Évaluation de l'influence polarisée dans un réseau multi-relationnel : Application à Twitter

Lobna Azaza [1,2], **Marinette Savonnet** [1], **Éric Leclercq** [1], **Sergey Kirgizov** [1], **Rim Faiz** [2]

1. *Laboratoire Le2i FRE2005, CNRS, Arts et Métiers*
   Univ. Bourgogne Franche-Comté, 9, Avenue Alain Savary, F-21078 Dijon - France
   Prénom.Nom@u-bourgogne.fr

2. *Laboratoire Larodec, Université de Carthage, Tunis, Tunisie*
   Rim.Faiz@ihec.rnu.tn

RÉSUMÉ. *L'étude de l'influence sur Twitter est un sujet de recherche intense, certains utilisateurs révèlent plus de capacité pour influencer d'autres personnes. Nous proposons une nouvelle approche pour une évaluation de l'influence polarisée dans les réseaux multi-relationnels tels que Twitter. Nous prenons en compte le contenu des tweets pour déterminer leur polarité en utilisant l'algorithme des forêts d'arbres décisionnels. Puis, nous fusionnons, au moyen des fonctions de croyance, les informations provenant des relations (retweet, mention ou répond, etc.) pour obtenir un degré d'influence pour chaque utilisateur. Nous expérimentons notre méthode sur les données collectées lors des élections européennes de 2014. Les résultats montrent que notre modèle est suffisamment flexible pour répondre aux besoins des spécialistes en sciences sociales et que l'utilisation de la théorie des fonctions de croyances est efficace pour traiter des relations multiples.*

ABSTRACT. *Influence in Twitter has become recently a hot research topic. Some users are more able than others to influence peers. In this study, we propose a new approach for polarized influence assessment in multi-relational networks such as Twitter. We take into account the content of the tweets using a random forest algorithm to deduce their polarity. After that, based on the belief functions theory, we merge information from different relations (e.g retweet, mention or reply, etc.) to deduce the influence degree of each user. We experiment our method on data gathered during the European Elections 2014. The results show that our model is flexible enough to consider social scientists needs and that the belief theory is accurate for information fusion in multi-relational networks.*

MOTS-CLÉS : *Réseau multi-relationnel, Influence, Analyse de sentiment, Fusion d'information, Fonctions de croyance.*

KEYWORDS: *Multi-relational network, Influence, Sentiment Analysis, Information fusion, Belief theory.*

Hermès Science Publication

1. Introduction

Les réseaux sociaux numériques tel que *Twitter* rassemblent les personnes et renforcent leurs relations avec de nouvelles formes de coopération et de communication. En raison de son immense popularité, *Twitter* produit de gros volumes de données, il offre des APIs qui permettent de collecter les données pour développer des applications ou effectuer des analyses. De nombreux travaux étudient les données issues de Twitter dans des domaines très différents : marketing, politique, catastrophes naturelles, etc.

Nous avons réalisé une plateforme SNFreezer (Basaille *et al.*, 2016) qui collecte, stocke et analyse les données de Twitter. Dans le cadre du projet TEE 2014, nous travaillons avec des chercheurs en sciences sociales. Cette collaboration interdisciplinaire a pour but de définir un ensemble d'outils d'analyses du discours politique. Les questionnements de ces chercheurs sont multiples : comment les candidats aux élections se sont appropriés Twitter (utilisations de mentions auto-promotion, de hashtags, d'URLs) ? Quelle est l'influence d'un évènement sur leur communication ? Quels types de relations se nouent entre les candidats politiques et les autres utilisateurs et en particulier l'influence. Notre contribution porte sur la dernière question, c'est-à-dire sur l'étude de l'influence polarisée des candidats politiques.

L'une des caractéristiques de *Twitter* est la diffusion d'information par l'utilisation d'opérateurs, tweeter, mentionner ou citer un utilisateur, utiliser un hashtag ou une URL par exemple. Les liens entre les utilisateurs déterminent le flux de l'information et conditionnent ainsi l'influence d'un utilisateur sur un autre. Certains utilisateurs, appelés influents, sont plus capables que d'autres de diffuser des informations à un grand nombre d'utilisateurs. Par conséquent, la détection des utilisateurs influents dans un réseau est une clé de succès pour parvenir à une diffusion d'information à large échelle et à faible coût.

L'influence sur *Twitter* est définie comme la capacité d'un utilisateur à provoquer une action chez un autre utilisateur (Leavitt *et al.*, 2009). Le terme "action" désigne les différentes interactions possibles entre les utilisateurs au moyen des opérateurs. Par conséquent, la mesure de l'influence sur *Twitter* est un problème complexe puisque *Twitter* offre plusieurs types d'opérateurs (*retweet*, *réponse*, *mention*, *suivre*), qui peuvent être combinés pour former différentes catégories d'interactions. Un utilisateur peut *suivre* un autre utilisateur, ce qui lui permet de voir les *tweets* et les informations de l'utilisateur qu'il suit. Il est également capable de *retweeter* un *tweet*, ce qui expose ce *tweet* à ses abonnés, qui peuvent à leur tour le *retweeter*. Un utilisateur peut *mentionner* un autre utilisateur en utilisant le préfixe "@" s'il veut lui adresser le *tweet*, ce même *tweet* pouvant être *retweeté* par un autre utilisateur. Enfin, un utilisateur peut *répondre* à un *tweet* et créer ainsi une conversation avec l'utilisateur du *tweet* initial. Au niveau des données, les relations induites par l'utilisation des opérateurs peuvent être représentées par un réseau multi-relationnel (Wu *et al.*, 2013a; Rodriguez, Shinavier, 2010).

L'évaluation de l'influence pose quatre défis principaux. Le premier est la polarité des *tweets*, il est important d'analyser le contenu des *tweets* afin de déduire si l'influence exercée est positive ou négative. Le second défi est la diversité des interactions sur lesquels nous pouvons baser les calculs de l'influence. Il faut combiner l'influence respective des différentes interactions afin d'établir une mesure générale d'influence qui prend en compte les différentes modalités d'interaction entre les utilisateurs selon leur contexte. Le troisième défi est la considération de l'influence indirecte. L'influence est indirecte lorsqu'elle atteint un utilisateur à travers des utilisateurs intermédiaires. Par exemple, un utilisateur peut *retweeter* un *tweet* d'un autre utilisateur indirectement à travers un utilisateur intermédiaire. Il est donc nécessaire de mesurer l'influence en tenant compte des interactions directes et indirectes dans le réseau. Le quatrième défi est relatif à l'incertitude lors de la combinaison d'interactions. Dans le cas des réseaux multi-relationnels, il est difficile d'attribuer des pondérations valuées aux différentes interactions avant de fusionner les données quantitatives.

Dans cet article, nous proposons une évaluation de l'influence polarisée. La polarité de l'influence indique si l'influence exercée, d'un certain utilisateur, est positive, négative ou neutre. Pour ce faire, nous combinons différentes interactions définies par des experts du domaine, en tenant compte de la polarité des *tweets* et de l'incertitude dans le processus de la mesure. La mesure peut être établie entre un couple d'utilisateurs au moyen des différentes interactions entre eux deux, mais aussi étendue à une mesure d'influence globale d'un utilisateur dans le réseau. Nous utilisons l'algorithme des forêts d'arbres décisionnels pour déterminer la polarité des *tweets*, il s'agit d'analyser les sentiments exprimés à travers les *tweets* pour déterminer s'ils ont une tendance positive, négative ou neutre. Ensuite, nous définissons un cadre théorique sur la base de la règle de combinaison conjonctive de la théorie des fonctions de croyance et la règle de Smets (Smets, 1997) pour la fusion des informations. Une évaluation à travers des expérimentations à été réalisée, elle s'appuie sur des données *Twitter* collectées dans le cadre du projet inter-disciplinaire TEE 2014 lors de la campagne pour les élections européennes de 2014.

Le reste de l'article est organisé comme suit. La section 2 présente un état de l'art. La section 3 décrit notre approche. La section 4 présente les résultats expérimentaux. Et enfin la section 5 conclut le papier et présente les perspectives dégagées des résultats expérimentaux et des analyses par les chercheurs en sciences sociales.

2. État de l'art

L'objectif de notre approche étant de mesurer l'influence polarisée, nous présentons, dans cette section, les travaux sur l'évaluation de l'influence et l'analyse de sentiments dans *Twitter* pour déterminer la polarité des *tweets*. Nous rappelons ensuite les concepts de base de la théorie des fonctions de croyance sur laquelle se fonde notre approche pour effectuer la fusion et la combinaison des informations.

2.1. L'influence dans Twitter

L'évaluation de l'influence dans *Twitter* a donné lieu à un grand nombre de recherches et différentes approches ont été proposées pour évaluer l'influence des utilisateurs (Riquelme, González-Cantergiani, 2016; Neves *et al*., 2015). Nous avons identifié trois grands types d'approches. Elles sont basées sur des **mesures de popularité**, sur la **topologie du réseau** incluant les algorithmes *PageRank* et *HITS* pour classer les utilisateurs les plus influents. Une autre famille étend les approches topologiques pour assurer la **fusion d'informations** issues des différentes interactions qui doivent être prises en compte dans l'évaluation de l'influence.

À côté de cette recherche académique, il existe également des outils disponibles en ligne pour estimer le score d'influence tels que Klout[1], Kred [2], SocialMention[3] et SocialBakers[4]. Ces outils restent des boîtes noires car ils n'exposent pas les méthodes utilisées pour évaluer l'influence ce qui ne permet pas à un utilisateur de comprendre comment l'influence a été calculée. Dans la suite, nous présentons les principaux travaux académiques pour chacune des approches précédentes.

Les méthodes basées sur les **mesures de popularité** exploitent le résumé statistique des actions et des attributs des utilisateurs. Dans *Twitter*, de nombreuses caractéristiques peuvent être prises en considération. Les auteurs dans (Leavitt *et al*., 2009) utilisent quatre caractéristiques pour mesurer l'influence : le nombre de *reply*, *retweets* et *mentions*, en plus du nombre de *followers*. Ils donnent des statistiques relatives à ces caractéristiques mais ne proposent pas un score global de l'influence se basant sur toutes les relations prises en compte. Cha et al. (Cha *et al*., 2010) définissent trois mesures d'influence dans *Twitter*, le nombre de *followers*, indiquant la taille de l'audience d'un utilisateur ou sa popularité, le nombre de *retweets*, indiquant la capacité d'un utilisateur à écrire du contenu à transmettre à d'autres et le nombre de *mentions*, indiquant sa capacité à engager avec les autres des conversations. Les auteurs calculent la valeur de chaque relation pour 6 millions d'utilisateurs puis ils les comparent. Pour ce faire, ils trient les utilisateurs en fonction de chaque relation, puis, ils quantifient comment le classement d'un utilisateur varie selon les différentes relations. La corrélation de Spearman est utilisée comme une mesure de la force d'association entre deux ensembles du classement. Ils ont constaté que le nombre de *followers* représente la popularité d'un utilisateur, mais qu'il n'est pas lié à d'autres relations telles que les *retweets* et les *mentions*. Leur résultat suggère que le nombre de *followers* seul révèle très peu sur l'influence d'un utilisateur. Cependant, la méthode ne fournit pas une mesure globale de l'influence. Lee *et al*. (2010) calculent le nombre cumulé d'utilisateurs, nommés lecteurs potentiels, qui ont vu un *tweet* et étudient comment ce nombre évolue avec le temps. Les utilisateurs influents

1. https://klout.com/home
2. http://home.kred/
3. http://www.socialmention.com
4. https://www.socialbakers.com

sont déterminés en se basant sur leur nombre de lecteurs potentiels. Suh et al. (Suh *et al*., 2010) ont analysé les facteurs qui ont un impact positif sur le nombre de *retweets* : les URLs, les hashtags, l'ancienneté du compte, le nombre de *followers/followings*, en revanche, ils constatent que le nombre de *tweets* précédemment émis est sans influence. Bakshy *et al*. (2011) utilisent les cascades de diffusion d'URLs raccourcis et considèrent que les utilisateurs, à la source des URLs qui produisent les cascades les plus longues sont les plus influents. Les résultats présentés sont obtenus à partir d'une enquête effectuée auprès de 1,6 million d'utilisateurs sur une période de deux mois en 2009. Dans ce travail, la définition de l'influence est limitée à la capacité d'être le premier à publier l'URL qui sera ensuite *retweetée* par les *followers*. Dans (Cossu *et al*., 2015), les utilisateurs examinent les critères qui peuvent être extraits de *Twitter* dans le but de classer les utilisateurs et de détecter les utilisateurs influents dans la vie réelle sur la base de leur profil *Twitter*. Ils citent de nombreux critères tels que les caractéristiques générales (par exemple le nombre de *followers*), les interactions entre les utilisateurs et les occurrences de termes (URL, ponctuation, etc.). Ils utilisent ensuite un classifieur linéaire, la régression logistique, ainsi que des classifieurs non-linéaires (SVM avec noyaux) où un utilisateur est représenté sous la formes de sacs de mots. Les expérimentations ont été menées sur le jeu de données CLEF RepLab 2014[5].

