Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien

Imane Guellil[*,**]**— Faical Azouaou**[*]**— Houda Saâdane**[***]**— Nasredine Semmar**[****]

** Laboratoire des Méthodes de Conception des Systèmes. École nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie. http://www.esi.dz*
*** École Supérieure des Sciences Appliquées d'Alger ESSA-Alger*
**** GEOLSemantics, 12 Avenue Raspail, 94250 Gentilly, France*
***** CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, 91191 Gif-sur-Yvette, France*
i_guellil@esi.dz ; i.guellil@essa-alger.dz ; f_azouaou@esi.dz ; houda.saadane@geolsemantics.com ; nasredine.semmar@cea.fr

RÉSUMÉ. *La plupart des outils d'analyse de sentiments traitent essentiellement l'arabe standard moderne (ASM), et peu d'entre eux ne prennent en considération les dialectes. À notre connaissance, aucun outil en libre accès n'est disponible concernant l'analyse de sentiments de textes écrits en dialecte algérien. Cet article présente un outil d'analyse de sentiments des messages écrits en dialecte algérien. Cet outil est fondé sur une approche combinant l'utilisation de lexiques ainsi qu'un traitement spécifique de l'agglutination. Nous avons évalué notre approche en utilisant deux lexiques annotés en sentiments et un corpus de test contenant 749 messages. Les résultats obtenus sont encourageants et montrent une amélioration continue après l'exécution de chaque étape de notre approche.*

ABSTRACT. *Most of the sentiment analysis tools process only Modern Standard Arabic (MSA). Indeed, few dialects are considered by the actual tools, in particular Algerian dialect where we do not identify any free tool carrying texts of this dialect. In this article we present a tool for sentiment analysis of messages written in Algerian dialect. This tool is based on an approach which uses both lexicons and specific treatment of agglutination. This approach was experimented using two sentiment lexicons and a test corpus containing 749 messages. The obtained results were encouraging and showing continuous improvement after each step of the considered approach.*

MOTS-CLÉS : *analyse de sentiments, dialecte algérien, lexique de sentiments, agglutination.*

KEYWORDS: *sentiment analysis, algerian dialect, sentiment lexicon, agglutination.*

1. Introduction

L'arabe est une langue riche et complexe utilisée par plus de 400 millions de locuteurs dans le monde (Siddiqui *et al.*, 2016). Cette langue est cependant dans un état de diglossie[1] dans les pays où elle est utilisée car elle coexiste avec vingts-deux dialectes (Sadat *et al.*, 2014). L'intérêt porté à l'arabe et ses dialectes a largement augmenté au cours de ces dernières années. Un intérêt principalement dû à la proportion des locuteurs arabes au sein des médias sociaux. Au cours de la dernière décennie, plusieurs travaux ont été menés sur l'arabe et ses dialectes. Une revue des méthodes et résultats pour le traitement de l'arabe dialectal a été réalisée par Shoufan et Alameri (2015). Quatre types de tâches y sont présentés : 1) l'analyse basique, 2) la construction de ressources, 3) l'identification du dialecte utilisé, et 4) l'analyse sémantique.

La richesse des médias sociaux en termes d'opinions, d'émotions et de sentiments a suscité l'intérêt de la communauté de recherche à se pencher beaucoup plus sur les problématiques reliées à l'analyse sémantique et plus particulièrement sur l'analyse de sentiments de l'arabe et ses dialectes. L'analyse de sentiments consiste à déterminer la valence (positive, négative ou neutre) d'un message donné. Plusieurs approches d'analyse ou de classification de sentiments ont vu le jour : 1) *l'approche supervisée :* utilisant les techniques d'apprentissage, elle est fondée sur les corpus annotés, 2) *l'approche non supervisée :* fondée sur l'existence d'un lexique de sentiments contenant un ensemble de termes, leur valence et leur intensité pouvant aller de -1 à $+1$ ou encore de -5 à $+5$, et 3) *l'approche hybride :* combinant les deux approches précédentes. Toutes ces approches ont été adoptées dans le cas de l'arabe et de ses dialectes. Cette adoption, ou adaptation, véhicule des problèmes liés à l'approche à laquelle sont ajoutés les problématiques de l'arabe et ses dialectes. Une présentation de ces problèmes est donnée dans (Guellil et Boukhalfa, 2015). Ces problèmes sont principalement liés au manque de corpus annotés et aux problèmes quantitatif et qualitatif des lexiques de sentiments.

Ces problématiques sont recensées dans la littérature comme suit : 1) les problématiques orthographiques liées à la diacritisation ainsi qu'aux différentes manières d'écrire chaque lettre arabe, et 2) les problématiques morphologiques liées à la dérivation, la flexion et l'agglutination. En plus de ces problématiques, chaque dialecte rajoute des problématiques supplémentaires. Prenons par exemple le dialecte algérien (qui est l'objet de ce travail), il s'agit d'un dialecte maghrébin utilisé par plus de 40 millions de locuteurs (ce qui représente 10 % de la population s'exprimant en arabe). Il souffre cependant d'un manque considérable de travaux, d'outils et de ressources. En plus des problématiques de ce dialecte partagées avec l'arabe standard, il dispose d'une très forte richesse du vocabulaire provenant de plusieurs autres langues. Dans (Meftouh *et al.*, 2012), les auteurs illustrent le fait que le dialecte algérien est composé de 65 % d'arabe, de 19 % de français et de 16 % de turque et de

1. Situation où sont en usage deux langues apparentées génétiquement et structurellement et dont les distributions fonctionnelles sont complémentaires (Fishman, 1967).

berbère. Ce dialecte dispose également d'une très grande richesse morphologique, par exemple, l'agglutination aux mots des pronoms personnels et les pronoms de compléments d'objet direct (COD), que nous trouvons aussi dans l'arabe standard, est étendue aux pronoms de compléments d'objets indirect (COI) ainsi que la négation.

Dans ce travail, nous présentons et implémentons une approche d'analyse de sentiments (AS) du dialecte algérien (DALG) combinant l'utilisation de lexiques de sentiments et le traitement de l'agglutination. Cette dernière se compose de deux principales étapes : 1) construction d'un lexique de sentiments, et 2) calcul de la valence et intensité du sentiment d'un message donné. La première étape consiste à construire un lexique de sentiments en dialecte algérien en disposant d'un lexique en anglais. Dans la seconde étape, nous procédons à tous les traitements morphologiques nécessaires dans le cadre de l'analyse de sentiments. Pour évaluer notre approche, nous utilisons deux lexiques de sentiments en anglais : SentiWordNet[2] et SOCAL[3]. Nous construisons par la suite SentiALG et SOCALALG (représentant la version algérienne de ces lexiques). Nous utilisons deux corpus de test : 1) une partie annotée du corpus multidialectal PADIC (Meftouh *et al.*, 2015), et 2) une partie contenant des messages (posts et commentaires) extraits du média social Facebook.

La suite de l'article est organisée comme suit : nous présentons d'abord dans la section 2 les principales caractéristiques de l'arabe et de ses dialectes, ensuite nous exposons dans la section 3 les principaux travaux connexes menés sur l'analyse de sentiments de cette langue et de ses dialectes. La section 4 décrit l'approche que nous proposons pour l'analyse. Nous consacrons la section 5 aux expériences menées ainsi qu'à la présentation des résultats obtenus. La section 6 conclut notre étude et présente nos travaux futurs.

2. Spécificités de l'arabe et ses dialectes : zoom sur le dialecte algérien

La langue arabe est une des langues les plus parlées et utilisées dans le monde. Elle est la langue officielle de plus de vingt-deux pays parlée par plus de 400 millions de locuteurs et elle est utilisée comme vecteur de transmission religieux pour tous les musulmans au nombre de un milliard et demi (Saâdane, 2015) à travers le monde. Elle constitue ainsi un élément principal dans la culture et la pensée d'une partie importante de l'humanité et du patrimoine mondial (Saâdane *et al.*, 2013). Elle est également la quatrième langue la plus utilisée d'Internet (Siddiqui *et al.*, 2016). L'arabe dispose de trois principales variétés qui coexistent côte à côte à savoir : 1) l'arabe classique utilisé dans le Coran, livre sacré des musulmans, 2) l'arabe standard moderne (ASM) utilisé par les locuteurs arabes instruits dans leurs écrits et dans les conversations formelles à savoir dans le système éducatif et littéraire. La plupart des travaux de recherche s'appuient sur cette variante, 3) l'arabe dialectal qui constitue le moyen de

2. http ://sentiwordnet.isti.cnr.it/
3. https ://github.com/sfu-discourse-lab/SO-CAL

communication de la vie quotidienne, employé dans les conversations informelles, interviews et la littérature orale.

