Des représentations continues de mots pour l'analyse d'opinions en arabe : une étude qualitative

Amira Barhoumi[1, 2] Nathalie Camelin[1] Yannick Estève[1]
(1) LIUM, Le Mans, France - amira.barhoumi.etu@univ-lemans.fr , prenom.nom@univ-lemans.fr
(2) MIRACL, Sfax, Tunisie - amirabarhoumi29@gmail.com

RÉSUMÉ

Nous nous intéressons, dans cet article, à la détection d'opinions dans la langue arabe. Ces dernières années, l'utilisation de l'apprentissage profond a amélioré des performances de nombreux systèmes automatiques dans une grande variété de domaines (analyse d'images, reconnaissance de la parole, traduction automatique, ...) et également celui de l'analyse d'opinions en anglais. Ainsi, nous avons étudié l'apport de deux architectures (CNN et LSTM) dans notre cadre spécifique. Nous avons également testé et comparé plusieurs types de représentations continues de mots (*embeddings*) disponibles en langue arabe, qui ont permis d'obtenir de bons résultats. Nous avons analysé les erreurs de notre système et la pertinence de ces *embeddings*. Cette analyse mène à plusieurs perspectives intéressantes de travail, au sujet notamment de la constitution automatique de ressources expert et d'une construction pertinente des *embeddings* spécifiques à la tâche d'analyse d'opinions.

ABSTRACT

Word embeddings for Arabic sentiment analysis : a qualitative study

In this paper, we are interested in Arabic sentiment analysis task. Recently, the use of deep learning improves many automatic systems in a wide variety of fields (image analysis, speech recognition, machine translation, ...), among others English sentiment analysis. Thus, we study the performance of two architectures (CNN and LSTM) in our specific framework. In addition, we investigated the use of several types of word embeddings publically available for Arabic, that achieve good results. Finally, the analysis of the errors of our system and the relevance of the different embeddings was also proposed. These analysis lead to several interesting perspectives : building expert resources (lexicon) and relevant task-specific embeddings.

MOTS-CLÉS : Analyse d'opinion, représentation continue de mot, apprentissage profond, langue arabe.

KEYWORDS: Sentiment analysis, word embeddings, deep learning, arabic language.

1 Introduction

Avec la montée d'internet et la révolution des réseaux sociaux, un grand nombre d'individus peuvent exprimer leurs points de vue et leurs sentiments sur des entités, des produits, des personnes, *etc.* Dans ce contexte, le domaine de l'analyse automatique d'opinions connait un intérêt croissant de la part des entreprises et de la communauté scientifique [1]. Par ailleurs, les avancées scientifiques récentes dans les techniques d'apprentissage profond ainsi que la croissance des puissances de calcul, a mené à l'amélioration significative des performances dans différents domaines tels que la reconnaissance

1. https://trends.google.com/trends/explore?date=all&q=sentiment\%20analysis

de la parole ou la traduction automatique. La recherche en analyse d'opinions a également tiré profit de l'apprentissage profond, et plusieurs travaux ont été réalisés avec ce type d'apprentissage.

Dans cet article, nous nous focalisons sur la détection d'opinions par des méthodes à base de réseaux de neurones pour la langue arabe. Nous effectuerons nos experiences sur le corpus *Large-scale Arabic Book Review* (LABR) qui est un corpus de critiques de livres en langue arabe. Nous présentons en section 2 un état de l'art du domaine. Nous proposons ensuite, en section 3, nos deux systèmes neuronaux. Le premier s'appuie sur un réseau de neurones convolutifs CNN et le second sur un réseau neuronal réccurent de type *Long Short-Term Memory* LSTM. Nous étudions particulièrement l'utilisation de plusieurs types de représentations continues de mots disponibles pour la langue arabe (section 4). Nous analysons, en section 5, les erreurs de nos systèmes puis menons une analyse afin d'évaluer la pertinence des embeddings pour la tâche spécifique de détection d'opinions. Nous concluons et exposons les perspectives en section 6.

2 Etat de l'art

L'analyse d'opinions consiste à identifier la subjectivité et la polarité (positive, négative, neutre) d'un énoncé donné (Pang *et al.*, 2008). On peut l'appliquer au niveau du document, de la phrase ou d'un groupe de mots (Wilson *et al.*, 2004).

