Détection d'influenceurs dans des médias sociaux

Kévin Deturck[1, 2]
(1) ERTIM, 2 rue de Lille, 75007 Paris, France
(2) Viseo, 4 avenue Doyen louis Weil, 38000 Grenoble, France
kevin.deturck@viseo.com

RÉSUMÉ

Les influenceurs ont la capacité d'avoir un impact sur d'autres individus lorsqu'ils interagissent avec eux. Détecter les influenceurs permet d'identifier les quelques individus à cibler pour toucher largement un réseau. Il est possible d'analyser les interactions dans un média social du point de vue de leur structure ou de leur contenu. Dans nos travaux de thèse, nous abordons ces deux aspects. Nous présentons d'abord une évaluation de différentes mesures de centralité sur la structure d'interactions extraites de Twitter puis nous analysons l'impact de la taille du graphe de suivi sur la performance de mesures de centralité. Nous abordons l'aspect linguistique pour identifier le changement d'avis comme un effet de l'influence depuis les messages d'un forum.

ABSTRACT

Influencer detection in social medias

There is an increasing interest in the detection of influencers on social medias. Different features may be used: the text of the messages and the structure of the network. The initial PhD works presented in this paper explore these two aspects. We evaluate the effectiveness of centrality measures from the state of the art in detecting Twitter influencers building graphs from interactions between Twitter users. In a second experimentation, we analyze the impact of the network size on the selection of the most appropriate centrality algorithms. We segment Twitter users according to the number of their followers and build their respective underlying "following graph". We then run the selected algorithms on the different graphs and evaluate their performance throughout the different graph sizes. We finally present a current work on linguistic features to detect opinion change as an influence effect.

MOTS-CLÉS : influence, média social, réseau, centralité, opinion

KEYWORDS: influence, social media, network, centrality, opinion

1 Introduction

Nos travaux de thèse visent à détecter les individus ayant une influence dans des médias sociaux. Nous entendons par "influence" le pouvoir que possède un individu qui parvient à mobiliser d'autres individus en faveur d'une action ou d'une opinion.

Détecter automatiquement les influenceurs dans des médias sociaux peut être vu comme une tâche pour un système de recherche d'information destiné à un utilisateur qui aurait besoin de connaitre les personnes meneuses et leurs propos dans un certain domaine d'activité. Les influenceurs constituent

aussi des points d'entrée efficaces pour la diffusion ciblée d'une information lors de campagnes de santé publique, pour la promotion d'un produit ou encore la gestion de réputation en ligne.

La pluralité des médias sociaux et la variété des informations qu'ils contiennent constituent autant de facettes possibles pour caractériser l'influence ouvrant ainsi de larges perspectives à l'élaboration d'un cadre formel permettant de la détecter. Il s'agit aussi d'une complexité supplémentaire pour modéliser l'information utile à la détection des influenceurs. Nous aurons à distinguer et à recouper la nature des informations disponibles à travers les différents médias sociaux pour les intégrer dans notre modèle. Par exemple, comparer la portée des interactions "Favori" dans Twitter et "J'aime" dans Facebook portant toutes les deux sur un contenu. Aussi, les textes que nous étudierons à travers les différents médias sociaux auront des formats tout à fait différents qui constitueront un corpus particulièrement hétérogène. Par exemple, nous ne pourrons pas aborder un texte de forum avec les mêmes modalités que celles utilisées pour un Tweet.

Les interactions entre les individus sont à la base de toute influence. Dans un média social, ces interactions peuvent être analysées d'après leur structure ou leur contenu. La structure peut porter une signification variée qui dépend à la fois des natures d'interaction considérées et de l'interprétation que nous leur prêtons. Nous nous situons dans le domaine du Traitement Automatique des Langues et nous focalisons donc nos travaux sur le contenu textuel des médias sociaux même si d'autres types de contenu comme l'image peuvent être utilisés par les influenceurs.

Identifier structurellement les influenceurs dans un groupe d'individus revient à trouver les clés de voûte de leurs interactions. Pour ce faire, nous devons trouver une représentation des interactions qui tienne compte de leurs natures variées. Par exemple, nous devons choisir comment inclure dans une même structure les interactions Twitter d'abonnement et de Retweet afin qu'elles soient les plus significatives possibles.

Nous avons à déterminer les phénomènes linguistiques qui permettent, du point de vue des influenceurs, de mettre en œuvre leur pouvoir et qui, du point de vue des individus cibles, font état de leurs réactions face aux influenceurs. Le principe est alors d'extraire des régularités expliquant sous différents types éventuels le phénomène d'influence par la langue écrite.

