OMAM at SemEval-2017 Task 4: Evaluation of English State-of-the-Art Sentiment Analysis Models for Arabic and a New Topic-based Model

Ramy Baly[1], **Gilbert Badaro**[1], **Ali Hamdi**[2]
Rawan Moukalled[1], **Rita Aoun**[1], **Georges El-Khoury**[1], **Ahmad El-Sallab**[3]
Hazem Hajj[1], **Nizar Habash**[4], **Khaled Bashir Shaban**[5], **Wassim El-Hajj**[6]
[1] Department of Electrical and Computer Engineering, American University of Beirut
[2] Faculty of Computing, Universiti Teknologi Malaysia
[3] Computer Engineering Department, Cairo University
[4] Computational Approaches to Modeling Language Lab, New York University Abu Dhabi
[5] Computer Science and Engineering Department, Qatar University
[6] Department of Computer Science, American University of Beirut
{rgb15,ggb05,rrm32,rra47,gbe03,hh63,we07}@mail.aub.edu,
nizar.habash@nyu.edu, ali@alihamdi.com
ahmad.elsallab@gmail.com, khaled.shaban@qu.edu.qa

Abstract

While sentiment analysis in English has achieved significant progress, it remains a challenging task in Arabic given the rich morphology of the language. It becomes more challenging when applied to Twitter data that comes with additional sources of noise including dialects, misspellings, grammatical mistakes, code switching and the use of non-textual objects to express sentiments. This paper describes the "OMAM" systems that we developed as part of SemEval-2017 task 4. We evaluate English state-of-the-art methods on Arabic tweets for subtask A. As for the remaining subtasks, we introduce a topic-based approach that accounts for topic specificities by predicting topics or domains of upcoming tweets, and then using this information to predict their sentiment. Results indicate that applying the English state-of-the-art method to Arabic has achieved solid results without significant enhancements. Furthermore, the topic-based method ranked 1st in subtasks C and E, and 2nd in subtask D.

1 Introduction

Sentiment Analysis (SA) is a fundamental problem aiming to allow machines to automatically extract subjectivity information from text (Turney, 2002), whether at the sentence or the document level (Farra et al., 2010). This field has been attracting attention in the research and business communities due to the complexity of human language, and given the range of applications that are interested in harvesting public opinion in different domains such as politics, stocks and marketing.

The interest in SA from Arabic tweets has increased since Arabic has become a key source of the Internet content (Miniwatts, 2016), with Twitter being one of the most expressive social media platforms. While models for SA from English tweets have achieved significant success, Arabic methods continue to lag. Opinion mining in Arabic (OMA) is a challenging task given: (1) the morphological complexity of Arabic (Habash, 2010), (2) the excessive use of dialects that vary significantly across the Arab world, (3) the significant amounts of misspellings and grammatical errors due to length restriction in Twitter, (4) the variations in writing styles, topics and expressions used across the Arab world due to cultural diversity (Baly et al., 2017), and (5) the existence of Twitter-specific tokens (hashtags, mentions, multimedia objects) that may have subjective information embedded in them. Further details on challenging issues in Arabic SA are discussed in (Hamdi et al., 2016).

In this paper, we present the different systems we developed as part of our participation in SemEval-2017 Task 4 on Sentiment Analysis in Twitter (Rosenthal et al., 2017). This task covers both English and Arabic languages. Our systems work on Arabic, but is submitted as part of the OMAM (Opinion Mining for Arabic and More) team that also submitted a system that analyzes sentiment in English (Onyibe and Habash, 2017).

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 594–601,
Vancouver, Canada, August 3 - 4, 2017.

