la fois sur une analyse structurelle de liens interpersonnels et sur une analyse linguistique du contenu de messages.

Concernant l'analyse structurelle, nous avons évalué différentes mesures de centralité pour détecter les influenceurs dans un ensemble d'utilisateurs Twitter, ayant extrait, à partir de cet ensemble, des données concernant deux types d'actions interpersonnelles : le retweet et le suivi.

Les expériences ont montré que certaines mesures de centralité, avec l'information de suivi, parvenaient à mettre en avant les influenceurs par rapport à un classement aléatoire, et de façon relativement correcte en comparaison avec des systèmes utilisant de l'apprentissage automatique. Elles ont aussi montré que les mesures de centralité avaient des sensibilités spécifiques à la taille du graphe.

En vue d'affiner l'analyse structurelle, avec des informations linguistiques relatives à l'influence individuelle, nous avons expérimenté différentes configurations de classifieurs

pour distinguer les messages qui contiennent une partie de nos traits discursifs. Les meilleures configurations ont été mises en exergue pour être utilisées dans le processus d'hybridation.

Nous avons fait différentes expériences afin d'évaluer l'apport de l'analyse linguistique lorsqu'elle est combinée avec l'analyse par centralité. Les résultats de cette approche, comparés à ceux d'une *baseline*, ont montré que l'utilisation d'informations sur nos traits discursifs permettait de mieux détecter les influenceurs. Cela semble valider la pertinence de notre approche hybride et de notre caractérisation linguistique des processus d'influence.

8.2. Limites

8.2.1. Modèle

Nous n'avons utilisé que le forum *Change my view* comme référentiel d'observation, ce qui limite la robustesse de notre modèle. Il est possible que dans d'autres types de médias sociaux, comme Instagram ou TikTok, les processus d'influence, en particulier du point de vue discursif, ne correspondent pas tout à fait à ceux que nous avons caractérisés.

Notre modèle est simplificateur, en particulier dans sa dimension psychologique. Il représente l'influence comme s'exerçant d'un individu sur un autre, or, l'influence peut aussi s'exercer par plusieurs personnes, qui vont conjointement engendrer, par exemple, un changement d'avis chez quelqu'un d'autre. Par ailleurs, nous ne considérons que le contenu des échanges entre les individus, mais pas les caractéristiques de ces derniers. Or, ils peuvent aussi être des facteurs d'influence, notamment dans le processus d'identification à autrui (Rosenthal & McKeown, 2016).

8.2.2. Système

Notre système est prévu pour une application à des données textuelles. Cela constitue une limite à la détection d'influenceurs dans d'autres formes de communication, comme l'image, qui est beaucoup utilisée, notamment sur le réseau social Instagram.

L'implémentation de notre modèle est limitée par la qualité et la quantité de nos données annotées. Nous avons vu que celles produites lors de la campagne d'annotation contiennent du bruit, des erreurs, qui ont un impact sur les performances de notre système. Quant à la quantité des annotations, elle était suffisante pour faire de l'apprentissage classique, mais elle est insuffisante pour expérimenter des méthodes d'apprentissage profond, qui ont fait leurs preuves en TAL.

L'évaluation de notre système a été faite sur des données issues du média social qui a servi à le spécifier. Cela permet de vérifier sa pertinence pour une première évaluation, mais les données ne sont pas assez diversifiées pour en évaluer la robustesse. En plus, cette ressource, le forum *Change my view*, est orientée vers la production de processus d'influence, avec un formalisme les rendant explicites. C'est donc un environnement favorable, ce qui constitue une limite à la pertinence de l'évaluation.

8.3. Perspectives

8.3.1. Amélioration de notre approche

La qualité de la détection des traits discursifs peut être améliorée en essayant d'autres approches par apprentissage automatique. Une approche dite *frugale* pourrait être pertinente pour composer avec la limitation imposée par la quantité de nos données annotées (Wang et al., 2020). Nous pourrions aussi hybrider différents algorithmes d'apprentissage pour obtenir un modèle plus performant (Jain et al., 2021). L'utilisation d'algorithmes par apprentissage non supervisé, comme ceux destinés au clustering, pourraient aider à la classification des messages (Zhang et al., 2015).