Les messages que nous devons analyser contiennent une expression informelle, un défi intéressant pour l'analyse linguistique parce que hors d'une grammaire standard de la langue. Pour l'analyse automatique, nous devrons mettre en œuvre un processus important de prétraitement permettant de ramener les textes des médias sociaux à une langue standard qui pourra être traitée par des outils de Traitement Automatique de la Langue. Le caractère elliptique des messages, en particulier dans les réseaux sociaux, peut empêcher l'identification de phénomènes linguistiques tels qu'ils sont traditionnellement décrits. Cela requiert une nouvelle modélisation des phénomènes qu'on veut repérer pour qu'ils soient applicables aux modes d'expression récents que nous souhaitons analyser. Prenons l'exemple d'un influenceur qui argumente pour changer l'opinion de quelqu'un. Pour caractériser son argumentation à l'intérieur d'un Tweet et la reconnaitre dans d'autres messages, nous ne pourrons certainement pas utiliser exactement les modèles traditionnels de l'argumentation.

2 État de l'art

2.1 Sur la détection structurelle des influenceurs

La centralité permet de mesurer l'importance structurelle d'un nœud dans un graphe. Nous pouvons analyser un média social par la structure de ses interactions représentées comme les arcs d'un graphe. Dans un réseau, l'influenceur est un utilisateur qui polarise les interactions, c'est donc un nœud central du graphe correspondant.

Le positionnement d'un individu dans un réseau est déterminé par ses connexions avec d'autres utilisateurs. Ces connexions sont données par les interactions de tous types entre les individus d'un média social. C'est plus particulièrement les réseaux sociaux qui sont à l'étude ici puisque la nature même du réseau rend essentielle l'analyse de ses liens. Un influenceur est alors identifié par un positionnement central à l'intérieur du réseau. Différentes mesures de centralité ont été proposées donnant lieu à des interprétations d'influence différentes. (Mariani et al., 2014) analysent les réseaux de citations pour mesurer l'influence scientifique. Ils appliquent une mesure de centralité qui prend en compte la proximité d'un utilisateur par rapport aux autres dans le graphe selon des liens que sont les citations des publications et les collaborations. La valeur de centralité pour chaque utilisateur indique son degré d'influence. (Sheikhahmadi et al., 2017) mettent en avant l'importance de la communauté dans les interactions des utilisateurs d'un réseau social. Les auteurs utilisent la tendance des utilisateurs à communiquer à l'intérieur d'un groupe constitué par exemple autour d'une thématique. Ils affirment donc que les influenceurs doivent être identifiés pour chaque communauté et divisent donc le graphe par détection de communauté avant d'utiliser la quantité d'interactions pour pondérer l'influence de chaque utilisateur dans sa communauté. (Khadangi, Bagheri, 2017) distinguent les interactions marquant une affinité entre les utilisateurs de celles qui reflètent l'activité propre de d'un utilisateur comme les « J'aime » sur Facebook, ces dernières étant moins étudiées pour l'influence.

Puisque les influenceurs ont la capacité de mobiliser d'autres individus pour des actions ou des opinions, nous pouvons aussi les identifier en analysant les dynamiques dans la structure des interactions. Un influenceur est alors un individu qui parvient à propager une attitude le plus rapidement et le plus longuement à travers un réseau. (Dave et al., 2011) cherchent à prédire quelle sera la dynamique de propagation d'une action à partir d'un certain utilisateur pour prédire son influence. (Jabeur et al., 2012) voient plus généralement les influenceurs comme ayant une capacité à propager une information. Les auteurs se concentrent ainsi, dans Twitter, sur les Retweets pour construire un graphe d'utilisateurs qu'ils analysent à la manière de Page Rank en considérant l'importance des utilisateurs Retweetant pour calculer l'influence de l'utilisateur Retweeté. (Gionis et al., 2013) s'intéressent à l'opinion des utilisateurs pour prédire ceux dont elle va le mieux se maximiser à travers un réseau d'après la constitution de leur voisinage respectif. Ces utilisateurs sont les clés du problème de maximisation d'influence à travers un réseau social. Le principe est de trouver l'ensemble d'utilisateurs qui permettra de diffuser une information le plus rapidement possible dans un réseau.

2.2 Sur la détection linguistique des influenceurs

Les approches qui analysent les textes issus de médias sociaux pour détecter les influenceurs, comme (Biran et al., 2012), essayent d'identifier les marqueurs d'un discours influent. L'influenceur

s'exprime souvent pour soutenir une revendication à propos d'un certain sujet. Cette revendication contient un point de vue, elle se caractérise donc essentiellement par une énonciation subjective. À ce propos, (Pak, Paroubek, 2010) cherchent à repérer automatiquement le point de vue adopté dans des Tweets par apprentissage automatique sur les catégories morphosyntaxiques en présence. Les auteurs mettent en avant l'utilisation d'adjectifs superlatifs pour renforcer l'expression des émotions et la forte présence de pronoms personnels indiquant une personnalisation du discours comme des traits particulièrement discriminants pour le discours subjectif. En général, l'influenceur essaye de faire adhérer d'autres personnes à son discours en utilisant deux grands procédés argumentatifs : persuader et convaincre.

La persuasion se caractérise par une implication fortement personnelle qui s'exprime avec des émotions et des opinions. La richesse des émotions exprimées peut, dans le cas de la persuasion, compenser l'absence de raisonnement. Persuader fait appel à une argumentation intuitive. La répétition d'un même propos entre ainsi dans le cadre de la persuasion.

