Construction et exploitation d'un corpus multilingue algérien pour l'analyse des opinions et des émotions

Leila Moudjari*, Karima Akli-Astouati**

*l.moudj11@gmail.com
**kakli@usthb.dz
Laboratoire RIIMA, USTHB, Alger, Algérie.

Résumé. Le contenu de ce papier prend en compte la nature linguistique informelle et mixte des langues de médias sociaux qui sont associées au dialecte algérien et utilisées comme moyen d'exprimer des opinions ou des sentiments.

Après avoir identifié les défis de ce type de recherche et mis en avant les spécificités du multilinguisme, une plateforme collaborative appelée TWIFIL (TWIter proFIL) pour l'annotation de données multilingues est proposée. Le résultat est un corpus de tweets annotés. Les premières informations recueillies ont permis d'enrichir les informations de chaque tweet. Des tests ont été realisés sur le corpus généré en utilisant les techniques d'apprentissage automatique.

1 Introduction

Avec plus de 4 milliards d'internautes, le nombre d'utilisateurs des médias sociaux en 2018 est estimé à 3,196 milliards, dont 9 sur 10 ont accès aux plateformes choisies via un appareil mobile. Environ 76% des utilisateurs de ces plateformes ont tendance à exprimer leurs sentiments en cliquant sur les boutons comme "*J'aime*", "*Je n'aime pas*", etc.. . 50% des internautes expriment leurs opinions et sentiments sur les médias sociaux à l'aide *d'émoticônes, d'emojis* ou de *smileys*. Concernant les personnes qui s'expriment en arabe, on retrouve du texte avec 30% de caractères en arabe, 26% de caractères en latin, pour exprimer principalement des idées en anglais ou en français, et environ 15% combinent les deux caractères (Salem, 2017).

Partant de ce constat, et du fait des nombreuses invasions qu'à connu l'Algérie, romaine, byzantine, arabe, turque, espagnole et française, où une réalité socio-linguistique assez complexe est constatée, nous nous intéressons aux posts exprimés dans le dialecte algérien (DALG) pour faire de l'analyse des opinions et des émotions. Nous devons prendre en compte sa diversité langagière où le multilinguisme est omniprésent dans la société algérienne, influençant ainsi le langage d'expression dans les réseaux sociaux. Dans les conversations usuelles, l'arabe dans sa variété soutenue n'est pas utilisé dans les conversations familiales , amicales, etc...

Plus de 99% des Algériens utilisent le tamazight et l'arabe algérien. Il s'agit des langues maternelles de la région. Environ 73% parlent l'arabe algérien et 27% une variante de tamazight [1].

1. https ://www.worldatlas.com/articles/what-languages-are-spoken-in-algeria.html

Ceci s'est répercuté sur l'usage du dialecte dans les échanges sur les réseaux sociaux. Ces dialectes ont moins de normalisation et de standardisation (Saadane et Habash, 2015). Exemple, "Bonjour, labas" est une expression du DALG qui combine un mot français et le mot "labas" écrit en latin et qui correspond phonétiquement au mot arabe "لاباس", signifiant "ça va ?".

Comme les études de l'analyse d'opinions et d'émotions (AOE) arabophones se concentrent principalement sur l'Arabe Standard Moderne (ASM) (Al-Smadi et al., 2017), (Baly et al., 2017), il existe peu de ressources (lexiques, ontologies, corpus, ...) pour le DALG.

Afin de pallier à cela, nous avons développé une plateforme ouverte pour l'annotation manuelle de tweets exprimés en DALG. Elle nous permettra d'abord de constituer un corpus annoté, pour ensuite créer un modèle de prédiction de polarité pour enrichir les métadonnées du modèle à entraîner.

Pour cela, nous avons organisé notre article comme suit. Dans la section 2, nous présentons certains travaux liés à notre problématique en précisant leurs limites. Puis, nous introduisons notre plateforme d'annotation multilingue TWIFIL dans la section 3 en mettant l'accent sur certaines spécificités et défis du DALG. Avant de conclure et rappeler quelques perspectives à notre travail, des tests sont réalisés dans la section 4.

2 Travaux connexes

Les approches proposées pour l'AOE en arabe se concentrent essentiellement sur l'ASM, et ne fournissent que la classification de la polarité des sentiments (neutre, positif et négatif). De plus, les recherches sur les dialectes nord africains sont si rares qu'on peut dire qu'ils sont inexistants. Les premières solutions ont appliqué les outils du Traitement Automatique du Langage Naturel (TALN) conçus pour l'ASM directement sur le DALG. Cependant les performances étaient très faibles. Ceci met l'accent sur la nécessité du développement de solutions et de ressources propres à l'analyse du DALG (Saadane et Habash, 2015).

