Identification Semi-Automatique de Mots-Germes pour l'Analyse de Sentiments et son Intensité

Amal Htait, Sébastien Fournier, Patrice Bellot

Identification Semi-Automatique de Mots-Germes pour l'Analyse de Sentiments et son Intensité

Amal Htait — Sébastien Fournier — Patrice Bellot

Aix Marseille University, CNRS, ENSAM, Toulon University, LSIS UMR 7296,13397, Marseille, France.
Aix-Marseille University, CNRS, CLEO OpenEdition UMS 3287, 13451, Marseille, France.
{amal.htait, sebastien.fournier, patrice.bellot}@openedition.org

RÉSUMÉ. *Dans le but d'exploiter les opinions dans les tweets, cet article présente une classification à partir du sentiment contenu au sein des tweets. Nous présentons une méthode d'identification de nouveaux mots-germes. Ils sont utilisés pour la prédiction de l'intensité de sentiments des mots en co-occurrence avec ces mots-germes. Ensuite, le calcul de similarités entre sentiments est appliqué en utilisant: la mesure de la similarité entre deux mots et l'utilisation de plongement de mots (e.g. word2vec, GloVE) couplé à la mesure cosinus. Les résultats montrent l'importance de l'utilisation de mots-germes adaptés aux tweets, ainsi que la taille et le prétraitement de corpus. Pour conclure, nous avons obtenu les meilleurs résultats grâce à l'application de la méthode utilisant le plongement de mots couplée à la mesure cosinus.*

ABSTRACT. *For the purpose of opinion exploring in tweets, this article presents a sentiment classification of tweets content. First, we present a method to identify new sentiment similarity seed words. These seed words are used for predicting sentiment intensity of other words and short phrases in co-occurrence. Then, for testing sentiment similarity, we use: Similarity Measures methods between words and cosine similarity measure between the word embedding representations (e.g. word2vec, GloVE). The experiments results highlight the importance of adapted for tweets seed words. In addition of the corpora size and its pre-treatement. As a conclusion, best results were achieved using cosine similarity measure between the word embedding representations.*

MOTS-CLÉS : *Mots-germes, Twitter, Mesure de la Similarité, Plongement de mot, Word2vec, GloVe.*

KEYWORDS: *Seed words, Twitter, Similarity Measures, Word Embedding, Word2vec, GloVe.*

1. Introduction

L'analyse de sentiments et l'analyse des opinions sont à un haut degré d'importance, au niveau de la recherche et de l'industrie, pour exploiter la grande quantité d'information disponible dans les textes des réseaux sociaux sur le WEB. Notre travail, dans cet article, concerne l'analyse de sentiments et plus particulièrement, la prédiction de l'intensité de sentiments reliée à un mot ou une phrase courte. Notre article présente une des méthodes de classification du sentiment basée sur l'utilisation de lexiques, introduite par Turney et Littman (2003). La méthode est inspirée des méthodes à base de similarité sémantique. Elle est appliquée au champ d'analyse du sentiment comme une mesure de similarité du sentiment entre les mots. La méthode de similarité entre sentiments est exploitée sur les mots, mais aussi sur le plongement de mots (Tang *et al.*, 2014), ce qui est appliqué dans notre travail en utilisant *word2vec* (Mikolov *et al.*, 2013) et *GloVE* (Pennington *et al.*, 2014).