Par contraste avec la persuasion, convaincre nécessite le développement d'une argumentation plus rationnelle pour servir la revendication exprimée. Il s'agit alors d'un discours moins direct fondé sur le raisonnement. (Rosenthal, 2014)

Détecter un texte argumentatif revient à faire de l'analyse du discours. L'argumentation est décrite en certaines relations rhétoriques présentées dans (Biran & Rambow, 2011) qui lient une partie revendicative à une partie justificative. Chacune de ces parties constitue une unité de discours à identifier. Une fois les unités de discours élémentaires identifiées, il est question de trouver les relations rhétoriques qu'il y a entre elles (Danlos, 2011). Ces relations permettent de structurer l'argumentation.

(Quercia et al., 2011) s'intéressent à l'usage de la langue chez les utilisateurs de Twitter dans une approche statistique pour identifier les influenceurs. Ils mettent en avant l'usage d'un champ lexical de l'émotion pour instaurer une intimité avec les lecteurs et ainsi faciliter l'impact du message à transmettre. Cette intimité est complétée par un usage de la troisième personne pour donner aux lecteurs l'impression d'appartenance à une même communauté qui se démarque des autres.

Pour repérer l'évolution d'une opinion sur Twitter, (Bifet et al., 2011) analysent la spécificité d'utilisation des termes en présence pour chaque Tweet d'un utilisateur à travers des fenêtres temporelles. Ils considèrent la polarité positive ou négative du lexique pour attribuer une valeur à l'opinion et ils utilisent les Hashtags pour créer des ensembles de Tweets portant sur une même thématique.

L'analyse linguistique peut aussi servir à orienter la détection d'influenceurs vers une thématique, nous avons mentionné précédemment l'importance de la communauté, notamment thématique, pour l'exercice d'une influence. (Hamzehei et al., 2017) analysent l'influence des utilisateurs de réseaux sociaux à l'aune des topiques détectés dans leurs messages.

3 Expérimentations réalisées

Nous avons, en section 1, défini l'influence comme un pouvoir, ce qui nous amène à caractériser les influenceurs selon deux aspects : les ressources qui leur permettent d'influencer et les effets de leur influence du point de vue des individus cibles. Nous allons rechercher les ressources comme les

effets des influenceurs aussi bien dans le contenu des messages échangés entre les influenceurs et leur audience que sur des caractéristiques topologiques concernant la structure et la nature de leurs interactions. Nos expérimentations comportent ainsi deux pans avec la relations d'un individu Nous en déduisons de nouvelles pistes afin de créer un système hybride entre les aspects de structure et de contenu souvent étudiés indépendamment.

3.1 Comparaison de mesures de centralité pour la détection d'influenceurs Twitter

Le but de cette expérimentation est d'évaluer des algorithmes mesurant chacun une certaine centralité. Nous pouvons distinguer deux types de centralité. La centralité locale qui prend en compte uniquement le voisinage d'un nœud pour lui attribuer une valeur de centralité. La centralité globale qui regarde le nœud par rapport à l'ensemble du graphe afin d'évaluer sa centralité. Pour chaque mesure de centralité, nous évaluons sa capacité à modéliser l'influence des individus représentés comme les nœuds d'un graphe qui représente la structure de leurs interactions.

Nous présentons six mesures de centralité qui nous permettent d'évaluer autant de types de centralité.

- **Degré entrant** : c'est une mesure de centralité locale puisqu'elle n'est basée que sur le calcul du nombre de liens entrants pour un nœud donné. En termes d'influence, c'est une polarisation directe envers un utilisateur (Freeman, 1978)

- **Intermédiarité** et **Proximité** : il n'y a pas de résultat pour Intermédiarité et Proximité parce qu'ils ont besoin de calculer les chemins les plus courts entre tous les nœuds du graphe, ce qui nécessite un graphe connecté, ce n'est pas le cas des graphes que nous avons extraits (Freeman, 1978) Nous les mentionnons tout de même parce qu'elles sont des mesures importantes qui ont bien été prises en compte dans nos expérimentations.

- **Hits** : apporte une sémantique supplémentaire à la centralité globale avec une distinction mutuelle entre un score d'autorité et un score de relais. L'autorité est importante parce que particulièrement écoutée par des relais qui sont eux-mêmes importants parce qu'ils écoutent beaucoup d'autorités (Kleinberg, 1999)

- **Page Rank** : calcule l'importance d'un nœud dans un graphe d'après la valeur de chaque lien pointant vers lui et l'importance du nœud à l'origine, avec une initialisation uniforme des poids et par itérations successives sur tous les nœuds du graphe jusqu'à ce que le poids de chacun soit à l'équilibre, cet algorithme donne des valeurs de centralité globale. La valeur de chaque se dilue avec le nombre de liens sortants de son nœud d'origine. Est ajoutée une probabilité uniforme pour tous les nœuds du graphe de se départir de la structure du réseau pour « sauter » d'un nœud à l'autre (Page et al., 1999)

- **Leader Rank** : entend améliorer la modélisation de Page Rank pour les réseaux sociaux en affirmant que la probabilité de passage d'un nœud à un autre sans utiliser les arcs du graphe ne doit pas être uniforme mais inversement proportionnelle au nombre de liens sortants qui sont disponibles

Ces algorithmes attribuent à chacun des nœuds du graphe un score de centralité permettant d'établir un classement d'utilisateurs pour un réseau considéré. Pour les évaluer, nous devons mesurer la qualité du classement produit à partir de chaque algorithme en fonction d'une référence qui donne l'influence de chaque utilisateur. Nous avons ajouté une Baseline fondée sur un classement au hasard des utilisateurs pour mieux appréhender les résultats des algorithmes évalués.