Les travaux effectués dans ce domaine peuvent être classés selon trois approches. La première est symbolique, elle utilise des lexiques et des règles linguistiques. La deuxième consiste en une approche statistique qui s'appuie sur des méthodes d'apprentissage automatique. Pour finir, il existe une approche hybride qui est une combinaison des deux précédentes : elle utilise à la fois des lexiques et des algorithmes d'apprentissage automatique. Jusqu'à récemment, les machines à vecteurs de supports SVM (Gaurangi *et al.*, 2014; Zainuddin & Selamat, 2014) et les classifieurs naïfs de Bayes NB (Tripathy *et al.*, 2015) représentaient les classifieurs les plus répandus dans ce domaine. Suivant la mouvance actuelle, les travaux récents font recours à l'apprentissage profond (Hassan, 2017; Deriu *et al.*, 2017; Zhou *et al.*, 2016).

Peu de travaux ont été réalisés pour l'analyse d'opinions en langue arabe. Ceci s'explique par le faible nombre de ressources développées et leur non disponibilité (Al-Kabi *et al.*, 2016). Nous citons quelques travaux existants selon leur catégorie. Suivant une approche linguistique, (Almas & Ahmad, 2007; Farra *et al.*, 2010) proposent une méthode s'appuyant sur un ensemble de patrons permettant d'extraire les polarités d'un document financier. Pour les travaux à base de lexiques, (Abdulla *et al.*, 2014a) construisent manuellement un lexique contenant 4815 mots. Leur système calcule le nombre de mots positifs et négatifs dans un texte afin de générer sa polarité globale. (Al-Kabi *et al.*, 2014) ont mis en place un outil qui détermine la subjectivité, la polarité d'une opinion et son intensité. Ils utilisent deux lexiques généraux et seize lexiques spécifiques. Suivant une approche statistique, (Abdulla *et al.*, 2014b) proposent un système de détection de subjectivité et de polarité dans les réseaux sociaux en utilisant des attributs morphologiques. (Bayoudhi *et al.*, 2015) comparent trois classifieurs : SVM, NB et un réseau de neurones simple. Pour finir, nous présentons les travaux à base de systèmes hybrides. (El-Halees, 2011) est le premier à avoir proposé un système hybride pour l'analyse d'opinions pour l'arabe. Il propose une hiérarchie séquentielle de classifications combinées. (Ibrahim *et al.*, 2015) utilise un lexique de 5244 adjectifs, un lexique de 3296 idiomes pour améliorer la classification de phrases avec un SVM. (Refaee & Rieser, 2016) appliquent une approche hybride pour la prédiction de l'intensité de la polarité dans les tweets. Ils ont utilisé particulièrement la régression logistique pour prédire les scores initiaux qui sont ajustés en appliquant des règles extraites à partir d'un lexique de polarité.

Plusieurs travaux récents appliquent des techniques d'apprentissage profond pour l'analyse d'opinion. (Barhoumi *et al.*, 2017) utilise les représentations continues de documents combinées avec un perceptron multicouche (PMC) tandis que (Dahou *et al.*, 2016) utilisent un CNN.

Nous détaillons dans la suite les différents systèmes que nous avons mis en place pour l'analyse d'opinions en arabe avec des méthodes d'apprentissage neuronal.

3 Systèmes d'analyse d'opinions pour l'arabe

Dans ce travail, nous nous intéressons à la classification selon leur polarité de critiques de livres en langue arabe. Nous avons implémenté deux systèmes : un CNN et un LSTM dont nous détaillons, dans la suite, les architectures. Nous décrivons également les différents types d'*embeddings* que nous avons utilisés.