Il est possible d'augmenter la quantité des traits discursifs détectés, en ajoutant ceux de la *stimulation*. Les résultats de notre système étant prometteurs sans, nous pensons qu'ils pourraient être meilleurs avec ces informations supplémentaires.

Notre référentiel d'observation des influenceurs pourrait être élargi par l'utilisation de nouvelles données, cela permettrait notamment d'enrichir la spécification linguistique de notre modèle. Ces nouvelles données constitueraient aussi un nouveau référentiel d'évaluation pour mettre à l'épreuve et améliorer la robustesse de notre système, en particulier avec des textes d'un autre genre que celui du forum *Change my view*. Pour élargir notre vision des influenceurs, nous pouvons aussi regarder des travaux connexes.

Il y a un courant de recherche sur la détection de point de vue (*stance detection*), qui consiste à identifier la position éventuelle d'un énonciateur sur un sujet (Küçük & Fazli, 2020). Cette problématique a notamment fait l'objet d'une tâche dans la campagne d'évaluation *SemEval 2016*, à propos de sujets de société (Mohammad et al., 2016). Elle est aussi étudiée pour la détection de fausses informations (Mrowca & Wang, 2017). Les influenceurs peuvent être importants sur ces thématiques, lorsqu'ils prennent une position qui aura tendance à être adoptée par leur audience.

Les résultats proposés par notre système peuvent être affinés : il pourrait être intéressant de voir qui sont les meilleurs influenceurs, c'est-à-dire ceux qui ont la plus

grande probabilité d'influencer les autres, ou encore d'indiquer les discussions, les messages, voire les traits discursifs qui ont contribué à détecter un influenceur.

Nous pourrions voir quels seraient les résultats de la détection des influenceurs avec un autre format que le classement pour distinguer les individus. Nous pensons en particulier à une classification binaire de chaque individu, selon qu'il est un influenceur ou pas. Il est possible, à cet égard, de créer un classifieur par apprentissage automatique avec, par exemple, les valeurs de centralité des individus comme traits distinctifs.

8.3.2. Scénarios d'application

Cette thèse a été réalisée en partenariat avec Viseo[46], une entreprise de services du numérique (ESN). Ses clients sont eux-mêmes des entreprises, qui peuvent utiliser notre système pour identifier des influenceurs dans leur secteur d'activité, afin de mener une veille stratégique ou une campagne de promotion d'un de leurs produits.

C'est justement pour faire du marketing numérique que des marques utilisent beaucoup les médias sociaux (Brown & Fiorella, 2014). Notre système pourrait alors être un moyen pour ces marques de se mettre rapidement en contact avec des influenceurs, afin de faire du marketing d'influence (Harrigan et al., 2021). Ces influenceurs communiqueraient ensuite avec des potentiels clients afin de les inciter à acheter des produits.

Le marketing numérique peut aussi s'appliquer à d'autres domaines, comme c'est le cas avec le marketing politique (Ayankoya et al., 2015). Notre système pourrait être un atout dans le cadre d'une campagne électorale : il permettrait de détecter les individus qui sont les plus à même d'avoir un impact sur le vote d'une population ciblée.

Les annotations de référence utilisées par notre système font qu'il est davantage orienté vers la détection de claims, que nous avons définis comme des affirmations sur des sujets concrets. Ainsi, ses domaines d'application les plus pertinents sont ceux en relation directe avec la société, comme la politique, la défense ou la santé. La composante dédiée à la détection du changement d'avis pourrait particulièrement s'appliquer aux cas d'influence pour des élections ou avec de la propagande.

En politique, il y a le cas des campagnes de communication en santé publique (Gough et al., 2017). Elles concernent notamment des sujets sensibles et urgents, comme la pandémie de la Covid-19 (Teichmann et al., 2020). De tels sujets nécessitent un discours efficace et adapté à la population ciblée. Il pourrait alors être intéressant d'identifier, sur un sujet de santé publique et concernant une certaine population, les individus qui ont un impact sur les points de vue et les comportements.

