A Characterization Study of Arabic Twitter Data with a Benchmarking for State-of-the-Art Opinion Mining Models

Ramy Baly,[1] **Gilbert Badaro,**[1] **Georges El-Khoury,**[1] **Rawan Moukalled,**[1] **Rita Aoun,**[1]
Hazem Hajj,[1] **Wassim El-Hajj,**[2] **Nizar Habash,**[3] **Khaled Bashir Shaban**[4]
[1] Department of Electrical and Computer Engineering, American University of Beirut
[2] Department of Computer Science, American University of Beirut
[3] Computational Approaches to Modeling Language Lab, New York University Abu Dhabi
[4] Department of Computer Science, Qatar University
{rgb15,ggb05,gbe03,rrm32,rra47}@mail.aub.edu
{hh63,we07}@aub.edu.lb, nizar.habash@nyu.edu
khaled.shaban@qu.edu.qa

Abstract

Opinion mining in Arabic is a challenging task given the rich morphology of the language. The task becomes more challenging when it is applied to Twitter data, which contains additional sources of noise, such as the use of unstandardized dialectal variations, the non-conformation to grammatical rules, the use of Arabizi and code-switching, and the use of non-text objects such as images and URLs to express opinion. In this paper, we perform an analytical study to observe how such linguistic phenomena vary across different Arab regions. This study of Arabic Twitter characterization aims at providing better understanding of Arabic Tweets, and fostering advanced research on the topic. Furthermore, we explore the performance of the two schools of machine learning on Arabic Twitter, namely the feature engineering approach and the deep learning approach. We consider models that have achieved state-of-the-art performance for opinion mining in English. Results highlight the advantages of using deep learning-based models, and confirm the importance of using morphological abstractions to address Arabic's complex morphology.

1 Introduction

Opinion mining, or sentiment analysis, aims at automatically extract subjectivity information from text (Turney, 2002) whether at sentence or document level (Farra et al., 2010). This task has attracted a lot of researchers in the last decade due to the wide range of real world applications that are interested in harvesting public opinion in different domains such as politics, stock markets and marketing.

Huge amounts of opinion data are generated, on a daily basis, in many forums, personal blogs and social networking websites. In particular, Twitter is one of the most used social media platforms, where users generally express their opinions on everything from music to movies to politics and all sort of trending topics (Sareah, 2015). Furthermore, Arabic language is the 5^{th} most-spoken language worldwide (UNESCO, 2014), and has recently become a key source of the Internet content with a 6,600% growth in number of users compared to the year 2000 (Stats, 2016). Therefore, developing accurate opinion mining models for Arabic tweets is a timely and intriguing problem that is worth investigating.

However, applying Natural Language Processing (NLP) and learning opinion models for Arabic Twitter data is not straightforward due to several reasons. Tweets contain large variations of unstandardized dialectal Arabic (DA), in addition to significant amounts of misspellings and grammatical errors, mainly due to their length restriction. They also contain "Arabizi", where Arabic words are written using Latin characters. Due to the cultural diversity across the Arab world, an opinion model that is developed for tweets in one region may not be applicable to extract opinions from tweets in another region. Finally, tweets usually contain special tokens such as hashtags, mentions, multimedia objects and URLs that need to be han-

Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 110–118,
Valencia, Spain, April 3, 2017.

dled appropriately, in order to make use of the subjective information they may implicitly carry.

In this paper, we present a characterization study of Twitter data collected from different Arab regions, namely Egypt, the Levant and the Arab Gulf. This study illustrates how the discussed topics, the writing style and other linguistic phenomena, vary significantly from one region to another, reflecting different usages of Twitter around the Arab world. We also evaluate the model that ranked first at SemEval-2016 Task 4 (Nakov et al., 2016) on "Sentiment Analysis in Twitter". This model is developed for opinion mining in English, and uses feature engineering to extract surface, syntactic, semantic and Twitter-specific features. Therefore, we extract an equivalent feature set for Arabic to train a model for opinion mining in Arabic tweets. We compare this model to another class of models that are based on deep learning techniques. In particular, we use recursive deep models that achieved high performances (Socher et al., 2013; Tai et al., 2015). Experimental results show the advantage of deep learning at learning subjectivity in Arabic tweets without the need for artificial features that describe the properties and characteristics of Twitter data.

