Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 117 (2017) 266–273

Procedia Computer Science

www.elsevier.com/locate/procedia

ELSEVIER

CrossMark

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5–6 November 2017, Dubai, United Arab Emirates

Comparative Evaluation of Sentiment Analysis Methods Across Arabic Dialects

Ramy Baly[a], Georges El-Khoury[a], Rawan Moukalled[a], Rita Aoun[a], Hazem Hajj[a], Khaled Bashir Shaban[b], Wassim El-Hajj[c]

[a]*Electrical and Computer Engineering Department, American University of Beirut, Beirut, Lebanon*
[b]*Computer Science and Engineering Department, Qatar University, Doha, Qatar*
[c]*Computer Science Department, American University of Beirut, Beirut, Lebanon*

Abstract

Sentiment analysis in Arabic is challenging due to the complex morphology of the language. The task becomes more challenging when considering Twitter data that contain significant amounts of noise such as the use of Arabizi, code-switching and different dialects that varies significantly across the Arab world, the use of non-textual objects to express sentiments, and the frequent occurrence of misspellings and grammatical mistakes. Modeling sentiment in Twitter should become easier when we understand the characteristics of Twitter data and how its usage varies from one Arab region to another. We describe our effort to create the first Multi-Dialect Arabic Sentiment Twitter Dataset (MD-ArSenTD) that is composed of tweets collected from 12 Arab countries, annotated for sentiment and dialect. We use this dataset to analyze tweets collected from Egypt and the United Arab Emirates (UAE), with the aim of discovering distinctive features that may facilitate sentiment analysis. We also perform a comparative evaluation of different sentiment models on Egyptian and UAE tweets. These models are based on feature engineering and deep learning, and have already achieved state-of-the-art accuracies in English sentiment analysis. Results indicate the superior performance of deep learning models, the importance of morphological features in Arabic NLP, and that handling dialectal Arabic leads to different outcomes depending on the country from which the tweets are collected.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

1. Introduction

Sentiment analysis is the automatic extraction of opinions from words, sentences or documents [28, 26]. It has attracted a lot of attention because of the wide range of applications that can benefit from harvesting the public opinion, and the huge amounts of opinionated data that are available online [37]. In particular, Twitter stands as one of the most used social media platforms, with around 500 million tweets being sent out daily, expressing opinions about personal or trending topics [48, 44]. Tweets are written in all languages including Arabic, which ranks 4th as both the most used language on Twitter [49, 1] and the most spoken language worldwide [35], hence becoming a key source

E-mail addresses: rgb15@mail.aub.edu (Ramy Baly)., gbe03@mail.aub.edu (Georges El-Khoury)., rrm32@mail.aub.edu (Rawan Moukalled)., rra47@mail.aub.edu (Rita Aoun)., hh63@aub.edu.lb (Hazem Hajj)., khaled.shaban@qu.edu.qa (Khaled Bashir Shaban)., we07@aub.edu.lb (Wassim El-Hajj).

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2017.10.118

of the Internet content. Given these facts, improving the performance of sentiment analysis models in Arabic tweets is a timely and intriguing problem.

Sentiment analysis generally involves natural language processing (NLP) and machine learning to model text semantics. These tasks are challenging when applied to Arabic tweets for several reasons. First, users tend to tweet using unstandardized Arabic dialects that vary significantly across the Arab world, or using Arabizi; writing Arabic words using Latin characters. They tend to commit misspellings to abide by the allowed space of 140 characters per tweet, and also use special tokens such as hashtags, mentions and URLs, which may express sentiments implicitly. Second, due to the cultural diversity in the Arab world, sentiment models trained on tweets from one region may not applicable to another region. For instance, the phrase *SbHAn Allh* ‘God is perfect’ is usually used in Levant countries to express positive sentiment, while it is used in Gulf countries to praise God with no sentiment implication. Finally, efforts to create sentiment twitter datasets were limited to specific dialects, mainly Egyptian [38], Gulf [11] and Jordanian [23]. Hence, more corpora are needed to support comparative evaluations of sentiment models across the different Arabic dialects.