2.1. *Spécificités de l'ASM*

L'ASM est classé dans le groupe des langues sémitiques contemporaines qui s'écrit de droite à gauche. L'alphabet arabe contient vingt-huit lettres dont vingt-cinq consonnes et trois voyelles. En plus des voyelles, l'arabe utilise également des marques diacritiques correspondant à des voyelles courtes. Prenons par exemple la lettre ب qui se prononce (b). Si nous mettons au-dessus de cette lettre la diacritique fatha, la lettre devient : بَ se prononçant : 'ba'. Si nous mettons la même diacritique mais au-dessous de la lettre (correspondant à la kasra), la lettre devient بِ et se prononce 'bi'. Mise à part la diacritisation, les lettres arabes ont en général quatre manières de s'écrire : 1) au début du mot, elle s'écrit بـ, prenons l'exemple du mot بئر *biŷr* [4] 'un puits', 2) au milieu du mot, par exemple ـبـ(حقيبة *Haqiybaħ* 'une valise'), 3) à la fin du mot en étant attachée à lettres précédente, par exemple ـب (قلب *qalb* 'un cœur'), et 4) à la fin du mot sans être attachée à la lettre précédente, par exemple ب (باب *baAb* 'une porte').

Un mot en ASM peut également présenter plusieurs aspects morphologiques dont la dérivation, la flexion et l'agglutination. La dérivation consiste à représenter chaque mot sous la forme de « lemme-schéma ». Par exemple, les trois lettres « ktb » est un lemme lié à « l'écriture ». Dans les schémas que nous utilisons, les lettres du lemme sont remplacées par les chiffres 1, 2 et 3 dans l'ordre. Si nous appliquons par exemple le schéma « 1a2a3a » au lemme, nous obtenons le mot كَتَبَ *kataba* 'Il a écrit'. La flexion représente les différentes variations grammaticales d'un mot pouvant être reliées à sa conjugaison, son passage au féminin, pluriel, etc. Le lemme « ktb » est donc conjugué au présent de la sorte : أَكْتُبُ *Âktubu* 'j'écris', تَكْتُبِين *taktubiyn* 'tu écris'/ féminin, etc. Ce verbe est conjugué différemment au passé. Par exemple la traduction de 'j'ai écrit' en arabe est كَتَبْتُ *katabtu*. L'agglutination consiste à rassembler un ensemble de mots, de pronoms (préfixes et suffixes) et de clitiques entre eux. Par exemple la forme agglutinée سيكتبونها *sayaktubuwnahaA* 'ils l'écriront' peut être séparée de la sorte : س+ ي + كتب + ون + ها. Le lemme étant كتب *(ktb)*. La lettre س étant la marque du futur. Les deux lettres ون séparées par le lemme représentent le pronom personnel : « Ils ». Le pronom ها représente le complément d'objet direct (COD). Il est à signaler que les spécificités que nous venons de décrire pour l'ASM

4. Translittération arabe présentée dans schème Habash-Soudi-Buckwalter (HSB) (Habash *et al.*, 2007).

sont aussi présentes dans ses dialectes. Les différences entre l'ASM et ses dialectes résident principalement dans 1) la richesse du vocabulaire des dialectes par rapport à l'ASM et 2) le changement des affixes (préfixes et suffixes) utilisés dans les dialectes. Pour illustrer ces différences, nous nous penchons sur le DALG. Ce dialecte souffre d'un manque considérable de ressources, d'outils et de travaux de recherche le traitant en le comparant aux autres dialectes arabes.

2.2. *Spécificités du dialecte algérien*

Le DALG est utilisé principalement pour la communication orale de tous les jours (dans la vie quotidienne, les séries télévisées en Algérie, etc.). Il n'est pas enseigné dans les écoles, et reste absent des communications écrites officielles. Néanmoins ces dernières années, ce dialecte prend une place plus importante à l'écrit avec les médias sociaux (Harrat *et al.*, 2017). Comme nous avons exposé les principales caractéristiques orthographiques et morphologiques du DALG partagées avec l'ASM au sein de la section 2.1, nous nous concentrons dans cette partie sur les caractéristiques propres au DALG.

2.2.1. *Spécificités orthographiques du dialecte algérien*

Le DALG fait appel à toutes les voyelles et consonnes utilisées par l'ASM. En plus de ces dernières, il fait appel aux trois lettres ڤ، ڨ، پ, se prononçant respectivement p, g, v (Meftouh *et al.*, 2015). Le DALG est enrichi par les langues des groupes ayant colonisé ou géré la population algérienne au cours de l'histoire du pays. Parmi les langues de ces groupes, citons le turc, l'espagnol, l'italien et plus récemment le français (Saâdane et Habash, 2015 ; Saâdane, 2015 ; Meftouh *et al.*, 2012). De ce fait, au sein du DALG, nous trouvons des mots tels que سلم *sallam* 'saluer' et ayant comme origine l'ASM (Harrat *et al.*, 2017). Nous pouvons également trouver un mot comme فرملي *Farmliy* 'infirmier' étant originaire du français, ou encore le mot بابور *baAbuwr* 'bateau', issu du turc, شلاغم *šlaAγam* 'moustaches' emprunté du berbère, زبلة *zablaħ* 'faute', issu de la langue italienne et سيمانة *siymaAnaħ* 'une semaine' emprunté de l'espagnol.

2.2.2. *Spécificités morphologiques du dialecte algérien*

Nous abordons au cours de cette partie quatre aspects importants reliés à la morphologie (agglutination) du DALG, à savoir : 1) la conjugaison, 2) la négation, 3) les noms et adjectifs, et 4) les compléments d'objet direct (COD) et les compléments d'objet indirect (COI).

2.2.2.1. Conjugaison en dialecte algérien

Comme au sein de n'importe quel langage, la conjugaison inclut l'ajout d'un ensemble de préfixes et de suffixes à un lemme donné. Ces affixes varient selon le

Verbe	Présent	Traduction	Passé	Traduction	Impératif	Traduction
حب	نحب	J'aime	حبيت	J'ai aimé	-------	-------
	تحب	Tu aimes	حبيت	Tu as aimé	حب	Aime
	تحبي	Tu aimes	حبيتي	Tu as aimé	حبي	Aime
	يحب	Il aime	حب	Il a aimé	-------	-------
	تحب	Elle aime	حبت	Elle a aimé	-------	-------
	نحبو	Nous aimons	حبينا	Nous avons aimé	نحبو	Aimons
	تحبو	Vous aimez	حبيتو	Vous avez aimé	حبو	Aimez
	يحبو	Ils aiment	حبو	Ils ont aimé	-------	-------

Tableau 1. *Conjugaison du verbe 'aimer' au présent, passé et impératif*

pronom utilisé. Ils sont généralement les mêmes pour tous les verbes. Nous faisons figurer dans le tableau 1 la conjugaison du verbe 'aimer' en DALG aux temps les plus utilisés, c'est-à-dire le présent et le passé composé de l'indicatif ainsi qu'à l'impératif. Notons qu'en arabe il existe deux formes pour la deuxième personne du singulier, selon le sexe de la personne à qui l'on s'adresse.

À partir du tableau 1, nous pouvons déjà conclure que les lettres ن ، ت ، ي représentent des préfixes et les lettres ي ، و ،ت ، يت ، ينا ، يتي ، بيتو représentent des suffixes pour le DALG. Il est également à noter que la conjugaison au futur ne figure pas sur ce tableau car cette dernière est la même que celle au présent associé à des mots du futur tels que اومبعد *Awmbʕd*, راح *raAH*, غدوة *γadwaħ*, etc, voulant respectivement dire (demain, aller, après) (Harrat *et al.*, 2016).

2.2.2.2. Négation en dialecte algérien

La négation du DALG peut se faire de deux manières principales, soit 1) à l'aide des lettres ما ش *maA…š* ou encore des lettres م ش *maA…š*, 2) à l'aide du mot ماشي *maAšiy*. Prenons l'exemple de la phrase « Je l'aime » qui devient en DALG نحبو *nHabuw* pour le masculin et نحبها *nHabhaA* pour le féminin. La négation de cette phrase étant « Je ne l'aime pas » qui devient مانحبوش *mAnHabuwš*. Donc pour faire une analogie avec le français, le ما *maA* joue le rôle du 'ne' et le «ش» *š* joue le rôle de 'pas'. Prenons un autre exemple avec la phrase : « Cette fille est bien » qui

COD	Exemple	Traduction	COI	Exemple	Traduction
ني	تحبيني	Tu m'aimes	لي	قولي	Tu me le dis
ك	كرهتك	Je t'ai détesté	لك	قولتلك	Je te l'ai dit
ه هو و	كرهتو حبيته	Je l'ai détesté Je l'ai aimé	لو	نقولو	Je lui dis
ها	كرهتها	Je l'ai détesté	لها	قولتولها	Je le lui ai dit
نا	كرهتونا	Vous nous avez détestés	لنا نا	قولتونا	Tu nous l'as dit
كم	حبيناكم	Nous vous avons aimés	لكم	قولتلكم	Je vous l'ai dit
هم	كرهتوهم	Vous les avez détestés	لهم	قوللهم	Dis-leur

Tableau 2. *Les pronoms COD et COI du dialecte algérien*

devient en DALG «هاد الطفلة مليحة» *haAd AlTuflaħ mliyHaħ*. Sa négation étant « Cette fille n'est pas bien » qui devient en DALG «هاد الطفلة ماشي مليحة» *haAd AlTuflaħ maAšiy mliyHaħ*. Nous constatons maintenant que pour exprimer la négation de مليحة *mliyHaħ*'bien', nous employons le terme ماشي *maAšiy*. Donc, pour résumer, nous observons que la séquence ما ش est utilisée avec les verbes et le terme ماشي est utilisé avec les noms et les adjectifs.