Les approches basées sur la **topologie du réseau** reposent sur l'analyse structurelle du réseau engendré par les données de *Twitter*. Il s'agit de considérer l'utilisateur comme un nœud et d'étudier la structure du réseau auquel il appartient. Les différentes mesures de centralité (centralité de degré, de proximité, d'intermédiarité) permettent d'identifier les utilisateurs influents (Sun, Tang, 2011). Chen *et al*. (2013) proposent une méthode de classement local, nommée Cluster Rank, prenant en considération le nombre de voisins et leur cœfficient de clustering[6]. Un des avantages de ces mesures est qu'elles sont faciles à mettre en œuvre, mais elles ne tiennent compte que d'une information locale correspondant à une partie du réseau. Brown et al. (Brown, Feng, 2011) partent du principe que la localisation d'un nœud dans le réseau peut déterminer son influence. Ainsi, un nœud, situé au centre du réseau et ayant peu de voisins très influents, peut avoir plus d'influence qu'un nœud ayant un plus grand nombre de voisins moins influents. Considérant ce fait, l'algorithme de décomposition *k-shell* peut être utilisé (Seidman, 1983). Son principe est d'attribuer un indice de référence ks pour chaque nœud tel que les nœuds ayant les valeurs les plus faibles sont situés à la périphérie du réseau tandis que les nœuds avec les valeurs les plus élevées se trouvent au centre du réseau, ce sont alors ces nœuds qui auront le plus d'influence. Bien qu'ayant adapté l'algorithme de décomposition *k-shell* aux caractéristiques du réseau *Twitter*, ils ont observé que leurs résultats sont fortement biaisés. Ainsi, ils proposent une modification de l'algorithme utilisant une échelle logarithmique, afin de produire des valeurs de *k-shell* moins nombreuses et plus significatives. L'approche proposée dans (Qasem *et al*., 2015) détecte les utilisateurs qui augmentent la taille du réseau

5. http://nlp.uned.es/replab2014/#dataset

6. En théorie des graphes, le cœfficient de clustering mesure à quel point les voisins d'un sommet sont connectés.

social en attirant de nouveaux utilisateurs dans le réseau, le nombre de *followers* est utilisé pour analyser la taille du réseau.

D'autres recherches proposent de classer les utilisateurs en utilisant des algorithmes basés sur le PageRank (Page *et al.*, 1999). L'idée principale du PageRank est que "Les pages les plus importantes (des sites Web) sont susceptibles de recevoir plus de liens à partir d'autres pages" ; dans le cadre de l'influence, l'hypothèse est qu'un utilisateur influent doit être en relation avec de nombreux voisins très influents. Plusieurs adaptations de l'algorithme PageRank ont été proposées afin de classer les utilisateurs influents dans *Twitter*, (Riquelme, González-Cantergiani, 2016) en recense 17. Dans la suite nous en présentons quelques-unes. Daniel Tunkelang a proposé TunkRank (Tunkelang, 2009) pour calculer l'influence d'un utilisateur à partir de l'influence de ses *followers* en prenant en compte les faits suivants :

– si i appartient aux *followers* de j, alors il y a $\frac{1}{|following(i)|}$ probabilité que i lise un *tweet* émis par j où $following(i)$ est l'ensemble des utilisateurs que i suit ;

– si i lit un *tweet* émis par j, il y a une probabilité de p pour que i le *retweete*.

L'influence de i est alors donnée par :

$$TunkRank(i) = \sum_{Y \in followers(i)} \frac{1 + p * TunkRank(j)}{|following(j)|}$$

(Kwak *et al.*, 2010; Ashwini, Sindhu, 2015) proposent des approches similaires mais travaillent sur la relation *retweet*.

NodeRanking (Pujol *et al.*, 2002) est une autre variante du PageRank avec deux différences : 1) il travaille sur des graphes pondérés, 2) le facteur de téléportation est calculé pour chaque nœud. Ghosh et al. (Ghosh *et al.*, 2012) avec CollusionRank, proposent aussi une approche basée sur PageRank mais ils initialisent avec une valeur négative les scores des nœuds identifiés comme spammeurs. Ainsi, un utilisateur est pénalisé pour avoir suivi des spammeurs et non pour être suivi par des spammeurs, le score CollusionRank d'un nœud est calculé en fonction du score de ses *followings* (et non de ses *followers*). Par conséquent, les utilisateurs qui suivent un plus grand nombre de spammeurs ou qui suivent ceux qui à leur tour suivent des spammeurs, reçoivent un score négatif et sont poussés vers le bas du classement. Dans (Lü *et al.*, 2011), les auteurs proposent l'algorithme LeaderRank qui travaille avec la relation *followers*. LeaderRank est basé sur PageRank mais le réseau est rendu fortement connecté par l'introduction d'un nœud g ayant deux arcs orientés e_{gi} et e_{ig} vers chaque nœud i du réseau d'origine, ce qui permet à l'algorithme de converger plus rapidement. LeaderRank donne de meilleurs résultats que PageRank en termes d'efficacité de classement et de robustesse contre les manipulations des données biaisées. Li et al. (Li *et al.*, 2014) améliorent LeaderRank en introduisant un mécanisme de pondération ; les nœuds pondérés avec leur différents nombres de *followers* obtiennent des rangs différents à partir du nœud g. Dans (Weng *et al.*, 2010), les auteurs proposent TwitterRank afin de mesurer l'influence des utilisateurs en tenant compte des thématiques (*via* les hashtags) associés aux tweets. Bien que l'idée soit prometteuse, les résultats

expérimentaux montrent qu'il y a des utilisateurs qui *suivent* d'autres utilisateurs sans présence de similarité de thématiques entre eux. La méthode a aussi ignoré d'autres critères importants tels que les *mentions* et les *réponses*.

Romero et al. (Romero *et al*., 2011) ont modifié l'algorithme HITS (Hyperlink-Induced Topic Search), un algorithme d'analyse de liens qui évalue les pages Web, développé par Jon Kleinberg (Kleinberg, 1999). HITS attribue deux scores pour chaque page : un score d'autorité qui estime la valeur du contenu de la page, et un score de hub qui estime la valeur de ses liens vers d'autres pages. Les auteurs considèrent l'influence comme le niveau de propagation du contenu dans le réseau (relation *retweets*). De plus, ils estiment que l'influence d'un utilisateur ne dépend pas seulement de la taille de son audience, mais aussi de sa passivité. La passivité d'un utilisateur est le fait qu'il ne transmet pas l'information au réseau. Cet algorithme a montré une meilleure précision que d'autres mesures d'influence tels que PageRank, le nombre de *followers* et le nombre de *mentions*. Bien que la passivité semble être un facteur à prendre en compte dans le calcul de l'influence, ce travail a ignoré d'autres relations importantes telles que la *réponse*.

Dans des travaux récents, la **fusion d'information** est considérée afin de contourner les limitations des approches précédentes. Dans (Simmie *et al*., 2013), les auteurs proposent la combinaison de deux modèles pour classer les utilisateurs influents : l'algorithme PageRank et un HMM (Hidden Markov Model). Le modèle permet d'observer l'évolution de l'influence à travers le temps et utilisent les trois relations *retweet*, *mention* et *réponse*. Le modèle est évalué sur une enquête considérée comme une réalité du terrain. Le modèle proposé diffère des autres par la combinaison de trois relations. Toutefois, puisque le but est de classer l'influence des utilisateurs, l'influence d'un utilisateur donné ne révèle pas d'informations sur son degré d'influence (forte ou faible influence), le résultat du modèle est utile uniquement pour le classement des utilisateurs. De plus, les auteurs n'offrent pas une mesure d'influence en exploitant la combinaison de critères avec leur incertitude inhérente.

L'inconvénient des algorithmes basés sur la topologie du réseau est de considérer les informations de l'utilisateur, c'est-à-dire les liens des nœuds, sans considérer les interactions complexes entre les utilisateurs à travers des séquences de liens. Les méthodes basées sur les algorithmes PageRank et HITS ont pour principale lacune qu'ils ne traitent que d'une seule relation, c'est-à-dire un seul type de liens, à la fois. Néanmoins, ces travaux nous aident à déterminer les critères à prendre en compte dans l'évaluation de l'influence. L'étude des différents travaux montre que les approches sont nombreuses et que la notion d'influence n'a pas de définition consensuelle mais qu'elle dépend fortement du domaine étudié. Ainsi, on n'utilisera pas les mêmes relations pour étudier l'influence des candidats politiques que pour étudier l'influence d'utilisateurs dans une campagne marketing. Par ailleurs, aucun travail de recherche existant ne prend en compte le contenu des *tweets* pour étudier la polarité de l'influence. De plus, la combinaison de plusieurs relations avec de l'incertitude n'a pas été considérée. Or, il nous parait important, pour mesurer l'influence, de tenir compte des degrés d'incertitude sur les poids attribués aux différentes interactions selon leur

importance. Dans cet objectif, dans des recherches récentes, la théorie des fonctions de croyance est exploitée pour mesurer l'influence dans des réseaux pondérés (Wei *et al*., 2013) et complexes (Mo *et al*., 2015) avec l'objectif commun de modifier les mesures de centralité existantes. Au meilleur de notre connaissance, c'est la première fois que la théorie des fonctions de croyance est exploitée pour mesurer l'influence sur le réseau *Twitter* avec des patterns d'interactions au lieu des mesures de centralité.

2.2. L'analyse de sentiments dans Twitter

L'analyse du langage subjectif a été largement appliquée à la classification des opinions et des émotions dans le texte (Wiebe *et al*., 2005). En effet, l'analyse du sentiment, qui vise à annoter le texte à l'aide d'une échelle mesurant le degré de sentiment négatif et positif dans le texte, est considérée comme l'un des axes de recherche les plus importants pour les chercheurs dans les domaines de recherche d'information, fouille de données et apprentissage automatique.