Le travail sur la communication peut avoir trait à la production de contre-discours. Au sein de l'entité recherche de Viseo, nous avons pris part à des projets de recherche

[46] https://www.viseo.com/fr/emea

collaboratifs. Parmi ceux-ci, il y avait le projet européen Trivalent[47], visant à mieux comprendre les causes de la radicalisation en Europe, afin de mieux la contrer.

Il y a un lien entre ce sujet de recherche et le nôtre, avec l'influence, dans les médias sociaux, d'individus radicalisés (Fernandez et al., 2018). C'est pour cette raison que nous avons travaillé sur un jeu de tweets d'individus étiquetés comme pro-État islamique. L'objectif était de voir si nous y retrouvions les traits discursifs de notre modèle, afin d'identifier les influenceurs parmi ces individus spécifiques.

Le scénario d'application contre la radicalisation est en lien avec la thématique de la propagande, et notamment la propagation de fausses informations (Yang et al., 2021). Les influenceurs peuvent être des manipulateurs d'informations. Il serait alors intéressant de pouvoir les détecter afin de concevoir une stratégie focalisée de contre-discours.

Nous voulons que le corpus de référence, ainsi que l'ensemble des annotations produites, soient disponibles en libre-accès afin de favoriser d'autres publications intégrant les traits discursifs que nous avons étudiés.

[47] https://trivalent-project.eu

Bibliographie

Ajzen, I. (1996). The social psychology of decision making. *Social Psychology: Handbook of Basic Principles*.

Alex, B., Grover, C., Shen, R., & Kabadjov, M. (2010). Agile corpus annotation in practice: An overview of manual and automatic annotation of CVs. *ACL 2010 - LAW 2010: 4th Linguistic Annotation Workshop, Proceedings*.

Amigó, E., Carrillo-De-Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., De Rijke, M., & Spina, D. (2014). Overview of RepLab 2014: Author profiling and reputation dimensions for online reputation management. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8685 LNCS*, 307–322. https://doi.org/10.1007/978-3-319-11382-1_24

Artstein, R. (2017). Inter-annotator Agreement. In *Handbook of Linguistic Annotation*. https://doi.org/10.1007/978-94-024-0881-2_11

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. In *Computational Linguistics*. https://doi.org/10.1162/coli.07-034-R2

Ayankoya, K., Calitz, A. P., & Cullen, M. (2015). A Framework for the use of Social Media for Political Marketing: An Exploratory Study. *Retrieved December*, *12*, 2018.

Bavelas, A. (1948). A Mathematical Model for Group Structures. *Human Organization*, *7*(3), 16–30. https://doi.org/10.17730/humo.7.3.f4033344851gl053

Bavelas, A. (1950). Communication Patterns in Task-Oriented Groups. *Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.1906679

Bigonha, C., Cardoso, T. N. C., Moro, M. M., Gonçalves, M. A., & Almeida, V. A. F. (2012). Sentiment-based influence detection on Twitter. *Journal of the Brazilian Computer Society*, *18*(3). https://doi.org/10.1007/s13173-011-0051-5

Binali, H., Wu, C., & Potdar, V. (2010). Computational approaches for emotion detection in text. *4th IEEE International Conference on Digital Ecosystems and Technologies - Conference Proceedings of IEEE-DEST 2010, DEST 2010*. https://doi.org/10.1109/DEST.2010.5610650

Bonacich, P. (1987). Power and Centrality: A Family of Measures. *American Journal of Sociology*, *92*(5), 1170–1182. https://doi.org/10.1086/228631

Borgs, C., Brautbar, M., Chayes, J., & Lucier, B. (2014). Maximizing social influence in nearly optimal time. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 946–957. https://doi.org/10.1137/1.9781611973402.70

Brandes, U. (2001). A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, *25*(2), 163–177. https://doi.org/10.1080/0022250X.2001.9990249