The first system extends English state-of-the-art feature engineering methods, and is based on training sentiment classifiers with different choices of surface, syntactic and semantic features. The second is based on clustering the data into groups of semantically-related tweets and developing a sentiment classifier for each cluster. The third extends recent advances in deep learning methods. The fourth is a topic-based approach for twitter SA that introduces a mechanism to predict the topics of tweets, and then use this information to predict their sentiment polarity. It further allows operating at the domain-level as a form of generalization from topics. We evaluate these models for message polarity classification (subtask A), topic-based polarity classification (subtasks B-C) and tweet quantification (subtasks D-E). Experimental results show that English state-of-the-art methods achieved reasonable results in Arabic without any customization, with results being in the middle of the group in subtask A. For the remaining subtasks, the topic-based approach ranked 2^{nd} in subtask D and 1^{st} in subtasks C and E.

The rest of this paper is organized as follows. Section 2 describes previous efforts on the given task. Section 3 presents the details of the Arabic OMAM systems. Section 4 illustrates the performances achieved for each subtask. We conclude in Section 5 with remarks on future work.

2 Related Work

SA models for Arabic are generally developed by training machine learning classifiers using different choices 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, 2013; 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 with syntactic features (root and part-of-speech *n*-grams) and stylistic features (digit and letter *n*-grams, word length, etc.) and achieved good performances after applying the Entropy-Weighted Genetic Algorithm for feature reduction (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, 2015)

A framework was developed for tweets written in Modern Standard Arabic (MSA) and containing Jordanian dialects, Arabizi (Arabic words written using Latin characters) and emoticons. This framework was realized by training different classifiers using features that capture the different 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 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; Mohammad et al., 2016).

We briefly mention the state-of-the-art performances achieved in English SA. 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 (Socher et al., 2013), the Gated Recurrent Neural Networks (Tang et al., 2015) and the Dynamic Memory Networks (Kumar et al., 2015). These models were only evaluated on reviews documents, and were never tested against the irregularities and noise that exist in Twitter data. A framework to automate the human reading process improved the performance of several state-of-the-art models (Baly et al., 2016; Hobeica et al., 2011).

3 OMAM Systems

In this section, we present the four OMAM systems that we investigated to perform the different subtasks of SemEval-2017 Task 4. These systems were explored during the development phase, and those that achieved best performances for each subtask were then used to submit the test results.

3.1 System 1: English State-of-the-Art SA

The state-of-the-art system selected from English was the winner of SemEval-2016 Subtask C "Five-point scale Tweet classification" in English (Ba-

likas and Amini, 2016). To apply it for Arabic, we derived an equivalent set of features to train a similar model for sentiment classification in Arabic tweets. The derived features are listed here:

- Word *n*-grams, where $n \in [1, 4]$. To account for the morphological complexity and sparsity of Arabic language, lemma *n*-grams are extracted since they have better generalization capabilities than words (Habash, 2010)
- Character *n*-grams, where $n \in [3, 5]$
- Counts of exclamation marks, question marks, and both marks
- Count of elongated words
- Count of negated contexts; a negated context is any phrase that occurs 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 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 added two additional binary features that indicate the presence of (1) user mentions and (2) URL or any other media content.

3.2 System 2: Cluster-based SA

This system is based on grouping semantically-related tweets, then training different sentiment classifiers for each group independently. At test time, each upcoming tweet is assigned to one of the pre-defined clusters, and the corresponding sentiment classifier is used to predict its polarity. Clusters are identified by applying the *k*-means algorithm to cluster the word embedding space that is generated using the skip-gram embedding model (Mikolov et al., 2013). Consequently, each cluster corresponds to a collection of semantically-related word vectors, and each tweet is assigned to the cluster whose word vectors are most similar (closest) to the tweet's words' vectors. Tweets that are assigned to the same cluster are used together to train a sentiment classifier using *n*-gram features. We trained several classifiers including the logistic regression, linear and non-linear SVM, Bernoulli Naive Bayes, Multinomial Bayes Naive. During model development, we only tuned the number of clusters k, whereas we used the default parameters of the different classifiers as implemented in scikit-learn (Pedregosa et al., 2011).