The rest of this paper is organized as follows. Section 2 describes previous work on opinion mining with particular focus on application to Twitter data. Section 3 presents the characterization study and highlights distinctive characteristics of tweets collected from different Arab regions. Section 4 describes the opinion models that we evaluate in this paper, and experimental results are presented in Section 5. Conclusion is provided in Section 6.

2 Related Work

Opinion Mining models for Arabic are generally developed by training machine learning classifiers using different types of features. The most common features are the word *n*grams features that were used to train Support Vector Machines (SVM) (Rushdi-Saleh et al., 2011; Aly and Atiya, ; Shoukry and Rafea, 2012), Naïve Bayes (Mountassir et al., 2012; Elawady et al., 2014) and ensemble classifiers (Omar et al., 2013). Word *n*grams were also used along with syntactic features (root and part-of-speech *n*-grams) and stylistic (letter and digit *n*grams, word length, etc.). These features performed well after reduction via the Entropy-Weighted Genetic Algorithm (EWGA) (Abbasi et al., 2008). Sentiment lexicons also provided an additional source of features that proved useful for the task (Abdul-Mageed et al., 2011; Badaro et al., 2014; Badaro et al., 2015).

Many efforts have been made to develop opinion models for Arabic Twitter data and creating annotated Twitter corpora (Al Zaatari et al., 2016). A framework was developed to handle tweets containing Modern Standard Arabic (MSA), Jordanian dialects, Arabizi and emoticons, by training different classifiers under different features settings of such linguistic phenomena (Duwairi et al., 2014). A distant-based approach showed improvement over existing fully-supervised models for subjectivity classification (Refaee and Rieser, 2014a). A subjectivity and sentiment analysis (SSA) system for Arabic tweets used a feature set that includes different forms of the word (lexemes and lemmas), POS tags, presence of polar adjectives, writing style (MSA or DA), and genre-specific features including the user's gender and ID (Abdul-Mageed et al., 2014). Machine translation was used to apply existing state-of-the-art models for English to translations of Arabic tweets. Despite slight accuracy drop caused by translation errors, these models are still considered efficient and effective, especially for low-resource languages (Refaee and Rieser, 2014b; Salameh et al., 2015).

A new class of machine learning models based on deep learning have recently emerged. These models achieved high performances in both Arabic and English, such as the Recursive Auto Encoders (RAE) (Socher et al., 2011; Al Sallab et al., 2015), the Recursive Neural Tensor Networks (RNTN) (Socher et al., 2013) and Generalized Regression Neural Networks (GRNN) (Hobeica et al., 2011; Baly et al., 2016).

Finally, we describe models that won SemEval-2016 on "Sentiment Analysis in Tweets" in English (Nakov et al., 2016). For three-way classification, the winner model is based on training two Convolutional Neural Networks, and combining their outputs (Deriu et al., 2016). These networks share similar architectures but differ in the choice of some parameters, such as the embeddings and the number of convolution filters. As for five-way classification, the winner model uses feature engineering. It extracts a collection of surface, syntactic, semantic and genre-specific features to train a SVM classifier.

3 Arabic Tweets Characterization

Developing opinion mining models requires understanding the different characteristics of the texts that they will be applied to. For instance, dealing with reviews requires different features and methods compared to comments or tweets, as each of these types have their own characteristics. Furthermore, when dealing with Arabic data, it is important to appreciate the rich cultural and linguistic diversity across the Arab region, which may translate into different challenges that need to be addressed during model development. First, we describe the general features of Twitter data, and then we present an analysis of three sets of tweets collected from main Arab regions: Egypt, the Arab Gulf and the Levant.

