Analyse des sentiments à partir des commentaires Facebook publiés en Arabe standard ou dialectal marocain par une approche d'apprentissage automatique

Abdeljalil Elouardighi*, Mohcine Maghfour*, Hafdalla Hammia*, Fatima-Zahra Aazi**

* Laboratoire de Modélisation Mathématiques et de Calculs Economiques
Faculté des Sciences Juridiques Economiques et sociales, Université Hassan 1er,
Km 3, route de Casablanca, B.P. : 784, Settat, Maroc.
abdeljalil.elouardighi@uhp.ac.ma, m.maghfour.@gmail.com, hhammia@gmail.com
** ESCA Ecole de Management,
7, Rue Abou Youssef El Kindy, 20 070 Casablanca, Maroc.
faazi@esca.ma

Résumé. L'analyse des sentiments est un processus pendant lequel la polarité (positive, négative ou neutre) d'un texte donné est déterminée. Nous nous intéressons dans ce travail à l'analyse des sentiments à partir des commentaires Facebook, réels, partagés en arabe standard ou dialectal marocain par une approche basée sur l'apprentissage automatique. Ce processus commence par la collecte des commentaires et leur annotation à l'aide du crowdsourcing suivi d'une phase de prétraitement du texte afin d'extraire des mots arabes réduits à leur racine. Ces mots vont être utilisés pour la construction des variables d'entrée en utilisant plusieurs combinaisons de schémas d'extraction et de pondération. Pour réduire la dimensionnalité, une méthode de sélection de variables est appliquée. Les résultats obtenus des expérimentations sont très prometteurs.

1 Introduction

L'analyse des sentiments (AS) devient un domaine d'étude très ouvert à la recherche. L'objectif étant d'analyser, à partir des textes partagés sur les réseaux sociaux, les opinions, les sentiments, les attitudes et les émotions des communautés sur différents sujets. En général, on distingue deux catégories d'approches pour l'AS des textes publiés sur les réseaux sociaux : la première est basée sur le lexique, consiste à utiliser une collection prédéfinie de mots et d'annoter chacun avec une valeur traduisant sa polarité (sentiment positif, négatif ou neutre). La deuxième est basée sur des techniques d'apprentissage automatique. L'AS dans ce cas peut être vu comme étant un problème de classification supervisée de texte. Sur les réseaux sociaux, comme Facebook, les commentaires partagés en arabe prend généralement la forme de l'Arabe Standard Moderne (ASM) ou l'Arabe dialectal (Duwairi et Qarqaz, 2014). Nous nous intéressons dans ce travail à l'AS à partir des commentaires Facebook écrits en ASM ou en Arabe Dialectal Marocain (ADM) en utilisant une approche basée sur l'apprentissage automatique.

Nos principales contributions dans ce travail consistent à : décrire les propriétés de la langue ASM et surtout l'ADM et leurs défis pour l'AS ; présenter un ensemble de techniques de prétraitement des commentaires Facebook écrits en ASM et en ADM pour l'AS et finalement construire et sélectionner des entités (mots ou séquence mots) des commentaires permettant d'obtenir le meilleur modèle de classification des sentiments.

Le reste de cet article est organisé comme suit : dans la section 2, nous décrivons le processus d'apprentissage automatique proposé et son application aux commentaires Facebook écrits en ASM et en ADM. Nous présentons également les méthodes de sélection et d'extraction des variables (mots ou séquences de mots) utilisées dans la phase de classification. Les résultats des expérimentations sont donnés dans la section 3. Une conclusion et quelques perspectives de ce travail sont présentées dans la section 4.

2 Processus d'apprentissage automatique appliqué aux commentaires Facebook en Arabe

L'AS des commentaires Facebook écrits en ASM ou en ADM selon une approche d'apprentissage automatique, nécessite l'implémentation de plusieurs étapes. La figure suivante (Fig. 1) fournit un aperçu global de ce processus. Nous décrirons dans les paragraphes suivants les principales tâches de chaque étape.