In this paper, we describe our efforts to create the Multi-Dialect Arabic Sentiment Twitter Dataset (MD-ArSenTD), which is composed of tweets retrieved from 12 Arab countries. Tweets are assigned sentiment labels according to a 5-point scale to incorporate information of both polarity and intensity. While previous corpora assumed that tweets collected from a specific country are written using that country’s dialect [38, 23], MD-ArSenTD contains country and region-level dialect annotation, since we observed that tweets can be written using either Modern Standard Arabic (MSA), the country’s dialect, the user’s dialect, or a foreign language. This dataset will help the Arabic NLP research community understand the specificities of Arabic tweets by providing insights into Twitter’s topics, dialects and writing styles in the different Arab countries. We focus on characterizing tweets from Egypt (from the Nile Basin) and the United Arab Emirates (from the Arabian Gulf). We highlight the differences in the dialects, the discussed topics and the expressed sentiments between tweets of both countries. We also present a comparative evaluation of advanced sentiment models that belong to two classes of machine learning. The first is based on training Support Vector Machines (SVM) using an engineered feature set tailored for Twitter data [22, 14], and the second is based on training deep learning models, namely the Long Short-term Memory networks (LSTM), using generic and dialect-specific embeddings. Results indicate the superiority of deep learning, the importance of using morphological features, and that accounting for the local dialect leads to different outcomes depending on the country from which the tweets are collected. To our knowledge, this is the first attempt to perform sentiment analysis in UAE tweets, and also the first attempt to perform sentiment analysis in Egyptian tweets on a 5-point scale.

The rest of the paper is organized as follows. Section 2 describes previous work on Twitter sentiment analysis. Section 3 describes the system to create the MD-ArSenTD, and presents a characterization study of tweets from Egypt and UAE. Section 4 describes the models used in the comparative evaluation and presents the results along with results analysis. Concluding remarks are provided in Section 5.

2. Related Work

Sentiment analysis in Arabic is usually achieved by training classifiers using different sets of ‘engineered’ features. The most common features are the word *n*-grams; simple but semantically shallow features that are used to train SVM [42, 10, 45], Naïve Bayes (NB) [32, 24] and ensemble classifiers [34]. Word *n*-grams were also used with syntactic features (root and part-of-speech (POS) tag *n*-grams) and stylistic features (letter and digit *n*-grams, word and document lengths and vocabulary richness), achieving good accuracies after reduction via the Entropy-Weighted Genetic Algorithm (EWGA) [2]. Sentiment lexica provided an additional source of deeper semantic features, which helped boosting the accuracy. Several lexica were developed for MSA (such as ArabSenti [5], ArSenL [12] and SLSA [25]) and for dialectal Arabic (such as SANA [4] and AraSenTi [8]).

Efforts were made to perform sentiment analysis in Arabic Twitter and to create Twitter corpora [9]. A framework was developed to handle Jordanian tweets by training different classifiers using features that capture the semantics of MSA, Jordanian dialect, Arabizi and emoticons [23]. An emoticon-based distant supervision approach improved sentiment classification compared to fully supervised models [39]. A subjectivity and sentiment analysis system for Arabic social media (SAMAR) exploited Arabic morphology using features such as stems, lemmas and POS tag *n*-grams. It also used features such as the presence of polar adjectives, dialect, user ID and gender [3]. Machine translation was also used to apply state-of-the-art models in English to automatic translations of Arabic tweets. Despite slight accuracy drop due to translation errors, these models can be considered efficient and effective, especially for low-resource languages [40, 43]. The winner of SemEval-2017 Task 4 on topic-based sentiment analysis task in Arabic Twitter [41] is based on first predicting the topic of upcoming tweets and then predicting their sentiment using topic-specific sentiment classifiers [15]. Furthermore, using features proposed by the winner of SemEval-2016 Task on sentiment analysis in English Twitter [22] has also achieved solid performance in Arabic.

Recently, deep learning has emerged based on advances in neural networks. In NLP, deep models are trained using embedding vectors that capture syntactic and semantic properties of the words [18, 30]. They proved successful in many NLP tasks, achieving state-of-the-art accuracies in sentiment analysis in English. Examples of such models include Recursive Auto Encoders (RAE) [46], Recursive Neural Tensor Networks (RNTN) [47], Long Short-term Memory (LSTM) networks [50], Gated Recurrent Neural Networks (GRNN) [51, 17] and Dynamic Memory Networks (DMN) [27]. A few efforts were made to explore deep learning for sentiment analysis in Arabic. Preliminary results indicated that RAE and LSTM performed better than baseline SVMs [7, 14]. Improving several aspects of RAE including (1) input using sentiment embeddings and (2) text representation through morphological tokenization, improved the performance drastically [6]. Further improvements were achieved by exploring the space of morphology and orthography. Training RNTN with stems, automatic diacritization and marked letter repetition achieved best results on ArSenTB; the first Arabic sentiment treebank [16].