Dans ce contexte, *Twitter* a constitué le terrain de jeu le plus utilisé pour les solutions d'analyse de sentiments, les entreprises et les scientifiques tentent de comprendre l'enthousiasme des utilisateurs partageant leurs opinions publiquement en ligne. Une des difficultés inhérentes dans l'analyse de sentiments est la traduction des données textuelles dans un format que l'ordinateur peut comprendre et traiter. Pour cette raison, un certain nombre de méthodes de Traitement Automatique de Langage (TAL) ont été développées au fil des ans. Les plus populaires sont le sac de mots et les N-grammes. Le sac de mots peut être considéré comme la méthode la plus simple. Selon cette approche, les phrases du document (ou texte) dont la machine a besoin pour juger le sentiment sont divisées en un ensemble de mots en utilisant l'espace ou les caractères de ponctuation (Pak, Paroubek, 2010). Un document ou un texte particulier est représenté par les occurrences des mots le composant. Les N-grammes sont très semblables au sac de mots mais la différence réside dans le fait que le texte est divisé en pseudo-mots consécutifs de longueur égale (Pang, Lee, 2008). La longueur N dépend de la nature des documents ou de textes d'entrée et du problème à résoudre. Généralement, 2-grammes, 3-grammes et 4-grammes sont utilisés. Les N-grammes permettent par exemple d'attacher la négation avec le mot qui la suit (par exemple : je n'aime pas). Une telle procédure permet d'améliorer l'analyse de sentiments dans des textes puisque la négation joue un rôle particulier dans une expression d'opinion et de sentiment.

La popularité des réseaux sociaux a rendu la tâche de l'analyse de sentiments difficile. En effet, les textes à analyser sont devenus courts, contenant de nombreuses abréviations, ainsi que de nombreuses erreurs de syntaxe et de grammaire. Alors, il est impératif de filtrer et de nettoyer les textes avant toute étape d'analyse de sentiments. Plusieurs possibilités s'offrent telles que l'élimination des mots vides (ou *stop words*), se sont des mots qui sont tellement communs qu'il est inutile de les utiliser dans une recherche. En français, des mots vides évidents pourraient être « le », « la », « de », « du », « ce ». Nous pouvons aussi faire la racinisation des mots (*stemming*),

c'est-à-dire la transformation/réduction des mots en leur racine qui correspond à la partie du mot restante une fois que l'on a supprimé son (ses) préfixe(s) et suffixe(s).

L'étape suivante est l'analyse des textes préparés et filtrés en utilisant différents algorithmes d'apprentissage automatique (classifieurs). Dans (Psomakelis *et al*., 2015), les auteurs présentent une revue des algorithmes les plus populaires de l'analyse de sentiments dans *Twitter*. Ils ont testé plusieurs algorithmes de classification en utilisant le sac de mots et les N-grammes. Les algorithmes utilisés sont : les machines à vecteurs de support, la classification naïve bayésienne, la régression logistique, le perceptron multi-couches et les arbres de décision. Les résultats ont montré la supériorité des performance de l'algorithme de régression logistique en utilisant 5-grammes.

Dans (Burnap, Williams, 2015), Burnap et al. ont proposé un classifieur de *tweets* par rapport au discours de haine. La méthode est basée sur le sac de mots, un dictionnaire de termes et phrases de discours de haine de Wikipedia et les dépendances typées (De Marneffe *et al*., 2006) afin de représenter les relations grammaticales entre les mots dans une phrase. Les auteurs utilisent les algorithmes des machines à vecteurs de support, les forêts d'arbres décisionnels et la régression logistique bayésienne. Dans les résultats des expérimentations, les performances des différents algorithmes utilisés étaient similaires et les critères les plus efficaces sont les N-grammes combinés avec le dictionnaire des termes relatifs à la haine. Dans (Burnap, Williams, 2016), les auteurs proposent un modèle de classification supervisé pour la détection de haine par rapport aux sujets : race, handicap et orientation sexuelle. Ils se sont basés sur quatre critères : le sac de mots, les N-grammes, des termes et phrases de discours de haine de Wikipedia et les dépendances typées. Deux algorithmes de classification ont été utilisés, les machines à vecteurs de support et les forêts d'arbres décisionnels. Les résultats ont montré que l'utilisation des dépendances typées est très intéressante contrairement aux termes de discours de haine qui sont des indicateurs faibles. Dans (Burnap *et al*., 2015), les auteurs développent une application de fouille d'opinion capable de classer les *tweets* publics en tenant compte du niveau de tension. Les algorithmes de machines à vecteurs de support et la classification naïve bayésienne ont été utilisés en se basant sur les N-grammes et sur un ensemble de mots préalablement classés en tant que mots qui expriment la tension. Les résultats indiquent que le dictionnaire de mots utilisé est un fort indicateur de tension.

Il existe également des sites disponibles en ligne pour analyser les sentiments dans *Twitter* [7] en introduisant le nom de la personne ou l'entité à propos de laquelle nous souhaitons connaître le sentiment. Ces outils sont des boîtes noires et ne montrent pas l'approche utilisée pour analyser les sentiments.

Ainsi, dans le domaine de l'analyse de sentiments, la méthode dépend du domaine étudié, certains algorithmes ou méthodes peuvent donner de bons résultats dans un domaine et échouer dans d'autres. Le principal paramètre à prendre en compte est

7. http://socialmouths.com/2010/03/31/6-tools-for-twitter-sentiment-tracking/

l'adéquation entre les caractéristiques retenues pour la modélisation et la question à laquelle on cherche à répondre.

2.3. Théorie des fonctions de croyance

La théorie des fonctions de croyance est considérée comme un outil général pour le raisonnement avec incertitude, et a été reliée à d'autres cadres tels que les théories des probabilités, des possibilités et des probabilités imprécises (Denoeux, Masson, 2012). La théorie des fonctions de croyance, aussi connue comme la théorie de l'évidence ou théorie de Dempster-Shafer, a d'abord été introduite par A. Dempster dans le contexte de l'inférence statistique, et a été développée plus tard par G. Shafer comme un outil général pour la modélisation de l'incertitude épistémique (Kotz, N. L. Johnson eds., 1982).

Dans les paragraphes suivants, nous allons rappeler les concepts de base de la théorie des fonctions de croyance. Soient Ω un ensemble fini de réponses à une question et 2^Ω l'ensemble de tous les sous-ensembles de Ω. Dans le contexte de la théorie des fonctions de croyance, Ω est souvent appelé un cadre de discernement. La masse de croyance m est une fonction $m : 2^\Omega \rightarrow [0, 1]$ tel que :

$$\sum_{X \in 2^\Omega} m(X) = 1 \text{ and } m(\emptyset) = 0 \tag{1}$$

La masse $m(X)$ exprime la part de la croyance qui accrédite le sous-ensemble X de Ω, $m(\emptyset) = 0$ car nous considérons que le cadre de discernement est exhaustif et exclusif c'est-à-dire que nous connaissons toutes les réponses.

La théorie des fonctions de croyance permet, non seulement la représentation de la connaissance partielle, mais aussi la fusion de l'information (Nimier, Appriou, 1995). En prenant différentes sources d'information, nous cherchons à les fusionner avec une seule masse de croyance. La fusion d'information est réalisée par la règle de combinaison conjonctive (Smets, 1997), elle suppose que toutes les sources sont fiables et consistantes. Considérant deux fonctions de masse m_1 et m_2, la règle de combinaison conjonctive est définie par :

$$(m_1 \textcircled{\cap} m_2)(C) = \sum_{A \cap B = C} m_1(A) m_2(B), \quad A, B, C \in 2^\Omega \tag{2}$$

Afin de prendre une décision, il est nécessaire de sélectionner l'hypothèse la plus probable, ce qui peut être difficile à réaliser directement avec les fonctions de croyance où les fonctions de masse sont données, non seulement pour les singletons, mais aussi pour les sous-ensembles du cadre de discernement. Il existe plusieurs solutions pour assurer la prise de décision au sein de la théorie des fonctions de croyance, la plus connue est la probabilité pignistique (Smets, 1989). Contrairement aux fonctions de masse qui sont définies sur 2^Ω, la probabilité pignistique est une mesure de probabilité définie sur Ω. La probabilité pignistique a été proposée dans le modèle des croyances

transférables (Smets, Kennes, 2008). Elle est basée sur deux niveaux : le "niveau crédal" où les croyances sont représentées par des fonctions de croyance et le "niveau pignistique" où les croyances sont utilisées pour prendre la décision et représentées comme des fonctions de probabilité appelées probabilités pignistiques et notées *bet* définies par :

$$\mathrm{bet}(x) = \sum_{x \in X \subseteq \Omega} \frac{m(X)}{|X|} + \frac{1}{1 - m(\emptyset)} \tag{3}$$

La probabilité pignistique consiste à répartir équitablement chaque masse de croyance entre les singletons de X.

3. Approche proposée

Afin de mesurer l'influence polarisée d'un utilisateur, nous utilisons : 1) l'algorithme des forêts d'arbres décisionnels pour analyser la polarité des *tweets* et 2) la théorie des fonctions de croyance pour effectuer la fusion des informations issues des différentes relations. La figure 1 donne un aperçu des étapes principales de l'approche proposée. Tout d'abord, les données de *Twitter* sont modélisées par un réseau multi-relationnel obtenu en sélectionnant les relations pertinentes. L'étape suivante est l'analyse de sentiments. En utilisant l'algorithme des forêts d'arbres décisionnels, nous analysons le contenu des *tweets* pour déduire leurs polarités : positif, neutre ou négatif. Nous en déduisons trois sous-réseaux, chacun représentant une polarité. Enfin, dans l'étape de l'évaluation de l'influence, nous effectuons le choix des relations pertinentes et des masses de croyance, ce choix dépend du domaine étudié. Ensuite, pour mesurer l'influence polarisée d'un certain utilisateur, nous combinons les différentes masses de croyance associées à chaque relation considérée pour obtenir la masse de croyance combinée relative à chaque sous-réseau. Après, nous combinons les masses de croyances résultantes de chaque sous-réseau pour déduire l'influence polarisée pour chaque utilisateur.

3.1. Modélisation du réseau Twitter

Les réseaux sociaux sont généralement modélisés comme un graphe[8] représenté par $G = (V, E)$ comprenant un ensemble V de sommets ou nœuds et un ensemble E d'arcs ou de liens (Barnes, 1969).

Dans le réseau *Twitter*, le graphe est hétérogène puisque nous avons différents liens (relations) entre les nœuds et différents types de nœuds. Par exemple, il peut exister un lien représentant la relation *suivre* entre deux utilisateurs, un lien *retweet* entre un

8. Nous utilisons la terminologie de réseau complexe pour désigner des données du monde réel et le terme de graphe pour désigner les objet mathématiques/théoriques associés.