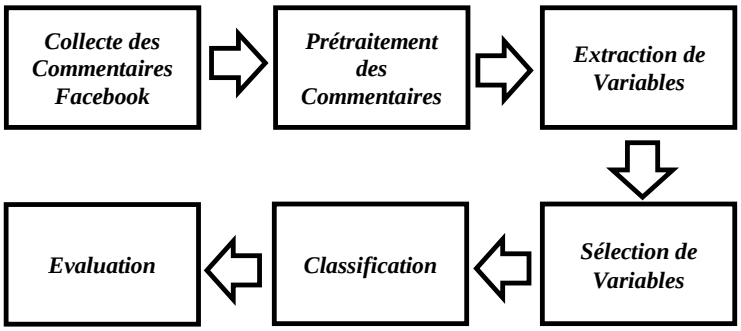

FIG. 1 – *Etapes du processus proposé pour l'analyse des sentiments.*

2.1 Collecte des commentaires Facebook

Le processus d'apprentissage automatique proposé est testé sur des commentaires Facebook écrits en ASM ou en ADM, sur les élections législatives marocaines ayant eu lieu le 7 octobre 2016. Nous avons ciblé des journaux marocains qui publient en ligne des commentaires en langue arabe (moderne ou dialectal). La collecte de ces commentaires a été effectuée utilisant l'API : "Facebook Graph API" sur une période de 70 jours et nous a permis de sélectionner 10254 commentaires. Pour une libre exploitation, nous avons publié la base de données dans (ElecMorocco, 2017).

2.2 Traitement des commentaires Facebook

Le traitement du texte des commentaires est une phase critique dans le processus d'analyse des sentiments étudié. L'objectif étant d'extraire des variables (sous forme de mots ou de séquences de mots) pour les utiliser dans la classification. La qualité du traitement a donc clairement un impact majeur sur les performances des modèles de classification et les résultats obtenus à la fin du processus.

On commence cette étape par un nettoyage et une normalisation du texte : suppression des signes, des symboles, des lettres répétées, des mots vides ou des mots qui ne fournissent aucune information sur le sujet étudié. La tâche suivante est l'opération de tokénisation par laquelle le texte du commentaire est divisé en unités lexicales (tokens). Dans un texte en Arabe moderne ou dialectal, ces unités sont plus complexes puisqu'elles sont composées souvent de plus d'un mot, d'où l'importance de la tâche de désuffixation ou racinisation. Pour développer et appliquer un racinisateur sur les commentaires collectés, nous nous sommes inspirés du racinisateur light10 de (Larkey et al., 2007) qui propose l'élimination des préfixes et suffixes les plus courants d'un token dans le langage arabe moderne, sur lequel nous ajoutons une extension pour l'arabe dialectal marocain .

Le tableau suivant (Table. 1) présente un exemple de prétraitement d'un commentaire.

Tâche	Résultat
Texte initial	الكلام الي كي يقولو هاذ السياسي ماشي معقوووول ! هههههه #السياسةـالمغربية Les discours de ce politicien ne sont pas raisonnables! hahahahaha #politique_marocaine
Nettoyage	الكلام الي كي يقولو هاذ السياسي ماشي معقوووول هههههه السياسة المغربية
Normalisation	الكلام الي كي يقولو هاد السياسي ماشي معقول هه السياسة المغربية
Tokénisation	'الكلام', 'الي', 'كي', 'يقولو', 'هاد', 'السياسي', 'ماشي', 'معقول', 'هه', 'السياسة', 'المغربية'
Supression des mots vides	'الكلام', 'يقولو', 'السياسي', 'ماشي', 'معقول', 'هه', 'السياسة', 'المغربية'
Désuffixation	'كلام', 'قول', 'سياس', 'ماش', 'معقول', 'هه', 'سياس', 'مغرب'

TAB. 1 – *Exemple de prétraitement d'un commentaire.*

2.3 Extraction et sélection de variables

L'AS par apprentissage supervisé nécessite la détermination au préalable de la polarité des opinions exprimées dans le texte pour construire la variable objectif. L'annotation des commentaires collectés a été effectuée en utilisant le crowdsourcing. Cette tâche a été confiée à un groupe de volontaires pour définir la polarité des commentaires, positifs ou négatifs. Au final, 6581 commentaires ont été annotés négatifs et 3673 positifs. Les variables d'entrée sont automatiquement extraites depuis le corpus formé à partir des commentaires prétraités utilisant les schémas d'extraction n-gram et de pondération TF / TF-IDF.

Des études telle que celle de (Pang et al., 2008) ont montré que la qualité des modèles de classification dépend des spécificités des données utilisées. Pour cela, nous avons testé six

combinaisons de schémas d'extraction et de pondération pour garantir la meilleure qualité des modèles développés.