Twitter is a micro-blogging website where people share messages that have a maximum length of 140 characters. Despite their small size, the tweets' contents are quite diverse and can be made up of text, emoticons, URLs, pictures and videos that are internally mapped into automatically shortened URLs, as in Table 1, example (a). Users tend to use informal styles of writing to reduce the length of the text while it can still be interpreted by others. Consequently, Twitter data become noisy as they contain significant amounts of misspellings, and do not necessarily follow the grammatical structure of the language, as shown in Table 1, example (b). Arabizi and code-switching are frequently used and observed in tweets, as shown in Table 1, example (c). Hashtags are very common and are used to highlight keywords, to track trending topics or events, to promote products and services, and for other personal purposes including fun and sarcasm. Also, "user mentions" are quite common and have different usages including tagging users in tweets to start a conversation, replying to someone's tweet and giving credit for some media or source. Table 1, example (d) shows how hashtags and mentions are used in Tweets. Finally, users can react to a tweet in three different ways, either using (1) "Like" by pressing the heart button, (2) "Reply" by mentioning the author and typing their comment in a new tweet, or (3) "Re-Tweet" by sharing it to their own followers.

We manually analyzed three sets of tweets that were retrieved from Egypt, the Arab Gulf and the Levant, using the Twitter4J API (Yamamoto, 2014). We refer to these sets of tweets as "EGY", "GULF" and "LEV", respectively, where each set contains 610 tweets. Examples of tweets written in each of the region's dialect are shown in Table 1, examples (e,f and g). We did not use a specific query as a keyword, in order to retrieve tweets covering the different topics being discussed in each region. We also did not use the API's language filter, in order to retrieve tweets that may be written in Arabizi. For each set, one of the authors analyzed the used languages, the discussed topics and the presence of emoticons, sarcasm, hashtags, mentions and elongation.

Table 2 shows the distribution of the different topics in each set. Table 3 shows the different writing styles and languages that are used in each set. Table 4 illustrates, for each set, the percentage of tweets that contain special Twitter tokens.

	EGY	LEV	GULF
Religion	20.0%	22.3%	32.1%
Personal	35.1%	58.9%	50.5%
Politics	3.6%	5.3%	4.4%
Sports	0.3%	6.9%	1.3%
Other news	2.9%	1.6%	1.9%
Spam	8.5%	3.4%	5.6%
Foreign	29.5%	1.6%	4.1%

Table 2: Topics discussed in each set.

	LEV	EGY	GULF
MSA	28.5%	40.7%	55.7%
Local dialect	18.4%	31.5%	28.5%
Arabizi	0.7%	1.9%	0.0%
English	13.4%	7.2%	4.1%
Foreign	31.8%	1.6%	4.4%
N/A	7.2%	7.1%	7.2%

Table 3: Languages and writing styles in each set.

Special tokens	EGY	LEV	GULF
User mentions	17.1%	31.6%	34.6%
Hashtags	7.5%	13.4%	13.7%
Emoticons	20.3%	30.9%	25.6%
Elongation	2.6%	8.2%	3.3%

Table 4: Use of special Twitter tokens in each set.

It can be observed that most of the tweets in "GULF" are written in MSA, and to a less extent using the local dialect. Compared to the other

(a)	☹ ☹ ☹ https://t.co/aszVLSZIpx
(b)	توصيات سينمائية اللّي ماتابع هذا المسلسل فاته دراما واكشن موطبيعي التقيـيم 9.0/10 *twSyAt synmA}yp Ally mA tAbE h∗A Almslsl fAth drAmA wAk$n mwTbyEy Altqyym 9.0/10* 'cinematic recommendations who did not follow this series has missed unreal drama and action assessment 9.0/10'
(c)	*mat2lysh alkalm dah 5lyna saktin* (example of Egyptian dialect Arabizi) 'don't say such a thing let's keep quiet'
(d)	@drkh189 @nogah015 @Almogaz #كفى كل اللّي حوالينا حروب بالوكالة *@drkh189 @nogah015 @Almogaz kl Ally HwAlynA Hrwb bAlwkAlp #kfY* '@drkh189 @nogah015 @Almogaz all what's happening around us are proxy wars #enough'
(e)	هو في حد لسه بيجيب فاكهة و حاجات كتير بالآسعار دي؟ (example of tweet in Egyptian Arabic) *hw fy Hd lsh byjyb fAkhp w HAjAt ktyr bAl>sEAr dy?* 'is there still anybody who brings fruits and many other stuff with these prices?'
(f)	@Mnallhfc علمي علمش بس شكله معروف بس انا وياش الّي ما نعرفه (example of tweet in Arab Gulf dialect) *Elmy Elm$ bs $klh mErwf bs AnA wyA$ Al∼y mA nErfh* 'I know the same as you know, but it seems he is known but we don't know him'
(g)	ومش مخلّين تويت مش عاملين في منشن عن القضية (example of tweet in Levantine) *wm$ mxl∼yn twyt m$ EAmlyn mn$n En AlqDyp* 'and they haven't left a tweet without a mention of the case'