(Saadane et Habash, 2015) ont présenté une orthographe conventionnelle pour le DALG qui peut être utilisée dans la plupart des applications de TALN, comme l'analyse des sentiments, où ils ont défini des règles phonétiques standards à suivre afin de faciliter la traduction automatique des variantes du DALG et de l'arabe classique.

(Mataoui et al., 2016) ont proposé une approche d'AOE basée sur le lexique pour le DALG. Un ensemble de ressources se résumant en lexique des mots de négation, lexique des mots d'intensification, une liste d'émoticônes avec leurs polarités assignées et un dictionnaire des phrases communes du DALG ont été proposés. Puis, le calcul de la polarité s'est basé sur un ensemble de données annotées manuellement et les ressources citées précédemment.

SIAAC : Sentiment Polarity Identification on Arabic Algerian Newspaper Comments est un corpus proposé par (Rahab et al., 2017). Il est dédié à la classification de polarité des textes recueillis sur le site Internet d'Echorouk (un journal algérien). Les classificateurs Support Vector Machines et Naïve Bayes ont donné des résultats satisfaisants en terme de précision dans les deux modèles. L'utilisation de bigrams a également augmenté leur précision.

Il est évident, d'après les travaux cités, que les ressources accessibles au public pour l'AOE du DALG sont rares. Et celles disponibles, comme celle de (Mataoui et al., 2016) ne donne que la polarité des textes sans aucune information sur l'émotion exprimée ou son émetteur.

Pour pallier à ce problème, nous avons utilisé l'opinion publique en créant une plateforme ouverte pour l'annotation de tweets que l'on présentera dans ce qui suit.

3 Plateforme d'annotation collaborative

Le DALG est moins normalisé que l'ASM. Il a un vocabulaire inspiré de l'arabe, modifié phonologiquement et morphologiquement (Meftouh et al., 2012). Nous devons tenir compte de ce genre de spécificités et bien d'autres dans le développement de notre plateforme.

3.1 Défis et spécificités de DALG

Nous présentons un certain nombre de spécificités et défis du DALG que nous allons considérer dans le traitement du dialecte pour TWIFIL :

Alternance codique, ou code-switching : il s'agit d'une alternance d'au moins deux codes linguistiques (langues ou dialectes) dans une conversation ou même dans une phrase. L'internaute algérien altèrne entre deux ou plusieurs langues, dans le contexte d'une même conversation. Par exemple, "آعطِيْنِي la serviette" emploie un mot arabe et un mot français, qui signifie "donne-moi la serviette". Cependant, le DALG est aussi formé par une transformation des mots des langues qui ont inspiré les Algériens à travers les âges. Prenons le mot "فِنجَال" qui s'inspire du mot arabe "فِنجَان" qui signifie "une tasse", où la dernière lettre est passée de "ن" à "ل".

Codage d'une langue en utilisant l'alphabet d'une autre langue : Il s'agit d'expressions arabes écrites en lettres latines ou "arabizi". Par exemple "khdma kbira", écrite en arabe comme "خدمَة كبِيرَة" signifie "tâche énorme". Ou bien, faire l'inverse, comme dans l'exemple "بَاي بَاي" qui correspond à l'expression anglaise "bye bye".

On peut aussi combiner le code-switching et le codage des langues dans différents alphabets.

Utilisation de chiffres au lieu de lettres ou de mots : Les chiffres ressemblant à certaines lettres et syllabes arabes ont été exploités par la jeunesse algérienne dans les réseaux sociaux. Par exemple, le 7 remplace la lettre "ح" et le 9 remplace le "ق".

Dérivés du DALG : en Algérie, les régions est et ouest ont des accents totalement différents. Prenons l'expression "femme" en arabe, à l'est c'est"مرَا : m'ra", à l'ouest ce sera "شِرَا : shiira".

Idiomes et expressions : Elles sont utilisées généralement à des fins sarcastiques ou pour passer un message indirectement. Par exemple "عُمَر.قَهوَة, ..." est une façon de demander un pot-de-vin. Le sens exact de l'expression "قَهوَة" est "café" et "عُمَر" est un nom commun.

Ces diversités linguistiques requièrent une attention particulière, c'est pourquoi les dialectes écrits sont des langues très riches et variées [2].