Afin de réduire la dimensionnalité et améliorer la qualité des modèles de classification, une méthode de sélection de variables a été utilisée. Il s'agit du score « somme des carrés intergroupe à intra-groupe » (BSS / WSS) utilisé dans (Dudoit et al., 2002; Sehgal et al., 2006), pour sélectionner les mots ou les séquences de mots les plus discriminants. Le score permet de classer les variables par ordre de pertinence. Une fois l'ordre établi, on choisit le sous ensemble optimal de mots par la méthode pas à pas de type forward.

2.4 Classification des commentaires

Pour classer les commentaires Facebook, nous avons appliqué trois algorithmes de classification supervisée(implémentés sur le logiciel R) : Naïve Bayes(NB), les Forêts Aléatoires (FA) et les Machines à Vecteurs Support (SVM).

3 Résultats et discussion

La combinaison des schémas d'extraction et de pondération nous a permis de tester six configurations différentes (six jeux de données), pour lesquelles nous avons appliqué le score de sélection de variables. Chaque jeu de données était devisé en trois sous-ensembles : 50% pour l'apprentissage, 25% pour la validation et 25% pour le test. Le tableau (Table. 2) résume les résultats des expérimentations menées.

Configurations	Classifieurs	Nombre de variables sélectionnées	TBC avec les variables sélectionnées sur l'échantillon de validation	TBC avec les variables sélectionnées sur l'échantillon de test	TBC calculé en présence de toutes les variables
1 - Unigram/ TF	SVM	186	0.74	0.75	0.76
	NB	56	0.71	0.73	0.39
	FA	149	0.75	0.75	0.74
2 - Unigram/ TF-IDF	SVM	198	0.77	0.78	0.78
	NB	55	0.72	0.72	0.42
	FA	56	0.76	0.76	0.75
3 - Bigram/ TF	SVM	195	0.72	0.73	0.72
	NB	20	0.69	0.68	0.35
	FA	175	0.73	0.72	0.73
4 - Bigram/ TF-IDF	SVM	199	0.72	0.72	0.72
	NB	20	0.67	0.67	0.36
	FA	198	0.72	0.73	0.72
5 - (Unigram+Bigram)/ TF	SVM	200	0.76	0.77	0.76
	NB	100	0.74	0.74	0.39
	FA	89	0.76	0.76	0.71
6 - (Unigram+Bigram)/ TF-IDF	SVM	199	0.77	0.77	0.78
	NB	50	0.72	0.71	0.56
	FA	148	0.76	0.75	0.73

TAB. 2 – *Taux de bon classement pour les configurations testées avec selection de variables.*

Pour chaque configuration, nous avons présenté le meilleur taux de bon classement (TBC), obtenu sur la base de l'échantillon de validation. Le graphique (Fig. 2) présente l'évolution des TBC de la configuration [Unigram+Bigram/TF], pour les trois algorithmes utilisés, en fonction du nombre de variables insérées dans l'ordre décroissant de pertinence. Le tableau (Table. 2)

présente aussi les TBC obtenus sur l'échantillon test ainsi que ceux calculés en présence de toutes les variables (sans sélection préalable de variable).