3.1 Architectures à base de réseaux de neurones

Les réseaux convolutifs CNN ont prouvé leurs performances dans l'analyse d'opinions pour l'anglais (Kim, 2014). Nous avons donc choisi cette architecture pour implémenter notre premier système et évaluons ses performances pour l'arabe. Le CNN prend en entrée une matrice d'*embeddings* de taille fixe et applique une convolution de filtres, dont la taille de la fenêtre est une des valeurs de l'ensemble$\{3, 5, 7\}$, pour extraire de nouveaux attributs à partir de la matrice d'*embeddings*. Puis, un *max_pooling* est appliqué sur la sortie de la couche de convolution dans le but de conserver uniquement les attributs les plus pertinents qui sont concaténés au niveau d'une couche entièrement connectée. Enfin, le CNN applique la fonction *sigmoid* à la couche de sortie pour générer la polarité du document fourni en entrée. Deux polarités sont possibles : positif ou négatif (il s'agit d'une classification binaire).

Motivés par les bons résultats d'un système à base de réseaux LSTM pour l'anglais (Hassan, 2017), nous avons également décidé d'implémenter cette architecture. Il s'agit d'un cas particulier de réseaux de neurones récurrents (RNN) dont l'avantage principal est d'être composé d'unités neuronales appropriées pour permettre au réseau d'*oublier* ou de *mémoriser* : certaines observations du passé auront plus de poids que d'autres si elles sont jugées plus pertinentes pour la classification lors de l'apprentissage. Notre LSTM utilise comme entrée la même matrice d'*embeddings* que le CNN. Il est constitué d'une couche réccurrente de type LSTM unidirectionnelle simple connectée à une couche finale activée par une fonction *sigmoid*, pour générer la prédiction.

3.2 Représentations continues de mots arabes

Dans ce travail, nous avons utilisé deux ressources d'*embeddings* (disponibles gratuitement) comme entrée de nos systèmes neuronaux. La première ressource est celle de (Dahou *et al.*, 2016). Ils ont entrainé le modèle word2vec (Mikolov *et al.*, 2013) de type Skip-gram et *continuous bag of words* (CBOW) sur des pages web. Leurs expériences ont montré que CBOW est plus performant, ils l'ont donc mis à disposition. La deuxième ressource (Soliman *et al.*, 2017) est plus riche : elle regroupe six modèles d'*embeddings* entrainés sur trois types de corpus différents : twitter, wikipédia et des pages web. Ils ont entrainé CBOW et Skip-gram sur les trois types de corpus, mettant ainsi à disposition six

ensembles d'embeddings. Il est important de signaler que tous les embeddings disponibles sont de dimension 300.

4 Expériences

4.1 Corpus LABR

Pour évaluer nos systèmes, nous avons utilisé le corpus LABR (Nabil *et al.*, 2014) qui contient 63k critiques de livres composées d'un commentaire et d'une note associée (nombre d'étoiles). Nous nous plaçons dans le cadre d'une classification binaire et regroupons les critiques comme proposé dans (Nabil *et al.*, 2014) : les commentaires associés à une ou deux étoiles composent la classe *négative* et ceux à quatre ou cinq étoiles composent la classe *positive*. Ainsi les commentaires neutres ne sont pas considérés et le corpus utilisé se réduit à un ensemble de 40845 commentaires (68% positifs) pour le corpus d'apprentissage et 10211 pour le corpus de test (69% positifs). Notons que 10% de l'ensemble d'apprentissage est utilisé comme corpus de développement. Le corpus que nous utilisons est ainsi composé de 51k critiques, soit plus de trois millions de mots sur un vocabulaire de taille 324k. Pour mieux comprendre la distribution des mots, il est intéressant de connaitre les quelques statistiques suivantes : Le nombre d'occurences du mot le plus fréquent est de 76855 quand il est à 319 pour le 1000e mot le plus fréquent ; Si on considère qu'un mot peu fréquent est un mot qui apparait moins de 5 fois dans le corpus, on couvre alors 86,5% du corpus avec 13% du vocabulaire.

4.2 Comparaison des différents systèmes de détection d'opinions en arabe

Cette section expérimentale présente dans un premier temps les résultats récents des travaux déjà parus sur le corpus LABR. Nous notons que les meilleurs résultats ont été obtenus par (Dahou *et al.*, 2016) avec l'utilisation d'un CNN. Or, ces résultats n'ont pas été obtenus avec la répartition officielle du corpus. Le code des auteurs étant disponible, nous avons testé ce système sur la répartition officielle et avons obtenu 77,39% d'exactitude [2]. Le deuxième meilleur système est celui de (ElSahar & El-Beltagy, 2015). Les bonnes performances de ce système s'expliquent notamment par l'utilisation de connaissances de type expert *a priori* relatives à la polarité par le moyen de lexiques, malheureusement non disponibles. Nous comparons donc les résultats de nos systèmes à l'exactitude de (Dahou *et al.*, 2016) sur corpus officiel, qui correspond au meilleur résultat obtenu sans connaissances *a priori* (**soit une baseline à 77,39%**).