3.3 System 3: Recursive Auto Encoders

We trained the RAE deep learning model that achieved high performances in both English (Socher et al., 2011) and Arabic (Al Sallab et al., 2015). Briefly, the RAE model derive a sentence representation by combining word embeddings, two at a time, following the structure of a syntactic parse tree. The sentence representation is then used to train a softmax sentiment classifier. We followed the setup proposed by (Al Sallab et al., in press 2017) by applying RAE to morphologically tokenized text which proved to improve the performance by reducing the lexical sparsity of the language. We also use a broader semantic representation of words by concatenating word embeddings trained using the skip-gram model (Mikolov et al., 2013) with sentiment embeddings trained using the ArSenL sentiment lexicon (Badaro et al., 2014).

3.4 System 4: Topic-based SA

This system is based on the assumption that tweets discussing a particular topic are likely to share some unique semantic features. Figure 1 shows the architecture of this system. It is composed of several modules; (A) unsupervised topic classifier, (B) supervised topic classifier, (C) supervised domain classifier, in addition to a (D) generic sentiment classifier. The idea behind this system is that, since the test tweets may belong to topics that are not present in the training set, the different modules attempt to predict the topic and then classify the tweet's sentiment given the predicted topic. Before running the system in Figure 1, topic-specific and domain-specific sentiment classifiers are trained offline. Tweets belonging to each topic or domain in the train set are used, along with their sentiment labels, to train sentiment classifiers that are specific to the corresponding topic or domain. These classifiers are used with the above-mentioned modules as follows.

($\mathbf{M}_1$) Unsupervised Topic Classification Since the topic of each new tweet is unknown and can be different from those in the training set, we aim to discover which of the training topics is closest

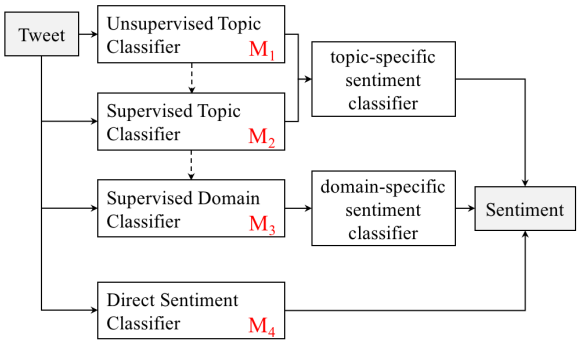

Figure 1: Architecture of the topic-based sentiment analysis system.

(or mostly related) to that of the tweet. This is achieved by training an embedding model, similar to that in System 2. Then, for each new tweet the system checks the similarity between the vector of each of the training topics and those of the tweet's words. The tweet is then assigned to the topic with the highest similarity, and its sentiment polarity is predicted using the sentiment classifier that is trained using instances of that particular topic.

(M_2) Supervised Topic Classification In many cases, all similarity values turn out to be small and close to 0. This is possible if the test tweet's topic is totally different from those in the train set, or if the tweet's words are implicitly related to the discussed topic. In such cases, we refer to a supervised topic classifier; a multi-class classifier, where the number of classes is equal to the number of topics in the training set. The topic classifier is trained using *n*-gram features extracted from all training tweets. Once the topic of the test tweet is predicted, its sentiment polarity is predicted using the sentiment classifier that is trained using instances of that particular topic.

(M_3) Supervised Domain Classification Some topics may not have sufficient instances to train an accurate sentiment classifier, therefore we introduce the concept of "domain"; a generalized form of the topic. A supervised domain classifier is a multi-class classifier, where the number of classes is equal to the number of domains in the training set. The domain classifier is trained using *n*-gram features extracted from all training tweets. Once the domain of the test tweet is predicted, its sentiment polarity is predicted using the sentiment classifier that is trained using instances of that particular domain.

(M_4) Direct Sentiment Classification In addition to the topic-specific and domain-specific classifiers, we also experiment with the direct sentiment classifier that ignores the topic information and is trained using all tweets in the training set.