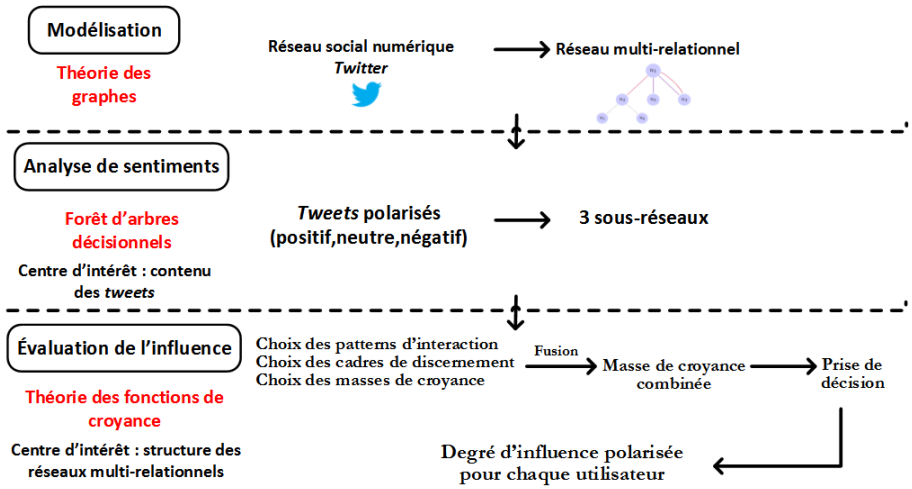

Figure 1. Étapes de l'approche proposée

tweet et un utilisateur. Afin de modéliser l'hétérogénéité des liens, un réseau multi-relationnel ou multi-couches ou multiplexes peut être utilisé (Kivelä *et al.*, 2014; Kanawati, 2015; Dai *et al.*, 2012). Dans un réseau multi-relationnel, l'ensemble des liens E est divisé en classes disjointes : $E = \bigcup_{r \in R} E_r$, où R est l'ensemble de relations possibles, l'ensemble des nœuds est homogène. De Domenico *et al.* (2013) ont transposé des outils courants comme les centralités de degré et de vecteur propre, les cœfficients de clustering, les algorithmes de marches aléatoires (*random walks*), la modularité sur un réseau multi-relationnel. (Wu *et al.*, 2013b) ont étudié la détection de communauté dans les réseaux multi-relationnels.

Nous définissons un *pattern* d'interaction p comme une séquence de relations, par exemple, un *retweet* contenant une *mention* ou un *retweet* d'une *réponse*. Soit P l'ensemble des *patterns* d'interaction possibles, notons par $R = R \cup P$ l'ensemble des relations y compris les *patterns* d'interaction. Par exemple, dans *Twitter*, nous pouvons considérer $R =$ {*retweet*, *mention*, *réponse*, *suivre*, *retweet* + *réponse*, *retweet* + *mention*, *mention* + *mention*}.

La figure 2 présente une modélisation possible du réseau *Twitter* $G = (V, E)$ où V est l'ensemble des nœuds de type utilisateurs et où chaque couche représente un type de relation avec $R =$ {*retweet*, *mention* et *réponse*}. Par exemple, dans la couche *mention* @, l'arc entre u_1 et u_2 signifie que l'utilisateur u_2 mentionne u_1 ; dans la couche *Réponse* **Rép** l'utilisateur u_5 a répondu à u_2. Les arcs présents dans une seule couche sont dits liens intra-couche ; les arcs présents entre deux différentes couches sont dit lien inter-couches, ils sont représentés par des flèches pointillées dans la figure 2, par exemple, l'arc **Rép** @ est un arc inter-couches, il signifie que u_5 a fait une réponse contenant une *mention* à u_1.

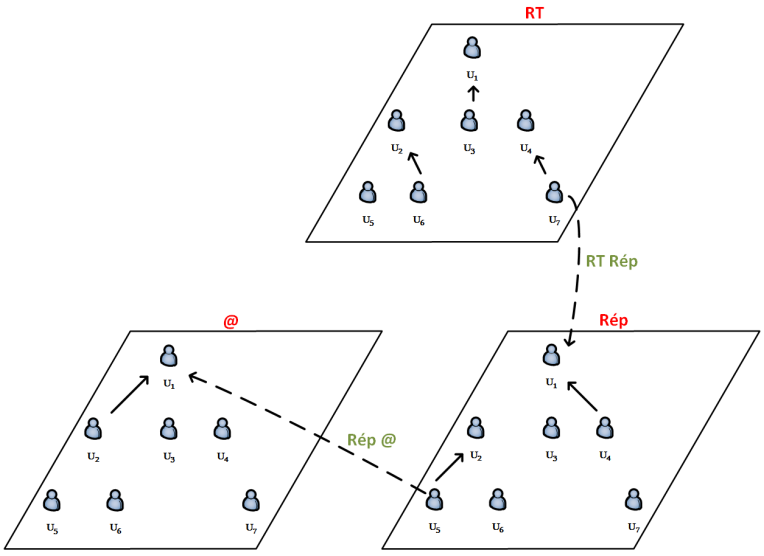

Figure 2. Un réseau multi-relationnel de Twitter

3.2. Analyse de sentiments des tweets

L'objectif de cette étape est d'analyser la polarité des *tweets*, c'est-à-dire déterminer si le contenu des *tweets* est positif, négatif ou neutre. Pour construire un modèle d'analyse de sentiments qui sera capable de classer les *tweets*, nous prenons un ensemble de *tweets* manuellement annotés par des experts selon leur polarité et nous essayons de déduire les critères, présents dans les *tweets*, qui permettent de déterminer leurs polarités. Cet ensemble de *tweets* est divisé en deux, le premier est un échantillon à partir duquel on construit le modèle de classification, ensuite le modèle construit est utilisé pour prédire la polarité du deuxième échantillon de *tweets*. Enfin, nous comparons les prédictions avec leur polarité spécifiée pour évaluer les performances du modèles construit.

La première étape de la construction du modèle est la préparation des données. Il s'agit de présenter les *tweets* à analyser sous une forme compréhensible par les différents algorithmes d'analyse de sentiments. Pour ceci, chaque *tweet* est transformé en N-grammes. Le processus d'obtention de N-grammes à partir d'un *tweet* est le suivant :

1. **Le filtrage :** nous supprimons les mots vides, les liens URL, les noms d'utilisateur Twitter (avec le symbole @ indiquant un nom d'utilisateur), les mots spéciaux de *Twitter* (tel que «RT»), les ponctuations et les émoticônes.

2. **Le sac de mots :** les *tweets* sont divisés en un ensemble de mots en utilisant l'espace entre les mots.

3. **Les N-grammes :** les *tweets* sont représentés aussi sous forme de N-grammes de mots consécutifs, ce qui permet par exemple d'attacher la négation avec le mot qui la suit.

L'étape suivante est l'utilisation d'un algorithme d'analyse de sentiments. Nous construisons un classifieur en utilisant l'algorithme des forêts d'arbres décisionnels. Nous avons aussi essayé les algorithmes des machines à vecteurs de support, la classification naïve bayésienne et la régression logistique mais nous avons choisi l'algorithme des forêts d'arbres décisionnels car il nous a permis d'obtenir les meilleurs résultats de classification. Les forêts d'arbres décisionnels ont été formellement proposées en 2001 par Leo Breiman (Breiman, 2001). Elles font partie des techniques d'apprentissage automatique. L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents. La proposition de Breiman vise à corriger plusieurs inconvénients connus de la méthode initiale des arbres de décision dont la principale limite est la dépendance des performances de l'échantillon de départ. Il s'agit précisément de l'*overfitting*, il se produit quand un modèle est excessivement complexe, comme avoir trop de paramètres par rapport au nombre d'observations. L'*overfitting* d'un modèle se traduit par de mauvaises performances prédictives sur des données autres que les jeux d'apprentissage ou de tests. Dans le principe des forêts d'arbres décisionnels, plutôt que d'avoir un modèle d'arbre décisionnel complexe, il s'agit de construire de nombreux modèles d'arbres décisionnels moins performants individuellement. Chaque modèle a sa vision du problème et fait au mieux pour le résoudre avec les données dont il dispose. Les modèles sont unis pour donner une vision globale du problème, ce qui rend les forêts d'arbres décisionnels très efficaces. Le nom **forêt** d'arbres vient du fait de la construction de nombreux arbres de décision. Pour éviter d'avoir des arbres semblables, chaque arbre utilise au hasard différentes observations et variables, on parle d'arbres décisionnels **aléatoires**.

Après avoir analysé le sentiment des *tweets*, le réseau multi-relationnel étudié est divisé en trois sous-réseaux, chaque sous-réseau représente une polarité : positif, négatif et neutre.

3.3. Évaluation de l'influence polarisée

Notre objectif est d'évaluer l'influence polarisée des utilisateurs dans *Twitter*. La propagation d'information et donc l'influence des utilisateurs est essentiellement due aux *retweets* et aux *mentions*. Les *retweets* permettent d'atteindre les *followers* de l'utilisateur, les *mentions* permettent en revanche d'atteindre n'importe qui directement et donc de rendre l'information plus visible en ciblant les utilisateurs les plus appropriés. Ainsi, les relations sont les critères de manifestation de l'influence d'un utilisateur.

L'influence d'un utilisateur est déterminée par l'importance des relations qui lui sont associées. Chaque relation est associée à un degré d'influence d_r pour $r \in R$, par exemple, la relation *retweet* est associée au degré d'influence $d_{retweet} =$ Très Faible.

Soit Ω_{Inf} l'ensemble de tous les degrés d'influence. Les fonctions de masse expriment un lien entre les différentes relations qui jouent sur l'influence d'un utilisateur, elles représentent l'importance des relations, une fonction de masse est associée pour chaque relation, les fonctions de masses sont définies comme suit : $m_r : \Omega_{Inf} \rightarrow [0, 1]$. Alors, pour chaque relation $r \in R$, en plus du degré d'influence d_r, une fonction de masse m_r est associée. d_r et m_r dépendent de la relation.

En se basant sur la théorie des fonctions de croyance présentée dans la section 2, nous fusionnons différentes fonctions de masse définies dans le réseau multi-relationnel. Afin d'estimer le degré d'influence polarisée d'un nœud spécifique u, nous prenons en compte la structure locale de chaque sous-réseau multi-relationnel autour du nœud u et nous combinons les fonctions de masses de croyance des liens incidents de chaque sous-réseau en utilisant une version modifiée de la règle de combinaison conjonctive (2), nous combinons ensuite le résultat de la fusion des masses de chaque sous-réseau pour déduire l'influence globale polarisée :

$$(m \otimes m')(z) = \sum_{y \circledcirc x = z} m(x)m'(y), \quad x, y, z \in \Omega_{Inf} \tag{4}$$

$\circledcirc$ est une opération qui donne l'intersection entre les différents degrés d'influence ou les polarités. Nous détaillons dans la section suivante (section 4) le processus de la combinaison en fonction de l'opération $\circledcirc$.

L'algorithme 1 formalise l'étape d'évaluation de l'influence polarisée, il requiert en entrée, les trois sous-réseaux multi-relationnels $g_n \subset G, n \in 1, 2, 3$, et l'initialisation des masses pour les différentes relations m_r avec $r \in R$. Pour chaque utilisateur, l'algorithme commence par mesurer l'influence relative à chaque sous-réseau, c'est-à-dire l'influence relative à chaque polarité, pour ceci, il compte le nombre d'occurrences pour chaque relation présente dans chaque sous-réseau. Ensuite, pour chaque relation de type r, en utilisant la formule (4), il calcule la combinaison des masses de croyances. La formule (4) est à nouveau utilisée pour combiner les masses de croyance pour toutes les relations, nous obtenons ainsi l'influence relative à chaque sous-réseau/polarité. Enfin, en utilisant la formule (4) mais avec une autre opération $\circledcirc$, l'algorithme calcule l'influence polarisée globale en fusionnant les mesures d'influence résultant de chaque sous-réseau. L'algorithme renvoie l'influence polarisée finale : le degré d'influence qui est le degré ayant la probabilité pignistique maximale ; la masse de croyance et la polarité.