Table 1: Samples of tweets, with their English translations and transliterations[2], highlighting the different linguistic phenomena that can be observed in Twitter data.

sets, a significant amount of these tweets discuss religious topics. It can also be observed that Arabizi and code switching do not appear, and that tweets written in English are rare. Regarding the "EGY" set, MSA is less common compared to "GULF", and a significant number of tweets are written using Egyptian Arabic. Most of the tweets discuss personal matters (nearly 59%). Also, Arabizi and code switching are rarely used. Finally, emoticons and user mentions are more frequently used compared to the other sets. As for the "LEV" set, it can be observed that both MSA and DA are used less than the other sets. Most of these tweets discuss personal matter, while religious topics are less discussed. A significant portion of the tweets are written in English, and many are written in foreign languages that pertain to neighboring countries (e.g., Cyprus and Turkey). Finally, it can be observed that elongation (letter repetition) is not common in the collected sets of tweets, and that Arabizi and code switching are infrequent as well.

This analysis confirms that Twitter is used differently (different characteristics, features and topics), across the Arab world. This implies that different opinion models are needed to account for the peculiarities of each region's tweets.

4 Opinion Mining Models

In this section, we describe two models that achieved state-of-the-art performances in opinion mining. The first model won the SemEval-2016 Task 4 on "Sentiment Analysis in Twitter" (English), and uses feature engineering to train an opinion classifier (Balikas and Amini, 2016). The second model is based on modeling compositionality using deep learning techniques (Socher et al., 2013). In this paper, we evaluate these models for opinion mining in Arabic tweets.

4.1 Opinion Mining with Feature Engineering

According to (Nakov et al., 2016; Balikas and Amini, 2016), training a SVM with a collection of surface, syntactic, semantic features achieved state-of-the-art results on opinion mining in English tweets. Below, we describe the equivalent

set of features that we extracted to train a similar model for opinion mining in Arabic tweets.

- Character *n*-grams; $n \in [3, 5]$.
- Word *n*-grams; $n \in [1, 4]$. To account for the complexity and sparsity of Arabic language, we extract lemma *n*-grams since lemmas have better generalization capabilities than raw words (Habash, 2010).
- Counts of exclamation marks, question marks, and both exclamation and question marks.
- Count of elongated words.
- Count of negated contexts, defined by phrases that occur between a negation particle and the next punctuation.
- Counts of positive emoticons and negative emoticons, in addition to a binary feature indicating if emoticons exist in a given tweet.
- Counts of each part-of-speech (POS) tag in the tweet.
- Counts of positive and negative words based on ArSenL (Badaro et al., 2014), AraSenti (Al-Twairesh et al., 2016) and ADHL (Mohammad et al., 2016) lexicons.

We also add to this set the two binary features indicating the presence of user mentions and URL or media content. Many of these features align with the factors that we single out in the characterization study presented in Section 3.

4.2 Opinion Mining with Recursive Neural Networks

Most deep learning models for opinion mining are based on the concept of compositionality, where the meaning of a text can be described as a function of the meanings of its parts and the rules by which they are combined (Mitchell and Lapata, 2010). In particular, the Recursive Neural Tensor Networks (RNTN) model has proven successful for opinion mining in English (Socher et al., 2013). Figure 1 illustrates the application of a RNTN to predict the sentiment of a three-word sentence $\{C_1, C_2, C_3\}$, where words are represented with vectors that capture distributional syntactic and semantic properties (Bengio et al., 2003; Collobert and Weston, 2008; Mikolov et al., 2013). Each sentence is represented in the form of a binary parse tree. Then, at each node of the tree, a tensor-based composition function combines the child nodes' vectors (e.g., C_1, C_2) and produces the parent node's vector (e.g., P_1). This process repeats recursively until it derives a vector for each node C_i in the tree, including the root node that corresponds to the whole sentence. These vectors are then used to train a softmax classifier to predict the opinion distribution $y^{C_i} \in \mathbb{R}^K$ for the text represented by the i^{th} node, where K is the number of opinion classes. Further details are available in (Socher et al., 2013).