Comme référence, nous avons choisi le jeu de données conçu pour la compétition RepLab 2014 (Amigó et al., 2014). Cette compétition proposait notamment une tâche consistant à classer des utilisateurs Twitter du plus influent au moins influent. Le jeu de données contient plus de 7000 comptes Twitter catégorisés selon qu'ils appartiennent au domaine de la banque, de l'automobile ou à des domaines différents des deux précédents.. Ils ont été annotés par les spécialistes de la réputation en ligne Llorente & Cuenca[1]. Cette annotation considère l'influence réelle (dans le monde) des utilisateurs, en indiquant s'ils sont influenceurs ou pas. Le jeu de données contient en moyenne 1/3 d'influenceurs. Nous avons construit un échantillon de 50 utilisateurs du domaine de la banque pour limiter la taille du graphe à construire et le temps d'extraction des informations depuis l'API Twitter[2], qui impose des limites à la quantité d'information extraite par intervalle de temps. Nous avons fait en sorte de conserver la proportion d'influenceurs originale (1/3) pour garder une certaine comparabilité avec les systèmes de la compétition.

À partir de cet ensemble d'utilisateurs initial, nous extrayons deux types d'interaction.

- Le **suivi** qui constitue une audience et donc un terreau d'utilisateurs pour l'exercice d'une influence

- Le **Retweet** ou la rediffusion d'un contenu constituant une réaction comme un effet d'influence

À partir des utilisateurs initiaux, nous extrayons ces deux types d'interaction et donc de nouveaux utilisateurs qui seront représentés dans un graphe pour chacun des deux types d'interaction.

Nous comparons la référence binaire aux classements issus des algorithmes avec la mesure Mean Average Precision (MAP) utilisée pour RepLab 2014. Elle est fondée sur l'intuition en Recherche d'Information que les résultats les plus pertinents, les influenceurs, doivent apparaitre au début. Nous avons calculé MAP par la formule qui suit :

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

avec N le nombre total d'utilisateur, n le nombre d'influenceurs correctement trouvés, $p(i)$ la précision au rang i (en ne considérant que les i premiers utilisateurs trouvés) et $R(i)$ est à 1 si l'utilisateur au rang i est un influenceur sinon à 0.

[1] www.llorenteycuenca.com/en/

[2] www.developer.twitter.com

Sur un graphe de suivi

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

Tableau 1 : caractéristiques du graphe de suivi construit

Algorithme	Baseline	Degré entrant	Page Rank	Leader Rank	Hits
MAP (%)	**38,67**	**43,49**	**44,28**	**44,53**	**51,68**

Tableau 2 : résultats des mesures de centralité sur le graphe de suivi

- On observe *Tableau 1* une faible densité du graphe, les utilisateurs du corpus initial se suivent peu, ce qui explique pourquoi Page Rank et Leader Rank, mesures de centralité globales, donnent *Tableau 2* des résultats similaires au degré entrant, locale

- Hits se démarque en distinguant les influenceurs comme étant des « autorités », ce qui permet de donner comme première caractéristique des influenceurs d'être suivis par des « relais » tels que nous les avons précédemment présentés.

Sur un graphe de Retweet

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

Tableau 3 : caractéristiques du graphe de Retweet construit

Algorithme	Baseline	Degré entrant	Page Rank	Leader Rank	Hits
MAP (%)	**38,67**	**40,91**	**40,91**	**40,91**	**40,91**

Tableau 4 : résultats des mesures de centralité sur le graphe de Retweet

- Même problème de connectivité que sur le graphe de suivi pour expliquer l'absence de résultat d'Intermédiarité et de Proximité

- À nouveau, la faible densité du graphe, signifiant ici que les utilisateurs se retweetent peu, empêche les mesures de centralité globales d'apporter de meilleurs résultats par rapport au degré entrant, local, qui ne prend en compte que le voisinage direct

- Le fait que Hits ne se distingue pas sur le graphe de Retweet peut indiquer que l'information de suivi est plus significative pour détecter les influenceurs.