Although a lot of research was done on Arabic sentiment analysis, most of these efforts did not account for the specificities of Twitter data. There is also a lack in understanding how the usage and the characteristics of Twitter differ from one country to another, which is valuable information to decide whether to rely on generic models or to develop models that are tailored to specific regions.

3. MD-ArSenTD: The Multi-Dialect Arabic Sentiment Twitter Dataset

We present the approach to create MD-ArSenTD and a characterization of Egyptian and UAE tweets to understand Twitter's usage, structure and distinctive features in both countries. The MD-ArSenTD was created as follows:

1. **Tweets Retrieval:** The Twitter4J API [52] was used to collect 470K tweets, starting from 3/1/2017 until 4/30/2017. Specific geo-locations were used to force retrieving tweets from 12 Arab countries in four regions; the Arabian Gulf, the Levant, Egypt and North Africa, following the classification in [53]. Table 1 shows the number of tweets that were retrieved for each country. It also ranks the countries according to Twitter popularity by dividing the number of streamed tweets in each country by the population of that country. This rank helps indicating the extent to which Twitter is used to share news and opinions, in each country. It can be observed that Twitter is mostly used in the Gulf region and least used in North Africa countries.

2. **Tweets Selection:** We set the target size of the MD-ArSenTD to 14,400 tweets, hence we need to select and annotate 1,200 tweets for each country. Duplicates and tweets with less than 30 characters are first removed. Then, the pre-trained sentiment model [15] that won SemEval-2017 task 4 [41] was applied to the remaining tweets. For each country, we selected the top 1,200 tweets that were predicted as positive, negative and neutral with highest confidence. This step decreases the likelihood of selecting tweets that are mostly neutral, and hence irrelevant to the task of sentiment analysis.

3. **Sentiment and Dialect Annotation:** We used CrowdFlower to annotate the selected tweets for both sentiment and dialect. For dialect annotation, annotators were instructed to identify the dialect's country and region, otherwise to select either the MSA or the Foreign language options. When annotating tweets from a specific region, only annotators from that region were allowed to participate in the task, to ensure quality annotation. For sentiment annotation, annotators were asked to assign a sentiment polarity on a 5-point scale: very negative, negative, neutral, positive and very positive. They were provided with detailed guidelines, and their performance was monitored using a gold set of 100 test tweets (per country), annotated by the authors. Only those who maintained a performance accuracy above 75% were allowed to remain on the task. Each tweet was randomly assigned to 3-4 annotators, and its final annotation was based on majority voting. In case of ties, labels assigned by annotators with higher accuracy are assigned higher weights.

Table 1: Number of streamed tweets per country starting from March 1, 2017 until April 30, 2017.

Region	Country	# of Streamed Tweets	Tweet per capita
Gulf	Kuwait	33,175	2nd
	KSA	176,155	3rd
	Qatar	31,458	1st
	UAE	49,639	4th
Levant	Jordan	17,813	7th
	Lebanon	18,563	6th
	Palestine	18,756	5th
	Syria	15,896	9th
North Africa	Algeria	21,151	11th
	Morocco	11,536	12th
	Tunisia	14,885	8th
Egypt	Egypt	60,148	10th

In this paper, we are interested in performing comparative analyzes and evaluations between tweets from Egypt and the UAE. First, we evaluate the annotation quality for tweets of both countries by 1) sampling 100 tweets from each country, and then 2) calculating Cohen's Kappa κ between the authors' (gold) annotation and the result of aggregating the CrowdFlower annotations of these tweets to measure the proportion of agreement above what would be expected by chance [21]. We obtained a kappa equals to 0.8 for region-level dialect annotation and 0.65 for sentiment annotation. The later increased to 0.8 when neglecting sentiment intensity, which indicates that a significant amount of disagreement lies in the sentiment intensity, and is not a fundamental confusion in polarity.