Brown, D., & Fiorella, S. (2014). Influence marketing: how to create, manage, and measure brand influencers in social media marketing. *Choice Reviews Online*, *51*(05), 51-2752-51–2752. https://doi.org/10.5860/choice.51-2752

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Chen, W., Wang, C., & Wang, Y. (2010). Scalable influence maximization for prevalent viral

marketing in large-scale social networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/1835804.1835934

Cossu, J. V., Gonzalo, J., Hajjem, M., Hamon, O., Latiri, C., & SanJuan, E. (2018). CLEF MC2 2018 Lab: Technical overview of cross language microblog search and argumentative mining. *CEUR Workshop Proceedings*.

Cunningham, H., Tablan, V., Roberts, A., & Bontcheva, K. (2013). Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Computational Biology*, *9*(2). https://doi.org/10.1371/journal.pcbi.1002854

Dave, K., Bhatt, R., & Varma, V. (2011). Identifying influencers in social networks. *Researchgate.Net*. https://www.researchgate.net/profile/Rushi_Bhatt3/publication/266329886_Identifying_Influencers_in_Social_Networks/links/5559747e08ae980ca6106ad6/Identifying-Influencers-in-Social-Networks.pdf

Deturck, K., Patel, N., Avouac, P.-A., Lopez, C., Nouvel, D., Partalas, I., & Segond, F. (2019). Detecting influencial users in social networks: Analysing graph-based and linguistic perspectives. In *IFIP Advances in Information and Communication Technology* (Vol. 571). https://doi.org/10.1007/978-3-030-29904-0_9

Deturck, Kévin. (2020). Détection d'influenceurs dans des médias sociaux. *COnference En Recherche d'Informations et Application, CORIA 2018 - 15th French Information Retrieval Conference, Proceedings*.

Deturck, Kévin. (2021). *Guide d'annotation en discours pour la détection d'influenceurs*. https://hal.archives-ouvertes.fr/hal-03154171

Deturck, Kévin, Goswami, P., Nouvel, D., & Segond, F. (2018). ERTIM@MC2: Diversified argumentative Tweets retrieval. *CEUR Workshop Proceedings*, *2125*.

Deturck, Kévin, Nouvel, D., & Segond, F. (2018). Évaluation comparative d'algorithmes de centralité pour la détection d'influenceurs. *Revue Des Nouvelles Technologies de l'Information*, *Extraction*, 369–370.

Devi, P., Gupta, A., & Dixit, A. (2014). Comparative Study of HITS and PageRank Link based Ranking Algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, *3*(2), 5749–5754.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*.

Dillard, J. P., & Wilson, S. R. (2014). Interpersonal influence. In *Interpersonal Communication*. https://doi.org/10.1515/9783110276794.155

Domingos, P., & Richardson, M. (2001). Mining the network value of customers. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/502512.502525

Eckle-Kohler, J., Kluge, R., & Gurevych, I. (2015). On the role of discourse markers for discriminating claims and premises in argumentative discourse. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d15-1267

Fernandez, M., Asif, M., & Alani, H. (2018). Understanding the Roots of Radicalisation on Twitter. *Proceedings of the 10th ACM Conference on Web Science - WebSci '18*, 1–10. https://doi.org/10.1145/3201064.3201082

Finucane, M. L., Peters, E., & Slovic, P. (2003). Judgment and Decision Making: The Dance

of Affect and Reason. In *Emerging Perspectives on Judgment and Decision Research*. https://doi.org/10.1017/cbo9780511609978.012

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. In *Sociometry* (Vol. 40, Issue 1, p. 35). https://doi.org/10.2307/3033543

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Friedkin, N. E., & Johnsen, E. C. (1990). Social influence and opinions. *The Journal of Mathematical Sociology*. https://doi.org/10.1080/0022250X.1990.9990069

Germesin, S., & Wilson, T. (2009). Agreement detection in multiparty conversation. *ICMI-MLMI'09 - Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interfaces*. https://doi.org/10.1145/1647314.1647319