Le code source est disponible sur github [9], il s'agit du code R général qui peut être spécialisé en fonction du réseau étudié et des relations utilisées.

9. https://github.com/kerzol/Influence-assessment-in-twitter

Algorithme 1 : Évaluation de l'influence polarisée

Input : $g_n \subset G, n \in 1, 2, 3$, les sous-réseaux multi-relationnels
L'ensemble des relations $R = r_1, r_2, \ldots$
Initialisation des masses $m_r, r \in R$
Output : Degré d'influence Inf_u, Masse de croyance M_u, Polarité Pol_u

for $u \in U$ **do**
  for $g_n \subset G$ **do**
    for $i \in [1..|R|]$ **do**
      ℓ_{u,r_i,g_n} := nombre de relations de type r_i pour l'utilisateur u, dans le sous-réseau g_n ;
      $M_{u,r_i,g_n} := m_{r_i}$;
      for $i \in [2..\ell_{u,r_i,g_n}]$ **do**
        $M_{u,r_i,g_n} := M_{u,r_i,g_n} \otimes m_{r_i}$;
      end
    end
    $M_{u,g_n} := M_{u,r_1,g_n}$;
    for $i \in [2..|R|]$ **do**
      $M_{u,g_n} := M_{u,g_n} \otimes M_{u,r_i,g_n}$;
    end
  end
  $M_{u1,2,3} := M_{u,g_1} \otimes M_{u,g_2} \otimes M_{u,g_3}$; // Notons que $\otimes$ dépend de l'opération @.
  Bet_u := Distribution de la probabilité pignistique ;
  Inf_u := Degré d'influence maximal;
  M_u := Masse de croyance correspondante au degré d'influence Inf_u ;
  Pol_u := Polarité de l'influence ;
end
return $\mathrm{Inf}_u, M_u, Pol_u, u \in U$;

4. Expérimentations et résultats

Dans cette section, nous commençons par la description des données utilisées dans les expérimentations, ensuite, nous appliquons les différentes étapes de l'approche décrite dans la section précédente sur les données. Nous présentons également des exemples d'illustration pour certaines étapes afin de mieux comprendre le fonctionnement de l'approche.

4.1. Description des données

Les travaux de recherche menés se déroulent dans le cadre du projet TEE 2014 dont l'intitulé exact est "*Twitter* aux élections européennes : une étude contrastive internationale des utilisations de *Twitter* par les candidats aux élections au Parlement

Européen en mai 2014". Ce projet international, mené par la Maison des Sciences de l'Homme (MSH) de Dijon, réunit près de 45 chercheurs (majoritairement des politologues, sociologues, chercheurs en communication) de 10 laboratoires de recherche répartis dans 6 pays européens (France, Allemagne, Belgique, Italie, Espagne et Royaume-Uni). L'objectif global de ce projet est d'observer et d'analyser la communication des politiques sur *Twitter* durant les élections européennes de mai 2014 dans les 6 pays couverts par l'étude.

Pour collecter les informations de *Twitter*, nous avons utilisé notre outil *SNFreezer*[10] (Leclercq *et al*., 2015). Trois types d'informations (généralisées sous le terme "source") peuvent être pris en paramètre dans cette collecte : des comptes utilisateurs, des *hashtags* et des mots ou phrases. Ces différentes sources ont été choisies par les politologues, et nous retrouvons parmi elles les noms des principaux candidats, leurs comptes *Twitter*, et les *hashtags* relatifs à ces candidats, leurs partis, ou plus généralement à l'élection étudiée. L'objectif de la collecte est de capter les *tweets* mentionnant les utilisateurs désignés, ceux contenant un certain *hashtag*, mot ou phrase, ou encore les *tweets* envoyés par les utilisateurs spécifiés. En plus, nous collectons les informations sur ces *tweets* tels que les *tweets retweetés*, les utilisateurs *mentionnés* dans les *tweets* et les *réponses* aux *tweets*. La collecte sur deux mois a produit 50 millions de *tweets* pour un volume de 50Go environ. Dans nos expérimentations, nous nous concentrons sur le corpus français. Le tableau 1 présente les paramètres du jeu de données utilisé.

Tableau 1. Paramètres des données relatives au corpus français

Nombre de *tweets*	4 593 665
Nombre d'utilisateurs	937 860
Nombre de candidats	616
Nombre de relations	2 922 566
Nombre de *retweets*	639 531
Nombre de *mentions*	1 945 773
Nombre de *réponses*	337 262

4.2. Modélisation

L'objectif de nos expérimentations est de mesurer l'influence polarisée des candidats sur *Twitter*. Dans l'étape de la modélisation, les données collectées sont représentées dans un réseau multi-relationnel composé de différentes couches où chaque couche représente un type de relation. La figure 3 montre une représentation visuelle partielle du réseau multi-relationnel correspondant à la couche *retweet* des candidats français. Afin de contourner la complexité visuelle de tout le graphe, nous n'utilisons que 1% de toutes les données du graphe. Les grands nœuds correspondent

10. https://github.com/SNFreezer

aux comptes des candidats, les petits nœuds représentent les autres utilisateurs. Les liens représentent la relation *retweet*.

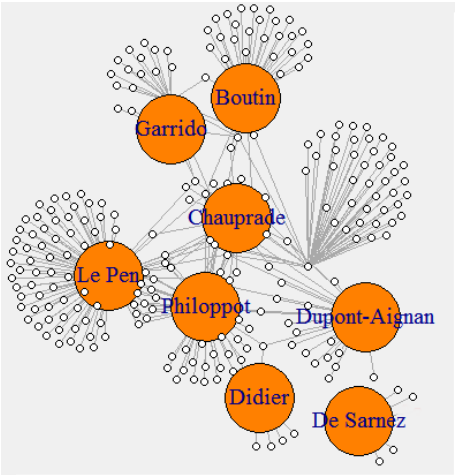

Figure 3. Un exemple de la couche retweet du réseau multi-relationnel TEE2014

4.3. Analyse de sentiments

Pour construire le modèle d'analyse de sentiments, nous avons choisi aléatoirement un échantillon de 2000 *tweets* composé de 1000 *tweets* contenant des *mentions* des candidats et 1000 *réponses* aux candidats. Ensuite, les *tweets* ont été manuellement annotés par des sociologues du projet TEE 2014 selon leurs polarités : positif, neutre ou négatif. La première étape de la construction du modèle d'analyse de sentiments est le filtrage des *tweets* et la préparation des données, tous les *tweets* sont transformés en sac de mots. Nous supprimons les mots vides, les liens URL, les noms d'utilisateur Twitter (y compris les candidats), les mots spéciaux de *Twitter* (tels que «RT»), les ponctuations et les émoticônes. Ensuite, les *tweets* sont divisés en un ensemble de mots en utilisant l'espace entre les mots. Nous avons aussi transformé les *tweets* en 3-grammes et 2-grammes mais nous avons constaté que les résultats des performances sont meilleures sans transformation en N-grammes (Voir tableau 2). Après le filtrage, les *tweets* sont nettoyés et prêts à être utilisés par l'algorithme d'analyse de sentiments. Nous avons obtenu 861 termes (mots) différents.

Nous procédons ensuite à une approche de validation croisée, une méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage. L'échantillon est divisé en deux, 1700 *tweets* comme ensemble de validation et les 300 *tweets* restants constitueront l'ensemble d'apprentissage. Nous appliquons l'algorithme des forêts d'arbres décisionnels pour construire le modèle de classification des *tweets* à partir de l'échantillon d'apprentissage. Le choix de cet algorithme est justifié par sa forte performance dans les travaux de l'état de l'art. En se basant sur leurs polarités réelles, l'algorithme cherche à identifier les termes présents dans les

tweets qui permettent d'identifier leurs polarités. Le nombre d'arbres générés est fixé à 500 arbres décisionnels, paramètre généralement utilisé dans les tâches de classification. Ensuite, le modèle construit est testé sur l'échantillon de validation (1700 *tweets*) pour étudier sa performance en comparant les prédictions du modèle avec leurs polarités réelles. La métrique d'évaluation utilisée est le F-mesure, basé sur la Précision P et le Rappel R calculés pour chaque polarité, ce qui est classique dans les tâches de classification. Le F-Score est calculé comme suit:

$$F = 2 \times \frac{(P \times R)}{P + R} \tag{5}$$

$P = VP/VP + FP$; $R = VP/VP + FN$ (où $VP =$ vrais positifs, $FP =$ faux positifs, et $FN =$ faux négatifs).

Le tableau 2 représente les différentes valeurs de F-mesures avec la variation des paramètres N-grammes et le pourcentage des termes sparses, il s'agit d'ignorer les termes qui ont une sparsité supérieure à un seuil donné (la sparsité = 1 - fréquence), ce qui peut aider à prévenir l'*overfitting*. Par exemple, si sparse = à 0.8, cela supprimera chaque terme qui apparaît dans moins de 20% de documents. Au contraire, si sparse = 0.01, seuls les termes qui apparaissent dans presque chaque document seront conservés. En langage naturel, des mots communs comme «le» sont susceptibles de se produire dans chaque texte et donc ne seront jamais sparses.

Tableau 2. F-mesure en fonction des variations des différents paramètres

N-grammes	3	2	Non	Non	**Non**
Termes sparses	0.997	0.997	0.991	0.993	**0.997**
Nombre de termes	59	67	20	45	**117**
Précision	0.66	0.76	0.78	0.78	**0.79**
Rappel	0.61	0.64	0.72	0.75	**0.76**
F-mesure	0.63	0.69	0.74	0.76	**0.77**

Nous constatons que les meilleures valeurs de F-mesure ont été obtenues en utilisant sparse = 0.997 et sans transformation en N-grammes. Le nombre de termes utilisés est 117. Ces paramètres constituent notre modèle qui a été appliqué ensuite sur tout le corpus TEE 2014. La figure 4 montre les 20 mots les plus utilisés dans les *tweets* filtrés. La taille des mots représentent l'importance de leurs utilisations dans les *tweets*, par exemple, le mot "pas" est le plus utilisé. Après application du modèle sur les *retweets*, *mentions* et *réponses* relatives aux candidats, nous avons obtenu 1981 *tweets* positifs, 53706 neutres et 3417 négatifs.

Figure 4. Nuage de mots des tweets filtrés

4.4. Évaluation de l'influence polarisée

Dans cette sous-section, nous commençons par 1) choisir les paramètres indispensables pour notre algorithme, 2) ensuite, nous effectuons l'évaluation de l'influence dans chaque sous-réseau, 3) et enfin, nous fusionnons l'influence de chaque sous-réseau pour obtenir l'influence polarisée. Pour les deux dernières étapes, nous expliquons la méthode utilisée puis nous montrons un exemple d'illustration pour mieux comprendre la démarche suivie et nous donnons enfin les résultats des expérimentations sur TEE 2014.