During the characterization and analysis of the 2,400 Egyptian and UAE tweets, we focused on highlighting the differences in the topics being discussed, the dialects being used and the sentiment distribution between tweets of both countries. Table 2 shows that the majority of Egyptian tweets discuss personal matters (e.g., love, mournings and blessings), and to a less extent religious matters (stating verses from the Quran or praising God تَسابيح *tsAbyH*). On the other hand, religion is the most discussed topic in UAE tweets, followed by personal matters (e.g., wishing well and replies to other tweets). It can also be observed that the topics discussed in Egyptian tweets are more versatile than those discussed in UAE tweets. For instance, we did not encounter, in our UAE sample, tweets discussing business or reviewing products and services. The only topic that was found in UAE and not in Egyptian tweets is the "check-in", where users tend to tweet about locations (restaurants, hotels, venues) they are visiting.

Table 2: Topic distribution

Topic	Egypt tweets	UAE tweets
Business	1.7%	–
Multimedia	9.5%	–
Personal	50.8%	40.4%
Politics	8.3%	11.7%
Product Reviews	3.1%	–
Religion	24.1%	43.0%
Sports	2.5%	3.2%
Check-ins	–	1.8%

This difference in topic distribution can be associated with a difference in the languages that are used to compose the tweets in each country, as shown in Table 3. It can be observed that the majority of Egyptian tweets are written in Egyptian dialect, as they mainly discuss personal matters that can be expressed in everyday language. On the other hand, the majority of UAE tweets are written in MSA, as they mainly discuss religious matters and contain Quranic verses and worship prayers that are written in MSA. In contrary to Egyptian tweets, there exist UAE tweets that are written in Foreign languages. This can be due to the high number of foreigners living in the UAE that has become a major international business hub. It can also be observed that, in both countries, only a small fraction of tweets are written using neither the country's local dialect nor the MSA. This suggests that it can be safe to assume that tweets in a country are generally written using either its local dialect or MSA to different extents, depending on the country and the topics being discussed.

Finally, Table 4 shows how sentiment is distributed across tweets of each country. Sentiment in Egyptian tweets is normally-distributed, with most tweets being neutral and very few have high sentiment intensities. On the other hand, UAE tweets generally express positive sentiment, with a significant amount of tweets praising God or wishing well and greeting such as صبَاح الخير... جزَاك الله خيرًا 'Good morning... God bless you.'

Table 3: Dialect Distribution

Dialect	Egypt tweets	UAE tweets
Egyptian	67.4%	2.1%
MSA	31.4%	60.5%
Gulf	0.8%	29.4%
Levant	0.3%	0.5%
North Africa	0.1%	0%
Foreign	0%	7.5%

Table 4: Sentiment Distribution

	Egypt tweets	UAE tweets
Very negative	1.41%	1.21%
Negative	20.88%	19.08%
Neutral	57.42%	34.60%
Positive	18.64%	42.20%
Very positive	1.65%	2.91%

4. Experiments and Results

In this section, we describe the machine learning models that are evaluated for sentiment analysis in Arabic Twitter. In particular, we focus on models that stem from two schools of machine learning; feature engineering and deep learning. Then, we describe details related to data preprocessing, model implementation and training. Finally, we present and discuss the experimental results, and provide useful insights.

4.1. Sentiment Models

The winner of SemEval-2016 Task 4 on Sentiment analysis in English Twitter [33] trained SVM with a collection of hand-crafted features covering surface, syntactic and semantic information [13]. We evaluated an equivalent model that was accustomed to Arabic by extracting the following features, as proposed by [14].

- Lemma n-grams; $n \in [1, 4]$ and character n-grams; $n \in [3, 5]$.
- Counts of exclamation marks, question marks, both exclamation and question marks and elongated words.
- Count of negated contexts, defined by phrases that occur between a negation particle and the next punctuation.
- Counts of pos/neg emoticons, and of pos/neg words based on ArSenL [12], AraSenti [8] and ADHL [31]
- Counts of each part-of-speech (POS) tag in the tweet.
- Presence of emoticons, user mentions and URL or media content.

We also evaluated LSTM; a class of deep recurrent neural networks (RNNs) that achieved high accuracies in English sentiment analysis [50]. The LSTM goes through the tweet's words sequentially. At each step, it combines the current word with the output representation of preceding words to produce an updated output using a cell composed of input, output, forget and memory gates that aim to capture long and short-term relations between the current word and previous ones. The output that is produced after going through the whole tweet is then used to train a softmax layer for sentiment classification. To train LSTM to classify variable-length tweets, all tweets are padded with zeros so they all become of equal lengths.