Notre premier système s'appuie sur une implémentation de CNN similaire à celle de (Dahou *et al.*, 2016). En plus des *embeddings* de (Dahou *et al.*, 2016), nous avons également testé les *embeddings* de (Soliman *et al.*, 2017) décrits dans la section 3.2. Notre second système s'appuie sur un LSTM et a été testé avec les différents *embeddings*.
Les performances de ces différentes combinaisons architecture/*embeddings* sont résumées dans la table 1. Elles nous permettent d'étudier de façon exploratoire l'impact de différentes constructions de représentations continues de mots sur la détection d'opinion. Nous notons ici que le CNN obtient de meilleurs résultats que le LSTM, et ce, quels que soient les *embeddings* utilisés. La meilleure performance est atteinte par un CNN appris sur les *embeddings* de (Soliman *et al.*, 2017) avec une approche CBOW appliquée sur un corpus issu du Web. Ce système noté *CNN_Soliman_CBOW_Web* sera analysé dans la section suivante.

2. En utilisant leurs partitions personnelles du corpus du LABR, nous retrouvons leurs résultats.

	(Dahou *et al.*, 2016)	(Soliman *et al.*, 2017)					
	Web	**Twitter**		**Wikipédia**		**Web**	
	CBOW	CBOW	Skip-gram	CBOW	Skip-gram	CBOW	Skip-gram
CNN	77,39%	77,41%	77,55%	77,51%	77,43%	77,56%	77,47%
LSTM	75,03%	74,87%	74,65%	74,92%	74,58%	74,74%	74,95%

TABLE 1 – Exactitudes des architectures CNN et LSTM sur LABR avec différents *embeddings*.

On remarque également que les *embeddings* obtiennent tous des résultats similaires malgré le fait que certains *embeddings* n'ont pas été appris avec la même approche ou le même type de corpus. Nous nous interrogeons ainsi sur la pertinence des représentations de mots disponibles pour la tâche spécifique de la détection d'opinions. Dans la section suivante, nous analysons dans un premier temps les erreurs de notre meilleur système puis proposons une première analyse des *embeddings* utilisés.

5 Analyse des résultats

5.1 Analyse des erreurs de prédiction

Nous avons calculé la matrice de confusion de notre meilleur système, *CNN_Soliman_CBOW_Web*. Le système prédit bien les commentaires positifs avec 80,34% de précision et 89,80% de rappel. Les exemples négatifs sont, quant à eux, plus difficiles à détecter avec 67,76% de précision et seulement 49,37% de rappel. Notre système montre donc une faiblesse dans la prédiction de la classe négative.

Pour analyser plus finement la composition des critiques, nous nous appuyons sur les mots issus du lexique *LABR_lex* de (ElSahar & El-Beltagy, 2015) qui regroupe 873 expressions [3] dont la polarité est connue. On dit que ce sont des mots *polarisés*. Les mots de ce lexique constituent 2,4% des occurrences de mots contenus dans les critiques positives ou négatives du corpus LABR. La majorité ($\geq$ 1,6%) de ces mots sont des mots positifs. La difficulté de classification des critiques négatives peut donc être due à l'utilisation de figures de styles comme l'humour ou l'ironie qui implique qu'une expression positive est utilisée alors que le sens se veut négatif. Une autre explication à l'apparition de ces mots positifs dans une critique négative est qu'ils sont utilisés en conjonction avec un terme de négation. Nous avons par exemple remarqué que parmi les vingt mots les plus fréquents, trois étaient des termes de négation. Nous pensons également que la difficulté de classification des critiques négatives peut être fortement liée à la pertinence des *embeddings* d'entrée pour la tâche donnée. Nous proposons dans la section suivante un protocole d'analyse afin d'étudier cette hypothèse.