2.2.2.3. COD et COI du dialecte algérien (clitiques pronominaux)

Les COD et COI sont également agglutinés aux verbes conjugués en DALG, au même titre que les pronoms personnels et la négation. Les pronoms COD et COI représentent des suffixes du verbe conjugué. Ces pronoms ont déjà été étudiés (Guellil et Azouaou, 2017 ; Saâdane et Habash, 2015 ; Harrat *et al.*, 2016). Il existe cependant un nombre plus important de pronoms que ceux cités dans ces travaux car l'agglutination des pronoms de base entre eux donne naissance à de nouveaux suffixes. Nous récapitulons dans le tableau 2, l'ensemble des COD et COI de base.

2.2.2.4. Noms et adjectifs dans le dialecte algérien

Comme pour toutes les langues, les noms et les adjectifs ont un genre et un nombre. Plusieurs auteurs (Harrat *et al.*, 2016 ; Harrat *et al.*, 2017 ; Guellil et Azouaou, 2016) attestent que pour former le féminin des noms et adjectifs en DALG la lettre ة doit être ajoutée comme suffixe. Par exemple le féminin de l'adjectif مليح *mliyH* 'bien' est مليحة *mliyHaħ*. Concernant le pluriel, tous les travaux étudiés s'accordent sur le fait que le masculin pluriel et le féminin pluriel sont formés à partir du nom ou de l'adjectif auxquels sont respectivement ajoutés les suffixes ين *yn* et ات *At*. Par exemple, le pluriel de l'adjectif فنيان *fanyaAn* 'fainéant' est فنيانين *fanyaAniyn* et le pluriel du nom شيخة *šiyxaħ* 'enseignante' est شيخات *šiyxaAt*.

3. Analyse de sentiments de l'arabe et ses dialectes : état de l'art

L'analyse de sentiments (AS) est un domaine interdisciplinaire se trouvant entre les domaines de traitement du langage naturel, de l'intelligence artificielle et la fouille de texte (Medhat *et al.*, 2014). L'AS s'effectue sur trois niveaux : documents, phrases et aspects. La richesse des médias sociaux en termes d'opinion et de sentiment a suscité l'intérêt de la communauté de recherche (Guellil et Boukhalfa, 2015). Cet intérêt est aussi important pour la langue arabe compte tenu du nombre massif des utilisateurs s'exprimant en arabe et ses dialectes sur Internet : 156 millions d'utilisateurs selon Siddiqui *et al.* (2016), soit 18,8 % de la population globale d'Internet (Korayem *et al.*, 2012). En nous fondons sur les caractéristiques de l'arabe et de ses dialectes présentées au sein de la section 2, nous concluons que les approches dédiées aux autres langues ne pourraient être appliquées dans notre cas (sauf avec modifications majeures). Au sein du présent travail, nous nous concentrons donc sur les travaux menés sur l'arabe et ses dialectes où nous nous fondons sur six états de l'art regroupant et analysant les travaux menés sur cette langue ainsi que ses dialectes (Kaseb et Ahmed, 2016 ; Biltawi *et al.*, 2016a ; Korayem *et al.*, 2012 ; Harrag, 2014 ; Assiri *et al.*, 2015 ; Alhumoud *et al.*, 2015). Après une analyse approfondie de ces études, nous concluons cependant que l'AS de l'arabe et de ses dialectes peut s'effectuer en suivant trois approches (comme pour toutes les autres langues) : supervisées, non supervisées et hybrides. Nous présentons, l'ensemble des travaux menés sur l'AS de l'arabe et ses dialectes tout en les regroupant par le type d'approche utilisé.

3.1. *Approches supervisées*

L'approche supervisée dépend de l'existence des données (documents, phrases, etc.) annotées comme positives, négatives ou neutres (Biltawi *et al.*, 2016b). L'approche supervisée, reconnue également par la classification supervisée, peut se faire en faisant appel à plusieurs algorithmes de classification tels les machines à vecteurs support (MVS), les classifieurs bayésiens naïfs (BN), les arbres de décision

(AD), etc. De nombreux travaux ont été réalisés pour analyser et classer les sentiments de l'arabe et ses dialectes en utilisant des approches supervisées. Citons notamment le travail de Cherif *et al.* (2015a) qui a fait appel à la technique MVS pour classer les sentiments de message écrit en arabe (ASM) en cinq classes allant de très bien à pas bien du tout. Pour réaliser cette tâche, les auteurs ont commencé par le prétraitement des phrases. Ils font également appel à un extracteur de lemmes présenté dans un travail précédent (Cherif *et al.*, 2015b) afin de supprimer les préfixes et suffixes des mots pour obtenir leurs radicaux. Il faut cependant noter que ces auteurs suppriment les préfixes et suffixes reliés à la conjugaison, au pluriel et aux pronoms. Il ne supprime cependant pas les affixes reliés à la négation qui pourraient affecter la qualité de l'analyse de sentiments.

Dans (Hadi, 2015) les auteurs ont utilisé les deux méthodes MVS et BN pour classifier un ensemble de messages en positif, négatif ou neutre. Pour ce faire, ils construisent un corpus arabe (ASM) contenant 3 700 messages extraits de Twitter. Chaque message a été annoté par trois locuteurs arabes natifs en positif, négatif et neutre. Nous enchaînons avec le système « SAMAR » analysant en parallèle la subjectivité d'un texte ainsi que ses sentiments (Abdul-Mageed *et al.*, 2014). Ce travail se focalise principalement sur le ASM ainsi que le dialecte égyptien. Les auteurs utilisent plusieurs corpus dont certains extraits des médias sociaux et d'autres ayant été utilisés dans d'autres travaux tels que (Diab *et al.*, 2010). Les auteurs de ce travail font appel à une variante de la MVS *« light »* proposée dans (Joachims, 2002). Ils se fondent cependant sur beaucoup de caractéristiques (*features*) dont l'analyse morphologique, la recherche des parties du discours, le lexique annoté, etc. Dans (Itani *et al.*, 2012), les auteurs ont exploité un modèle BN pour classifier automatiquement les sentiments des posts Facebook écrits en plusieurs dialectes arabes. Ils se concentrent sur les dialectes syriens, égyptiens, irakiens et libanais. Nous finissons cette partie avec le travail de Mdhaffar *et al.* (2017) qui combinent plusieurs classificateurs pour analyser le sentiment de messages écrits en dialecte tunisien. Parmi les principaux classificateurs utilisés la MVS et le BN. Dans ce travail, les auteurs présentent également la construction du corpus TSAC qui est un corpus tunisien dédié à l'analyse de sentiments.

Tous les travaux présentés au sein de cette catégorie se fondent sur un corpus annoté pour pouvoir effectuer la classification des sentiments. La construction de ce corpus est, dans la majorité des cas, manuelle, ce qui est très consommateur de temps et d'effort, et amène les auteurs à construire souvent des corpus réduits influant négativement les résultats. Du fait que notre approche s'appuie sur un lexique de sentiments, nous avons fondé les travaux présentés dans notre cas sur les différents prétraitements proposés.

3.2. *Approches non supervisées*

L'approche non supervisée est une approche qui se fond sur un lexique de sentiments. Plusieurs travaux ont également été menés en faisant appel à cette

approche. Nous commençons par le travail de Al-Ayyoub *et al.* (2015) qui a permis de construire un lexique de 120 000 termes arabes (ASM). Pour aboutir à ce dernier, les auteurs ont commencé par collecter des lemmes en arabe. Ils les ont traduits en anglais en utilisant Google traduction. Ils ont supprimé ensuite les mots répétés. Ces auteurs ne prennent pas en considération le contexte du lemme dans le processus de traduction. Ils utilisent ensuite un lexique de sentiments anglais pour déterminer leur valence et intensité.

Dans la même perspective, un autre lexique contenant 157 969 synonymes et 28 760 lemmes a été construit dans (Badaro *et al.*, 2014). Pour aboutir à ce dernier, les auteurs ont dû combiner plusieurs ressources de l'arabe. Dans (Mohammad et Turney, 2013), les auteurs ont développé un lexique de sentiments contenant 14 182 unigrammes anglais classés en positif ou négatif à l'aide du Amazon Mechanical Turk[5]. Ce lexique a ensuite été traduit en quarante langues dont l'ASM. L'auteur dans (AL-Khawaldeh, 2015) a construit également un lexique de sentiments, mais en traitant aussi de la négation. Ces auteurs se consacrent sur l'arabe (ASM) et ont également défini un ensemble de règles pour capturer la morphologie de la négation. Abdulla *et al.* (2014a) ont commencé par la construction d'un corpus contenant 4 000 commentaires textuels collectés à partir de Twitter et Yahoo Maktoob[6]. Un lexique a alors été construit, il ne contient que 300 mots.