We evaluated the following sequences of these modules: [M_1 →M_2 →M_3], [M_2 →M_3], [M_3] or [M_4]. For instance, in the first sequence, the tweet's topic is predicted using the unsupervised module (M_1), and then its polarity is predicted using the sentiment classifier for that topic. If no similarity was detected, we proceed to module (M_2) to predict the tweet's topic using the topic classifier, and then predict its sentiment using the sentiment classifier for that topic. If the topic is rare and no sentiment classifier exists for that topic, we proceed to module (M_3) to predict the tweet's domain using the domain classifier, and then predict its sentiment using the sentiment classifier for that domain.

4 Experiments and Results

In this section, we describe the experiments and results we achieved as part of our participation in SemEval-2017 Task 4. We describe the datasets we used, the preprocessing steps we applied and the performance of the different systems for each subtask. Table 1 illustrates the design of the evaluation experiments, highlighting the systems that were evaluated for each subtask. The system that achieved the best evaluation results, for each subtask, was then used to submit the test results.

Subtask	Systems
Message Polarity Classification (A)	Systems 1, 2, 3
Topic-based Polarity Classification (B-C)	Systems 1, 4
Tweet Quantification (D-E)	Systems 1, 4

Table 1: Design of evaluation experiments.

4.1 Datasets and Preprocessing

To run our experiments, we used datasets provided by the task organizers (Rosenthal et al., 2017) as follows. During evaluation, we trained our models on the TRAIN set, and evaluated our different systems on the DEV set. During testing, the system that achieved the best development results is trained using the combination of TRAIN and DEV sets, and tested the model on the TEST set.

For the English state-of-the-art approach (System 1), tweets are preprocessed by (1) replacing mentions and URLs with special tokens, (2) extracting emoticons and emojis and replacing them with special tokens using the emojis sentiment lexicon (Novak et al., 2015) and a in-home emoticons lexicon, (3) normalizing hashtags by removing the # symbol and the underscores that connect words in composite hashtags, and (4) normalizing letter repetitions (elongations). Then features are extracted by performing 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). We only included n-grams that occurred more than a pre-defined threshold t, where $t \in [3, 5]$ is tuned on the "DEV" set.

For the cluster-based SA approach (System 2), we trained the skip-gram word embedding model using a collection of datasets including the TRAIN and the DEV tweets provided by the organizers, the Qatar Arabic Language Bank (QALB) (Zaghouani et al., 2014) and several Arabic Twitter corpora from (Nabil et al., 2015; Refaee and Rieser, 2014b). We also used the k-means algorithm to cluster the embedding space into k clusters, with k ranging between 1 (no clustering) and 12. Best results during development were obtained using $k = 4$ and 5.

For the RAE approach (System 3), tweets are processed similar to System 1. We used MADAMIRA v2.1 to perform morphological tokenization following the ATB scheme (Habash and Sadat, 2006). We also used the Stanford parser (Green and Manning, 2010) to generate the syntactic parse trees. Since the resulting trees are not necessarily binary, and hence cannot be used to train recursive models, we used left-factoring to transform the trees to the Chomsky Normal Form (CNF) grammar that only contains unary and binary production rules.

For the topic-based approach (System 4), tweets are preprocessed by applying normalization and stemming using the NLTK ISRI stemmer (Taghva et al., 2005) and stopword removal. Then, n-grams are extracted using SKlearn TFiDFvectorizer (Pedregosa et al., 2011), with a variance threshold for feature reduction. The tweets in the training set that is provided by the task organizers pertain to 34 topics. We came up with a list of 8 generic domains that correspond to these topics, as shown in Table 2.