5.2 Analyse des *embeddings*

Dans un premier temps, nous proposons de calculer la couverture des mots du corpus LABR par les projections existantes dans l'un des 7 espaces d'*embeddings* considérés. Pour ce faire, nous avons considéré d'une part tous les mots puis d'autre part les mots les plus fréquents (nombre d'occurrences >5), et calculé les couvertures d'une part sur le vocabulaire du corpus LABR (Table 3) et sur le corpus lui-même (Table 2).

Nous remarquons que la couverture du corpus par les différents espaces d'*embeddings* se situe aux alentours de 60% quels que soient l'espace considéré. La couverture augmente de six à huit points si on ne considère que les mots fréquents. Au niveau du vocabulaire, plus de 55% des mots fréquents

3. Une expression dans le lexique peut être constituée d'un ou plusieurs mots.

corpus LABR	(Dahou *et al.*, 2016)	(Soliman *et al.*, 2017)					
	Web	**Twitter**		**Wikipédia**		**Web**	
	CBOW	CBOW	Skip-Gram	CBOW	Skip-Gram	CBOW	Skip-Gram
tous	67,33%	60.27%	60.87%	61.53%	61.53%	60.35%	60.19%
occur > 5	71,07%	66.04%	66.32%	68.06%	68.06%	66.23%	66.07%

TABLE 2 – Couverture du corpus LABR par les différents modèles d'embeddings.

vocabulaire LABR	(Dahou *et al.*, 2016)	(Soliman *et al.*, 2017)					
	Web	**Twitter**		**Wikipédia**		**Web**	
	CBOW	CBOW	Skip-Gram	CBOW	Skip-Gram	CBOW	Skip-Gram
tous	40,97%	22.11%	24.33%	19.16%	19.16%	21.45%	21.30%
occur > 5	64,89%	57.00%	58.44%	53.48%	53.48%	57.76%	57.38%

TABLE 3 – Couverture du vocabulaire de LABR par les différents modèles d'embeddings.

sont couverts alors que la couverture du vocabulaire chute à 20% si on considère tous les mots. Ceci indique que la grande majorité des mots du corpus LABR n'ayant pas d'*embeddings* dans les modèles disponibles sont des mots peu fréquents. Ainsi, bien que la couverture ne soit pas très grande elle semble suffisante pour la classification.

Dans un second temps, afin d'évaluer la pertinence dans le cadre spécifique de la tâche d'analyse d'opinions des représentations de mots dans un espace continu, nous proposons d'étudier la polarité des mots voisins, en considérant leur *embeddings* dans chacun des espaces, pout les mots polarisés. Pour chaque expression, son ensemble des n plus proches mots polarisés voisins (Top$_n$) dans l'espace d'embeddings, est considéré selon la similarité cosinus. Nous calculons alors un ratio de *positivité* des mots de polarisés associés à une polarité positive ($lexique^+$) (voir équation 1).

$$\%^+_{Top_n} = 100 \times \frac{\sum_{mot_i \in \{lexique^+\}} \#mot^{lexique^+}_{i,Top_n}}{n \times \#lexique^+} \quad (1)$$

avec : n le nombre de mots voisins considérés ; $\#mot^{lexique^+}_{i,Top_n}$ le nombre de mots positifs parmi les n plus proches voisins du mot i du corpus $lexique^+$; $\#lexique^+$ le nombre de mots positifs dans *lexique*.

Nous calculons également un ratio de *négativité* selon la même formule en ne considérant que les mots négatifs. Nous considérons qu'une représentation pertinente des mots dans un espace continu pour la tâche de détection d'opinions projetterait les mots positifs dans la même zone et les mots négatifs dans une autre zone. On observerait alors un ratio proche de 100%.