Nous enchaînons avec les travaux de Abdulla *et al.* (2014b) qui se focalisent sur trois techniques de construction de lexiques avec une manuelle et deux autres automatiques. Pour la partie automatique, les auteurs se focalisent sur la traduction du lexique de sentiments anglais SentiStrength[7] en utilisant Google traduction. Nous terminons avec le travail de Mataoui *et al.* (2016) qui est le seul à étudier l'AS du DALG. Au sein de ce travail, les auteurs ont construit manuellement un lexique de sentiments en commençant par un lexique arabe et égyptien existant. Pour répondre aux caractéristiques morphologiques de cette langue et de ce dialecte, les auteurs utilisent l'outil de lemmatisation nommé « Khoja »[8] (Khoja et Garside, 1999).

Nous constatons donc que l'approche non supervisée est essentiellement fondée sur la construction de lexiques. Pour aboutir à cette construction, les auteurs tendent vers trois techniques : 1) construction manuelle (dans ce cas-là ce n'est pas purement supervisé car les mots sont manuellement annotés), 2) combinaison entre plusieurs ressources existantes, 3) traduction d'une ressource existante. Pour notre cas et vu que l'approche de construction manuelle est très consommatrice de temps (en plus, elle a déjà été présentée dans (Mataoui *et al.*, 2016)), que la combinaison de plusieurs ressources est impossible dans le cas du dialecte algérien, souffrant d'un manque considérable de ressources, nous optons pour une construction à base de traduction. Pour ce faire, nous nous appuyons sur les différents travaux de Abdulla *et al.* (2014a)

5. https ://www.mturk.com/mturk/welcome
6. L'édition arabe de yahoo.
7. http ://sentistrength.wlv.ac.uk/
8. https ://github.com/motazsaad/khoja-stemmer-command-line

et de Abdulla *et al.* (2014b), qui se focalisent sur des lexiques de sentiments existants pour faire la traduction vers l'ASM. En ce qui concerne les travaux sur le dialecte algérien, l'unique travail recensé est celui de Mataoui *et al.* (2016). Néanmoins, ce travail fait appel à l'outil « Khoja » pour la phase de lemmatisation. Cependant, cet outil est dédié à l'ASM et ne peut donc pas être utilisé pour le DALG. L'une des problématiques majeures reliées aux dialectes arabes est que les outils dédiés à l'ASM ne donnent pas de bons résultats pour ses dialectes (Harrat *et al.*, 2014). En plus, les auteurs utilisent cet outil sans y apporter aucune modification. Pour illustrer l'incapacité de cet outil à traiter le DALG, nous avons lemmatisé un ensemble de phrases en DALG et nous avons constaté que cet outil ne traitait en aucun cas les mots agglutinés, tels que مانسساهاش *mAnnsAhAš* 'je ne l'oublierai pas'. Il change aussi le sens de certains mots, comme le mot مليح *mliyH* 'bien' qu'il lemmatise en ملح *mlH* 'sel'. Ceci est principalement dû aux spécificités et suffixes du DALG qui ne sont pas partagés avec l'ASM. Un tel outil ne peut donner de bons résultats que pour des messages pouvant être classés comme des messages partagés entre l'ASM et le DALG, par exemple كرهت حياتي *kraht HyaAtiy* 'j'en ai marre de ma vie'.

3.3. *Approches hybrides*

L'approche hybride consiste à combiner les méthodes utilisées dans l'approche supervisée et non supervisée. Par exemple, dans le travail de Hedar et Doss (2013), les auteurs ont utilisé un classificateur MVS. Pour faire cette classification, les auteurs utilisent un lexique contenant 1 300 mots dont 600 sont positifs et 700 négatifs. Ce travail s'appuie sur l'argot égyptien. Les résultats expérimentaux ont montré que l'utilisation du lexique améliore considérablement les résultats. Dans (Khalifa et Omar, 2014), les auteurs ont proposé une méthode hybride fondée sur un lexique et un classificateur BN en même temps. La méthode proposée est précédée d'une phase de prétraitement (normalisation, segmentation, etc.). Le lexique intervient pour remplacer les mots avec leurs synonymes. Ces auteurs se focalisent sur l'arabe standard (ASM).

L'approche hybride est une approche qui donne de très bons résultats, mais nous ne pouvons l'appliquer à ce stade puisque nous ne disposons pas de données annotées.

Dans le but de synthétiser tous les travaux présentés, nous les classifions dans le tableau 3, par rapport à la langue étudiée (ASM, dialecte arabe en donnant le type du dialecte), ainsi que l'approche de classification et la méthode utilisée pour le travail présenté. En ce focalisant sur le tableau 3, nous constatons qu'un seul travail a été mené sur l'AS du DALG. Nous précisons cependant que le DALG est un dialecte peu étudié en général. Peu de travaux se sont focalisés sur ce dernier (Meftouh *et al.*, 2015 ; Harrat *et al.*, 2017 ; Guellil *et al.*, 2017b ; Saâdane et Habash, 2015 ; Guellil *et al.*, 2017a ; Harrat *et al.*, 2016)

Approches et méthodes utilisées		ASM	Dialecte arabes	
			Les travaux	Le dialecte étudié
Approche supervisée	MVS	(Cherif *et al.*, 2015)	(Abdul-Mageed *et al.*, 2014)	Égyptien
	BN		(Itani *et al.*, 2012)	Syrien Égyptien Irakien libanais
	MVS + BN	(Hadi, 2015)	(Mdhaffar *et al.*, 2017)	Tunisien
Approche non supervisée		(Al-Ayyoub *et al.*, 2015) (Abdulla *et al.*, 2014b) (Abdulla *et al.*, 2014a) (Badaro *et al.*, 2014)	(Abdulla *et al.*, 2014a)	Jordanien
		(Mohammad *et al.*, 2013) (AL-Khawaldeh, 2015) (Mataoui *et al.*, 2016)	(Mataoui *et al.*, 2016)	Algérien
Approche hybride	MVS + lexique		(Hedar *et al.*, 2013)	L'argot égyptien
	NB + lexique	(Khalifa *et al.*, 2014)		

Tableau 3. *Classification et synthèse des travaux étudiés*

4. Analyse de sentiments de textes écrits en dialecte algérien

Dans cette partie, nous définissons et implémentons une approche non supervisée, fondée sur un lexique de sentiments, pour déterminer la valence (positive, négative) et l'intensité (1,54, $-\ 2,87$, etc.) d'un message donné écrit en DALG. Notre approche reçoit en entrée un message écrit en DALG (en caractères arabes) ainsi qu'un lexique de sentiments en DALG préalablement construit en traduisant un lexique de sentiments anglais existant. Elle retourne en sortie la valence du message ainsi que son intensité. Pour ce faire, notre approche est constituée de deux étapes principales 1) construction d'un lexique de sentiments en DALG, 2) calcul de la valence et de l'intensité d'un message (en appelant le lexique créée à l'étape 1). La figure 1 illustre l'architecture générale de notre approche.

4.1. *Étape 1 : construction d'un lexique de sentiments en dialecte algérien*

Cette étape reçoit en entrée un lexique de sentiments en anglais (nous choisissons l'anglais car c'est la langue qui a bénéficié de plus de travaux sur l'AS (Guellil et

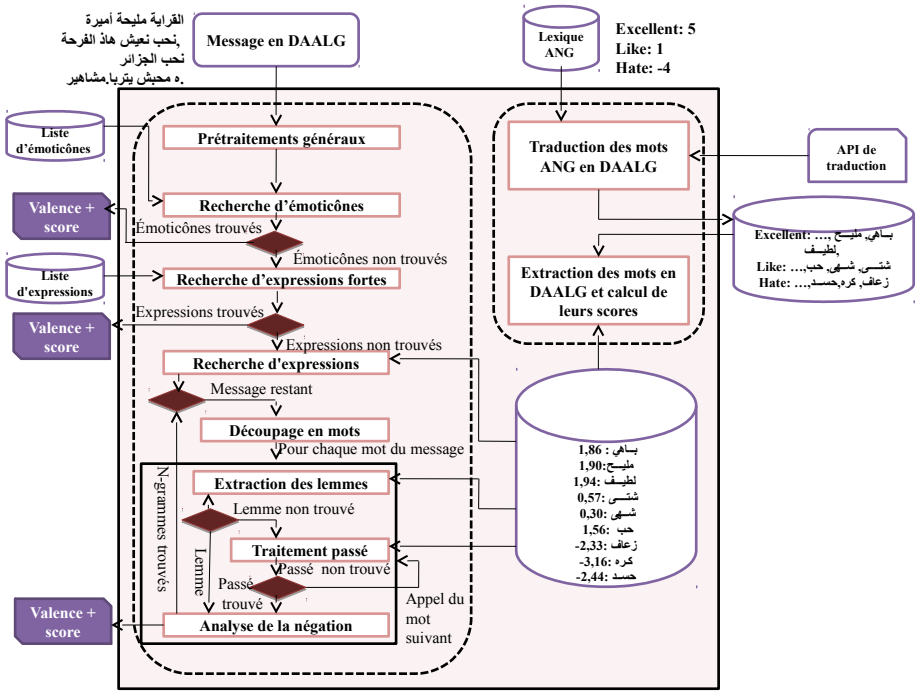

Figure 1. *Architecture générale de notre approche*

Boukhalfa, 2015)). Chaque mot de ce lexique est traduit en appelant une API de traduction. Après la phase de traduction, un lexique de sentiments est construit en extrayant chaque terme en DALG et calculant son score. Cette étape contient donc deux sous-étapes : 1) traduction des mots du lexique anglais en DALG et 2) extraction des mots en DALG et calcul de leurs scores.