Il faut souligner que l'efficacité de la méthode de similarité entre sentiments est corrélée avec la taille du corpus, puisque plus un corpus est grand, plus la probabilité de trouver les mots cherchés en co-occurrences avec des mots résents dans le lexique augmente. Certaines approches utilisent les pages Web (Turney et Littman, 2003) ou même Wikipedia (Gabrilovich et Markovitch, 2007) comme corpus. Cependant, selon certaines études, seulement 60% des co-occurrences dans une même page Web indiquent la même orientation du sentiment. La solution serait, selon Feng (2013), d'utiliser une méthode basée sur les tweets comme corpus pour atteindre une meilleure performance que ceux basés sur Google et Wikipedia. En effet, contrairement à une page Web, un tweet, va dans la grande majorité des cas à n'exprimer qu'une seule et unique polarité et non deux (Pak et Paroubek, 2010). De plus, du point de vue de l'analyse de sentiments, Twitter est considérée comme une source riche en opinions et en sentiments. Toutefois, les tweets contiennent des mots d'argot et des phrases informelles, par conséquent, les mots-germes classiques suggérés par Turney (2003) ne seront pas les plus appropriés. Le Tableau 1 contient les mots-germes suggérés par Turney et leurs nombres d'occurrences dans la ressource Sentiment140 (Go *et al.*, 2009) (une collection de 1,6 million de tweets étiquetés positifs ou négatifs selon les émoticônes) et dans notre collection de 300 millions tweets en langue anglaise provenant d'archive de Twitter[1], ces valeurs montrent la différence entre le taux d'utilisation de ces mots dans les tweets. Par exemple, le mot *Superior* n'est trouvé que 46 fois dans Sentiment140, contrairement au mot *Nice* qui est trouvé 20897 fois. Notre travail consiste à chercher de nouveaux mots-germes, plus adaptés aux tweets, et à les utiliser dans l'analyse de sentiment et de son intensité.

Notre contribution est présentée en deux sous-tâches, illustrée par la Figure 1 :

1) Définir une méthode d'expansion des mots-germes (seed words) positifs et négatifs classiques, et l'appliquer à la langue anglaise à l'aide d'un corpus de tweets annotés avec la polarité.

1. https ://archive.org/details/twitterstream

Mots-germes Positif	**good**	**nice**	**excellent**	**positive**	**fortunate**	**correct**	**superior**
Sent140	77018	20897	837	1140	3451	1079	46
300M	7322831	1652764	104033	309420	91670	234448	27922
Mots-germes Négatif	**bad**	**nasty**	**poor**	**negative**	**unfortunate**	**wrong**	**inferior**
Sent140	27152	862	6248	364	2234	6388	31
300M	2503681	161241	308587	148656	74761	1168136	8323

Tableau 1. *Les mots-germes suggérés par* ***Turney*** *(2003) et leurs nombres d'occurrences dans* ***Sentiment140*** *et dans nos* ***300M tweets****.*

2) Évaluer l'efficacité des nouveaux mots-germes ainsi que plusieurs mesures de similitude entre mots telles que Jaccard (Jaccard, 1901) et entre le plongement de mots comme word2vec (Mikolov *et al.*, 2013) couplé à la mesure cosinus, en mesurant la similarité du sentiment entre les données de test et les mots-germes positifs/négatifs.