Gionis, A., Terzi, E., & Tsaparas, P. (2013). Opinion maximization in social networks. *Proceedings of the 2013 SIAM International Conference on Data Mining*, 387–395. https://doi.org/10.1137/1.9781611972832.43

Gough, A., Hunter, R. F., Ajao, O., Jurek, A., McKeown, G., Hong, J., Barrett, E., Ferguson, M., McElwee, G., McCarthy, M., & Kee, F. (2017). Tweet for behavior change: Using social media for the dissemination of public health messages. *JMIR Public Health and Surveillance*, *3*(1). https://doi.org/10.2196/publichealth.6313

Harrigan, P., Daly, T. M., Coussement, K., Lee, J. A., Soutar, G. N., & Evers, U. (2021). Identifying influencers on social media. *International Journal of Information Management*, *56*. https://doi.org/10.1016/j.ijinfomgt.2020.102246

Heidemann, J., Klier, M., & Probst, F. (2010). Identifying key users in online social networks: A pagerank based approach. *ICIS 2010 Proceedings - Thirty First International Conference on Information Systems*, 1–21. https://doi.org/10.1016/j.infsof.2008.09.005

Hidi, S., & Baird, W. (1986). Interestingness-A neglected variable in discourse processing. *Cognitive Science*. https://doi.org/10.1016/S0364-0213(86)80003-9

Jain, S., Jain, A. K., & Singh, S. P. (2021). Building a Machine Learning Model for Unstructured Text Classification: Towards Hybrid Approach. *Advances in Intelligent Systems and Computing*, *1187*. https://doi.org/10.1007/978-981-15-6014-9_51

Katsimpras, G., Vogiatzis, D., & Paliouras, G. (2015). Determining influential users with supervised random walks. *Dl.Acm.Org*. https://dl.acm.org/citation.cfm?id=2742472

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39–43. https://doi.org/10.1007/BF02289026

Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the spread of influence through a social network. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *pages*, 137–146. https://doi.org/10.1145/956750.956769

Khadangi, E., & Bagheri, A. (2017). Presenting novel application-based centrality measures for finding important users based on their activities and social behavior. *Computers in Human Behavior*, *73*, 64–79. https://doi.org/10.1016/J.CHB.2017.03.014

Kiss, C., & Bichler, M. (2008). Identification of influencers—measuring influence in customer networks. *Elsevier*. https://www.sciencedirect.com/science/article/pii/S0167923608001231

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *The Structure and Dynamics of Networks*, *9781400841*(5), 514–542.