4.4.1. Choix des paramètres : cadres de discernement, relations et masses de croyance

Soit Ω_{Inf} les différentes réponses possibles à notre question "Quel est le degré d'influence d'un certain utilisateur ?", $\Omega_{Inf} = \{$Très Faible, Faible, Assez Moyenne, Moyenne, Assez forte, Forte, Très Forte, Extrêmement Forte$\}$ est l'ensemble ordonné des degrés d'influence possibles. Dans la théorie des fonctions de croyance, 2^{Ω} est utilisé comme domaine des fonctions de masse, dans nos expérimentations, nous n'avons besoin que d'un sous-ensemble Ω_I de $2^{\Omega_{Inf}}$, à savoir : $\Omega_I = \{$Très Faible, Faible, Assez Moyenne, Moyenne, Assez Forte, Forte, Très Forte, Extrêmement Forte, Ω_{Inf} $\}$. Nous étudions aussi la polarité de l'influence mesurée. Ω_{Pol} représente les différentes réponses possibles à notre question : "Quelle est la polarité de l'influence d'un certain utilisateur ?" Soit Ω_{Pol} l'ensemble des polarités d'influence possibles : $\Omega_{Pol} = \{$Positive, Neutre, Négative$\}$. De même, nous utilisons un sous-ensemble Ω_P de $2^{\Omega_{Pol}}$, précisément : $\Omega_P = \{$Positive, Neutre, Négative, $\Omega_{Pol}\}$

Afin d'étudier l'influence polarisée des candidats, nous affectons des masses aux relations considérées, puis pour chaque candidat, nous combinons les masses de croyances en itérant sur le nombre d'occurrences des relations. Les relations et *patterns* considérés dans nos expérimentations sont : $R =\{$*retweet*, *mention*, *réponse*, *retweet* + *réponse*, *retweet* + *mention*$\}$. Il existe d'autres patterns tels que *retweet* d'un *retweet*

ou *réponse* d'un *retweet* mais les APIs de *Twitter* ne les donnent pas, par exemple, une *réponse* à un *retweet* est stockée comme une *réponse* au *tweet* initial et non pas au *retweet*.

Le choix et l'affectation des masses dans l'étape d'initialisation sont une question importante lorsque nous traitons de données réelles. Dans certains domaines tels que la politique, les utilisateurs ont un très grand nombre de relations. Avec une initialisation des masses de valeurs importantes (par exemple 0,4 pour chaque relation), l'influence, étant le résultat de la combinaison de ces différentes masses, converge pour tous les candidats vers le plus haut degré d'influence Extrêmement Forte avec la valeur de masse maximale 1 après un nombre d'itérations restreint ($\simeq$ 45 itérations), ce qui ne nous permet pas de pouvoir comparer l'influence des candidats. Pour régler cette question, nous effectuons une mise à l'échelle et nous utilisons les affectations de masses suivantes :

$$\textit{retweet} \mapsto \begin{cases} m_{\text{retweet}}(\text{T.Faible}) = 0.55 \cdot 10^{-3} \\ m_{\text{retweet}}(\Omega_{Inf}) = 1 - 0.55 \cdot 10^{-3} \end{cases}$$

$$\textit{mention} \mapsto \begin{cases} m_{\text{mention}}(\text{T.Faible}) = 0.45 \cdot 10^{-3} \\ m_{\text{mention}}(\Omega_{Inf}) = 1 - 0.45 \cdot 10^{-3} \end{cases}$$

$$\textit{réponse} \mapsto \begin{cases} m_{\text{réponse}}(\text{T.Faible}) = 0.45 \cdot 10^{-3} \\ m_{\text{réponse}}(\Omega_{Inf}) = 1 - 0.45 \cdot 10^{-3} \end{cases}$$

$$\textit{retweet + réponse} \mapsto \begin{cases} m_{\text{retweet+réponse}}(\text{T.Faible}) = 0.75 \cdot 10^{-3} \\ m_{\text{retweet+réponse}}(\Omega_{Inf}) = 1 - 0.75 \cdot 10^{-3} \end{cases}$$

$$\textit{retweet + mention} \mapsto \begin{cases} m_{\text{retweet+mention}}(\text{T.Faible}) = 0.65 \cdot 10^{-3} \\ m_{\text{retweet+mention}}(\Omega_{Inf}) = 1 - 0.65 \cdot 10^{-3} \end{cases}$$

4.4.2. Fusion des masses dans chaque sous-réseau

Pour mesurer l'influence d'un candidat, nous commençons par calculer le nombre d'occurrences de chaque relation $r \in R$ dans chaque sous-réseau. Ensuite, en utilisant la formule (4), nous combinons les masses de croyance des relations de chaque sous-réseau, nous obtenons ainsi l'influence relative à chaque sous-réseau/polarité. Le tableau 3 représente l'opération $\textcircled{@}_1$ utilisée dans cette étape de fusion, elle donne les intersections entre les différents degrés d'influence. Cette fonction assure notre hypothèse : plus nous combinons des relations relatives à un utilisateur, plus il devient influent.

Exemple d'illustration :

Afin de mieux comprendre l'étape de la fusion des masses dans chaque sous réseau, nous présentons un exemple d'illustration dans lequel nous considérons

Tableau 3. Définition de l'opération $@_1$

$@_1$	**T.Faible**	**Faible**	**A.Moyenne**	**Moyenne**	**A.Forte**	**Forte**	**T.Forte**	**E.Forte**	Ω_{Inf}
T.Faible	Faible	A.Moyenne	Moyenne	A.Forte	Forte	T.Forte	T.Forte	E.Forte	T.Faible
Faible	A.Moyenne	A.Moyenne	Moyenne	A.Forte	Forte	T.Forte	T.Forte	E.Forte	Faible
A.Moyenne	Moyenne	Moyenne	A.Forte	Forte	T.Forte	T.Forte	T.Forte	E.Forte	A.Moyenne
Moyenne	A.Forte	A.Forte	Forte	Forte	T.Forte	T.Forte	T.Forte	E.Forte	Moyenne
A.Forte	Forte	Forte	T.Forte	T.Forte	T.Forte	T.Forte	T.Forte	E.Forte	A.Forte
Forte	T.Forte	T.Forte	T.Forte	T.Forte	T.Forte	E.Forte	E.Forte	E.Forte	Forte
T.Forte	T.Forte	T.Forte	T.Forte	T.Forte	T.Forte	E.Forte	E.Forte	E.Forte	T.Forte
E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte
Ω_{Inf}	T.Faible	Faible	A.Moyenne	Moyenne	A.Forte	Forte	T.Forte	E.Forte	Ω_{Inf}

les fonctions de masse suivantes associées à la relation *retweet* et au pattern *retweet+mention* :

$$\text{retweet} \mapsto \begin{cases} m_{\text{retweet}}(\text{Faible}) = 0.4 \\ m_{\text{retweet}}(\Omega_{Inf}) = 0.6 \end{cases} \qquad \text{retweet+mention} \mapsto \begin{cases} m_{\text{retweet+mention}}(\text{Moyenne}) = 0.7 \\ m_{\text{retweet+mention}}(\Omega_{Inf}) = 0.3 \end{cases}$$

Les masses de croyance $m_{\text{retweet}}(\Omega_{Inf})$ et $m_{\text{retweet+mention}}(\Omega_{Inf})$ représentent l'ignorance partielle. Pour cet exemple, nous n'utilisons pas les masses initialisées dans la sous-section 4.4.1 car elles sont destinées à être utilisées pour les données TEE 2014 où il y a un grand nombre de relations à traiter. Ici nous combinons uniquement deux relations. Nous avons affecté une masse de croyance plus importante au pattern d'interaction *retweet+mention* car nous considérons que l'existence de ce pattern est très significative en terme d'influence. Pour effectuer la combinaison d'un *retweet* avec le pattern *retweet+mention*, nous utilisons d'abord l'opération $@_1$ donnant les correspondances entre les degrés d'influence (tableau 3), après nous calculons la combinaison conjonctive en utilisant la formule (4). La fonction de masse combinée des deux relations est donnée dans le tableau 4:

Tableau 4. Combinaison d'un retweet avec un retweet d'une mention

$\otimes_{@_1}$	Faible 0.4	Ω_{Inf} 0.6
Moyenne 0.7	Assez Forte 0.28	Moyenne 0.42
Ω_{Inf} 0.3	Faible 0.12	Ω_{Inf} 0.18

Nous obtenons : $m(\text{Faible}) = 0.12 \quad m(\text{Moyenne}) = 0.42 \quad m(\text{Assez Forte}) = 0.28$
$m(\Omega_{Inf}) = 0.18$

Application sur TEE 2014 :

Après avoir présenté un exemple du fonctionnement de l'étape fusion des masses dans un sous-réseau, nous présentons maintenant les résultats obtenus après application de cette étape sur les données TEE 2014. Le tableau 5 montre les résultats dans le sous-réseau contenant des *tweets* de polarité positive pour les trois candidats français Marine Le Pen, Florian Philippot et Jean-Luc Mélenchon. Nous pouvons conclure que le degré d'influence dans le sous-réseau positif de *Twitter* pour la candidate Marine Le Pen est Très Forte avec une masse de croyance 0.294735272.

Tableau 5. Résultats pour 3 candidats dans le sous-réseau positif

	M. Le Pen	**F. Philippot**	**J.L. Mélenchon**
Ω_{Inf}	0.004484713	0.05811933	0.223688674
Très Faible	0.024371473	0.000011065	0.336662752
Faible	0.066098388	0.12563168	0.251646905
Assez Moyenne	0.119288674	0.18062197	0.124552509
Moyenne	0.161160203	0.19430546	0.045920874
Assez Forte	0.161160203	0.16682792	0.013451569
Forte	0.156003603	0.11908255	0.003260986
Très Forte	0.294735272	0.14199894	0.000815730
Extrêmement Forte	0	0	0

À la fin de cette étape, pour chaque polarité $p \in P$, chaque candidat est représenté par un degré d'influence et une masse de croyance. Par exemple, le candidat Marine Le Pen est représenté de la manière suivante :

$$\text{Marine Le Pen} \left\{ \begin{array}{r} \text{Positive, T.Forte, m = } 0.294735272 \\ \text{Neutre, T.Forte, m = } 0.979944 \\ \text{Négative, T.Forte, m = } 0.503692449 \end{array} \right\}$$

4.4.3. Fusion des masses des trois sous-graphes/polarités

Dans cette étape, les masses de croyances associées à chaque polarité sont fusionnées en utilisant la formule (4) afin d'obtenir l'influence globale polarisée de chaque candidat. D'abord, la fonction $\textcircled{@}_2$ (Tableau 6) est utilisée pour donner l'intersection entre les polarités, ensuite, la fonction $\textcircled{@}_3$ présentée dans le tableau 7 est utilisée pour donner les intersections entre les différents degrés d'influence, ici, nous n'utilisons pas l'opération $\textcircled{@}_1$ comme dans l'étape précédente car $\textcircled{@}_1$ assure l'hypothèse que plus nous combinons, plus l'influence est importante. Mais dans cette étape, les relations sont déjà combinées dans les trois sous-réseaux et l'objectif est d'avoir le degré d'influence global.