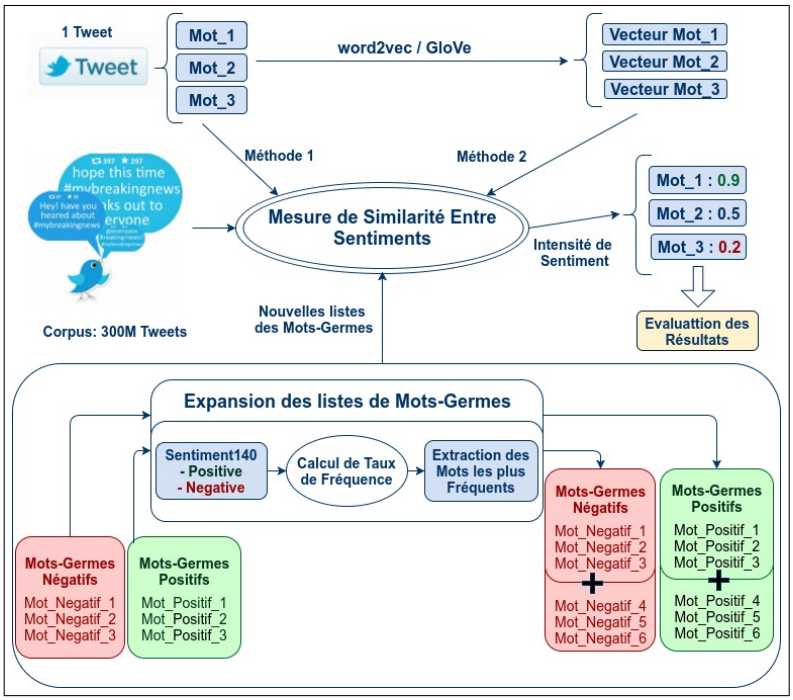

Figure 1. *Une description des sous-tâches de notre travail.*

2. Travaux connexes

Les mots qui portent un sentiment positif ou négatif sont fondamentaux pour l'analyse de sentiment. La polarité de ces mots peut être extraite à l'aide d'une approche basée sur une liste de mots-germes de sentiment connus grâce à différentes approches de similarité appliquée au sentiment. Mais comment extraire les mots-germes ? Des

approches ont sélectionné manuellement les mots-germes en se basant sur leur faible sensibilité au contexte (Turney et Littman, 2003) (Ju *et al.*, 2012). Ici, une méthode d'extraction de mots-germes est suggérée pour minimiser le taux de travail manuel, et pour offrir des mots-germes plus adaptés aux tweets.

En ce qui concerne l'utilisation des mots-germes, Ju et al. (2012) ont travaillé sur une méthode semi-supervisée de classification des sentiments qui vise à former un classifieur avec un petit nombre de données pré-annotées (mots-germes). Mais toute méthode supervisée ou semi-supervisée nécessite de grands corpus pré-annotés, d'autres expériences utilisent des méthodes non supervisées, comme Turney (2003) et Kaji (2007) selon des mesures statistiques tel que le point-wise mutual information (PMI) pour la mesure d'association entre les mots ou les phrases. Kanayama et Nasukawa (2006) supposent que les mots de même polarité apparaissent dans le texte consécutivement ; ainsi ces mots pourraient être trouvés dans le contexte des mots-germes. La même idée a été utilisée par Kiritchenko et al. (2014), dans le but de créer un ensemble de lexiques de mots portant des notions liées aux sentiments basé sur Twitter.

Plusieurs travaux récents ont démontré des améliorations significatives en termes de précision du modèle d'apprentissage grâce au paradigme de représentations vectorielles continues des mots (Collobert *et al.*, 2011), (Socher *et al.*, 2013). Cette approche a également été utilisée par Dos Santos et Gatti (2014), Tang et al. (2014) et d'autres en analyse de sentiments. Notre travail est basé sur les travaux précédents, en se basant sur le calcul de la similarité entre sentiments, mais aussi sur les plongement de mots, en utilisant les mots-germes, et toujours grâce à des méthodes non-supervisées.

Notre but est de prédire l'intensité des sentiments liés aux mots ou à de petits groupes de mots. Bien que de nombreuses expériences ont soutenu l'efficacité des méthodes qui utilisent des corpus pour la classification de sentiments (Turney, 2002), (Velikovich *et al.*, 2010), peu ont été réalisés pour la mesure de l'intensité du sentiment (Lenc *et al.*, 2016). À noter que certains travaux du domaine se différencient d'une simple classification positif, négatif et neutre, en se basant sur une classification à partir d'une échelle de 1 à 5 (Thelwall *et al.*, 2011). Des travaux sur la prédiction de l'intensité du sentiment à partir de textes ont été réalisés par Taboada et al. (2011), qui ont construit manuellement un lexique et ont présenté une méthode basée sur les mots couplée à l'utilisation des adjectifs, des intensifieurs et de la négation. Un autre exemple concernant la prédiction de l'intensité du sentiment est le Stanford Sentiment Treebank[2] (Socher *et al.*, 2013). Leur travail est basé sur un nouveau type de réseau de neurones récursif (Recursive Neural Network) construit sur les structures grammaticales des phrases. En outre, l'atelier international *SemEval*[3] (Évaluation sémantique) a enrichi ce domaine de recherche avec de nombreuses expérimentations et des campagnes d'évaluation, où la plupart des méthodes sont basées sur l'appren-