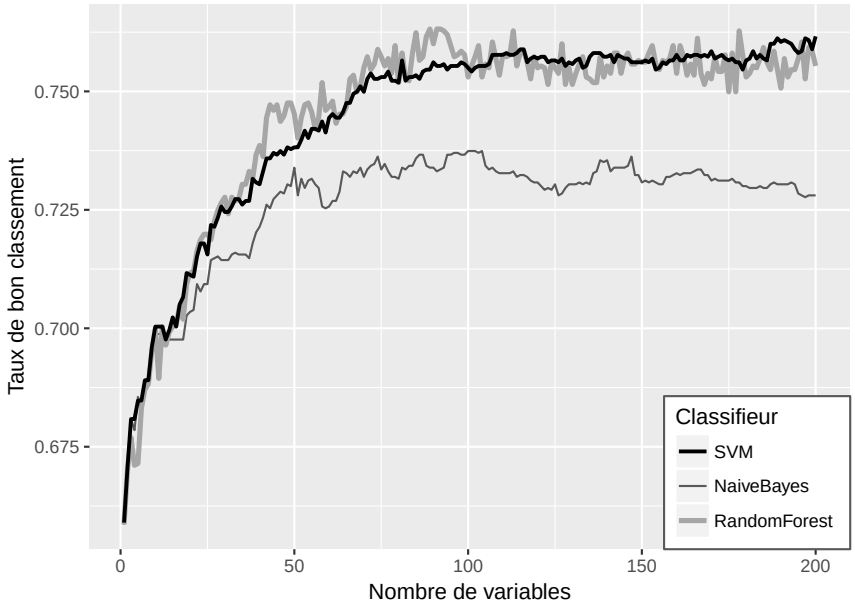

FIG. 2 – *Evolution du taux de bon classement en fonction des variables introduites pour la configuration : Unigram + Bigram / TF.*

En général, ces résultats montrent que les meilleures performances ont été obtenues avec les combinaisons [Unigram/TF-IDF] et [Unigram + Bigram/TF-IDF] quel que soit l'algorithme utilisé. Considérant l'impact de la sélection de variables pour les différentes configurations, nous pouvons conclure que la deuxième configuration, avec l'extraction Unigram et la pondération TF-IDF, est la plus efficace en terme du ratio : Taux de bon classement / nombre de variables.

4 Conclusion

Ce travail a porté sur l'AS en utilisant les commentaires Facebook écrits et partagés en ASM ou ADM. Le processus proposé est appliqué à des données relatives aux élections législatives marocaines de 2016. Plusieurs combinaisons de schémas d'extraction (n-gram) et de pondération (TF / TF-IDF) pour la construction des variables ont été testées pour garantir les meilleures performances des modèles de classification développés. Les résultats ont montré que la qualité des modèles dépend des sous-ensembles de variables constitués à partir de la combinaison des schémas d'extraction et de pondération. L'application d'une méthode de sélection de variables nous a permis de réduire les dimensions tout en gardant un niveau de performance similaire ou meilleur.

La taille de l'ensemble de données utilisé dans ce travail est relativement réduite. Pour avoir des conclusions plus solides, nous prévoyons de construire une base de commentaires plus importante et d'implémenter notre approche dans un environnement distribué (en utilisant le Framework Hadoop (Nodarakis et al., 2016b) ou Spark (Nodarakis et al., 2016a) par exemple) et de développer une méthode d'annotation automatique des commentaires basée sur un lexique.

Références

Dudoit, S., J. Fridlyand, et T. P. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association 97*(457), 77–87.

Duwairi, R. M. et I. Qarqaz (2014). Arabic sentiment analysis using supervised classification. In *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, pp. 579–583. IEEE.

ElecMorocco (2017). Dataset of facebook comments about moroccan elections 2016, https ://github.com/sentiprojects/elecmorocco2016.

Larkey, L. S., L. Ballesteros, et M. E. Connell (2007). Light stemming for arabic information retrieval. In *Arabic computational morphology*, pp. 221–243. Springer.

Nodarakis, N., S. Sioutas, A. K. Tsakalidis, et G. Tzimas (2016a). Large scale sentiment analysis on twitter with spark. In *EDBT/ICDT Workshops*, pp. 1–8.

Nodarakis, N., S. Sioutas, A. K. Tsakalidis, et G. Tzimas (2016b). Mr-sat : A mapreduce algorithm for big data sentiment analysis on twitter. In *WEBIST (1)*, pp. 140–147.

Pang, B., L. Lee, et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval 2*(1–2), 1–135.

Sehgal, M. S. B., I. Gondal, et L. Dooley (2006). Missing value imputation framework for microarray significant gene selection and class prediction. In *International Workshop on Data Mining for Biomedical Applications*, pp. 131–142. Springer.

Summary

Sentiment analysis is a process during which the semantic orientation or polarity (i.e. positive, negative or neutral) of a given text is determined. This work deals with the sentiment analysis for Facebook's comments written in Arabic Modern Standard or Moroccan Dialectal from a Machine Learning perspective. The process starts by collecting and preparing the Arabic Facebook comments that we have annotated using crowdsourcing. Then, several combinations of extraction and weighting schemes for features construction was conducted to ensure the highest performance of the developed classification models. In addition, to reduce the dimensionality and improve the classification performance, a features selection method is applied. Our Machine Learning approach was implemented with the purpose of analysing the Facebook comments, written in Modern Standard Arabic or in Moroccan Dialectal Arabic, on the real data.