4.1.1. *Traduction des mots du lexique anglais en dialecte algérien*

Pour faire cette traduction, nous faisons appel à l'API glosbe. Cette dernière prend en entrée un mot en anglais et retourne un ensemble de mots en DALG. La spécificité de cette API est que la traduction est faite par des utilisateurs ordinaires natifs du DALG. Nous traduisons chacun des mots de notre lexique de sentiments anglais en faisant appel à cette API. Nous affectons à tous les mots récoltés le même score que le mot en anglais. Prenons par exemple le mot anglais 'excellent'[9] voulant dire 'excellent' et ayant un score égal à $+5$. Sa traduction en DALG donne les mots :

9. https ://glosbe.com/en/arq/excellent

باهي *baAhiy*, لطيف *lTiyf*, مليح *mliyH*, etc. Tous ces mots ont donc un score égal à $+5$, de même que le mot 'excellent'. Nous nous inspirons dans cette partie de plusieurs travaux : 1) le travail de Elarnaoty *et al.* (2012) qui fait la traduction de la version anglaise du lexique MPQA (Multi-Perspective Question Answering) en arabe, 2) le travail de El-Halees *et al.* (2011) qui fait la traduction de Sentistrength en arabe également, 3) le travail de Al-Ayyoub *et al.* (2015) qui traduit des lemmes arabes en anglais.

4.1.2. *Extraction des mots en dialecte algérien et calcul de leurs scores*

Après la réalisation de la première phase (section 4.1.1), nous nous sommes rendu compte qu'un mot en DALG est associé à plusieurs mots en anglais et peut donc avoir plusieurs scores. Prenons par exemple le mot مليح *mliyH* 'bien'. Ce mot peut être associé aux mots anglais : *excellent, best, generous,* etc. Nous observons donc que les mots anglais auxquels est associé le mot مليح *mliyH* peuvent avoir différents scores ('excellent' $(+5)$, 'generous' $(+2)$, etc). Nous extrayons donc, au sein de cette partie, tous les mots en DALG et sans répétition et calculons leurs scores. Pour le calcul du score, nous prenons la moyenne des scores de tous les mots anglais auxquels notre mot en DALG est associé. Nous obtenons ainsi notre lexique en DALG contenant par exemple le mot مليح *mliyH* avec un score égal à 1,90, le mot كره *krah* avec un score égal à $-3,16$, etc.

4.2. *Étape 2 : calcul de la valence d'un message écrit en dialecte algérien*

Cette étape reçoit en entrée un message écrit en DALG et retourne en sortie sa valence et son intensité. Si nous prenons, par exemple, la phrase مليحـــــة القرررراية # أميرة *mliyHaħ AlqrrrraAyaħ # Âmiyraħ* traduite en 'C'est bien d'étudier Amira'. Si nous nous appuyons sur le lexique construit à l'étape 1, cette phrase est reconnue positive avec une intensité globale égale à 1,69. Nous constatons cependant que pour arriver à un tel résultat, un ensemble de traitements doivent être appliqués à ce message dont 1) différents prétraitements allant de la suppression des lettres répétées à la recherche des émoticônes et expression fortes, 2) recherche des n-gramme du message présents dans le lexique, 3) extraction du lemme de chaque mot du message non identifié comme n-gramme, 4) traitement du passé, et enfin 5) analyse de la négation.

4.2.1. *Prétraitement d'un message en dialecte algérien*

Nous nous inspirons des différents travaux de Cherif *et al.* (2015a) pour proposer plusieurs prétraitements nécessaires au traitement de l'arabe et de ses dialectes. Nous nous inspirons également des différents travaux de Guellil et Azouaou (2017), Harrat *et al.* (2016) et Saâdane et Habash (2015) se focalisant sur les caractéristiques propres au dialecte algérien. Nous proposons donc l'ensemble des prétraitements suivants :

– suppression des blancs, des lettres longues (reconnus par tatweel), par exemple

مليحـــــة devient مليحة *mliyHaħ* ;

– suppression des exagérations, par exemple القرررررراية *AlqrrrrraAyaħ* est transformé en القراية *AlqraAyaħ* ;

– suppression de certaines ponctuations telles que le # et espacements de certains points ('., !, ?') attachés au mots ;

– remplacement des caractères arabes par leurs Unicodes pour traiter le phénomène relié à la présence de différentes lettres selon leurs emplacements (traités au sein de la section 2.1) ;

– recherche d'émoticônes et d'expressions fortes tels ,نموت على *nmuwt ςlaý* 'j'adore' ou encore ينعل الله *yanςal Allah* 'Que Dieu maudisse', etc. Le but étant d'attribuer directement au message la valence de l'émoticône ou de l'expression trouvée. Dans le cas où il y a plusieurs émoticônes ou expressions, nous ne prenons en considération que la première ;

– prétraitements reliés à l'opposition exprimée en dialecte algérien à l'aide du mot بصح *baSaH* 'mais', notre système ne prend en considération que la partie du message qui se trouve après بصح. Nous procédons ainsi, car nous avons constaté que le mot بصح annulait le sentiment de la partie qui le précède. Prenons l'exemple du message : حبيت نخرج نلعب نفرح بصح راني مريضة *Habiyt nuxruj nalςb nafraH baSaH raAniy mriyĎaħ* traduit en 'je voulais sortir jouer être heureuse mais je suis malade'. Ce message à beau contenir plus de mots positifs que négatifs, il reste négatif parce que c'est la partie qui est après l'opposition qui détermine véritablement le sentiment.

4.2.2. *Recherche des expressions du message dans le lexique de sentiments*

Notre lexique de sentiments peut contenir des séquences d'un mot tel que مليح 'bien', de deux mots tels que عالي بصوت *bSawt ςaAliy* 'à haute voix', de trois mots tels que مرا مرا على *ςlaŷ maraA maraA* 'des fois' ou encore de quatre mots tels que برك وحدة مرة يستعمل *yastaςmal marаħ waHdaħ bark* 'il utilise une seule fois seulement'. Pour rechercher la présence d'une séquence de mots du message dans le lexique, nous formons d'abord l'ensemble de ces séquences, par exemple القراية مليحة أميرة *Âmiyraħ mliyHaħ AlqraAyaħ*, ce dernier contient trois mots individuels (القراية ، مليحة ، أميرة), deux séquences de deux mots (القراية مليحة ، مليحة أميرة) et une seule séquence de trois mots (القراية مليحة أميرة). Une fois les séquences de mots du message construites, nous commençons leur recherche dans le lexique. Dans le cas de l'exemple présenté, un seul unigramme مليحة *mliyHaħ* a pu être trouvé dans le lexique. Un traitement de la négation est ensuite indispensable pour déterminer le score des séquences retrouvées (que nous présenterons dans la section 4.2.5).

4.2.3. *Extraction du lemme des mots en dialecte algérien*

Dans cette étape, nous extrayons de chaque mot (non reconnu par l'étape 4.2.2) son lemme (à l'aide du lexique de sentiments). Nous procédons comme suit : 1) nous recherchons tous les mots du lexique inclus dans notre mot, 2) nous sélectionnons les mots ayant le plus grand nombre de lettres et 3) nous vérifions que le mot sélectionné peut être découpé en *préfixe* + *lemme* + *suffixe* et nous enlevons ainsi les préfixes et suffixes pour ne garder que le lemme et ce, en se fondant sur le travail de Cherif *et al.* (2015a). Pour cela nous définissons une liste globale des préfixes et suffixes du DALG. Nous illustrons cette étape en réutilisant l'exemple القراية مليحة أميرة. Rappelons juste que l'unigramme مليحة *mliyHaħ* a été identifié au sein de l'étape 4.2.2. Nous aurons donc à traiter les deux mots restants : القراية *AlqraAyaħ* et أميرة *Âmiyraħ*. Aucune partie du mot أميرة n'a été retrouvée dans notre lexique de sentiments. Néanmoins le mot قراي *qraAy* a été détecté comme faisant partie du mot القراية. Ce mot se présente donc sous cette forme : ال+قرايـ+ة *Al+qraAy+ħ*. Comme le mot ال (déterminant) est un préfixe reconnu du DALG et la lettre ة est un suffixe reconnu également, le lemme قراي *qraAy* est donc validé. La négation doit ensuite être traitée (se référer à la section 4.2.5).