Algorithme	UTDBRG	Lys	LIA	UAMCLYR	ORM_UNED
MAP	**0,41**	**0,52**	**0,45**	**0,49**	**0,32**

Tableau 5 : Résultats des système de RepLab sur la banque

Pour information, nous ajoutons *Tableau 5* les résultats des systèmes ayant participé à la compétition RepLab et diffusés dans (Amigó et al., 2014). Nous avons pris le meilleur essai pour chaque système sur l'ensemble du corpus et sur le même domaine que notre étude. Les meilleurs systèmes ont surtout utilisé des métadonnées des profils comme le texte de présentation, le nombre de suiveurs (Lys, UAMCLYR), la présence d'une image de profil et le statut de vérification du compte (Lys).

3.2 Étude de l'impact du nombre de suiveurs pour distinguer des influenceurs Twitter par des mesures de centralité

Nous nous concentrons pour cette expérimentation sur l'information de suivi qui nous a semblé plus significative. Nous restreignons notre sélection de mesures de centralité à Page Rank, Leader Rank et Hits qui n'ont pas besoin d'un graphe connexe pour calculer un résultat. Ces trois algorithmes opèrent itérativement pour calculer un score approximé tolérant pour la convergence une certaine variation entre deux itérations. Nous cherchons à la fois à analyser le comportement des mesures de centralité en faisant évoluer la taille du graphe de suivi et aussi à déterminer dans quelle mesure le nombre de suiveurs est significatif en l'utilisant dans des intervalles de valeur différents.

Nous utilisons le même corpus que précédemment en créant cette fois-ci des échantillons de 50 utilisateurs suivant les intervalles de nombre de suiveurs : [1000-5000], [5000-10,000], [10,000-50,000], [50,000-100,000], [100,000-500,000]. Nous faisons en sorte que chaque échantillon comprenne la même proportion d'influenceurs : 1/3 (proportion d'influenceurs du corpus d'origine). Nous construisons un graphe de suivi pour chacun des segments ce qui donne cinq graphes de suivi aux tailles différentes sur lesquelles nous appliquons les mesures de centralité précédemment sélectionnées.

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	10,521,923
Densité	10^{-6}	10^{-6}	10^{-6}	10^{-7}	10^{-7}

Tableau 6 : Caractéristiques des graphes de suivi selon les segments

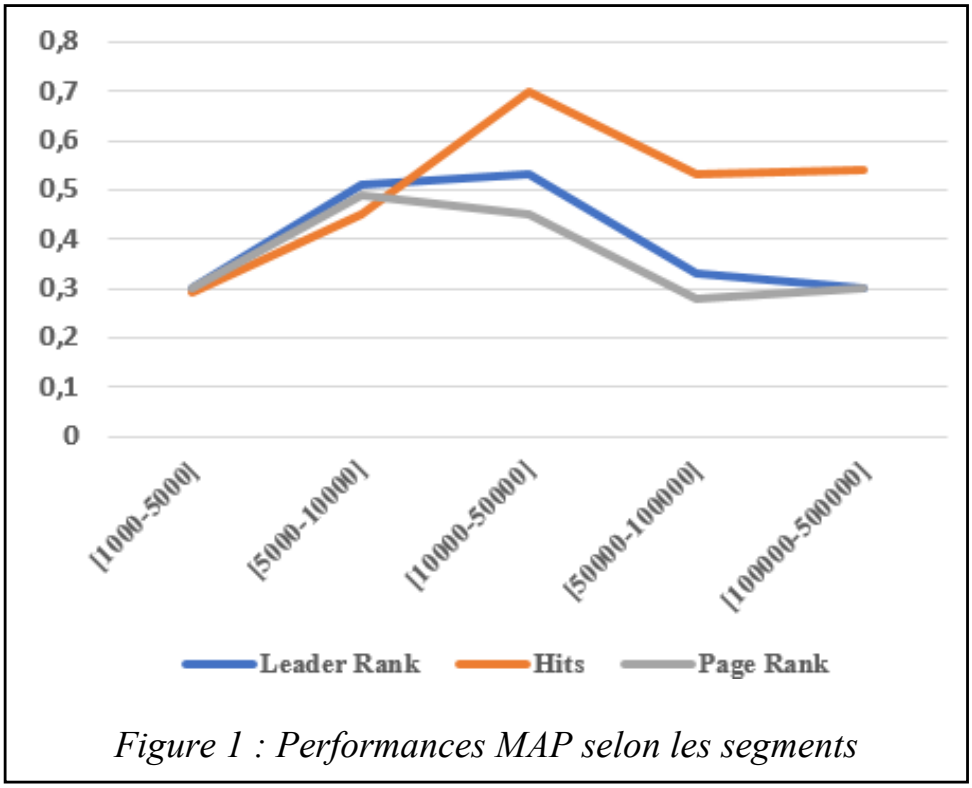

Figure 1 : Performances MAP selon les segments

La taille des segments augmentant, l'écart possible en nombre de suiveurs entre les utilisateurs augmente aussi, ce qui accroît potentiellement le pouvoir discriminant de ce critère pour distinguer les influenceurs. Le fait que la performance des trois algorithmes augmente jusqu'à un certain point avec la taille des segments montre que l'information de suivi est effectivement significative pour la détection des influenceurs. Nous expliquons globalement la baisse des performances en disant que le taille du graphe devient trop importante (*Tableau 3*, facteur 3 d'agrandissement du graphe entre deux segments) pour que l'approximation calculée par les trois algorithmes suffise à représenter correctement les forces en présence dans tout le graphe.