2. http ://nlp.stanford.edu/sentiment/treebank.html
3. http ://alt.qcri.org/semeval2016/

tissage automatique ou sur des méthodes supervisées. Cependant, notre travail ici est basé sur une comparaison de plusieurs méthodes non-supervisées avec l'introduction de nouvelles listes de mots-germes plus adaptés aux tweets.

3. Identification des mots-germes

Dans cet article, nous cherchons de nouveaux mots-germes plus adaptés aux tweets, afin de prédire l'intensité du sentiment d'autres mots ou phrases courtes. Dans ce but, nous considérons les étapes suivantes :

1) Pour trouver les mots-germes positifs, la fréquence d'apparition de chaque mot est calculée dans le corpus Sentiment140 (Go *et al.*, 2009) de tweets positifs, et les top 100 mots les plus fréquents sont sélectionnés après avoir éliminé les mots outils. La même procédure est appliquée pour les mots négatifs dans le corpus des tweets négatifs.

2) Un filtrage est appliqué aux listes, pour éliminer les mots neutres, et trouver les mots qui conservent leur polarité indépendamment du contexte. Dans ce but, une sélection des 38 mots les plus pertinents est appliquée pour chaque polarité, et la liste de Turney (2003) est ajoutée à cette liste. Les nouvelles listes sont présentées dans le Tableau 2.

Positif	Groupe 1	love, like, good, win, happy, lol, hope, best, thanks, great, funny, haha.
	Groupe 2	god, amazing, kind, fun, beautiful, nice, cute, laugh, cool, perfect, sweet.
	Groupe 3	awesome, okay, special, hopefully, haven, glad, congrats, dance, wonderful, dreams, sunshine.
	Groupe 4	hehe, yay, positive, fantastic, enjoying, correct, fabulous, excellent, fortunate, relaxing, superior.
Négatif	Groupe 1	ill, fucking, shit, need, fuck, hate, bad, break, suck, hell, sucks, cry.
	Groupe 2	damn, sad, wrong, tired, stupid, dead, pain, sick, alone, wtf, lost.
	Groupe 3	worst, fail, evil, bored, scared, hurts, missed, poor, afraid, upset, broken.
	Groupe 4	died, stuck, boring, crap, unfortunately, horrible, negative, nasty, sore, unfortunate, inferior.

Tableau 2. *Les nouvelles listes et sous-listes de mots-germes.*

4. Expérimentations

4.1. *Données*

Afin de construire les modèles basés sur word2vec et Glove, un corpus de tweets de grande taille est essentiel. Nous avons utilisé environ 300 millions tweets en langue anglaise provenant d'archive de Twitter[4] au format JSON de thématique général.

Dans un but de comparaison, le modèle word2vec de tweets réalisé par Godin et al. (2015) a été utilisé. Il s'agit d'un modèle avec une taille de vecteur égale à 400, et basé sur 400 millions de tweets. Le modèle GloVE basé sur les tweets et réalisé par Pennington et al. (2014) a aussi été utilisé. La taille du vecteur pour ce modèle est de 200. Le modèle a été construit à partir d'un milliard de tweets dont les caractères comme le Hashtag et les parenthèses ont été éliminés.

4. https ://archive.org/details/twitterstream

Dans le but d'évaluer l'analyse de sentiments, les données de la campagne d'évaluation de SemEval2016 Task7[5] sont utilisés. Ces données sont formées d'une liste de 1069 mots et phrases courtes, de polarité mixte, extraits par les organisateurs des tweets grâce au *Sentiment Composition Lexicon for Opposing Polarity Phrases* [6] et annotée par une valeur d'intensité du sentiment.