Domains	Topics
technology	ويندوز ١٠، جوجل، بوكيمون،أندرويد، غوغل، ايفون، آبل
shopping	غوتشي، امازون
sports	برشلونة، ريال مدريد، ميسي، فيديرر
media	هاري بوتر، جستن بيبر، بيونسيه
religion	رمضان، الاسلام
politics_isis	داعش، الارهاب
politics_me	بشار، اردوغان، إيران، سوريا، سورية، حلب، الأسد العراق، السعودية، سيسي
politics_us	هيلاري كلنتون، دونالد ترامب، باراك أوباما، اوباما

Table 2: The list of 8 generalized domains corresponding to the 34 topics in the training dataset.

4.2 Message Polarity Classification (A)

For this subtask, we evaluated the English state-of-the-art approach (System 1), the cluster-based SA approach (System 2) and RAE (System 3). The development and test results are illustrated in Table 3. It can be observed that System 1 achieved the best development results, and hence was used at the test phase. System 2 achieved slightly lower recall and higher accuracy, which indicates the potential benefits of training different sentiment classifiers for different clusters. Also, the inferior performance produced by System 3 can be due to its reliance on Arabic NLP tools, such as the syntactic parser, that are trained on MSA data, whereas the evaluation data are tweets that are likely to be noisy in terms of containing significant amounts of misspellings and grammatical errors.

	Model	Avg-R	Avg-F1	Acc.
DEV	*Sys 1*	**0.458**	**0.434**	**0.453**
	Sys 2	0.455	0.401	0.477
	Sys 3	0.424	0.394	0.410
TEST	*Sys 1*	0.438	0.422	0.430

Table 3: Results for subtask A (rank: #5/8).

4.3 Topic-based Polarity Classification (B-C)

For these subtasks, we evaluated the English state-of-the-art approach (System 1) and the different configurations of the topic-based SA approach (System 4) as discussed in subsection 3.2. The development and testing results for the 2-point and

the 5-point scale predictions are illustrated in Table 4 and 5, respectively.

	System	Avg-F1	Avg-R	Acc.
DEV	*Sys 1*	0.551	0.611	0.654
	Sys 4 [$M_1 \rightarrow M_2 \rightarrow M_3$]	0.473	0.536	0.554
	Sys 4 [$M_2 \rightarrow M_3$]	0.487	0.553	0.569
	Sys 4 [CM_3	0.495	0.576	0.569
	Sys 4 [M_4]	**0.581**	**0.640**	**0.690**
TEST	*Sys 4* [M_4]	0.678	0.687	0.679

Table 4: Results for subtask B (rank: #4/4).

	System	MAE_M	MAE_μ
DEV	*Sys 1*	0.410	0.568
	Sys 4 [$M_1 \rightarrow M_2 \rightarrow M_3$]	**0.387**	**0.551**
	Sys 4 [$M_2 \rightarrow M_3$]	0.414	0.648
	Sys 4 [M_3]	0.436	0.665
	Sys 4 [M_4]	0.422	0.647
TEST	*Sys 4* [$M_1 \rightarrow M_2 \rightarrow M_3$]	0.943	0.646

Table 5: Results for subtask C (rank: #1/2).

For Subtask B, it can be observed that ignoring the topic and domain information achieves highest performances. It can also be observed that generalizing from topics to domains in System 4 achieves better results than working at the topic-level only. As for Subtask C, results indicate that using topic-specific sentiment classifiers, and backing them with domain-specific sentiment classifiers, achieves the best performance in the competition on that subtask.

4.4 Tweet Quantification (D-E)

For these subtasks, we evaluated the English state-of-the-art approach (System 1) and the different configurations of the topic-based SA approach (System 4). The development and testing results for the 2-point and the 5-point scale quantifications are illustrated in Table 6 and 7, respectively.

	System	KLD	AE	RAE
DEV	*Sys 1*	**0.277**	**0.316**	**2.442**
	Sys 4 [$M_1 \rightarrow M_2 \rightarrow M_3$]	0.240	0.257	2.125
	Sys 4 [$M_2 \rightarrow M_3$]	0.319	0.668	2.783
	Sys 4 [M_3]	0.258	0.298	2.322
	Sys 4 [M_4]	0.581	0.640	0.690
TEST	*Sys 1*	0.202	0.238	4.835

Table 6: Results for subtask D (rank: #2/3).