Nous constatons *Figure* 1 que deux algorithmes sont plus stables : Page Rank et Leader Rank. Leur comportement similaire peut s'expliquer par le fait que le second est une modification du premier. Aussi, ces deux algorithmes ont la particularité d'autoriser le fait de ne pas suivre la structure du graphe pour passer d'un nœud à l'autre, cela explique pourquoi ils sont moins sensibles aux variations de taille du graphe.

Pour Hits, la dynamique est plus forte (50% de progression relative entre [1000-5000] et [10,000-50,000] et 15% de baisse absolue entre [10,000-50,000] et [100,000-500,000]). Contrairement aux deux algorithmes précédents, Hits suit strictement la structure du graphe : pas de « saut » d'un nœud à l'autre ni de dilution de la valeur des liens. Il est donc plus sensible à la modification de la taille du graphe. Le fait que l'augmentation de marge en nombre de suiveurs entre les utilisateurs aide l'algorithme qui se base le plus sur la structure montre aussi la significativité de l'information de suivi.

La modélisation « nœud de terre » dans Leader Rank tient à mieux modéliser le comportement des utilisateurs d'un réseau social lorsque le nombre de liens sortants augmente et donc lorsque la taille du graphe augmente. Or, nous pouvons constater *Figure 1* que Leader Rank se détache positivement de Page Rank lorsque la taille des segments croît. C'est un signal que la modélisation de Leader Rank est effectivement meilleure que celle de Page Rank.

3.3 Détection linguistique du changement d'avis comme un effet de l'influence dans un corpus "Change My View[3]"

Notre parti pris est de détecter les influenceurs par leur influence effective là où des travaux précédents émettent des hypothèses sur l'influençabilité de critères linguistiques. Ainsi, nous voyons la détection du changement d'avis comme une première étape à la détection d'influenceurs en tant qu'un effet d'influence. Il s'agirait pour la suite de repérer la source du changement d'avis pour identifier l'influenceur. L'objectif est d'obtenir des influenceurs réels plutôt que des potentiels influenceurs.

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
Pré-filtré	10,743	128,901	1,529	20,883
Final	3,191	42,776	672	9463

Tableau 7 : Statistiques générales des corpus jusqu'au filtrage final

Nous avions besoin d'une ressource contenant des manifestations textuelles de changement d'avis. Nous avons choisi le forum en anglais « Change My View », précédemment étudié pour des tâches connexes par (Tan et al., 2016). Le principe de ce forum est que l'auteur initial d'un fil de discussion expose son point de vue sur une thématique puis il demande aux lecteurs de le faire changer d'avis. Les autres participants au fil de discussion vont alors faire en sorte de développer une argumentation qui contredise assez bien le point de vue de l'auteur initial pour modifier son point de vue. Lorsque l'auteur initial reconnait qu'un message a eu l'effet escompté, il lui attribue un « delta », ce qui revient à une annotation « ad hoc » des messages gagnants. C'est cette annotation que nous utiliserons comme référence. L'auteur initial doit en plus justifier cette récompense dans un message qui explicite son changement d'avis.

Nous utilisons le corpus extrait par (Tan et al., 2016) qui essayent notamment de prédire les arguments qui vont créer un changement d'avis. Toujours dans l'optique de détecter un influenceur effectif, nous adoptons une approche inverse puisque nous partons des réponses à ces arguments pour détecter ceux qui ont eu un impact. Les auteurs ont extrait plus de 20,000 fils de discussion depuis la création du forum en 2013 jusqu'en 2015 (les statistiques du corpus initial sont dans le *Tableau 6*). Cela donne plus de 1,000,000 de messages pour environ 80,000 participants uniques, donc beaucoup d'habitués. Le corpus est déjà divisé entre une partie pour l'entrainement (90%) et une partie pour l'évaluation (10%). Nous partons du corpus que les auteurs ont filtré pour leur tâche sur la résistance à la persuasion (statistiques générales du corpus pré-filtré en *Tableau 6*). Ils posent des contraintes de pertinence sur une participation minimale pour l'auteur initial et pour les autres à l'intérieur d'un fil de discussion. Puisque nous cherchons à détecter le changement d'avis pour les auteurs initiaux, nous analysons leurs messages seulement dans les fils de discussion où ils changent d'avis au moins une fois (les statistiques générales du corpus final en *Tableau 6*). Les discussions qui contiennent au moins un changement d'avis représentent 30% du corpus pré-filtré par les auteurs (*Tableau 6*). Dans le corpus final, le taux d'exemples positifs (messages d'un auteur initial exprimant un changement d'avis) est de 10%, ce qui réduit légèrement le déséquilibre entre les

[3] www.reddit.com/r/changemyview/

classes (2% initialement). Tous les messages sont analysés sans distinction d'auteur. Nous avons supprimé des messages les marques liées à l'attribution d'une récompense comme « delta » pour ne pas biaiser la classification.