4.2. *Sous-groupe de mots-germes*

Après l'extraction de listes de mots positifs et négatifs, l'efficacité de ces mots comme mots-germes est testée pour obtenir une mesure de similarité appliquée au sentiment. Dans un premier temps, ces listes ont été divisées en sous-listes, afin de mesurer l'effet de changement des mots-germes ainsi que de l'impact du nombre de mots-germes. Donc, et pour chaque mot-germe, la fréquence de co-occurrence est calculée avec le reste des mots-germes de même polarité. Par exemple, le mot *love* apparaît plus que 5 millions fois avec les autres mots-germes positifs. Alors que le mot *superior* n'apparaît que 8 milles fois. Les mots-germes sont ordonnés suivant leur valeur de co-occurrence avec d'autres mots-germes, et groupés en 4 sous-listes. Les quatre sous-listes obtenues sont indiquées dans le Tableau 2.

4.3. *Modèles de similarité sémantique et d'association utilisés*

En utilisant les listes de mots-germes extraites et les sous-listes du Tableau 2, notre méthode est testée, en utilisant les approches : PMI, Dice et Jaccard. Après le calcul de similarité entre le mot candidat *w* avec les listes des mots-germes positifs et négatifs, l'orientation sentimentale (SO) de ce mot est calculée sur la base de la différence entre son association avec les mots-germes positifs et les mots-germes négatifs, comme le montre l'Équation 1.

$$SO(w) = sim(w, se_p) - sim(w, se_n) \quad [1]$$

4.4. *Résultats*

Les coefficients de corrélation de Kendall et Spearman sont utilisés pour évaluer le degré de similarité entre notre propre liste de résultats ordonnés et celle donnée par les organisateurs de SemEval Task7.

Les résultats de nos expérimentations sont présentés dans les Tableaux 3, 4 et 5. En premier, les résultats de chacune des quatre sous-listes de mots-germes spécifiés dans le Tableau 2, ensuite les résultats de deux sous-listes de mots-germes combinées, puis les résultats de l'utilisation des mots-germes fournis par Turney (Table 1), et ceux de l'utilisation d'un seul mot-germe pour chaque polarité : "Excellent" pour le positif et "Poor" pour le négatif (Turney, 2002). Puis, ces tableaux présentent les résultats de l'utilisation des nouveaux mots-germes sans, puis avec, ceux de Turney. Comme

5. http ://alt.qcri.org/semeval2016/task7/
6. http ://www.saifmohammad.com/WebPages/SCL.htmlOPP

dernier test, des émoticônes ont été ajoutés à nos mots-germes, deux pour chaque polarité (les plus usités dans les tweets), :) et :-) aux mots-germes positifs et :(et :'(aux mots-germes négatifs.

	PMI		**Dice**		**Jaccard**	
	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman
1 Mot	0.128	0.181	0.079	0.114	0.079	0.114
Turney's	0.228	0.335	0.13	0.195	0.22	0.337
Groupe_1	0.235	0.335	0.194	0.294	0.102	0.155
Groupe_2	0.306	0.438	0.248	0.374	0.248	0.374
Groupe_3	0.212	0.306	0.221	0.332	0.223	0.335
Groupe_4	0.202	0.291	0.164	0.246	0.170	0.253
Groupe_1+2	0.309	0.439	0.156	0.234	0.247	0.371
Groupe_3+4	0.247	0.358	0.234	0.353	0.239	0.358
All-Turney	0.317	0.449	0.0256	0.039	10.246	0.369
All	0.343	0.485	0.199	0.296	0.26	0.389
All+Emoticon	0.358	**0.503**	0.202	0.301	0.264	0.395

Tableau 3. *Résultats des méthodes de similarité entre sentiments des **Mots** avec différents groupements des mots-germes.*