https://doi.org/10.1145/324133.324140
Krippendorff, K. (1995). On the Reliability of Unitizing Continuous Data. *Sociological Methodology*, *25*, 47. https://doi.org/10.2307/271061
Krippendorff, K. (2004). Reliability in Content Analysis. *Human Communication Research*, *30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x
Küçük, D., & Fazli, C. A. N. (2020). Stance detection: A survey. In *ACM Computing Surveys* (Vol. 53, Issue 1). https://doi.org/10.1145/3369026
Leavitt, H. J. (1951). Some effects of certain communication patterns on group performance. *The Journal of Abnormal and Social Psychology*, *46*(1), 38–50. https://doi.org/10.1037/h0057189
Li, Y., Ma, S., Zhang, Y., Huang, R., & Kinshuk. (2013). An improved mix framework for opinion leader identification in online learning communities. *Knowledge-Based Systems*, *43*, 43–51. https://doi.org/10.1016/j.knosys.2013.01.005
Lü, L., Zhang, Y. C., Yeung, C. H., & Zhou, T. (2011). Leaders in social networks, the delicious case. *PLoS ONE*, *6*(6). https://doi.org/10.1371/journal.pone.0021202
Lundberg, G. A., & Lewin, K. (1939). The Conceptual Representation and the Measurement of Psychological Forces. *American Sociological Review*. https://doi.org/10.2307/2083779
Majnik, M., & Bosnić, Z. (2013). ROC analysis of classifiers in machine learning: A survey. In *Intelligent Data Analysis*. https://doi.org/10.3233/IDA-130592
Mariani, J., Paroubek, P., Francopoulo, G., & Hamon, O. (2014). Rediscovering 15 years of discoveries in language resources and evaluation: The LREC anthology analysis. *Researchgate.Net*. https://www.researchgate.net/profile/Gil_Francopoulo2/publication/265791639_Rediscovering_15_Years_of_Discoveries_in_Language_Resources_and_Evaluation_The_LREC_Anthology_Analysis/links/541bee130cf2218008c4d299.pdf
Mason, W. A., Conrey, F. R., & Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review*. https://doi.org/10.1177/1088868307301032
Mathet, Y. (2017). The agreement measure γcat a complement to γ focused on categorization of a continuum. *Computational Linguistics*, *43*(3), 661–681. https://doi.org/10.1162/COLI_a_00296
Mathet, Y., & Widlöcher, A. (2016). Évaluation des annotations: Ses principes et ses pièges. *TAL Traitement Automatique Des Langues*, *57*(2), 73–98.
Mathet, Y., Widlöcher, A., & Métivier, J.-P. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, *41*(3), 437–479. https://doi.org/10.1162/COLI_a_00227
Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). SemEval-2016 task 6: Detecting stance in tweets. *SemEval 2016 - 10th International Workshop on Semantic Evaluation, Proceedings*. https://doi.org/10.18653/v1/s16-1003
Mrowca, D., & Wang, E. (2017). Stance detection for fake news identification. *Eliaswang.Com*.
Nguyen, V. A., Boyd-Graber, J., Resnik, P., Cai, D. A., Midberry, J. E., & Wang, Y. (2014). Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, *95*(3), 381–421. https://doi.org/10.1007/s10994-013-5417-9
Nouvel, D., Deturck, K., Segond, F., & Patel, N. (2019). L'influence sur les réseaux, une proposition de modélisation. *Recherche d'information, Document et Web Sémantique*,

3(1). https://doi.org/10.21494/iste.op.2020.0465
Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. *World Wide Web Internet And Web Information Systems*. https://doi.org/10.1.1.31.1768
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*.
Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.3115/v1/d14-1162
Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2011). In the mood being influential on twitter mood. *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, 307–314. https://doi.org/10.1109/PASSAT/SocialCom.2011.27
Rosa, H., Carvalho, J. P., Astudillo, R., & Batista, F. (2018). Page rank versus katz: Is the centrality algorithm choice relevant to measure user influence in twitter? In *Studies in Computational Intelligence*. https://doi.org/10.1007/978-3-319-74681-4_1
Rosenthal, S., & McKeown, K. (2016). *Social Proof: The Impact of Author Traits on Influence Detection*. https://doi.org/10.18653/v1/w16-5604
Rosenthal, S., & McKeown, K. (2017). Detecting influencers in multiple online genres. *ACM Transactions on Internet Technology*, *17*(2), 1–22. https://doi.org/10.1145/3014164
Saurí, R., & Pustejovsky, J. (2012). Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text. *Computational Linguistics*. https://doi.org/10.1162/COLI_a_00096
Shaw, M. E. (1954). Group Structure and the Behavior of Individuals in Small Groups. *Journal of Psychology: Interdisciplinary and Applied*. https://doi.org/10.1080/00223980.1954.9712925
Song, X., Chi, Y., Hino, K., & Tseng, B. L. (2007). Identifying opinion leaders in the blogosphere. *International Conference on Information and Knowledge Management, Proceedings*, *07*(January 2007), 971–974. https://doi.org/10.1145/1321440.1321588
Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. *25th International World Wide Web Conference, WWW 2016*, 613–624. https://doi.org/10.1145/2872427.2883081
Teichmann, L., Bridgman, A., Nossek, S., Loewen, P. J., Owen, T., Ruths, D., & Zhilin, O. (2020). Public health Communication and Engagement on Social Media during the COVID-19 Pandemic (Preprint). *OSF Preprints*.
Travers, J., & Milgram, S. (1969). *An experimental study of the small world problem*.
Turner, J. C., & Oakes, P. J. (1986). The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence. *British Journal of Social Psychology*, *25*(3), 237–252. https://doi.org/10.1111/j.2044-8309.1986.tb00732.x
Voormann, H., & Gut, U. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, *4*(2). https://doi.org/10.1515/CLLT.2008.010