4.2.4. *Traitement du passé*

Certains verbes conjugués au passé ne peuvent pas être traités comme la plupart des mots cités dans la section 4.2.1. Nous ne pouvons donc pas directement extraire les affixes de ces verbes car nous devons d'abord faire des transformations sur leurs lemmes. Nous citons par exemple, les deux verbes نسى *nsaý* 'oublier' et بكى *bkaý* 'pleurer'. La conjugaison de ces deux verbes au passé est : نسيت *nsiyt* et بكيت *bkiyt* (pour la première personne du singulier). La lettre ى *ý* est donc supprimée afin de rajouter les suffixes nécessaires. Pour pouvoir former le lemme de ces verbes, il faut supprimer les suffixes et ajouter la lettre ى *ý*. Pour réaliser cette étape, nous procédons comme suit : 1) identification de la liste des suffixes passés (par exemple يتو، ينا، يت,etc.) et suppression de ces derniers dans les mots analysés (à cette étape نسيت *nsyt* est transformé en نس *ns*, puisque le suffixe يت est enlevé), 2) ajout de la lettre ى à la fin des mots récoltés (à cette étape نس est transformé en نسى *nsý*) et 3) recherche du lemme obtenu dans le lexique de sentiments, si le lemme obtenu est trouvé, le mot est validé et la négation est ensuite traitée (voir la section 4.2.5).

4.2.5. *Analyse de la négation*

L'analyse de la négation représente un défi de recherche important concernant l'AS et pas seulement pour l'arabe, mais pour toutes les langues. Néanmoins ce

défi est accentué dans le cas de l'arabe et ses dialectes où la négation s'agglutine le plus souvent au mot, au même titre que les différents pronoms. Pour plus de détails sur la négation, nous vous invitons à vous référer à la section 2.2.2.2. Les utilisateurs peuvent faire appel à la négation de différentes manières, par exemple le mot مانحبكمش *mAnHabkumš* 'je ne vous aime pas' peut s'écrire ما نحبكمش *mA nHabkumš*, ما نحبكم ش *mAnHabkum* š ou encore نحبكم ش *mA nHabkum š*. Nous constatons que la négation peut être agglutinée aux termes comme elle peut être séparée de ces derniers. Nous traitons dans ce travail deux sortes de négations : 1) la négation agglutinée au mot, et 2) la négation séparée du mot. Pour les deux cas, nous définissons une liste de préfixes et de suffixes reliés à la négation. Nous avons cependant constaté que, dans la plupart des cas, la négation n'influe pas seulement sur le mot qu'elle précède, mais sur le reste de la phrase également. Une fois qu'un préfixe ou un suffixe de négation est détecté, nous inversons le score des mots succédant à cette négation (multipliant le score par (-1)).

5. Étude expérimentale

Pour développer notre solution nous nous sommes inspirés du programme élaboré par Taboada *et al.* (2011). Le programme et les différents lexiques de ses auteurs sont librement téléchargeables[10]. La différence entre cette solution et la nôtre réside au niveau de l'identification des parties du discours faite par ces auteurs et non réalisée par notre système. Ces auteurs utilisent un outil très répandu pour l'identification des parties du discours[11], qui ne peut pas être utilisé pour le DALG. Nous définissons donc un seul lexique regroupant les quatre parties grammaticales : adjectifs, verbes, noms et adverbes. Pour illustrer les résultats de l'AS du DALG, nous présentons les quatre parties : 1) l'environnement expérimental, 2) les résultats expérimentaux et leurs analyse, 3) l'analyse des cas d'erreurs, et 4) perspective d'extension de notre approche à l'ASM.

5.1. *Environnement expérimental*

Nous présentons dans cette section l'ensemble des données et des paramètres utilisés dans nos expérimentations : 1) les lexiques, 2) les corpus de test, et 3) les différents types d'expérimentations effectuées.

5.1.1. *Lexiques utilisés*

Pour la construction des lexiques, nous avons fait appel à deux lexiques anglais. D'une part, le SOCAL, qui est utilisé dans (Taboada *et al.*, 2011), et d'autre part, le SentiWordNet qui est utilisé dans (Baccianella *et al.*, 2010). Concernant SOCAL, nous

10. https ://github.com/sfu-discourse-lab/SO-CAL
11. Stanford CoreNLP : https ://stanfordnlp.github.io/CoreNLP/

avons d'abord fusionné les lexiques d'adjectifs, de verbes, de noms et d'adverbes. Nous avons ainsi obtenu 6 769 termes dont le sentiment est étiqueté entre -1 et -5 pour les termes négatifs et entre $+1$ et $+5$ pour les termes positifs. Après que l'intégralité des termes a été envoyée à l'API de traduction, 3 952 termes en anglais ont été reconnus et traduits. Notre lexique final, que nous nommons SOCALALG, contient 2 375 termes en DALG dont 1 363 termes négatifs, 948 termes positifs et 64 termes neutres (avec un sentiment égal à 0). Pour SentiWordNet, nous avons commencé par construire un lexique contenant l'ensemble des termes de SentiWordNet avec la moyenne du sentiment de chaque terme. Pour ce faire, nous avons fait appel à l'API JAVA de Petter Tönberg fourni dans le site officiel de SentiWordNet[12]. Comme les termes de SentiWordNet ont un sentiment étiqueté entre -1 et $+1$, nous avons multiplié tous les sentiments par 5 (pour l'aligner au lexique SOCAL). Le lexique obtenu contient 39 885 termes étiquetés entre $+0,05$ et $+5$ pour les termes positifs et entre $-0,05$ et -5 pour les termes négatifs. Une fois envoyés à l'API de traduction, 12 780 termes en anglais ont été reconnus et traduits. Notre lexique final que nous nommons SentiALG, contient 3 408 termes en DALG dont 1 856 négatifs, 1 539 positifs et 13 neutres.

5.1.2. *Corpus de test utilisés*

Nous avons utilisé dans ce travail deux corpus de test : 1) le premier corpus contient 323 messages (phrases) pris du corpus PADIC, dont 157 sont positifs et 166 négatifs, PADIC étant le seul corpus parallèle multidialectal, contenant également le DALG (Meftouh *et al.*, 2015), 2) le deuxième corpus contient 426 messages extraits du média social Facebook[13], dont 220 sont positifs et 206 sont négatifs. Les statistiques relatives à ces corpus sont présentées dans le tableau 5.1.2. Ces corpus ont été annotés par deux annotateurs natifs du dialecte algérien (un des auteurs de ce travail étant l'un des annotateurs. L'accord entre annotateurs (Kappa) est de 0,954 (0,956 pour le corpus PADIC et 0,952 pour le corpus Facebook). Les annotateurs ont reçu les instructions suivantes :

– les messages doivent être annotés en deux classes (positive ou négative). Les messages objectives ou neutres ne doivent pas être pris en considération ;

– les annotateurs ne doivent en aucun cas se référer à leur opinion personnelle mais plutôt au sentiment général de la phrase ;

– prendre en considération les exagérations et les émoticônes qui pourraient accentuer le sentiment ;

– dans le cas où un message contiendrait du texte et des émoticônes de valences différentes, il faudrait privilégier le sentiment porté par le texte pour l'annotation.

5.1.3. *Types d'expérimentations effectuées*

Nos expérimentations ont été effectuées sur 749 messages répartis en deux corpus et sur les deux lexiques SOCALALG et SentiALG. En plus de cela, nous avons mené

12. http ://sentiwordnet.isti.cnr.it/

13. https ://fr-fr.facebook.com/policy.php

Corpus	PADIC			Facebook		
	Pos.	Nég.	Tout	Pos.	Nég.	Tout
Nbre messages	157	166	323	220	206	426
Nbre mots	849	952	1 802	1 711	1 735	3 446
Nbre mots/message	5,41	5,73	5,57	7,78	8,42	8,1
Nbre caractères/message	21,9	24,0	23,0	33,3	35,9	34,6
Nbre messages avec émoticônse	0	0	0	38	19	57

Tableau 4. *Statistiques relatives aux corpus de test*

	Lexique utilisé	PADIC			Facebook		
		P	R	F1	P	R	F1
n-gramme (1)	SOCALALG	0,71	0,45	0,55	0,68	0,36	0,47
	SentiALG	0,73	0,45	0,56	0,67	0,38	0,48
n-gramme + prétraitement (2)	SOCALALG	0,72	0,46	0,56	0,74	0 ,42	0,53
	SentiALG	0,74	0,47	0,57	0,71	0,42	0,53
N-gramme + prétraitement + lemme (3)	SOCALALG	0,70	0,69	0,70	0,70	0,63	0,67
	SentiALG	0,75	0,74	0,74	0,69	0,63	0,66
n-gramme + prétraitement + lemme + passé (4)	SOCALALG	0,70	0,70	0,70	0,72	0,64	0,67
	SentiALG	0,75	0,74	0,74	0,69	0,64	0,66
n-gramme + prétraitement + lemme + passé + négation (5)	SOCALALG	0,75	0,74	0,75	0,68	0,61	0,64
	SentiALG	0,78	0,78	0,78	0,67	0,61	0,64

Tableau 5. *Résultats expérimentaux avec les deux lexiques utilisés*

pour chaque corpus avec chaque lexique cinq expérimentations : 1) n-gramme, 2) n-gramme + prétraitement, 3) n-gramme + prétraitement + lemme, 4) n-gramme + prétraitement + lemme + passé, et 5) n-gramme + prétraitement + lemme + passé + négation. Nous souhaitons montrer à travers ces expérimentations l'impact de chaque étape de notre approche sur les résultats obtenus.