For both subtasks, it can be observed that ignoring the topic and domain information achieves the best performances. For subtask D, using the features from System 1 achieved best development results, and ranked 2^{nd} in the competition. On the other hand, for subtask E, it turns out that using the simple n-gram features for direct sentiment classification ranked 1^{st} in the competition.

	System	EMD
DEV	*Sys 1*	0.436
	Sys 4 [$M_1 \rightarrow M_2 \rightarrow M_3$]	0.473
	Sys 4 [$M_2 \rightarrow M_3$]	0.474
	Sys 4 [M_3]	0.458
	Sys 4 [M_4]	**0.426**
TEST	*Sys 4* [M_4]	0.548

Table 7: Results for subtask E (rank: #1/2).

5 Conclusion

In this paper, we evaluated the application of recent state-of-the-art English models for sentiment analysis in Arabic tweets. These systems were used to perform all Arabic-related subtasks in SemEval-2017 Task 4.

In some cases, such as for message polarity classification (subtask A), the feature-based approach outperformed a RAE deep learning approach and another system that is based on creating semantic clusters for the tweets and training a sentiment classifier for each cluster.

For topic-based polarity classification (subtasks B and C) and topic-based tweet quantification (subtasks D and E), we evaluated a system that predicts the topic of upcoming tweets, and then predicts their sentiment using topic-specific sentiment classifiers. We allow this system to generalize from topics to domains. Results indicate that ignoring the topic and the domain information achieves better performances, with an exception for subtask C, where using topic-specific classifiers and backing them with domain-specific classifiers performs better.

As part of our future work, we will focus on developing SA models for different Arabic dialects, and also to perform cross-regional evaluations to confirm whether different models are needed for different regions and dialects, or a general model can work for any tweet regardless of its origins.

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

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*. Association for Computational Linguistics, pages 587–591.

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.

Ahmad A Al Sallab, Ramy Baly, Gilbert Badaro, Hazem Hajj, Wassim El Hajj, and Khaled B Shaban. in press 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* .

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.

Mohamed A Aly and Amir F Atiya. 2013. Labr: A large scale arabic book reviews dataset. In *ACL (2)*. pages 494–498.

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.

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.

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

Ramy Baly, Gilbert Badaro, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Wassim El-Hajj, Nizar Habash, and Khaled Bashir Shaban. 2017. A characterization study of arabic twitter data with a benchmarking for state-of-the-art opinion mining models. *ANLP 2017* .

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.

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*. IEEE, pages 1–6.

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.

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

Spence Green and Christopher D Manning. 2010. Better arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pages 394–402.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, pages 49–52.

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

Ali Hamdi, Khaled Shaban, and Zainal Anazida. 2016. A review on challenging issues in arabic sentiment analysis. *Journal of Computer Science* 12(9):471–481.

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*. IEEE, pages 75–80.

Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. 2015. Ask me anything: Dynamic memory networks for natural language processing. *CoRR, abs/1506.07285* .

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.

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

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*. IEEE, pages 3298–3303.

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

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.

Chukwuyem J. Onyibe and Nizar Habash. 2017. OMAM at SemEval-2017 task 4: English sentiment analysis with conditional random fields. In *Proceedings of the 11th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Vancouver, Canada, SemEval '17.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(Oct):2825–2830.

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.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 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, SemEval '17.

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.

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

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*. Association for Computational Linguistics, pages 151–161.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, 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, volume 1631, page 1642.

Kazem Taghva, Rania Elkhoury, and Jeffrey Coombs. 2005. Arabic stemming without a root dictionary. In *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*. IEEE, volume 1, pages 152–157.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*. pages 1422–1432.

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*. Association for Computational Linguistics, pages 417–424.

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.