The input features to LSTM are pre-trained word embeddings generated using the skip-gram model from word2vec [30]. To evaluate the impact of dialects on sentiment analysis, we trained two types of embeddings: generic and dialect-specific. Generic embeddings are learned from all unlabeled 469K tweets that were retrieved to create the MD-ArSenTD, in addition to the newswire data used to create the Arabic Treebank [29], which is written in MSA. On the other hand, the dialect-specific embeddings are learned from the portion of tweets that pertain to the country under consideration. We account for the morphological complexity of Arabic by training LSTM under different morphological forms, namely stems and lemmas, which proved to improve performance [16, 3]. For this purpose, we created lemma and stem embeddings by training the SKIP-GRAM model on lemmatized and stemmed versions of the unlabeled tweets, generated using MADAMIRA [36].

4.2. Experiments and Results

The Egyptian and UAE tweets are preprocessed to improve the quality of the input text. We performed tokenization to separate digits and punctuation from words. We also applied normalization to 1) repeated characters (word elongation), 2) emoticons by replacing them with global happy/sad tokens, 3) parentheses by replacing their different forms with the square brackets.

We used libSVM [19] to train and evaluate SVM, and Keras [20] with TensorFlow libraries to train and evaluate the LSTM. Each set of tweets is split into training (80%) and testing (20%) sets using stratified sampling to maintain similar sentiment distributions in both subsets. Hyper-parameters are tuned by performing 10-fold crossvalidation within the training set. Then, the final model is trained on the full training set using the parameters that achieved the highest accuracies, and this model is evaluated on the unseen testing set. For SVM, we tuned both the gamma γ (width of the RBF kernel) and C (misclassification cost) parameters. For LSTM, we tuned the number of layers, number of neurons per layer, learning rate (α), dropout, batch size and number of epochs. Table 5 show the results reported in terms of accuracy and weighted F1-score, where the weight of each class is based on its frequency in the dataset.

Table 5: Results of training the SVM model [22] and the LSTM model [50] on the Egyptian and UAE tweets.

Model	Features		Egypt Tweets		UAE Tweets	
			accuracy	*weighted F1*	*accuracy*	*weighted F1*
SVM	hand-crafted features [14]		60.6	60.6	51.1	49.4
LSTM	word embeddings	dialect	64.6	66.0	60.4	61.7
		generic	64.6	65.9	62.4	63.3
	lemma embeddings	dialect	**70.0**	**69.1**	**63.7**	**64.8**
		generic	65.5	66.7	63.4	64.5
	stem embeddings	dialect	67.1	67.3	63.2	63.9
		generic	65.0	66.0	62.8	63.9

Results indicate a clear advantage of LSTM over SVM with feature engineering. By looking at the LSTM results, we can see that abstracting away from words consistently improve the performance, especially through lemmas, which align with the findings in [16]. It can also be observed that using dialect-specific embeddings is better than using the generic embeddings, especially in Egyptian tweets that are mostly written in Egyptian dialect. On the other hand, the gap between the generic and the dialect-specific embeddings is smaller in the UAE tweets since these tweets are evenly distributed between MSA and UAE dialect, as shown in Table 3. Finally, we can observe that both SVM and LSTM models performed significantly better on the Egyptian tweets. This can be explained by the difficulty in predicting the sentiment of religious tweets, which constitute the majority of the UAE tweets. While many of these tweets contain highly subjective terms or expressions, their sentiment is considered neutral since users tweet Quranic verses or 'worship prayers' to spread and strengthen their faith, not to express sentiment. For instance, a tweet such as سبحَان اللّٰه العظيم عدد خلقه ورضَا نفسه وزنة عرشه ومدَاد كلمَاته *sbHAn Allh AlE.zym Edd xlqh wrDA nfsh wznp Er$h wmdAd klmAth* is considered neutral despite many positive terms it contains.

5. Conclusion

In this paper we presented the approach to create the Multi-Dialect Arabic Sentiment Twitter Dataset (MD-ArSenTD); a corpus of tweets collected from 12 different Arab countries and annotated for both sentiment and dialects. This is the first comprehensive sentiment Twitter dataset that comes with dialect annotation, allowing cross-dialect comparative evaluations, and with five-scale sentiment annotation, allowing for modeling sentiment intensity.

We present a detailed comparative characterization of tweets from Egypt and the UAE. The study shows the difference in topics discussed on Twitter in both countries. For instance, users in Egypt tend to discuss personal matters (e.g., relationships), whereas users in Egypt tend to discuss religious matters, mainly stating verses from the Quran or praising God. The study also shows that Egyptian tweets are mostly written in Egyptian dialect, whereas UAE tweets are mostly written in MSA. Finally, the majority of tweets in Egyptian tweets are neutral, whereas the majority of UAE tweets are positive due to a significant amount of tweets praising God, wishing well and sending greetings.