3.2 TWIFIL

TWIFIL (TWIter et proFIL) est une plateforme publique accessible à tous via le web [3] ou mobile [4]. Elle a été conçue pour le DALG afin de faciliter la génération de corpus et la construction d'un dictionnaire dialectal algérien. Les lexiques (L1, L2) proposés par (Mataoui et al., 2016) ont été utilisés puis enrichis.

Annotation des données et enrichissement du dictionnaire : les annotateurs contribuent à l'enrichissement de la platforme en donnant :

2. Tous les mots cités du DALG ont été donnés par les auteurs, qui sont des utilisateurs réguliers du dialecte et des médias sociaux.
3. twifil.com
4. https ://play.google.com/store/apps/details ?id=com.leila.kinmokusu.twifil

— la polarité du texte partagé ;
— l'émotion ressentie en lisant le tweet ;
— le domaine du texte ;
— une estimation de l'âge et du genre de l'auteur ;
— la polarité et l'émotion d'un mot ou d'un idiome du dialecte algérien.
— Les utilisateurs peuvent également contribuer à enrichir l'aspect lexical du dialecte, en ajoutant de nouveaux mots, différentes orthographes des mots et différents mots liés, ainsi que des idiomes (ces mots/idiomes sont validés par un administrateur).

TWIFIL a été rendu public le 13/09/2018. Nous avons récolté 7000 tweets (2623 positifs, 2468 négatifs et 1909 neutres). Ils ont été validés par les administrateurs et annotés par 24 personnes de différents domaines et différents âges. Les annotations ont été exploitées comme suit :

— Pour la *polarité*, nous avons exporté la moyenne des valeurs données de l'intervalle -10 (Très négatif) à 10 (Très positif). Avec 0 pour la polarité neutre.
— Pour l'*émotion*, nous avons choisi l'émotion la plus ressentie du tweet. Même chose pour *le contexte et le genre*.
— Pour l'*âge* de l'auteur du tweet, nous avons considéré la moyenne de la médiane des classes d'âges choisies par les annotateurs (de douze à soixante-six ans organisés en classe de trois ans par classe (12-15)...).

Le corpus et le dictionnaire utilisé : un exemple est donné dans la table 1. Le corpus obtenu est disponible sur demande auprès des auteurs.

Texte	lang	Polarité	Âge	Émotion	Genre
had lyamat raho ydor kima linstitut italien ya dra 3leche	dz	-2	23	curiosité	Masculin

TAB. 1 – *Un extrait du corpus généré par TWIFIL*

Table 2 montre un exemple du dictionnaire utilisé pour la translitération des tweets contenant des mots du DALG.

Expression	Différentes écritures	Polarité	Émotions
Hayla : super	هَايلَا, هَايلَى, هَايلَة	5	admiration, joie

TAB. 2 – *Un extrait du dictionnaire généré par TWIFIL*

Le corpus nous a servi à faire de la prédiction/classification de polarité. Cependant il peut être utilisé pour plusieurs applications dans le cadre de l'AOE. On peut faire de la prédiction/classification des sentiments et même il est possible de prédire le genre et/ou l'âge des utilisateurs, etc.

Prétraitement des données : les algorithmes d'apprentissage nécessitent généralement des données numériques. Un prétraitement des tweets bruts collectés et annotés est nécessaire. Le processus que nous avons suivi est décrit par la figure 1. Le vecteur de mots (bag of words) est remplacé par un vecteur booléen de 3000 entrées (nombre de mots les plus courants dans le corpus) où la présence d'un terme du vocabulaire dans le texte est spécifiée.