Nous pouvons voir ce travail comme une tâche de classification binaire puisque pour chaque message, nous devons dire s'il exprime un changement d'avis ou pas. Nous utilisons un classifieur par régression logistique qui convient particulièrement à ce genre de classification. Le corpus original comme notre échantillon d'entrainement contiennent seulement 2% de messages exprimant un changement d'avis. Cela constitue un biais pour l'apprentissage du classifieur qui pourrait simplement annoter les messages avec la classe majoritaire (pas de changement d'avis) afin d'obtenir un résultat correct. Pour contrebalancer ce biais, nous mesurons la performance de notre classifieur en utilisant la mesure Area Under ROC Curve qui prend en compte le taux de vrais positifs par rapport au taux de faux positifs.

Dans cette expérimentation, toujours en cours, nous avons commencé par déterminer les descripteurs les plus pertinents en les utilisant séparément. Dans la poursuite de nos travaux, nous travaillerons sur la combinaison de ces traits pour obtenir le meilleur résultat possible.

Descripteur	Nombre de mots (référentiel)	Sac de mots	POS	Style	Passé
Score AUC (%)	**51,38**	**82,11**	**60,97**	**64,70**	**57,15**

Tableau 8 : Résultats de descripteurs pour la détection de changement d'avis

Le trait le plus simple, qui constitue le référentiel de notre évaluation, consiste à utiliser le nombre de mots du message. Nous obtenons un résultat assez neutre pour un référentiel à 51,38% (cf. *Tableau 8*).

Nous nous sommes demandé s'il y avait un emploi de termes particulier pour les messages exprimant un changement d'avis. Nous utilisons une représentation en sac de mots des tokens présents dans chaque message. Nous obtenons le meilleur résultat avec 82,11% pour ce trait seul (*Tableau 8*). Nous avons analysé les termes les plus discriminants pour donner du sens à ce résultat chiffré. Pour la classe positive (changement d'avis), le terme au poids le plus important (10,4) est « convinced », qui est central pour exprimer l'impact d'un argument gagnant comme dans « This one finally convinced me » (message *t1_cna7wg7*). La classe positive contient des termes de concession comme « concede », « still » qui marquent un tournant. Nous observons aussi un poids très important (8,1) pour « hadn » qui semble marquer une remise en cause du point de vue passé comme dans « I hadn't considered the obvious fact that (…) » (message *t1_cnbkumq*). Toujours dans ce modèle avec un terme au passé à polarité négative, nous observons l'émergence du terme « forgot » (6,5). Enfin, il y a un ensemble de termes ayant trait à la clairvoyance tels que « guesss » ou « realize » (7), qui dénotent la compréhension d'une nouvelle vision des choses comme dans « I think this post really helped me realize (…) » (message *t1_cni1gsy*). Pour la classe négative (messages sans expression d'un changement d'avis), le terme le plus fort (4,7) est « talking », qui peut à la fois montrer que la discussion n'est pas résolue et qu'il y a un malentendu comme dans « I am talking about (…) » (message *t1_cngt9pv*). Ce malentendu transparait aussi dans les messages de la classe négative avec le terme « clarify » (3,6) ou « referring » (3,5) comme dans « I'm referring to (…) » (message *t1_cndtf0x*). Au contraire d'un changement d'avis, les messages de classe négative peuvent dénoter la réaffirmation d'un propos avec le terme « already » (3,6) comme dans « I already said in different comments that (…) » (message *t1_cninoeb*).

Le descripteur POS fondé sur la fréquence des catégories morphosyntaxiques dans les messages donne 60,97%. Il utilise des traits de surface qui nécessitent tout de même une analyse linguistique plus approfondie que le sac de mots qui est remarquablement meilleur (cf. *Tableau 8*). Les traits discriminants pour la classe positive sont majoritairement les pronoms (3,0) pouvant dénoter la subjectivité d'un changement d'avis tandis que la classe négative est forte en coordonnants (0,8) et interjections (1,0) qui sont caractéristiques des débats.

Nous avons aussi essayé un trait un peu plus fondé sur la sémantique des messages en émettant l'hypothèse que lorsqu'ils expriment un changement d'avis, les auteurs font une sorte de bilan avec un retour sur le passé. La seule détection de toute forme du passé dans les messages donne un score de 57,15% (*Tableau 8*), ce qui semble valider notre hypothèse.