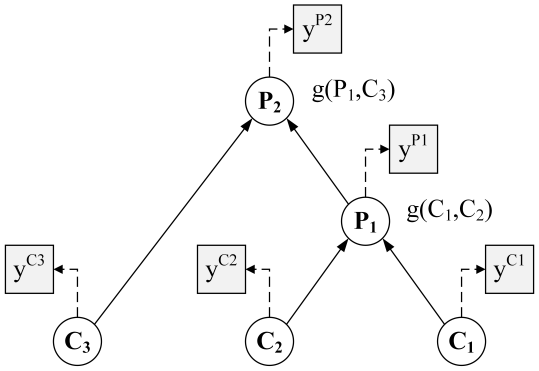

Figure 1: The application of RNTN for opinion prediction in a three-word sentence.

Training a RNTN model requires a sentiment treebank; a collection of parse trees with sentiment annotations at all levels of constituency. For English, the Stanford sentiment treebank was developed to train the RNTN (Socher et al., 2013). For Arabic, we developed the Arabic Sentiment Treebank (ArSenTB) by annotating ∼123K constituents pertaining to 1,177 comments extracted from the Qatar Arabic Language Bank (QALB) (Zaghouani et al., 2014).

5 Experiments and Results

In this section, we evaluate the performance of the feature engineering and deep learning-based models for opinion mining in Arabic tweets. We focus on the task of three-way opinion classification, where each tweet should be classified as positive, negative or neutral.

5.1 Dataset and Preprocessing

For our experiments, we use the Arabic Sentiment Twitter Data (ASTD) (Nabil et al., 2015) that consists of 10,006 tweets belonging to Egyptian Twit-

ter accounts. These tweets are annotated with four labels: positive (799), negative (1,684), neutral (832) and objective (6,691). Due to the highly skewed distribution of the classes, and since our focus is to perform opinion classification rather than subjectivity classification, we excluded the objective tweets, reducing the size of the data to 3,315 tweets with reasonable class distribution: 24% (positive), 51% (negative) and 25% (neutral). This data is split into a train set (70%), a development set (10%) and a test set (20%).

Each tweet is preprocessed by (1) replacing user mentions and URLs with special "global" tokens, (2) extracting emoticons and emojis using the "emoji" java library (Vdurmont, 2016) and replacing them with special tokens (for this we used the emojis sentiment lexicon from (Novak et al., 2015), and prepared our own emoticons lexicon), (3) normalizing hashtags by removing the "#" symbol and the underscores that are used to separate words in composite hashtags, and (4) normalizing word elongations (letter repetitions).

To extract features for the SVM classifier, we performed lemmatization and POS tagging using MADAMIRA v2.1, the state-of-the-art morphological analyzer and disambiguator in Arabic (Pasha et al., 2014), that uses the Standard Arabic Morphological Analyzer (SAMA) (Maamouri et al., 2010). Since the evaluation corpus is in Egyptian Arabic, we used MADAMIRA in the Egyptian mode. It is worth noting that some recent efforts have added Levantine to MADAMIRA, but it is not public yet (Eskander et al., 2016).

5.2 Experimental Setting

We only included *n*-grams that occurred more than a pre-defined threshold t, where $t \in [3, 5]$. Preliminary experiments showed that using the radial basis function (RBF) kernel is better than using the linear kernel. We used the development set to tune the model's parameters, namely the cost of misclassification and γ the width of the kernel, Then, the model with the parameters that achieved the best results is applied to the unseen test set.

As for the RNTN model, we generated word embeddings of size 25 by training the skip-gram embedding model (Mikolov et al., 2013) on the QALB corpus, which contains nearly 500K sentences. We train RNTN using ArSenTB, and then apply the trained model to perform opinion classification in tweets. We alleviate the impact of sparsity by training RNTN using lemmas, which is similar to our choice of training SVM using lemma *n*-grams.

Finally, the different models are evaluated using accuracy and the F1-score averaged across the different classes.