Furthermore, we performed a comparative evaluation of sentiment analysis models on the tweets of both Egypt and the UAE. We evaluated two models coming from two classes of machine learning; the first is based on training a sentiment classifier using an 'engineered' set of features tailored for Twitter data, and the other is based on the Long Short-term Memory (LSTM) deep learning model. Results indicate the superior performance of deep learning, the importance of using morphological features in Arabic, and that accounting for the local dialect leads to different outcomes depending on the country from which the tweets are collected from. For instance, using dialect-specific word embeddings to classify Egyptian tweets is more helpful than using them to classify UAE tweets, which are generally written in MSA.

Future work includes completing the comparative characterization and evaluation for tweets of all countries in the MD-ArSenTD. We are also willing to explore how dialect labels, in addition to other features extracted during tweets characterization can be used as features to further improve sentiment classification.

Acknowledgements

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

[1] , . https://www.statista.com/statistics/348508/most-tweeted-language-world-leaders.

[2] Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Transactions on Information Systems (TOIS) 26, 12.

[3] Abdul-Mageed, M., Diab, M., Kübler, S., 2014. Samar: Subjectivity and sentiment analysis for arabic social media. Computer Speech & Language 28, 20–37.

[4] Abdul-Mageed, M., Diab, M.T., 2014. Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis., in: LREC, pp. 1162–1169.

[5] Abdul-Mageed, M., Diab, M.T., Korayem, M., 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, Association for Computational Linguistics. pp. 587–591.

[6] Al-Sallab, A., Baly, R., Hajj, H., Shaban, K.B., El-Hajj, W., Badaro, G., 2017. Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 16, 25.

[7] Al Sallab, A.A., Baly, R., Badaro, G., Hajj, H., El Hajj, W., Shaban, K.B., 2015. Deep learning models for sentiment analysis in arabic, in: ANLP Workshop 2015, p. 9.

[8] Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A., 2016. Arasenti: Large-scale twitter-specific arabic sentiment lexicons. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics , 697–705.

[9] Al Zaatari, A., El Ballouli, R., ELbassouni, S., El-Hajj, W., Hajj, H., Shaban, K.B., Habash, N., 2016. Arabic corpora for credibility analysis. Proceedings of the Language Resources and Evaluation Conference (LREC) , 4396–4401.

[10] Aly, M.A., Atiya, A.F., 2013. Labr: A large scale arabic book reviews dataset., in: ACL (2), pp. 494–498.

[11] Assiri, A., Emam, A., Al-Dossari, H., . Saudi twitter corpus for sentiment analysis. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering 10, 272–275.

[12] Badaro, G., Baly, R., Hajj, H., Habash, N., El-Hajj, W., 2014. A large scale arabic sentiment lexicon for arabic opinion mining. ANLP 2014 165.

[13] Balikas, G., Amini, M.R., 2016. Twise at semeval-2016 task 4: Twitter sentiment classification. arXiv preprint arXiv:1606.04351 .

[14] Baly, R., Badaro, G., El-Khoury, G., Moukalled, R., Aoun, R., Hajj, H., El-Hajj, W., Habash, N., Shaban, K.B., 2017a. A characterization study of arabic twitter data with a benchmarking for state-of-the-art opinion mining models. WANLP 2017 (co-located with EACL 2017) , 110.

[15] Baly, R., Badaro, G., Hamdi, A., Moukalled, R., Aoun, R., El-Khoury, G., Al Sallab, A., Hajj, H., Habash, N., Shaban, K., et al., 2017b. Omam at semeval-2017 task 4: Evaluation of english state-of-the-art sentiment analysis models for arabic and a new topic-based model, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, Association for Computational Linguistics. pp. 601–608.

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

[17] Baly, R., Hobeica, R., Hajj, H., El-Hajj, W., Shaban, K.B., Al-Sallab, A., 2016. A meta-framework for modeling the human reading process in sentiment analysis. ACM Transactions on Information Systems (TOIS) 35, 7.

[18] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. Journal of machine learning research 3, 1137–1155.

[19] Chang, C.C., Lin, C.J., 2011. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2, 27.

[20] Chollet, F., 2015. keras. `https://github.com/fchollet/keras`.