La Table 4 montre les résultats du ratio de positivité calculé sur le lexique *LABR_lex*. Nous constatons que plus le voisinage considéré est large, plus le ratio de positivité est grand. Ceci signifie que les mots positifs sont de plus en plus entourés par des mots positifs du lexique. En revanche, pour le ratio de négativité, calculé également à l'aide du lexique *LABR_lex*, nous constatons que plus le voisinage est large, moins le mot négatif est entouré de mots négatifs. Etant donné que seuls les mots polarisés sont considérés, ceci signifie que les mots négatifs sont de plus en plus entourés par des mots

		(Dahou *et al.*, 2016)	(Soliman *et al.*, 2017)					
		Web	**Twitter**		**Wikipédia**		**Web**	
		CBOW	CBOW	Skip-G	CBOW	Skip-G	CBOW	Skip-G
$\%^{+}_{Top_n}$	n=2	43,13	41,79	41,48	38,83	38,83	42,59	41,15
	n=5	68,10↗	63,28 ↗	64,88 ↗	58,75 ↗	58,21 ↗	66,66 ↗	63,23 ↗
	n=10	73,46↗	68,43 ↗	69,92 ↗	63,39 ↗	62,14 ↗	72,07 ↗	70 ↗
$\#LABR\_lex^{+}$		153	134	135	112	112	135	130
$\%^{-}_{Top_n}$	n=2	34,88	38,75	42,42	31,77%	37,85%	39,09%	40,83%
	n=5	13,95↘	15,5↘	16,96↘	12,71↘	15,14↘	15,63↘	16,33↘
	n=10	6,97↘	7,75↘	8,84↘	6,35↘	7,57↘	7,81↘	8,16↘
$\#LABR\_lex^{-}$		172	160	165	107	107	133	131

TABLE 4 – Ratios de *positivité* (respectivement *négativité*) des mots positifs (respectivement négatifs) dont l'embedding existe à la fois dans *LABR_lex* et le corpus d'*embeddings* considéré.

positifs du lexique. Ces observations se vérifient pour les différents espaces d'*embeddings*. La polarité négative semble donc diffusée dans l'espace de représentations utilisé. Ceci appuie notre hypothèse d'un espace continu non adapté au cadre de la détection d'opinions, notamment pour représenter les mots négatifs. Cette observation explique les mauvais résultats en classification d'opinions des commentaires négatifs.

6 Conclusion et perspectives

Dans cet article, nous avons étudié l'utilisation de techniques d'apprentissage profond dans le cadre de l'analyse d'opinion pour l'arabe en étudiant sept ensembles d'*embbeddings* différents comme entrées de réseaux CNN et LSTM.

Nos expériences ont montré que l'architecture CNN est plus performante que l'architecture LSTM, quelque soit le modèle d'*embeddings* utilisé. Notre meilleur système ($CNN\_Soliman\_CBOW\_Web$) obtient une exactitude de 77,56% améliorant légèrement le meilleur système publié qui n'utilise pas de connaissances *a priori* (77,39% pour (Dahou *et al.*, 2016) appliqué sur la répartition officielle).

Nous proposons trois pistes d'amélioration de ces premiers travaux : (i) utilisation de formes fléchies des mots. En effet, plus de 80% des mots sont peu fréquents, une lemmatisation permettrait d'éviter la dispersion du vocabulaire ; (ii) création automatique de lexiques de mots polarisés. Les meilleurs résultats ont été obtenus avec des connaissances *a priori* coûteuses à obtenir. Nous souhaitons étudier la traduction de ressources existantes afin de créer un ensemble de connaissances *a priori* pour l'arabe ; (iii) création d'*embeddings* spécifiques. Notre analyse des *embeddings* génériques disponibles a montré que ceux-ci n'étaient pas forcément pertinents pour notre tâche. En nous appuyant sur les travaux de (Yu *et al.*, 2017) où des *embeddings d'opinions* sont construits pour l'anglais, nous souhaitons étudier la transposition de ces travaux pour l'arabe en nous appuyant sur les lexiques de mots polarisés que nous aurons construits.

Références

ABDULLA N. A., AHMED N. A., SHEHAB M. A., AL-AYYOUB M., AL-KABI M. N. & AL-RIFAI S. (2014a). Towards improving the lexicon-based approach for arabic sentiment analysis. *International Journal of Information Technology and Web Engineering (IJITWE)*, **9**(3), 55–71.

ABDULLA N. A., AL-AYYOUB M. & AL-KABI M. N. (2014b). An extended analytical study of arabic sentiments. *International Journal of Big Data Intelligence 1*, **1**(1-2), 103–113.