5.3 Results

Table 5 illustrates the performances achieved with the state-of-the-art models in feature engineering ($\text{SVM}_{all,lemmas}$) and deep learning (RNTN_{lemmas}). We compare to the following baselines: (1) the majority baseline that automatically assigns the most frequent class in the train set, and (2) the SVM trained with word *n*-grams ($\text{SVM}_{baseline}$), which has been a common approach in the Arabic opinion mining literature. To emphasize the impact of lemmatization, we include the results of SVM trained with features from (Balikas and Amini, 2016) and using word instead of lemma *n*-grams ($\text{SVM}_{all,words}$). We also include the results of RNTN trained with raw words (RNTN_{words}).

	Accuracy	**Average F1**
Majority	51.0%	22.5%
$\text{SVM}_{baseline}$	55.7%	29.0%
$\text{SVM}_{all,words}$	49.5%	41.6%
$\text{SVM}_{all,lemmas}$	51.7%	43.4%
RNTN_{words}	56.2%	51.1%
RNTN_{lemmas}	58.5%	53.6%

Table 5: Performance of the different models for opinion mining, evaluated on the ASTD dataset.

Results in Table 5 show that augmenting SVM with the different features from (Balikas and Amini, 2016) achieved significant performance improvement compared to the baseline SVM. It can also be observed that using the lemma feature to represent raw words contributes to this high performance, and confirms the importance of lemmas at reducing the lexical sparsity of Arabic language. Finally, the RNTN achieves best performance although it was trained on a dataset that is different from the tweets that are used for testing. We expect the performance of RNTN to further increase when it is trained on Twitter data. These results confirm the advantage of recursive deep learning that model semantic compositionality, over models that rely on feature engineering.

6 Conclusion

In this paper, we described the main challenges of processing Arabic language in Twitter data. We presented a characterization study that analyzes tweets collected from different Arab regions including Egypt, the Arab Gulf and the Levant. We showed that Twitter have different usages across these regions.

We report the performance of two state-of-the-art models for opinion mining. Experimental results indicate the advantage of using deep learning models over feature engineering models, as the RNTN achieved better performances although it was trained using a non-Twitter corpus. Results also indicate the importance of lemmatization at handling the complexity and lexical sparsity of Arabic language.

Future work will include evaluating opinion mining models on tweets from different Arab regions and covering different topics. Also, we intend to apply an automatic approach for analyzing tweet characteristics instead of the manual approach. We will exploit existing tools and resources for automatic identification of dialects in tweets.

We aim to perform cross-region evaluations to confirm whether different opinion models are needed for different regions and dialects, or a general model can work for any tweet regardless of its origins. This effort involves the collection and annotation of Twitter corpora for the different regions analyzed above.

Acknowledgments

This work was made possible by NPRP 6-716-1-138 grant from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12.

Muhammad Abdul-Mageed, Mona T. Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 587–591. Association for Computational Linguistics.

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. Samar: Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, 28(1):20–37.

Ahmad A. Al Sallab, Ramy Baly, Gilbert Badaro, Hazem Hajj, Wassim El Hajj, and Khaled B. Shaban. 2015. Deep learning models for sentiment analysis in arabic. In *ANLP Workshop 2015*, page 9.

Nora Al-Twairesh, Hend Al-Khalifa, and AbdulMalik Al-Salman. 2016. Arasenti: Large-scale twitter-specific arabic sentiment lexicons. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 697–705.

Ayman Al Zaatari, Reem El Ballouli, Shady ELbassouni, Wassim El-Hajj, Hazem Hajj, Khaled Bashir Shaban, and Nizar Habash. 2016. Arabic corpora for credibility analysis. *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 4396–4401.

Mohamed A. Aly and Amir F Atiya. Labr: A large scale arabic book reviews dataset.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale arabic sentiment lexicon for arabic opinion mining. *ANLP 2014*, 165.

Gilbert Badaro, Ramy Baly, Rana Akel, Linda Fayad, Jeffrey Khairallah, Hazem Hajj, Wassim El-Hajj, and Khaled Bashir Shaban. 2015. A light lexicon-based mobile application for sentiment mining of arabic tweets. In *ANLP Workshop 2015*, page 18.

Georgios Balikas and Massih-Reza Amini. 2016. Twise at semeval-2016 task 4: Twitter sentiment classification. *arXiv preprint arXiv:1606.04351*.