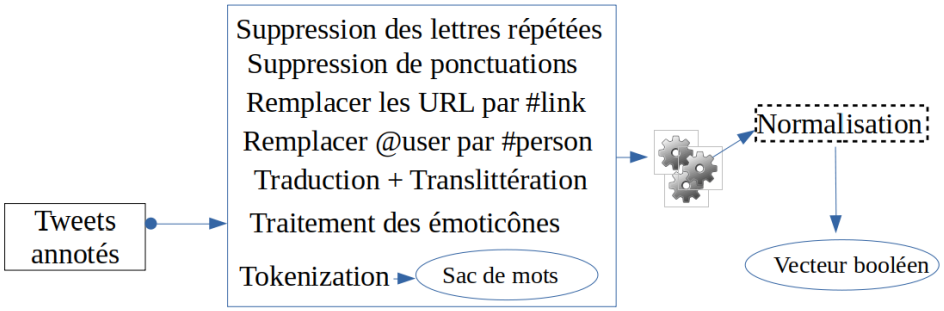

FIG. 1 – *Prétraitement des données*

3.3 Tests

En guise d'illustration de l'emploi du corpus, nous nous sommes focalisés sur la tâche de classification automatique d'opinion en nous basant sur les réseaux de neurones. Il faut noter que cette tâche ne fait pas appel à la totalité des informations offertes par le corpus.
Nous avons aussi effectué une série de tests pour choisir la meilleure longueur du vecteur booléen, ainsi que la meilleure architecture du réseau de neurones, obtenue grâce à la fonction "softmax" en 50 itérations et 100 neurones par couches interne. Les résultats sont prometteurs avec une faible perte (Figure 3) et une précision de 79% (Figure 2).

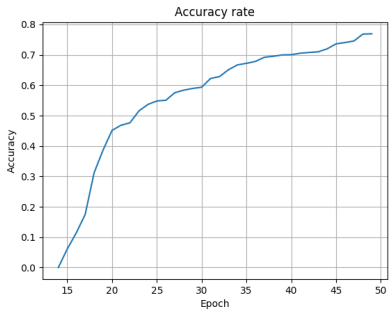

FIG. 2 – *Taux de précision*

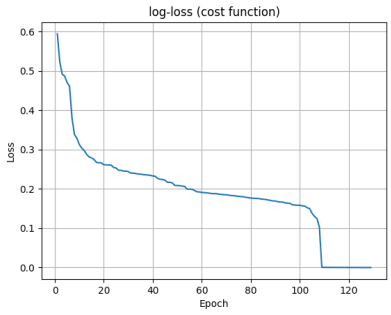

FIG. 3 – *Taux d'erreur (fonction de coût)*

4 Conclusion

L'analyse de l'opinion et de l'émotion du DALG est difficile en raison de la morphologie complexe de la langue. Certaines spécificités liées à ce dialecte ont été présentées. Nous avons mis en avant le manque de ressources accessibles au public pour l'analyse des sentiments.

Ceci, nous a amené à développer une plateforme ouverte pour l'annotation publique de Tweets en DALG "TWIFIL", créant ainsi un corpus qui a servi à faire de la prédiction/classification

de polarité de tweets. Nous avons pu produire une resource qui sera à la disposition de la communauté. Elle constitue un point de depart utile pour ceux qui développent des outils d'analyse de l'opinion et des émotions pour le DALG. Une analyse de l'opinion et des émotions sur les données annotées a été réalisée à un coût relativement faible.

A l'avenir, nous prévoyons de poursuivre ce travail et de relever les défis qui subsistent, afin de développer davantage de ressources et d'outils. Il serait aussi intéressant d'exploiter toutes les informations du corpus pour faire de la prédiction/classification de sentiment, du genre et/ou de l'âge des utilisateurs.

Références

Al-Smadi, M., O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, et B. Gupta (2017). Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotelsfeviews. *Journal of Computational Science*.

Baly, R., H. Hajj, N. Habash, K. B. Shaban, et W. El-Hajj (2017). A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in arabic. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 16*(4), 23.

Mataoui, M., O. Zelmati, et M. Boumechache (2016). A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic. *Research in Computing Science 110*, 55–70.

Meftouh, K., N. Bouchemal, et K. Smaïli (2012). A study of a non-resourced language : an algerian dialect. In *Spoken Language Technologies for Under-Resourced Languages*.

Rahab, H., A. Zitouni, et M. Djoudi (2017). Siaac : Sentiment polarity identification on arabic algerian newspaper comments. In *Proceedings of the Computational Methods in Systems and Software*, pp. 139–149. Springer.

Saadane, H. et N. Habash (2015). A conventional orthography for algerian arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pp. 69–79.

Salem, F. (Feb 5, 2017). Social media and the internet of things towards data-driven policymaking in the arab world : Potential, limits and concerns. *The Arab Social Media Report, Dubai : MBR School of Government, Vol. 7, 2017. Available at SSRN : https ://ssrn.com/abstract=2911832*.

Summary

This paper deals with the problem of the lack of resources in the opinion and emotion analysis related to north african dialects in general and the algerian dialect in particular. A collaborative platform "TWIFIL" for the annotation of multilingual public data is proposed.

The result is a human generated corpus of extracted tweets. The purpose of this action is two-fold. The first, it addresses the shortage of relevant data for algerian dialect's opinion and emotion analysis. Second, it provides a more reliable (the appreciation of not just one person) annotated corpus. We also report on a number of evaluations, we have performed to test the generated corpus.