Une fois que nous avons la masse de croyance de l'influence fusionnée d'un certain candidat, nous utilisons une version modifiée de la probabilité pignistique définie dans la formule (3) afin de prendre la décision à propos du degré d'influence globale polarisée. Dans notre cas, les masses de croyance sont définies sur Ω_I et Ω_P et la

Tableau 6. Définition de l'opération $@_2$

$@_2$	Positive	Neutre	Négative	Ω_P
Positive	Positive	Positive	Neutre	Positive
Neutre	Positive	Neutre	Négative	Neutre
Négative	Neutre	Négative	Négative	Négative
Ω_P	Positive	Neutre	Négative	Ω_P

Tableau 7. Définition de l'opération $@_3$

$@_3$	T.Faible	Faible	A.Moyenne	Moyenne	A.Forte	Forte	T.Forte	E.Forte	Ω_{Inf}
T.Faible	T.Faible	T.Faible	Faible	A.Moyenne	A.Moyenne	Moyenne	Moyenne	A.Forte	T.Faible
Faible	T.Faible	Faible	Faible	A.Moyenne	A.Moyenne	Moyenne	Moyenne	A.Forte	Faible
A.Moyenne	Faible	Faible	A.Moyenne	A.Moyenne	Moyenne	Moyenne	A.Forte	Forte	A.Moyenne
Moyenne	A.Moyenne	A.Moyenne	A.Moyenne	Moyenne	Moyenne	A.Forte	A.Forte	Forte	Moyenne
A.Forte	A.Moyenne	A.Moyenne	Moyenne	Moyenne	A.Forte	A.Forte	Forte	Forte	A.Forte
Forte	Moyenne	Moyenne	Moyenne	A.Forte	A.Forte	Forte	Forte	T.Forte	Forte
T.Forte	Moyenne	Moyenne	A.Forte	A.Forte	Forte	Forte	T.Forte	T.Forte	T.Forte
E.Forte	A.Forte	A.Forte	Forte	Forte	Forte	T.Forte	T.Forte	E.Forte	E.Forte
Ω_{Inf}	T.Faible	Faible	A.Moyenne	Moyenne	A.Forte	Forte	T.Forte	E.Forte	Ω_{Inf}

probabilité pignistique est calculée en répartissant uniformément la masse de Ω_{Inf} et Ω_{Pol} sur tous les autres éléments de Ω_P, Ω_I :

$$\mathrm{bet}(x) = m(x) + \frac{m(\Omega_{Inf}, \Omega_{Pol})}{|\Omega_P|}, \quad x \in \Omega_P, \Omega_I \tag{6}$$

Exemple illustration :

Pour illustrer cette étape, nous présentons un exemple d'un utilisateur ayant les masses des polarités suivantes :

$$\text{Utilisateur} \left\{ \begin{array}{r} \text{Positive, Moyenne, m = 0.3} \\ \text{Neutre, A.Moyenne, m = 0.7} \\ \text{Négative, Faible, m = 0.9} \end{array} \right\}$$

Le tableau 8 montre la combinaison de masses des deux polarités Positive et Neutre. Nous commençons par combiner les degrés de polarités en utilisant $@_2$, par exemple, l'intersection entre les polarités Positive et Neutre est la polarité Positive. Ensuite, nous effectuons l'intersection entre les degrés d'influence en utilisant l'opération $@_3$.

Nous déduisons ainsi le degré et la masse correspondants à chaque polarité, par exemple, pour la polarité Positive, nous avons obtenu deux degrés d'influence avec des masses différentes dans le tableau 8, pour les fusionner, nous regardons l'intersection

Tableau 8. Combinaison des masses des polarités Positive et Neutre

$\otimes_{@_2,@_3}$	Neutre, A.Moyenne 0.3	$\Omega_{Pol}, \Omega_{Inf}$ 0.7
Positive, Moyenne 0.7	Positive, A.Moyenne 0.21	Positive, Moyenne 0.49
$\Omega_{Pol}, \Omega_{Inf}$ 0.3	Neutre, A.Moyenne 0.09	$\Omega_{Pol}, \Omega_{Inf}$ 0.21

entre les deux degrés A.Moyenne et Moyenne dans $@_3$, puis nous effectuons la somme des deux masses 0.21 + 0.49 = 0.7. Nous obtenons alors :

$$\left\{\begin{array}{r} \text{Positive, A.Moyenne, m = 0.7} \\ \text{Neutre, A.Moyenne, m = 0.09} \\ \Omega_{Pol}, \Omega_{Inf} \text{ , m = 0.21} \end{array}\right\}$$

Maintenant, nous combinons les masses de la fusion des polarités Positive et Neutre avec les masses de la polarité Négative, les résultats sont donnés dans le tableau 9.

Tableau 9. Combinaison des masses des trois polarités

$\otimes_{@_2,@_3}$	Positive, A.Moy 0.7	Neutre, A.Moy 0.09	$\Omega_{Pol}, \Omega_{Inf}$ 0.21
Négative, Faible 0.9	Positive, Faible 0.63	Négative, Faible 0.081	Négative, Faible 0.189
$\Omega_{Pol}, \Omega_{Inf}$ 0.1	Positive, A.Moy 0.07	Positive, A.Moy 0.09	$\Omega_{Pol}\Omega_{Inf}$ 0.021

Nous obtenons alors : $\left\{\begin{array}{r} \text{Positive, A.Moyenne, m = 0.07} \\ \text{Neutre, Faible, m = 0.639} \\ \text{Négative, Faible, m = 0.27} \\ \Omega_{Pol}, \Omega_{Inf} \text{ , m = 0.021} \end{array}\right\}$

Finalement, pour prendre la décision sur le degré d'influence polarisée, nous calculons la probabilité pignistique en utilisant la formule (6) (Tableau 10). Par exemple, pour la polarité Neutre, nous procédons comme suit pour obtenir la probabilité pignistique :

$$\text{bet(Neutre, Faible)} = m(\text{Neutre, Faible}) + \frac{m(\Omega_{Inf}, \Omega_{Pol})}{|\Omega_P|} = 0.639 + \frac{0.021}{3} = 0.646$$

Nous pouvons conclure que l'influence polarisée est Positive avec le degré A.Moyenne et la probabilité pignistique 0.077. L'influence est positive car cette polarité a le degré d'influence le plus élevé (A.Moyenne).

Tableau 10. Probabilité pignistique

Positive, A.Moyenne	Neutre, Faible	Négative, Faible
0.077	0.646	0.277

Application sur TEE 2014 :

Nous avons appliqué notre approche sur le corpus français comprenant 616 candidats et 4 millions de *tweets*. Les résultats pour 3 candidats sont présentés dans le tableau 11. Les résultats fournissent non seulement le degré d'influence d'un candidat, mais aussi indiquent sa polarité et donnent par les masses une indication de la croyance que nous avons dans les résultats donnés.

Tableau 11. Résultats de l'influence polarisée pour 3 candidats

Candidat	**Polarité**	**Degré d'influence**	**Masse de croyance**
M. Le Pen	Neutre	Très Forte	0.4923
F. Philippot	Neutre	Forte	0.4592
J.L. Mélenchon	Neutre	Assez Forte	0.6879

4.5. Classement de l'influence des utilisateurs

L'approche proposée peut être aussi exploitée pour classer les utilisateurs selon leur influence. Afin de classer les candidats selon leur influence et en partant de nos résultats, nous procédons comme suit :

1. D'abord, nous classons les candidats selon leurs polarités en utilisant l'ordre suivant : Négative $<$ Neutre $<$ Positive.

2. Ensuite, pour les candidats ayant la même polarité, nous les classons selon les degrés d'influence en suivant l'ordre suivant : T.Faible $<$ Faible $<$ A.Moyenne $<$ Moyenne $<$ A.Forte $<$ Forte $<$ T.Forte $<$ E.Forte .

3. Enfin, pour les candidats ayant la même polarité et le même degré d'influence, nous les classons selon les masses de croyances qu'ils ont sur leurs degré d'influence. Les résultats sont présentés dans le tableau 12.

Tableau 12. Candidats français les plus influents selon notre approche

Classement	**Candidats**	**Polarité**	**Degré d'influence**	**Masse de croyance**
1	Marine Le Pen	Neutre	Très Forte	0.4923
2	Florian Philippot	Neutre	Forte	0.4592
3	Jean-Luc Mélenchon	Neutre	Assez Forte	0.6879
4	Christine Boutin	Neutre	Moyenne	0.7892
5	Aymeric Chauprade	Neutre	Moyenne	0.5489
6	Nicolas Dupont-Aignan	Neutre	Moyenne	0.4871
7	Geoffroy Didier	Neutre	Moyenne	0.4029
8	José Bové	Neutre	Moyenne	0.39845
9	Marielle De Sarnez	Neutre	Moyenne	0.36514
10	Raquel Garrido	Neutre	Assez Moyenne	0.796314

Le tableau 13 présente le classement obtenu en utilisant les critères utilisés par (Cha *et al.*, 2009). Ces critères sont le nombre de *retweets, mentions et réponses*. Le

tableau 14 présente le classement des candidats selon l'algorithme du HITS en utilisant les relations *réponse, retweet et mention*. Les résultats présentés ne montrent pas l'influence globale dans le réseau puisque nous trouvons différents classements pour chaque type de relation. Alors que notre méthode (Tableau 12) nous permet d'avoir un classement unique qui tient compte de toutes les relations considérées ainsi que le contenu des *tweets*. La dernière colonne du tableau 13 montre le classement des candidats selon leur degré de centralité calculé en utilisant le nombre de voisins de chaque candidat dans les trois sous-réseaux. Le degré de centralité permet, pour chaque candidat, d'avoir un classement global sans indication sur leur degré d'influence contrairement à nos résultats présentés dans le tableau 12, l'influence mesurée est globale en prenant en compte de la polarité et des relations possibles dans la même mesure. Dans (Azaza *et al*., 2016), nous avons testé notre approche sur un autre jeu de données : CLEF RepLab 2014[11].

Tableau 13. Candidats français les plus influents selon les différentes relations et le degré de centralité

Classement	***Retweet***	***Mention***	***Réponse***	***Degré de centralité***
1	Marine Le Pen	Marine Le Pen	Christine Boutin	Marine Le Pen
2	Florian Philippot	Christine Boutin	Marine Le Pen	Christine Boutin
3	Jean-Luc Mélenchon	Jean-Luc Mélenchon	Florian Philippot	Florian Philippot
4	Aymeric Chauparde	Florian Philippot	Jean-Luc Mélenchon	Jean-Luc Mélenchon
5	François Asselineau	Nicolas Dupont-Aignan	Louis de Gouyon Matigon	Nicolas Dupont-Aignan
6	Corinne Morel-Darleux	José Bové	Nicolas Dupont-Aignan	Aymeric Chauparde
7	Nicolas Dupont-Aignan	Aymeric Chauparde	Jean-Sébastien Herpin	José Bové
8	Louis Aliot	Raquel Garrido	Julien Rochedy	Geoffroy Didier
9	Denis Payre	Jérome Lavrilleux	Geoffroy Didier	Raquel Garrido
10	Yannick Jadot	Marielle de Sarnez	Louis Aliot	Yannick Jadot

Tableau 14. Candidats français les plus influents selon l'algorithme du HITS

Classement	***HITS-Réponse***	***HITS-Retweet***	***HITS-Mention***
1	M. Le Pen	M. Le Pen	M. Le Pen
2	C. Boutin	A. Chauprade	A. Chauparde
3	F. Filippot	B. Monot	F. Philippot
4	J. Mélenchon	F. Philippot	J.M. Le Pen
5	N. Dupont-Aignan	N. Bay	L. Aliot
6	A. Chauparde	B. Gollnisch	B. Monot
7	J. Bove	A. Guibert	G. Didier
8	G. Didier	G. Lebreton	J. Rochedy
9	R. Garrido	J.M. Le Pen	G. Lebreton
10	Y. Jadot	K. Ouchikh	B. Gollnisch