Ramy Baly, Roula Hobeica, Hazem Hajj, Wassim El-Hajj, Khaled Bashir Shaban, and Ahmad Al-Sallab. 2016. A meta-framework for modeling the human reading process in sentiment analysis. *ACM Transactions on Information Systems (TOIS)*, 35(1):7.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Jan Deriu, Maurice Gonzenbach, Fatih Uzdilli, Aurelien Lucchi, Valeria De Luca, and Martin Jaggi.

2016. Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. *Proceedings of SemEval*, pages 1124–1128.

RM Duwairi, Raed Marji, Narmeen Sha'ban, and Sally Rushaidat. 2014. Sentiment analysis in arabic tweets. In *Information and communication systems (icics), 2014 5th international conference on*, pages 1–6. IEEE.

Rasheed M. Elawady, Sherif Barakat, and Nora M. Elrashidy. 2014. Different feature selection for sentiment classification. *International Journal of Information Science and Intelligent System*, 3(1):137–150.

Ramy Eskander, Nizar Habash, Owen Rambow, and Arfath Pasha. 2016. Creating resources for dialectal arabic from a single annotation: A case study on egyptian and levantine. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3455–3465, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Noura Farra, Elie Challita, Rawad Abou Assi, and Hazem Hajj. 2010. Sentence-level and document-level sentiment mining for arabic texts. In *2010 IEEE International Conference on Data Mining Workshops*, pages 1114–1119. IEEE.

Nizar Y. Habash. 2010. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Roula Hobeica, Hazem Hajj, and Wassim El Hajj. 2011. Machine reading for notion-based sentiment mining. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 75–80. IEEE.

Mohamed Maamouri, Dave Graff, Basma Bouziri, Sondos Krouna, Ann Bies, and Seth Kulick. 2010. Standard arabic morphological analyzer (sama) version 3.1. *Linguistic Data Consortium, Catalog No.: LDC2010L01*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *J. Artif. Intell. Res.(JAIR)*, 55:95–130.

Asmaa Mountassir, Houda Benbrahim, and Ilham Berrada. 2012. An empirical study to address the problem of unbalanced data sets in sentiment classification. In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, pages 3298–3303. IEEE.

Mahmoud Nabil, Mohamed A. Aly, and Amir F. Atiya. 2015. Astd: Arabic sentiment tweets dataset. In *EMNLP*, pages 2515–2519.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. *Proceedings of SemEval*, pages 1–18.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one*, 10(12):e0144296.

Nazlia Omar, Mohammed Albared, Adel Qasem Al-Shabi, and Tareq Al-Moslmi. 2013. Ensemble of classification algorithms for subjectivity and sentiment analysis of arabic customers' reviews. *International Journal of Advancements in Computing Technology*, 5(14):77.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T. Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC*, volume 14, pages 1094–1101.

Eshrag Refaee and Verena Rieser. 2014a. Can we read emotions from a smiley face? emoticon-based distant supervision for subjectivity and sentiment analysis of arabic twitter feeds. In *5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data, LREC*.

Eshrag Refaee and Verena Rieser. 2014b. Subjectivity and sentiment analysis of arabic twitter feeds with limited resources. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, page 16.

Mohammed Rushdi-Saleh, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, and José M. Perea-Ortega. 2011. Oca: Opinion corpus for arabic. *Journal of the American Society for Information Science and Technology*, 62(10):2045–2054.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on arabic social media posts. In *HLT-NAACL*, pages 767–777.

Faiza Sareah. 2015. Interesting statistics for the top 10 social media sites. *Small Business Trends*.

Amira Shoukry and Ahmed Rafea. 2012. Sentence-level arabic sentiment analysis. In *Collaboration Technologies and Systems (CTS), 2012 International Conference on*, pages 546–550. IEEE.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the*

conference on empirical methods in natural language processing, pages 151–161. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Internet World Stats. 2016. Internet world users by language. http://www.internetworldstats.com/stats7.htm.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

UNESCO. 2014. World arabic language day. http://bit.ly/2lwRFYt.

Vdurmont. 2016. The missing emoji library for java. https://github.com/vdurmont/emoji-java.

Yusuke Yamamoto. 2014. Twitter4j-a java library for the twitter api.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large scale arabic error annotation: Guidelines and framework. In *LREC*, pages 2362–2369.