Tw-StAR at SemEval-2017 Task 4: Sentiment Classification of Arabic Tweets

Hala Mulki [1], Hatem Haddad [2], Mourad Gridach [3] and Ismail Babaoglu [1]

[1] Department of Computer Engineering, Selcuk University, Konya, Turkey
[2] Department of Computer and Decision Engineering, Université Libre de Bruxelles, Belgium
[3] High Institute of Technology, Ibn Zohr University, Agadir, Morocco
halamulki@selcuk.edu.tr,Hatem.Haddad@ulb.ac.be
m.gridach@uiz.ac.ma,ibabaoglu@selcuk.edu.tr

Abstract

In this paper, we present our contribution in SemEval 2017 international workshop. We have tackled task 4 entitled "Sentiment analysis in Twitter", specifically subtask 4A-Arabic. We propose two Arabic sentiment classification models implemented using supervised and unsupervised learning strategies. In both models, Arabic tweets were preprocessed first then various schemes of bag-of-N-grams were extracted to be used as features. The final submission was selected upon the best performance achieved by the supervised learning-based model. Nevertheless, the results obtained by the unsupervised learning-based model are considered promising and evolvable if more rich lexica are adopted in further work.

1 Introduction

Social media is literally shaping decision making processes in many aspects of our daily lives. Exploring online opinions is therefore becoming the focus of many analytical studies. Twitter is one of the most popular microblogging systems that enables a real-time tracking of opinions towards ongoing events (Saif et al., 2016). Hence, it provides the needed feedback information for analytical studies in several domains such as politics and targeted advertising (El-Makky et al., 2014). Sentiment analysis plays an essential role in performing such studies as it can extract the sentiments out of the opinions and classify them into polarities (Tang et al., 2015). Arabic language has recently been considered as one of the most growing languages on Twitter with more than 10.8 million tweets per day (Alhumoud et al., 2015). Yet, Arabic is remarkably less tackled in the research of Sentiment Analysis (Nabil et al., 2015; ElSahar and El-Beltagy, 2015). With more resources and tools for Arabic Natural Language Processing (NLP) becoming available, and with the recent developed sentiment lexica for Modern Standard Arabic (MSA) and dialectal Arabic, this year, SemEval contest offers the opportunity to apply sentiment classification on Arabic tweets through subtask 4A-Arabic (Rosenthal et al., 2017). Analyzing Arabic tweets is significantly challenging due to the complex nature and morphology of the Arabic language. Furthermore, Arabic tweets are mostly informal and written in different dialects in which same words or expressions may have drastically different sentiments. For example, يعطيك العافية is a compliment of a positive sentiment that means "May GOD grant you health" in the Levantine dialect while it has an aggressive meaning of "burn in fire" in the Moroccan and Tunisian dialects (El-Makky et al., 2014). Additionally, tweeters tend to use abbreviations, neologisms, emoji and sarcasm frequently (Maas et al., 2011; Rajadesingan et al., 2015), and sometimes in the same 140-characters tweet (Maas et al., 2011).

Here, we describe our participation in Task 4, subtask 4A-Arabic of SemEval 2017 under the team name "Tw-StAR" (Twitter-Sentiment analysis team for ARabic). The task requires classifying the sentiment of single Arabic tweets into one of the classes: positive, negative or neutral (Rosenthal et al., 2017). To accomplish this mission, we have used two classification models:

- Supervised learning-based model: bag-of-N-grams features of different schemes have been adopted to train the model. Support Vector Machines (SVM) and naïve Bayes (NB) algorithms have been used as classification algorithms.

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 664–669,
Vancouver, Canada, August 3 - 4, 2017.

- Unsupervised learning-based (lexicon-based) model: in which a merged MSA/multi-dialectal sentiment lexica along with the constant weighting strategy have been employed to classify the tweets' sentiment.

The remainder of the paper is organized as follows: in Section 2, we describe the preprocessing step. In Section 3, we identify the extracted feature sets. Section 4 introduces the learning strategies used in the presented models. Results are reviewed and discussed in Section 5 while Section 6 concludes the study and future work.