Les résultats présentés dans le Tableau 3 montrent que : chaque groupe de mots-germes a un effet différent. Par exemple, en comparant les résultats, le Groupe_2 montre de meilleurs résultats que le Groupe_1 dans toutes les méthodes testées. Et cela grâce aux mots fortement positifs présents dans le Groupe_2 comme : "amazing", et "perfect", alors que le Groupe_1 contient quelques mots qui peuvent être ambigus. Bien que positif, ces mots peuvent exister dans des expressions négatives (e.g. Why do mostly *good* people die young ?). En plus, les résultats varient suivant le choix de la méthode. Le Tableau 3 montre que les meilleurs résultats ont été obtenus en utilisant la méthode PMI indépendamment de la combinaison des mots-germes. En conclusion, le meilleur résultat, dans le Tableau 3, est obtenu en utilisant tous les mots-germes, avec les émoticônes, en appliquant la mesure PMI et cela avec une valeur de Spearman égale à 0.503.

	vecteur_200		**vecteur_400**		**Godin_400**	
	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman
1 Mot	0.226	0.327	0.22844	0.332	0.185	0.270
Turney's	0.312	0.44724	0.325	0.464	0.295	0.425
Groupe_1	0.308	0.43989	0.316	0.450	0.306	0.435
Groupe_2	0.374	0.52922	0.379	0.536	0.393	0.556
Groupe_3	0.354	0.50069	0.352	0.499	0.370	0.522
Groupe_4	0.332	0.47007	0.334	0.478	0.411	0.576
Groupe_1+2	0.365	0.51779	0.370	0.524	0.401	0.562
Groupe_3+4	0.350	0.49780	0.351	0.499	0.434	0.603
All-Turney	0.378	0.53616	0.382	0.540	0.446	0.616
All	0.373	0.52874	0.382	0.540	**0.464**	**0.638**
All+Emoticon	0.383	0.54150	0.387	0.546	**0.467**	**0.642**

Tableau 4. *Résultats des modèles **word2vec** avec différents groupements des mots-germes.*

Les résultats des modèles **word2vec** et **GloVE** présentés par les Tableau 4 et 5. Pour les modèles word2vec, les meilleurs résultats sont obtenus en utilisant le modèle de Godin et al. (2015), un modèle avec une taille de vecteur égale à 400, basé sur un corpus de 400 Millions tweets. Pour les modèles GloVE, les meilleurs résultats

	vecteur_200		vecteur_300		Stanford_200	
	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman
1 Mot	0.257	0.371	0.250	0.361	0.250	0.361
Turney's	0.346	0.495	0.347	0.497	0.346	0.495
Groupe_1	0.310	0.446	0.305	0.438	0.340	0.482
Groupe_2	0.410	0.577	0.416	0.584	0.416	0.582
Groupe_3	0.398	0.557	0.391	0.550	0.398	0.557
Groupe_4	0.406	0.570	0.405	0.568	0.406	0.570
Groupe_1+2	0.396	0.557	0.399	0.568	0.404	0.564
Groupe_3+4	0.418	0.583	0.418	0.584	0.418	0.583
All-Turney	0.416	0.582	0.418	0.584	0.422	0.587
All	0.422	0.58	0.423	0.590	**0.426**	**0.591**
All+Emoticon	**0.427**	**0.594**	**0.427**	**0.595**	**0.426**	**0.591**

Tableau 5. *Résultats des modèles **GloVE** avec différents groupements des mots-germes.*

sont relativement proches. Ils sont obtenus par les modèles créés avec une taille de vecteur égale à 200 et 300. Il est remarquable que les résultats du modèle de Stanford (Pennington *et al.*, 2014), avec une taille de vecteur égale à 200, n'ait pas pu dépasser les résultats de nos modèles, malgré qu'il soit basé sur 1 Milliard de tweets alors que les nôtres sont basés sur 300 Millions de tweets seulement. Nous pensons que la raison est la méthode de pré-traitement des tweets utilisé pour la création des modèles. En effet, d'après Godin et al. (2015), l'élimination des Hashtags a un effet négatif sur les résultats, alors que Stanford ont éliminé les Hashtags et tous les caractères spéciaux des tweets, même les émoticônes.