[21] Cohen, J., 1960. Kappa: Coefficient of concordance. Educ. Psych. Measurement 20.

[22] Deriu, J., Gonzenbach, M., Uzdilli, F., Lucchi, A., De Luca, V., Jaggi, M., 2016. Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. Proceedings of SemEval , 1124–1128.

[23] Duwairi, R., Marji, R., Sha'ban, N., Rushaidat, S., 2014. Sentiment analysis in arabic tweets, in: Information and communication systems (icics), 2014 5th international conference on, IEEE. pp. 1–6.

[24] Elawady, R.M., Barakat, S., Elrashidy, N.M., 2014. Different feature selection for sentiment classification. International Journal of Information Science and Intelligent System 3, 137–150.

[25] Eskander, R., Rambow, O., 2015. Slsa: A sentiment lexicon for standard arabic., in: EMNLP, pp. 2545–2550.

[26] 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, IEEE. pp. 1114–1119.

[27] Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R., 2016. Ask me anything: Dynamic memory networks for natural language processing, in: International Conference on Machine Learning, pp. 1378–1387.

[28] Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5, 1–167.

[29] Maamouri, M., Bies, A., Buckwalter, T., Mekki, W., 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus, in: NEMLAR conference on Arabic language resources and tools, pp. 466–467.

[30] 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, pp. 3111–3119.

[31] Mohammad, S.M., Salameh, M., Kiritchenko, S., 2016. How translation alters sentiment. J. Artif. Intell. Res.(JAIR) 55, 95–130.

[32] Mountassir, A., Benbrahim, H., Berrada, I., 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, IEEE. pp. 3298–3303.

[33] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V., 2016. Semeval-2016 task 4: Sentiment analysis in twitter. Proceedings of SemEval , 1–18.

[34] Omar, N., Albared, M., Al-Shabi, A.Q., Al-Moslmi, T., 2013. Ensemble of classification algorithms for subjectivity and sentiment analysis of arabic customers' reviews. International Journal of Advancements in Computing Technology 5, 77.

[35] Paolillo, J.C., Das, A., 2006. Evaluating language statistics: The ethnologue and beyond. Contract report for UNESCO Institute for Statistics .

[36] Pasha, A., Al-Badrashiny, M., Diab, M.T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., Roth, R., 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic., in: LREC, pp. 1094–1101.

[37] Ravi, K., Ravi, V., 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems 89, 14–46.

[38] Refaee, E., Rieser, V., 2014a. An arabic twitter corpus for subjectivity and sentiment analysis., in: LREC, pp. 2268–2273.

[39] Refaee, E., Rieser, V., 2014b. 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.

[40] Refaee, E., Rieser, V., 2014c. 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, p. 16.

[41] Rosenthal, S., Farra, N., Nakov, P., 2017. SemEval-2017 task 4: Sentiment analysis in Twitter, in: Proceedings of the 11th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Vancouver, Canada.

[42] Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A., Perea-Ortega, J.M., 2011. Oca: Opinion corpus for arabic. Journal of the American Society for Information Science and Technology 62, 2045–2054.

[43] Salameh, M., Mohammad, S., Kiritchenko, S., 2015. Sentiment after translation: A case-study on arabic social media posts., in: HLT-NAACL, pp. 767–777.

[44] Sareah, F., 2015. Interesting statistics for the top 10 social media sites. Small Business Trends .

[45] Shoukry, A., Rafea, A., 2012. Sentence-level arabic sentiment analysis, in: Collaboration Technologies and Systems (CTS), 2012 International Conference on, IEEE. pp. 546–550.

[46] Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D., 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions, in: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics. pp. 151–161.

[47] Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C., et al., 2013. Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Citeseer. p. 1642.

[48] Statistics, T., 2017. Socialbakers. https://www.socialbakers.com/statistics/twitter/.

[49] Stats, I.W., . Internet world users by language: Top 10 languages. https://www.internetworldstats.com/stats7.htmml, year=2017.

[50] Tai, K.S., Socher, R., Manning, C.D., 2015. Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075 .

[51] Tang, D., Qin, B., Liu, T., 2015. Document modeling with gated recurrent neural network for sentiment classification., in: EMNLP, pp. 1422–1432.

[52] Yamamoto, Y., 2014. Twitter4j-a java library for the twitter api.

[53] Zaidan, O.F., Callison-Burch, C., 2014. Arabic dialect identification. Computational Linguistics 40, 171–202.