2 Data Preprocessing

In this step, we have first cleaned the tweets from the unsentimental content such as URLs, Username, dates, hashtags, retweet symbols, punctuation, emotions and non-Arabic characters to get the Arabic text only as in (Shoukry and Rafea, 2012; Al-Osaimi and Badruddin, 2014). Secondly, the input data has been filtered from the words that do not affect the text meaning, the so called stopwords (El-Makky et al., 2014). Since our data contains several dialects we had to use an already built stopwords list of 244 words for MSA and Egyptian dialect used in (Shoukry and Rafea, 2012) merged with a manually-built list of 12 words from the Levantine and Gulf dialects such as فين, شو which mean"where" and "what" in the Gulf and Levantine dialects respectively. Furthermore, MSA/dialectal negation words such as ما, مش that mean "not" in Levantine and Egyptian dialects respectively, have been excluded from the used stopwords lists, as they may reverse the polarity of a tweet (Duwairi et al., 2014). Thus, a tweet such as "https://t.co/wPg3KEz4bW-ما يدور في رأس دونالد ترامب" which means "what is going on in Trump's mind" becomes "ما يدور رأس دونالد ترامب" after preprocessing. Lastly, for the ulexicon-based model, we have subjected each tweet to tokenization then to stemming to facilitate the words lookup process in the lexica. Stemming has been carried out using the Information Science Research Institute's (ISRI) Arabic stemmer provided by NLTK library (Bird, 2006). ISRI is a root-extraction stemmer that can provide a normalized form of unstemmed words rather than leaving them unchanged. Moreover, being a context-sensitive stemmer prevents ISRI from producing insensible and invalid roots (Dahab et al., 2015).

3 Feature Extraction

Bag-of-N-grams features have been adopted to be used in both of the presented models (Shoukry and Rafea, 2012; Abdulla et al., 2013; Ahmed et al., 2013). N-grams represent a sequence of adjoining N items collected from a given corpus. Extracting N-grams can be thought of as exploring a large piece of text through a window of a fixed size (Pagolu et al., 2016). Features selection has been performed using NLTK module FreqDist which gives a list of the distinct words ordered by their frequency of appearance in the corpus (Bird, 2006). A specific number of features was defined (equals to 40100 for the combination of unigrams+bigrams+trigrams) in order to be selected from the FreqDist's list . The feature extraction pipeline is illustrated in Figure 1. For a certain

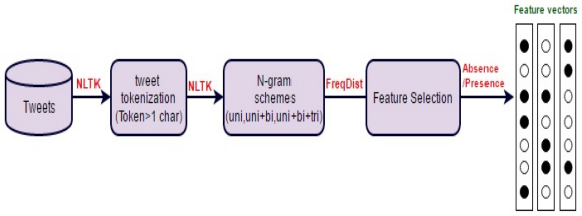

Figure 1: Feature extraction pipeline

N-grams scheme, a tweet's feature vector is constructed via examining the presence/absence of the N-grams features among the tweet's tokens. Consequently, the feature vector's values are identified as True (presence) or False (absence).

4 Learning Strategies

In this section, we describe the learning strategies adopted by the presented models. The mechanism of each strategy is briefly reviewed, in addition to an introduction of the python[1] supported tools used by these strategies to build the classification models.

4.1 Supervised learning

Supervised learning requires a labeled corpus to train the classifier on the text polarity prediction (Biltawi et al., 2016). In our case, a polarity labeled dataset of (3355) Arabic tweets provided by SemEval 2017 has been used such that 2684 tweets were dedicated to train the model while 671 tweets were used to tune it. The learning process

[1] https://www.python.org

has been carried out by inferring that a combination of specific features of a tweet yields a specific class (Shoukry and Rafea, 2012). We have used Naïve Bayes (NB) from Scikit-Learn (Pedregosa et al., 2011) since it is as powerful as Logistic Regression (Räbigera et al., 2016) and has proved its efficiency in classifying sentiment of multi-dialectal datasets (Itani et al., 2012). Additionally, linear SVM from LIBSVM was employed for its robustness and ease of implementation (Chang and Lin, 2011). Regarding used features, and as higher-order N-grams performed better compared to unigrams (Rushdi-Saleh et al., 2011). We have adopted N-grams schemes ranging from unigrams up to trigrams.

4.2 Unsupervised learning (lexicon-based)

In this strategy, neither labeled data nor training step are required to train the classifier. The polarity of a word or a sentence is determined using a sentiment lexicon or lexica that can be either pre-built or manually-built (Abdulla et al., 2013). Sentiment lexica usually contain subjective words along with their polarities (positive, negative). For each polarity, a sentiment weight is assigned using one of these weighting algorithms:

- Sum method: adopts the constant weight strategy to assign weights to the lexicon's entries, where negative words have the weight of -1 while positive ones have the weight of 1. The polarity of a given text is thus calculated by accumulating the weights of negative and positive terms. Then, the total polarity is determined by the sign of the resulted value (Abdulla et al., 2016).