Nous avons utilisé des traits sur le style de l'expression dans les messages qui pourraient mettre en évidence un changement d'avis de leur auteur. Puisque le changement d'avis est lié à la psychologie, nous avons utilisé deux ressources lexicales construites empiriquement par des psychologues et utilisées aussi par (Tan et al., 2016) notamment pour détecter la malléabilité d'un auteur initial d'après son message introductif. Chaque lemme considéré reçoit un score selon qu'il évoque la gaieté, le contrôle et la passion dans (Warriner et al., 2013) et la factualité dans (Brysbaert et al., 2014). Nous calculons un score pour chaque message et pour chaque trait en faisant une moyenne sur les lemmes en présence. Nous ajoutons l'emploi des pronoms personnels de première personne pour l'aspect subjectif de l'expression d'un changement d'avis. Ce descripteur particulièrement complexe donne un score à 64,70%, ce qui est notablement moins bon que le sac de mots (cf. *Tableau 8*). Pour la classe positive, il y a un emploi caractéristique de la première personne du singulier avec le plus grand coefficient du modèle à 7,3, ce qui peut se rapporter à l'autocritique exercée par l'auteur qui explique son changement d'avis. Il y a aussi la présence particulière d'un lexique de contrôle avec un coefficient à 6,1 dénotant l'aspect sage et mesuré du bilan que constitue l'expression d'un changement d'avis. La classe négative est quant à elle particulièrement factuelle (2,3), ce que nous pouvons comprendre comme le soutien à une argumentation, et passionnée (2,6) puisque ce sont des messages qui sont dans la dynamique du débat.

4 Conclusion et Perspectives

La partie sur l'analyse de la structure montre une problématique réelle sur la faible connectivité d'utilisateurs pris pourtant dans un même domaine. Nous devons essayer d'augmenter la densité du graphe que nous construisons afin que les mesures de centralité puissent gagner en pertinence. Nous pourrions essayer de « résumer » le graphe obtenu en supprimant les nœuds à faible degré qui n'apportent pas vraiment d'information ou partir d'une communauté identifiée en amont pour travailler sur son réseau.

Nous allons essayer d'affiner la détection du changement d'avis en combinant les critères les plus pertinents. Nous tenterons ensuite d'identifier linguistiquement la source d'un changement d'avis. Nous pourrons utiliser un critère de similarité textuelle en partant de l'hypothèse que le message qui exprime le changement d'avis reprend les éléments déterminants du message qui a créé ce changement d'état.

En lien avec le changement d'avis comme un effet de l'influence, nous allons étudier ce qui provoque le changement d'avis, l'argumentation en partant des théories sur l'argumentation. Nous allons nous intéresser à la structure d'un argument pour voir s'il y a une composition « gagnante ».

Références

AMIGÓ, E., CARRILLO-DE-ALBORNOZ, J., CHUGUR, I. (2014). Overview of replab 2014: author profiling and reputation dimensions for online reputation management. Actes de *International Conference of the Cross-Language Evaluation Forum for European Languages*, 307-322

BIFET A., HOLMES G., PFAHRINGER B. (2011). Detecting sentiment change in Twitter streaming data. Actes de *2nd Workshop on Applications of Pattern Analysis* 17, 5-11

BIRAN O., ROSENTHAL S., ANDREAS J., MCKEOWN K., RAMBOW O. (2012). Detecting influencers in written online conversations. Actes de *Second Workshop on Language in Social Media*, 37-45

BRYSBAERT M., WARRINER A. B., KUPERMAN V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods* 46, 904-911

DANLOS L. (2011). Analyse discursive et informations de factivité. Actes de *TALN 2011*

DAVE K., BHATT R., VARMA V. (2011). Identifying influencers in social networks. Actes de *5th International Conference on Weblogs and Social Media*, 1-9

FREEMAN, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks* 1-3, 215-239

GIONIS, A., TERZI, E., et TSAPARAS, P. (2013). Opinion maximization in social networks. Actes de *2013 SIAM International Conference on Data Mining*, 387-395

HAMZEHEI, A., JIANG, S., KOUTRA, D., WONG, R., CHEN, F. (2017). Topic-based Social Influence Measurement for Social Networks. *Australasian Journal of Information Systems*, 21

JABEUR, L. B., TAMINE L., BOUGHANEM M. (2012). Active microbloggers: Identifying influencers, leaders and discussers in microblogging networks. *International Symposium on String Processing and Information Retrieval*, 111-117

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*

KLEINBERG J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 604-632

MARIANI J., PAROUBEK, P., FRANCOPOULO, G., HAMON O. (2014). Rediscovering 15 years of discoveries in language resources and evaluation: The LREC anthology analysis. Actes de *9th International Conference on Language Resources and Evaluation*

PAGE, L., BRIN, S., MOTWANI, R., WINOGRAD T. (1999). The PageRank citation ranking: Bringing order to the web. *Stanford InfoLab*

PAK A., PAROUBEK P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Actes de *LREC* 10

QUERCIA D., ELLIS, J., CAPRA, L., CROWCROFT J. (2011). In the mood for being influential on twitter. Actes de *IEEE Third Inernational Conference on Social Computing (SocialCom)*, 307-314

SHEIKHAHMADI A., NEMATBAKHSH, M. A.,ZAREIE, A (2017). Identification of influential users by neighbors in online social networks. *Physica A: Statistical Mechanics and its Applications*

TAN, C., NICULAE, V., DANESCU-NICULESCU-MIZIL C., LEE L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. Actes de *25th international conference on world wide web*, 613-624

WARRINER A. B., KUPERMAN V., BRYSBAERT M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45, 1191-1207