5.2. *Résultats expérimentaux*

Nous illustrons nos résultats à l'aide de trois métriques : la précision (P), le rappel (R) et la F-mesure (F1). Nous présentons dans le tableau 5, les résultats obtenus (P, R et F1) sur les deux corpus de test utilisés (PADIC et Facebook) avec les deux lexiques SOCALALG et SentiALG.

D'après le tableau 5, nous constatons que les résultats évoluent positivement après l'exécution de chaque étape de notre approche et ce pour nos deux lexiques (sauf pour le traitement de la négation où il y a une légère régression des résultats). L'évolution la plus importante en termes de F1 a été observée au sein du corpus PADIC où les résultats initiaux étaient de 0,56 pour finir avec un score de 0,78.

5.3. *Analyse des cas d'erreurs*

Après une analyse approfondie des messages qui sont mal classés par notre système, nous présentons dans le tableau 6 les principales erreurs de classification du sentiment du DALG. D'après le tableau 6, nous constatons qu'il existe sept principales erreurs 1) les pluriels irréguliers, 2) les mots non existants dans nos lexiques, 3) les mots ayant une valence différente dans nos lexiques, 4) les mots en ASM, non utilisés en DALG, 5) les mots ayant une intensité trop élevée, 6) le non-traitement des intensificateurs, 7) les lemmes retrouvés dans le lexique par erreur. Les pluriels irréguliers signifient que la formation du pluriel de certains mots ne suit pas les règles que nous avons présentées dans la section 2.2.2.4. Prenons par exemple le mot مليح *mliyH* signifiant 'bien', son pluriel n'est pas مليحين *mliyHiyn* selon la règle, mais plutôt le mot ملاح *mlaAH*. Nous avons également constaté que plusieurs mots importants ne sont pas présents dans l'un de nos lexiques ou parfois même dans les deux. Parmi ces derniers, nous citons :كادو *kaAduw* 'un cadeau', زعف *zʕaf* 'Il est en colère', etc. D'autres mots tels que عجيب *ʕjiyb* 'bizarre' ont une valence différente d'un lexique à un autre. Certains mots tels que مطبوعة 'imprimé' *maTbuwʕaħ* ou encore احسن *AHsan* 'le meilleur' sont des mots en ASM et donc non reconnus en DALG. Certains mots tels que واش *waAš* 'Comment ?' ont une intensité trop élevée dans les lexiques, ce qui peut fausser le calcul du score final. D'autres mots comme بزاف *bzaAf* 'très', représentant des intensificateurs, ne sont pas reconnus comme tels. Leurs traitements amélioraient le score final. Enfin, les mots comme بكينا *bkiynaA* 'Nous avons pleuré' dont un lemme a été reconnu par erreur, avant même de faire appel au traitement du passé.

Afin de résoudre ces problématiques, nous proposons les améliorations suivantes de notre système :
– proposition d'un lemmatiseur propre au dialecte algérien dédié à la segmentation et à l'analyse des pluriels irréguliers ;
– extension de notre lexique pour qu'il ait un champ plus étendu. Nous prévoyons de faire cet extension à l'aide des techniques du « word Embedding » ;
– intégration du contexte dans l'annotation du lexique ;
– enrichissement de notre lexique en DALG avec un autre lexique en ASM.

Messages mal classés	Traduction	Cause principale de l'erreur	Annotation _cible	Annotation _système
الحمد لله انا نعرف غير ***الملاح*** AlHmd llh AnA nçrf γyr AlmlAH	Dieu soit loué je ne connais que des gens bien	Le mot 'الملاح' qui est le pluriel de 'مليح' n'a pas été reconnu	Positif	Négatif
هاديك لي جابتلي ***كادو*** في اللخر تاع العام HAdyk ly jAbtly kAdw fy Allxr tAç AlçAm	Celle qui m'a ramené un cadeau en fin d'année	Le mot 'كادو' n'existe pas dans nos lexiques	Positif	Négatif
قولك تقرا تلقى***يواش***	Qu'est-ce qu'on dit tu étudies tu trouveras	Le mot 'واش' ayant une intensité négative élevée	Positif	Négatif
القراية في وقتنا عادت ***بزاف*** صعيبة AlqrAyħ fy wqtnA çAdt bzAf Sçybħ	Les études en notre temps sont devenues très difficiles	Le non-traitement des intensificateurs tels que 'بزاف'	Négatif	Positif
بكينا! bkiynaA. !	On a pleuré !	Le mot 'كي' est retrouvé ne donnant pas la chance au mot 'بكى' d'être recherché	Négatif	Positif
الف مرة ***المطبوعة احسن*** AlmTbwçħ AHsn Alf mrħ	L'imprimerie est mieux mille fois.	Contient des mots en ASM qui n'ont pas été reconnus	Positif	Négatif
عجيب! çjiyb	Merveilleux	Le mot 'عجيب' ayant une valence positive dans un lexique et négative dans un autre	Positif	Négatif

Tableau 6. *Les principales erreurs de classification de notre système*

5.4. *Perspective d'extension de notre approche à l'ASM*

Afin de comparer notre approche avec les travaux existants, nous proposons l'extension de cette dernière à l'ASM. Pour ce faire, nous construisons en premier lieu nos lexiques : SOCAL_ASM et Senti_ASM en suivant la même méthode que celle présentée dans la section 4.1. Nous obtenons ainsi 5 190 termes pour SOCAL_ASM et 15 838 pour Senti_ASM. Afin de pouvoir appliquer notre approche sur les deux lexiques obtenus, nous utilisons le corpus ASM utilisé dans (Altowayan et Tao, 2016). Ce corpus contient 4 294 messages, dont 2 147 positifs et 2 147 négatifs. Il a été construit en combinant plusieurs autres corpus présentés au sein de la littérature. En ajoutant certains affixes à notre approche, dédiés à l'ASM, nous obtenons un F1-score égal à 0,58 pour Senti_ASM et 0,62 pour SOCAL_ASM. Nous constatons que la mise à l'échelle de notre approche sur l'ASM donne des résultats compétitifs dans le cadre d'une approche fondée sur les lexiques. Les résultats fournis par SOCAL_ASM (F1 égale à 0,62) ne s'éloignent pas trop des résultats que nous avons présentés dans le tableau 5 pour le corpus Facebook. Ceci est tout à fait justifié car le corpus ASM que nous avons utilisé est essentiellement alimenté avec des données provenant des médias sociaux.

6. Conclusion et perspectives

Dans cet article, nous avons proposé et implémenté une approche d'AS de messages écrits en DALG. Cette approche s'appuie sur la construction et l'utilisation de lexiques de sentiments en DALG. Elle est fondée également sur l'agglutination qui est une problématique très importante dans le traitement de l'ASM et de ses dialectes. Nous avons évalué cette approche à l'aide des deux lexiques de sentiments construits, SOCALALG et SentiALG, ainsi qu'un corpus de test annoté manuellement contenant 747 messages. Les résultats expérimentaux indiquent une amélioration continue après l'exécution de chaque étape de notre approche atteignant une précision de 0,78, un rappel de 0,78 et une F-mesure de 0,78. Ces résultats pourraient cependant être améliorés en prenant en considération plusieurs facteurs étudiés dans la section 5.3. Nos travaux futurs s'orientent vers une proposition d'intégrer l'ensemble des critères suivants :

– l'analyse des pluriels irréguliers et proposition d'une liste de changements et d'affixes qui pourraient traiter ces pluriels ;

– la fusion des deux lexiques SOCAL_ALG et Senti_ALG ainsi que la fusion (ASM et DALG), car plusieurs utilisateurs utilisent l'alternance codique entre l'ASM et le DALG. Il faudrait également procéder à l'intégration d'autres lexiques de sentiments anglais tels que MPQA ou SentiStrenght ;

– la définition d'une méthode combinant le traitement des lemmes et du passé en même temps ;

– le traitement de l'intensification ;

– la revue manuelle des lexiques utilisés pour pouvoir y intégrer la notion de contexte ainsi que l'enrichissement du lexique obtenu à l'aide des techniques de l'apprentissage profond (*deep learning*).

Enfin, nous comptons également étendre cette approche aux corpus annotés pour pouvoir exploiter les techniques de classification usuelles. Néanmoins ces approches requièrent des corpus annotés. La construction de ces corpus est très coûteuse en termes de temps et d'efforts. Nous prévoyons donc de proposer une approche de construction automatique ou semi-supervisée de ces corpus.

7. Remerciement

Les premiers auteurs sont soutenus par l'École nationale supérieure d'informatique (ESI) à Alger ainsi que l'École supérieure des sciences appliquées d'Alger (ESSAA). Le troisième auteur est soutenu par la DGE (ministère de l'Industrie de France) et par la DGE (ministère de l'Économie de France) : projet « DRIRS », référencé par le numéro 172906108. Nous tenons à remercier Billel Gueni pour sa collaboration et ses précieux retours.

8. Bibliographie

Abdul-Mageed M., Diab M., Kübler S., « SAMAR : Subjectivity and sentiment analysis for Arabic social media », *Computer Speech & Language*, vol. 28, n° 1, p. 20-37, 2014.