- Double Polarity (DP) method: assigns both a positive and a negative weight for each term in the lexicon. For example, if a positive term in the lexicon has a weight of 0.8, then its negative weight will be: -1+0.8 = -0.2. Similarly, a negative term of -0.6 weight has a 0.4 positive weight. Polarity is calculated by summing all the positive weights and all the negative weights in the input text. Consequently, the final polarity is determined according to the greater absolute value of the resulted sum (El-Makky et al., 2014).

Having the MSA/dialectal combination of our training dataset defined by manual annotation (see Table 1), we have adopted a merged of pre-built and manually-built sentiment lexica with 6587 total entries of single and compound terms.

Arabic Type	Number of Tweets
MSA	2643
Egyptian	247
Levantine	69
Gulf	393
Moroccan	3
Total	3355

Table 1: Dataset MSA/Dialectal combination.

The pre-built lexica included NileULex (El-Beltagy, 2016) for MSA and Egyptian, Arabic Emotion Lexicon (Salameh et al., 2015) for emojis and Arabic Hashtag Lexicon (Salameh et al., 2015; Mohammad et al., 2016) for MSA/multiple dialects. Levantine and Gulf dialects were targeted through two manually-built lexica. Table 2 lists the used lexica and their sizes.

Sentiment Lexicon Used	Size
NileULex	5953
Arabic Emotion Lexicon (seeds)	23
Arabic Hashtag Lexicon (seeds)	230
Levantine Lexicon (manually-built)	281
Gulf Lexicon (manually-built)	100

Table 2: Used lexica.

As in (Abdulla et al., 2013; El-Beltagy and Ali, 2013), we have used the Sum method to determine the tweets' polarity. The polarity calculation procedure involved looking for entries that match the tweet's unigrams or bigrams in the lexica. Besides, we have provided the ability to look for the stemmed word if the unstemmed one could not be found (Al-Horaibi and Khan, 2016). Stopwords and negation words were kept to increase the possibility of matching a tweet's token with the compound terms of the merged lexica. Thus, for a tweet such غوغل مبدعة شي خرافي means "Google is incredibly creative" the polarity is calculated by summing the polarity values of its tokens "google+incredibly+creative= 0+1+1=+2 >0 " hence, it is positive.

5 Results and Discussion

The provided dataset consists of three parts: TRAIN (2684 tweets) for training models, DEV (671 tweets) for tuning models, and TEST (6100

tweets) for the official evaluation. Data preprocessing involved using regular expressions recognition and substitution provided by the re Python module[2]. N-grams feature schemes (unigrams+bigrams+trigrams) have been generated via NLTK [3]. Having the data preprocessed and the features extracted, we have trained the supervised learning-based model then classified the sentiment of the DEV set. The used classification algorithms were SVM from LIBSVM [4] and NB from Scikit-Learn [5]. Table 4 lists the results of these two classification algorithms. Considering the baseline results reviewed in Table 3, it can be observed that a slight improvement was achieved by NB compared to the baseline. While SVM outperformed both the baseline and NB by achieving an average F-score (AVG F1) of 0.384 and an average Recall (AVG R) of 0.459.

	AVG F1	**AVG R**
Baseline	0.249	0.333

Table 3: The baseline results for DEV set.

Algorithm	**AVG F1**	**AVG R**
SVM	**0.384**	**0.459**
NB	0.284	0.418

Table 4: The performance of the supervised learning-based model on DEV dataset.

For the lexicon-based model, the tweet's tokens (unigrams+bigrams) have been looked up in the lexica to calculate the tweet's polarity using the Sum method. The lookup process involved looking for the stemmed token if the unstemmed one is not found in the lexica. In Table 5, we notice that when stemming assists the lookup process, the performance degraded from 0.342 to 0.309 in terms of F-score value. This is because dialectal words may not be stemmed correctly by ISRI stemmer[6] (Dahab et al., 2015). For example, the term أبغى means "I want" in the Gulf dialect and has a neutral sentiment, while its stem using ISRI is بغى means "injustice" that has a negative polarity. However, the experiment in which stemmer was not used achieved quite a close performance to that of the supervised model as it yielded 0.448 and 0.342 for average recall and average F-score respectively. This is due to the fact that MSA/Egyptian and Gulf/Levantine dialects were efficiently supported by the used lexica.