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Computing Surveys*, *53*(3). https://doi.org/10.1145/3386252

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*. https://doi.org/10.3758/s13428-012-0314-x

Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM 2010)*, 261–270. https://doi.org/10.1145/1718487.1718520

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.3115/1220575.1220619

Wyer, R. S., & Shrum, L. J. (2015). The Role of Comprehension Processes in Communication and Persuasion. *Media Psychology*. https://doi.org/10.1080/15213269.2014.912584

Xu, W. W., Sang, Y., Blasiola, S., & Park, H. W. (2014). Predicting Opinion Leaders in Twitter Activism Networks: The Case of the Wisconsin Recall Election. *American Behavioral Scientist*. https://doi.org/10.1177/0002764214527091

Yang, K. C., Pierri, F., Hui, P. M., Axelrod, D., Torres-Lugo, C., Bryden, J., & Menczer, F. (2021). The COVID-19 Infodemic: Twitter versus Facebook. *Big Data and Society*, *8*(1). https://doi.org/10.1177/20539517211013861

Zhang, W., Tang, X., & Yoshida, T. (2015). TESC: An approach to TExt classification using Semi-supervised Clustering. *Knowledge-Based Systems*, *75*. https://doi.org/10.1016/j.knosys.2014.11.028

Kévin DETURCK

DÉTECTION DES INFLUENCEURS DANS DES MÉDIAS SOCIAUX

Résumé

Dans cette thèse, nous présentons la conception et l'évaluation d'un système pour détecter automatiquement les personnes influentes dans les médias sociaux, à partir des manifestations de leur action d'influence dans les communications interpersonnelles.Les approches pour la détection des influenceurs utilisent généralement, soit la structure de la communication entre les individus, soit l'analyse de son contenu. Le cadre théorique retenu dans notre thèse a la particularité de combiner ces deux types d'approches pour leur complémentarité. Nous caractérisons l'action des influenceurs à l'échelle d'un individu cible, depuis sa mise en œuvre jusqu'à ses effets, par des traits discursifs relevant aussi bien des messages envoyés par les influenceurs que de ceux envoyés par les individus influencés. La détection automatique de ces traits discursifs est faite avec des méthodes en traitement automatique des langues, basées sur des règles linguistiques et des modèles par apprentissage automatique. À l'échelle d'un groupe, l'action des influenceurs est caractérisée par leur position centrale dans un graphe social qui représente des actions interpersonnelles ayant eu cours à l'intérieur de ce groupe. L'hybridité de notre système consiste en l'utilisation des informations linguistiques sur les traits discursifs d'influence, extraits automatiquement depuis les messages textuels échangés entre individus, afin de construire les graphes sociaux.

Mots clés : influenceurs, influence, traitement automatique des langues, réseaux sociaux, médias sociaux, analyse de graphe, mesures de centralité, apprentissage automatique

Résumé en anglais

In this thesis, we present the design and evaluation of a system to automatically detect influencers in social media, based on the manifestations of their influencing action in interpersonal communications. Approaches to influencer detection generally use either the structure of communication between individuals, or the analysis of its content. The theoretical framework chosen in our thesis has the particularity of combining these two types of approach for their complementarity. We characterise the action of influencers at the level of a target individual, from its means to its effects, by discursive features of both the messages sent by the influencers and those sent by the influenced individuals. The automatic detection of these discursive features in social media messages is done with methods in natural language processing, based on linguistic rules and machine learning models. At the group level, the action of influencers is characterised by their central position in a social graph, that represents interpersonal actions within the group. The hybridity of our system consists in the use of linguistic information, automatically extracted in discussions, to construct the social graphs whose structure will be analysed.

Keywords: influencers, influence, social media, social networks, natural language processing, machine learning, graph analysis, centrality measures