Enfin, la Figure 2 montre une comparaison entre les résultats des meilleurs représentations vectorielles. En se concentrant sur l'utilisation de tous les mots-germes (avec ou sans émoticônes), il semble évident que les résultats de notre modèle GloVE ont dépassé ceux de notre modèle word2vec. Alors que le modèle réalisé par Godin a pu dépasser tous les autres modèles. Ces résultats soulignent l'importance du bon pré-traitement du corpus suivant le domaine d'utilisation. Ce résultat (Spearman = 0,642) est très proche du meilleur résultat obtenu à SemEval2016 Task7 (Spearman = 0,674). À noter aussi que l'ajout d'émoticones aux mots-germes augmente les valeurs de Kendall et de Spearman, mais la différence n'est pas statistiquement significative selon le Test-t de Student (p-value = 0,986).

5. Conclusion et Travaux Futurs

Dans cet article, nous avons présenté une méthode d'identification de nouveaux mots-germes, qui ont été utilisés pour la prédiction de l'intensité de sentiments des mots et de phrases. Nos données sont des tweets, et pour le calcul des similarités entre sentiments, deux méthodes ont été appliquées : la mesure de similarité et d'association entre deux mots, comme PMI, et en utilisant le plongement de mots comme dans le cas de word2vec et GloVE couplé à la mesure cosinus. Les résultats des tests ont mis en évidence l'importance du choix des mots-germes, ceux ayant une forte intensité de sentiment donnent les meilleurs résultats. En ce qui concerne les méthodes utilisées, la représentation vectorielle des mots avec word2vec ou GloVE couplé à la

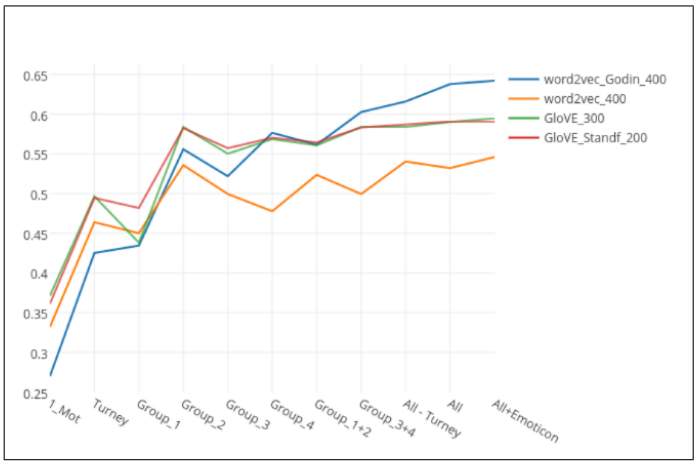

Figure 2. *La comparaison des résultats utilisant le modèle word2vec et le modèle GloVE créé, le modèle word2vec de Godin et le modèle GloVE de Stanford.*

mesure cosinus a atteint les meilleurs résultats pour chaque groupe de mots-germes testé, comparativement aux autres méthodes testées. Les meilleurs résultats atteints en utilisant le modèle word2vec de Godin et al. (2015) basé sur un corpus de 400 Millions de tweets. En se basant sur un même corpus de 300 millions de tweets, l'utilisation de GloVE a donné de meilleurs résultats que word2vec.

Afin d'améliorer ces résultats, nous pensons qu'il est d'abord nécessaire d'améliorer l'approche d'extraction et de groupement des mots-germes. Ensuite, il est nécessaire de collecter un corpus dépassant les 300 millions de tweets.