Abdulla N. A., Ahmed N. A., Shehab M. A., Al-Ayyoub M., Al-Kabi M. N., Al-rifai S., « Towards improving the lexicon-based approach for arabic sentiment analysis », *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 9, n° 3, p. 55-71, 2014a.

Abdulla N., Mohammed S., Al-Ayyoub M., Al-Kabi M. *et al.*, « Automatic lexicon construction for arabic sentiment analysis », *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, IEEE, p. 547-552, 2014b.

Al-Ayyoub M., Essa S. B., Alsmadi I., « Lexicon-based sentiment analysis of arabic tweets », *International Journal of Social Network Mining*, vol. 2, n° 2, p. 101-114, 2015.

AL-Khawaldeh F. T., « A Study of the Effect of Resolving Negation and Sentiment Analysis in Recognizing Text Entailment for Arabic. », *World of Computer Science & Information Technology Journal*, 2015.

Alhumoud S. O., Altuwaijri M. I., Albuhairi T. M., Alohaideb W. M., « Survey on arabic sentiment analysis in twitter », *International Science Index*, vol. 9, n° 1, p. 364-368, 2015.

Altowayan A. A., Tao L., « Word embeddings for Arabic sentiment analysis », *Big Data (Big Data), 2016 IEEE International Conference on*, IEEE, p. 3820-3825, 2016.

Assiri A., Emam A., Aldossari H., « Arabic sentiment analysis : a survey », *International Journal of Advanced Computer Science and Applications*, vol. 6, n° 12, p. 75-85, 2015.

Baccianella S., Esuli A., Sebastiani F., « Sentiwordnet 3.0 : an enhanced lexical resource for sentiment analysis and opinion mining. », *LREC*, vol. 10, p. 2200-2204, 2010.

Badaro G., Baly R., Hajj H., Habash N., El-Hajj W., « A large scale Arabic sentiment lexicon for Arabic opinion mining », *ANLP 2014*, 2014.

Biltawi M., Etaiwi W., Tedmori S., Hudaib A., Awajan A., « Sentiment classification techniques for Arabic language : A survey », *Information and Communication Systems (ICICS), 2016 7th International Conference on*, IEEE, p. 339-346, 2016a.

Biltawi M., Etaiwi W., Tedmori S., Hudaib A., Awajan A., « Sentiment classification techniques for Arabic language : A survey », *7th International Conference on Information and Communication Systems (ICICS),*, IEEE, p. 339-346, 2016b.

Cherif W., Madani A., Kissi M., « Towards an efficient opinion measurement in Arabic comments », *Procedia Computer Science*, vol. 73, p. 122-129, 2015a.

Cherif W., Madani A., Kissi M., « Towards an efficient opinion measurement in Arabic comments », *Procedia Computer Science*, vol. 73, p. 122-129, 2015b.

Diab M., Habash N., Rambow O., Altantawy M., Benajiba Y., « COLABA : Arabic dialect annotation and processing », 2010.

El-Halees A. *et al.*, « Arabic opinion mining using combined classification approach », 2011.

Elarnaoty M., AbdelRahman S., Fahmy A., « A machine learning approach for opinion holder extraction in Arabic language », *arXiv preprint arXiv :1206.1011*, 2012.

Fishman J. A., « Bilingualism with and without diglossia ; diglossia with and without bilingualism », *Journal of social issues*, vol. 23, n° 2, p. 29-38, 1967.

Guellil I., Azouaou F., « Arabic dialect identification with an unsupervised learning (based on a lexicon). application case : Algerian dialect », *Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 2016 IEEE Intl Conference on*, IEEE, p. 724-731, 2016.

Guellil I., Azouaou F., « ASDA : Analyseur Syntaxique du Dialecte Alg {\'e} rien dans un but d'analyse s {\'e} mantique », *arXiv preprint arXiv :1707.08998*, 2017.

Guellil I., Azouaou F., Abbas M., « Comparison between Neural and Statistical translation after transliteration of Algerian Arabic Dialect », *WiNLP : Women & Underrepresented Minorities in Natural Language Processing (co-located withACL 2017)*, p. 1-5, 2017a.

Guellil I., Azouaou F., Abbas M., Fatiha S., « Arabizi transliteration of Algerian Arabic dialect into Modern Standard Arabic », *Social MT 2017/First workshop on Social Media and User Generated Content Machine Translation*, p. 1-8, 2017b.

Guellil I., Boukhalfa K., « Social big data mining : A survey focused on opinion mining and sentiments analysis », *Programming and Systems (ISPS), 2015 12th International Symposium on*, IEEE, p. 1-10, 2015.

Habash N., Soudi A., Buckwalter T., « On arabic transliteration », *Arabic computational morphology*, Springer, p. 15-22, 2007.

Hadi W., « Classification of Arabic Social Media Data », *Advances in Computational Sciences and Technology*, vol. 8, n° 1, p. 29-34, 2015.

Harrag F., « Estimating the sentiment of arabic social media contents : A survey », *5th International Conference on Arabic Language Processing*, 2014.

Harrat S., Meftouh K., Abbas M., Hidouci W.-K., Smaili K., « An algerian dialect : Study and resources », *International journal of advanced computer science and applications (IJACSA)*, vol. 7, n° 3, p. 384-396, 2016.

Harrat S., Meftouh K., Abbas M., Smaili K., « Building resources for algerian arabic dialects », *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Harrat S., Meftouh K., Smaïli K., « Machine translation for Arabic dialects (survey) », *Information Processing & Management*, 2017.

Hedar A. R., Doss M., « Mining social networks arabic slang comments », *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2013.

Itani M. M., Zantout R. N., Hamandi L., Elkabani I., « Classifying sentiment in arabic social networks : Naive search versus naive bayes », *Advances in Computational Tools for Engineering Applications (ACTEA), 2012 2nd International Conference on*, IEEE, p. 192-197, 2012.

Joachims T., *Learning to classify text using support vector machines : Methods, theory and algorithms*, vol. 186, Kluwer Academic Publishers Norwell, 2002.

Kaseb G. S., Ahmed M. F., « Arabic Sentiment Analysis approaches : An analytical survey », 2016.

Khalifa K., Omar N., « A hybrid method using lexicon-based approach and naive Bayes classifier for Arabic opinion question answering », *Journal of Computer Science*, vol. 10, n° 10, p. 1961, 2014.

Khoja S., Garside R., « Stemming arabic text », *Lancaster, UK, Computing Department, Lancaster University*, 1999.

Korayem M., Crandall D., Abdul-Mageed M., « Subjectivity and sentiment analysis of arabic : A survey », *International Conference on Advanced Machine Learning Technologies and Applications*, Springer, p. 128-139, 2012.

Mataoui M., Zelmati O., Boumechache M., « A proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic », *Research in Computing Science*, vol. 110, p. 55-70, 2016.

Mdhaffar S., Bougares F., Esteve Y., Hadrich-Belguith L., « Sentiment Analysis of Tunisian Dialect : Linguistic Resources and Experiments », *WANLP 2017 (co-located with EACL 2017)*, 2017.

Medhat W., Hassan A., Korashy H., « Sentiment analysis algorithms and applications : A survey », *Ain Shams Engineering Journal*, vol. 5, n° 4, p. 1093-1113, 2014.

Meftouh K., Bouchemal N., Smaïli K., « A Study of a Non-Resourced Language : The Case of one of the Algerian Dialects », *The third International Workshop on Spoken Languages Technologies for Under-resourced Languages-SLTU'12*, 2012.

Meftouh K., Harrat S., Jamoussi S., Abbas M., Smaili K., « Machine translation experiments on padic : A parallel arabic dialect corpus », *The 29th Pacific Asia conference on language, information and computation*, 2015.

Mohammad S. M., Turney P. D., « Crowdsourcing a word–emotion association lexicon », *Computational Intelligence*, vol. 29, n° 3, p. 436-465, 2013.

Saâdane H., Le traitement automatique de l'arabe dialectalisé : aspects méthodologiques et algorithmiques, PhD thesis, Grenoble Alpes, 2015.

Saâdane H., Guidere M., Fluhr C., « La reconnaissance automatique des dialectes arabes à l'écrit », *colloque international «Quelle place pour la langue arabe aujourd'hui*, p. 18-20, 2013.

Saâdane H., Habash N., « A conventional orthography for Algerian Arabic », *Proceedings of the Second Workshop on Arabic Natural Language Processing*, p. 69-79, 2015.

Sadat F., Kazemi F., Farzindar A., « Automatic identification of arabic dialects in social media », *Proceedings of the first international workshop on Social media retrieval and analysis*, ACM, p. 35-40, 2014.

Shoufan A., Alameri S., « Natural language processing for dialectical Arabic : A Survey », *Proceedings of the Second Workshop on Arabic Natural Language Processing*, p. 36-48, 2015.

Siddiqui S., Monem A. A., Shaalan K., « Sentiment analysis in Arabic », *International Conference on Applications of Natural Language to Information Systems*, Springer, p. 409-414, 2016.

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.