11. http://nlp.uned.es/replab2014/#dataset

5. Conclusion

Dans cet article, nous avons proposé une approche pour une évaluation de l'influence polarisée sur un réseau multi-relationnel obtenu à partir du réseau social *Twitter*. Cette approche répond à des limites des systèmes existants tels que la prise en compte de la combinaison des relations et l'incertitude engendrée par la fusion d'informations. Dans notre approche, nous traitons un réseau multi-relationnel de l'influence en considérant plusieurs relations ainsi que les séquences possibles de relations. En utilisant l'algorithme des forêts d'arbres décisionnels, nous avons classé les *tweets* selon leurs polarités. Ensuite, en se basant sur la théorie des fonctions de croyance, nous avons établi une mesure d'influence polarisée globale pour les utilisateurs par combinaison des différentes relations. Nous avons expérimenté l'approche sur des données *Twitter* collectées dans le cadre du projet de recherche TEE 2014. Pour renforcer l'approche proposée, nous souhaitons enrichir le modèle et développer d'autres patterns d'interaction en collaboration avec les politologues du projet TEE 2014. Nous souhaitons intégrer dans notre approche les *hashtags* et les *favoris* et ainsi pouvoir traiter des patterns plus complexes. Cependant, l'extraction des patterns du jeu de données nécessite des structures de données adaptées pour concevoir une application utilisable par les politologues et permettant d'évaluer et d'affiner leur modèle de l'influence. En outre, nous allons appliquer l'approche proposée sur d'autres mesures qui nécessitent la fusion de l'information comme l'estimation de la crédibilité des utilisateurs et la catégorisation des styles de diffusion dans *Twitter*. Nous étudierons aussi l'applicabilité de l'approche dans d'autres réseaux sociaux (*facebook, instagram, forums...*) et pour d'autres domaines que la politique. Et enfin, nous souhaitons adapter la méthode pour prendre en compte les aspects temporels, à savoir l'évolution de l'influence plutôt que l'influence à un instant donné.

Bibliographie

Ashwini S. S., Sindhu M. (2015). Profile Ranking Using User Influence and Content Relevance with Classification Using Sentiment Analysis. *International Journal of Computer Science and Mobile Computing*, vol. 4, n° 6, p. 1075–1080.

Azaza L., Kirgizov S., Savonnet M., Leclercq É., Gastineau N., Faiz R. (2016). Information fusion-based approach for studying influence on twitter using belief theory. *Computational Social Networks*, vol. 3, n° 1, p. 5–31.

Bakshy E., Hofman J. M., Mason W. A., Watts D. J. (2011). Everyone's an Influencer: Quantifying Influence on Twitter. In *Proceedings of the fourth acm international conference on web search and data mining*, p. 65–74. New York, NY, USA, ACM.

Barnes J. A. (1969). Graph Theory and Social Networks: A Technical Comment on Connectedness and Connectivity. *Sociology*, p. 215-232.

Basaille I., Kirgizov S., Leclercq E., Savonnet M., Cullot N. (2016). Towards a twitter observatory: A multi-paradigm framework for collecting, storing and analysing tweets. In *Tenth IEEE international conference on research challenges in information science, (RCIS)*, p. 1–10.

Breiman L. (2001). Random forests. *Machine Learning*, vol. 45, n° 1, p. 5–32.

Brown P. E., Feng J. (2011). Measuring user influence on twitter using modified k-shell decomposition. In *Fifth international aaai conference on weblogs and social media*, p. 18-23.

Burnap P., Rana O. F., Avis N., Williams M., Housley W., Edwards A. *et al.* (2015). Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change*, vol. 95, p. 96–108.

Burnap P., Williams M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet*, vol. 7, n° 2, p. 223–242.

Burnap P., Williams M. L. (2016). Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, vol. 5, n° 1, p. 11.

Cha M., Haddadi H., Benevenuto F., Gummadi P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *Icwsm*, vol. 10, n° 30, p. 10-17.

Cha M., Mislove A., Gummadi K. P. (2009). A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In *Proceedings of the 18th international conference on world wide web*, p. 721–730.

Chen D.-B., Gao H., Lü L., Zhou T. (2013). Identifying Influential Nodes in Large-Scale Directed Networks: The Role of Clustering. *PLoS ONE*, vol. 8, n° 10, p. 1–10.

Cossu J., Labatut V., Dugué N. (2015). A review of features for the discrimination of twitter users: Application to the prediction of offline influence. *CoRR*, vol. abs/1509.06585.

Dai B. T., Chua F. C. T., Lim E.-P. (2012). Structural Analysis in Multi-Relational Social Networks. In *Proceedings of the 2012 siam international conference on data mining*, p. 451-462.

De Domenico M., Solé-Ribalta A., Cozzo E., Kivelä M., Moreno Y., Porter M. A. *et al.* (2013, Dec). Mathematical formulation of multilayer networks. *Phys. Rev. X*, vol. 3, p. 041022.

De Marneffe M.-C., MacCartney B., Manning C. D. *et al.* (2006). Generating typed dependency parses from phrase structure parses. , vol. 6, n° 2006, p. 449–454.

Denoeux T., Masson M.-H. E. (2012). Belief Functions: Theory and Applications. In *Proceedings of the 2nd international conference on belief functions*, p. 9–11.

Ghosh S., Viswanath B., Kooti F., Sharma N. K., Korlam G., Benevenuto F. *et al.* (2012). Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on world wide web*, p. 61–70.

Kanawati R. (2015). Multiplex network mining: a brief survey. *IEEE Intelligent Informatics Bulletin*, vol. 16, n° 1, p. 24–27.

Kivelä M., Arenas A., Barthelemy M., Gleeson J. P., Moreno Y., Porter M. A. (2014). Multilayer networks. *Journal of complex networks*, vol. 2, n° 3, p. 203–271.

Kleinberg J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal ACM*, vol. 46, n° 5, p. 604–632.

Kotz S., N. L. Johnson eds. W. (1982). Belief functions. *Encyclopedia of Statistical Sciences 1 209*.

Kwak H., Lee C., Park H., Moon S. (2010). What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th international conference on world wide web*, p. 591–600.

Leavitt A., Burchard E., Fisher D., Gilbert S. (2009). The Influentials: New Approaches for Analyzing Influence on Twitter. *Webecology Project*.

Leclercq E., Savonnet M., Grison T., Kirgizov S., Basaille I. (2015). SNFreezer: a Platform for Harvesting and Storing Tweets in a Big Data Context. In A. Frame, A. Mercier, G. Brachotte, C. Thimm (Eds.), *Twitter and the european parliamentary elections: researching political uses of microblogging*, p. 1–16. DE, Peter Lang.

Lee C., Kwak H., Park H., Moon S. (2010). Finding influentials based on the temporal order of information adoption in twitter. In *Proceedings of the 19th international conference on world wide web, www '10*, p. 1137–1138.

Li Q., Zhou T., Li L., Chen D. (2014). Identifying influential spreaders by weighted LeaderRank. *Physica A: Statistical Mechanics and its Applications*, vol. 404, p. 47–55.

Lü L., Zhang Y.-C., Yeung C. H., Zhou T. (2011). Leaders in social networks, the delicious case. *PloS one*, vol. 6, n° 6, p. e21202.

Mo H., Gao C., Deng Y. (2015). Evidential method to identify influential nodes in complex networks. *Systems Engineering and Electronics, Journal of*, vol. 26, n° 2, p. 381-387.

Neves A., Vieira R., Mourão F., Rocha L. (2015). Quantifying Complementarity among Strategies for Influencers' Detection on Twitter. *Procedia Computer Science*, vol. 51, p. 2435-2444.

Nimier V., Appriou A. (1995). Utilisation de la théorie de dempster-shafer pour la fusion d'informations. *GRETSI, Groupe d'Etudes du Traitement du Signal et des Images*, p. 137-140.

Page L., Brin S., Motwani R., Winograd T. (1999). The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th international world wide web conference*, p. 161–172.

Pak A., Paroubek P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. , vol. 10, n° 2010, p. 1320–1326.

Pang B., Lee L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, vol. 2, n° 1-2, p. 1–135.

Psomakelis E., Tserpes K., Anagnostopoulos D., Varvarigou T. A. (2015). Comparing methods for Twitter Sentiment Analysis. *CoRR*, vol. abs/1505.02973.

Pujol J. M., Sangüesa R., Delgado J. (2002). Extracting Reputation in Multi Agent Systems by Means of Social Network Topology. In *Proceedings of the first international joint conference on autonomous agents and multiagent systems: Part 1*, p. 467–474. ACM.

Qasem Z., Jansen M., Hecking T., Hoppe H. (2015). On the detection of influential actors in social media. In *Signal-image technology and internet-based systems (sitis), 11th international conference on*, p. 421–427.

Riquelme F., González-Cantergiani P. (2016). Measuring user influence on twitter: A survey. *Information Processing & Management*, vol. 52, n° 5, p. 949–975.

Rodriguez M. A., Shinavier J. (2010). Exposing multi-relational networks to single-relational network analysis algorithms. *Journal of Informetrics*, vol. 4, n° 1, p. 29–41.

Romero D. M., Galuba W., Asur S., Huberman B. A. (2011). Influence and passivity in social media. In *Proceedings of the 20th international conference companion on world wide web*, p. 113–114.

Seidman S. B. (1983). Network structure and minimum degree. *Social Networks*, vol. 5, n° 3, p. 269 - 287.

Simmie D., Vigliotti M., Hankin C. (2013). Ranking twitter influence by combining network centrality and influence observables in an evolutionary model. In *International conference on signal-image technology internet-based systems (sitis)*, p. 491–498.

Smets P. (1989). Constructing the pignistic probability function in a context of uncertainty. , vol. 89, n° 1989, p. 29–40.

Smets P. (1997). Imperfect Information: Imprecision and Uncertainty. In A. Motro, P. Smets (Eds.), *Uncertainty management in information systems*, p. 225-254.

Smets P., Kennes R. (2008). The transferable belief model. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, p. 693-736.

Suh B., Hong L., Pirolli P., Chi E. H. (2010). Want to Be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In *Proceedings of the 2010 ieee second international conference on social computing*, p. 177–184. Washington, DC, USA, IEEE Computer Society.

Sun J., Tang J. (2011). A Survey of Models and Algorithms for Social Influence Analysis. In *Social network data analytics*, p. 177–214.

Tunkelang D. (2009). *A Twitter Analog to PageRank.* http://thenoisychannel.com/2009/01/13/a-twitter-analog-to-pagerank.

Wei D., Deng X., Zhang X., Deng Y., Mahadevan S. (2013). Identifying influential nodes in weighted networks based on evidence theory. *Physica A: Statistical Mechanics and its Applications*, vol. 392, n° 10, p. 2564 - 2575.

Weng J., Lim E.-P., Jiang J., He Q. (2010). TwitterRank: Finding Topic-sensitive Influential Twitterers. In *Proceedings of the third acm international conference on web search and data mining*, p. 261–270. New York, NY, USA, ACM.

Wiebe J., Wilson T., Cardie C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, vol. 39, n° 2, p. 165–210.

Wu Z., Yin W., Cao J., Xu G., Cuzzocrea A. (2013a). Community detection in multi-relational social networks. In *Web Information Systems Engineering – WISE 2013*, p. 43-56.

Wu Z., Yin W., Cao J., Xu G., Cuzzocrea A. (2013b). Web information systems engineering – wise 2013: 14th international conference, nanjing, china, october 13-15, 2013, proceedings, part ii. In X. Lin, Y. Manolopoulos, D. Srivastava, G. Huang (Eds.),, p. 43–56. Berlin, Heidelberg, Springer Berlin Heidelberg.