Remerciements

Ce travail a été soutenu par le programme Français Investissements d'Avenir Equipex "A digital library for open humanities" d'OpenEdition.org.

6. Bibliographie

Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P., « Natural language processing (almost) from scratch », *JMLR*, vol. 12, p. 2493-2537, 2011.

dos Santos C. N., Gatti M., « Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. », *COLING*, p. 69-78, 2014.

Feng S., Zhang L., Li B., Wang D., Yu G., Wong K.-F., « Is Twitter a better corpus for measuring sentiment similarity ? », *EMNLP*, p. 897-902, 2013.

Gabrilovich E., Markovitch S., « Computing semantic relatedness using Wikipedia-based explicit semantic analysis. », *IJcAI*, vol. 7, p. 1606-1611, 2007.

Go A., Bhayani R., Huang L., « Twitter sentiment classification using distant supervision », *CS224N Project Report, Stanford*, vol. 1, p. 12, 2009.

Godin F., Vandersmissen B., De Neve W., Van de Walle R., « Multimedia Lab@ ACL W-NUT NER Shared Task : Named Entity Recognition for Twitter Microposts using Distributed Word Representations », *ACL-IJCNLP*p. 146, 2015.

Jaccard P., *Étude comparative de la distribution florale dans une portion des Alpes et des Jura*, Bulletin del la Société Vaudoise des Sciences Naturelles, 1901.

Ju S., Li S., Su Y., Zhou G., Hong Y., Li X., « Dual word and document seed selection for semi-supervised sentiment classification », ACM, p. 2295-2298, 2012.

Kaji N., Kitsuregawa M., « Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. », *EMNLP-CoNLL*, p. 1075-1083, 2007.

Kanayama H., Nasukawa T., « Fully automatic lexicon expansion for domain-oriented sentiment analysis », *EMNLP*, Association for Computational Linguistics, p. 355-363, 2006.

Kiritchenko S., Zhu X., Cherry C., Mohammad S., « NRC-Canada-2014 : Detecting aspects and sentiment in customer reviews », *SemEval*, 2014.

Lenc L., Král P., Rajtmajer V., « UWB at SemEval-2016 Task 7 : Novel method for automatic sentiment intensity determination », *SemEval*, 2016.

Mikolov T., Chen K., Corrado G., Dean J., « Efficient estimation of word representations in vector space », *arXiv*, 2013.

Pak A., Paroubek P., « Twitter as a Corpus for Sentiment Analysis and Opinion Mining. », *LREC*, vol. 10, p. 1320-1326, 2010.

Pennington J., Socher R., Manning C. D., « GloVe : Global Vectors for Word Representation », *Empirical Methods in Natural Language Processing (EMNLP)*, vol. 14, p. 1532-1543, 2014.

Socher R., Perelygin A., Wu J. Y., Chuang J., Manning C. D., Ng A. Y., Potts C., « Recursive deep models for semantic compositionality over a sentiment treebank », *EMNLP*, 2013.

Taboada M., Brooke J., Tofiloski M., Voll K., Stede M., « Lexicon-based methods for sentiment analysis », *Computational linguistics*, vol. 37, n° 2, p. 267-307, 2011.

Tang D., Wei F., Yang N., Zhou M., Liu T., Qin B., « Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. », *ACL*, p. 1555-1565, 2014.

Thelwall M., Buckley K., Paltoglou G., Cai D., Kappas A., « Sentiment strength detection in short informal text », *JASIST*, 2011.

Turney P. D., « Thumbs up or thumbs down ? : semantic orientation applied to unsupervised classification of reviews », ACL, p. 417-424, 2002.

Turney P. D., Littman M. L., « Measuring praise and criticism : Inference of semantic orientation from association », *ACM*, vol. 21, n° 4, p. 315-346, 2003.

Velikovich L., Blair-Goldensohn S., Hannan K., McDonald R., « The viability of web-derived polarity lexicons », ACL, p. 777-785, 2010.