AL-KABI M., AL-AYYOUB M., ALSMADI I. & WAHSHEH H. (2016). A prototype for a standard arabic sentiment analysis corpus. *Int. Arab J. Inf. Technol.*, **13**(1A), 163–170.

AL-KABI M. N., GIGIEH A. H., ALSMADI I. M., WAHSHEH H. A. & HAIDAR M. M. (2014). Opinion mining and analysis for arabic language. *International Journal of Advanced Computer Science and Applications (IJACSA), SAI Publisher*, **5**(5), 181–195.

ALMAS Y. & AHMAD K. (2007). A note on extracting 'sentiments' in financial news in english, arabic & urdu. In *The Second Workshop on Computational Approaches to Arabic Script-based Languages*, p. 1–12.

BARHOUMI A., ESTÈVE Y., ALOULOU C. & BELGUITH L. H. (2017). Document embeddings for arabic sentiment analysis. In *Proceedings of the First Conference on Language Processing and Knowledge Management, LPKM 2017, Kerkennah (Sfax), Tunisia, September 8-10, 2017.*

BAYOUDHI A., GHORBEL H. & BELGUITH L. H. (2015). Sentiment classification of arabic documents : Experiments with multi-type features and ensemble algorithms. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, p. 196–205.

DAHOU A., XIONG S., ZHOU J., HADDOUD M. H. & DUAN P. (2016). Word embeddings and convolutional neural network for arabic sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 2418–2427.

DERIU J., LUCCHI A., DE LUCA V., SEVERYN A., MÜLLER S., CIELIEBAK M., HOFMANN T. & JAGGI M. (2017). Leveraging large amounts of weakly supervised data for multi-language sentiment classification. *International World Wide Web Conference Committee (IW3C2).*

EL-HALEES A. (2011). Arabic opinion mining using combined. *Proceeding the International Arab Conference On Information Technology*.

ELSAHAR H. & EL-BELTAGY S. R. (2015). Building large arabic multi-domain resources for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, p. 23–34 : Springer.

FARRA N., CHALLITA E., ASSI R. A. & HAJJ H. (2010). Sentence-level and document-level sentiment mining for arabic texts. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, p. 1114–1119 : IEEE.

GAURANGI P., VARSHA G., VEDANT K. & KALPANA D. (2014). Sentiment analysis using support vector machine. *International Journal of Innovative Research in Computer and Communication Engineering*.

HASSAN A. (2017). Sentiment analysis with recurrent neural network and unsupervised neural language model.

IBRAHIM H. S., ABDOU S. M. & GHEITH M. (2015). Sentiment analysis for modern standard arabic and colloquial. *International Journal on Natural Language Computing (IJNLC)*, **4**(2).

KIM Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv :1408.5882*.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, p. 3111–3119.

NABIL M., ALY M. & ATIYA A. (2014). Labr : A large scale arabic sentiment analysis benchmark. *arXiv preprint arXiv :1411.6718*.

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

REFAEE E. & RIESER V. (2016). ilab-edinburgh at semeval-2016 task 7 : A hybrid approach for determining sentiment intensity of arabic twitter phrases. *Proceedings of SemEval-2016*, p. 474–480.

SOLIMAN A. B., EISSA K. & EL-BELTAGY S. R. (2017). Aravec : A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, **117**, 256–265.

TRIPATHY A., AGRAWAL A. & RATH S. K. (2015). Classification of sentimental reviews using machine learning techniques. *3rd International Conference on Recent Trends in Computing (ICRTC-2015)*, p. 821—829.

WILSON T., WIEBE J. & RWA R. (2004). Just how mad are you ? finding strong and weak opinion clauses. *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence*, p. 761–769.

YU L.-C., WANG J., LAI K. R. & ZHANG X. (2017). Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 534–539.

ZAINUDDIN N. & SELAMAT A. (2014). Sentiment analysis using support vector machine. *International Conference on Computer, Communications, and Control Technology (I4CT)*.

ZHOU P., QI Z., ZHENG S., XU J., BAO H. & XU B. (2016). Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. *arXiv preprint arXiv :1611.06639*.