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