Stemming	**AVG F1**	**AVG R**
Available	0.309	0.367
Not available	**0.342**	**0.448**

Table 5: Lexicon-based model performance on DEV dataset.

Considering the results in Table 4 and Table 5, the supervised learning-based model with SVM algorithm achieved the best average F-score and Recall values compared to the lexicon-based model. So, we selected it to provide the TEST set classification results for the final submission. Table 6 reviews the scores and the ranking of our system in the official evaluation.

Metric	**Value**	**Ranking**
Average F1	0.416	7/8
Average R	0.431	7/8
Average Accuracy	0.454	5/8

Table 6: Final submission evaluation of supervised learning-based model for the TEST dataset.

6 Conclusion and Future work

We have investigated sentiment classification of Arabic tweets via two classification models of various features and two learning strategies. Relatively, satisfying results were obtained by the supervised and lexicon-based models. For the final submission, we selected the supervised learning-based model, as it achieved the best average F-score and Recall values. However, the lexicon-based model has also yielded good results when the lookup process was not assisted by stemming. We noticed that MSA/multi-dialectal content has been efficiently handled by the merged lexica. Further improvement can be obtained in the future if Levantine/Gulf dialects are more efficiently supported by using their current lexica entries as seeds to produce a richer lexicon.

[2] https://docs.python.org/3/library/re.html

[3] http://www.nltk.org/

[4] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

[5] http://scikit-learn.org/stable/modules/naive_bayes.html

[6] http://www.nltk.org/_modules/nltk/stem/isri.html

References

Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, and Mahmoud Al-Ayyoub. 2013. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*. pages 1–6.

Nawaf A Abdulla, Nizar A Ahmed, Mohammed A Shehab, Mahmoud Al-Ayyoub, Mohammed N Al-Kabi, and Saleh Al-rifai. 2016. Towards improving the lexicon-based approach for arabic sentiment analysis. In *Big Data: Concepts, Methodologies, Tools, and Applications*. IGI Global, pages 1970–1986.

Soha Ahmed, Michel Pasquier, and Ghassan Qadah. 2013. Key issues in conducting sentiment analysis on Arabic social media text. In *2013 9th International Conference on Innovations in Information Technology (IIT)*. IEEE, pages 72–77.

Lamia Al-Horaibi and Muhammad Badruddin Khan. 2016. Sentiment Analysis of Arabic Tweets Using Semantic Resources. *International Journal of Computing & Information Sciences* 12(2):149.

Salha Al-Osaimi and Khan Muhammad Badruddin. 2014. Role of Emotion icons in Sentiment classification of Arabic Tweets. In *Proceedings of the 6th international conference on management of emergent digital ecosystems*. Association for Computing Machinery (ACM), pages 167–171.

Sarah O Alhumoud, Mawaheb I Altuwaijri, Tarfa M Albuhairi, and Wejdan M Alohaideb. 2015. Survey on Arabic Sentiment Analysis in Twitter. *International Science Index* 9(1):364–368.

Mariam Biltawi, Wael Etaiwi, Sara Tedmori, Amjad Hudaib, and Arafat Awajan. 2016. Sentiment classification techniques for Arabic language: A survey. In *2016 7th International Conference on Information and Communication Systems (ICICS)*. IEEE, pages 339–346.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, pages 69–72.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* 2(3):1–27.

Mohamed Y Dahab, Al Ibrahim, and Rihab Al-Mutawa. 2015. A comparative study on arabic stemmers. *International Journal of Computer Applications* 125(8):38–47.

R M Duwairi, Raed Marji, Narmeen Sha'ban, and Sally Rushaidat. 2014. Sentiment analysis in arabic tweets. In *2014 5th International Conference on Information and Communication Systems (ICICS)*. IEEE, pages 1–6.

Samhaa R. El-Beltagy. 2016. Nileulex: A phrase and word level sentiment lexicon for egyptian and modern standard arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Samhaa R El-Beltagy and Ahmed Ali. 2013. Open issues in the sentiment analysis of Arabic social media: A case study. In *9th the International Conference on Innovations and Information Technology (IIT2013)*. IEEE, pages 215–220.

Nagwa El-Makky, Khaled Nagi, Alaa El-Ebshihy, Esraa Apady, Omneya Hafez, Samar Mostafa, and Shimaa Ibrahim. 2014. Sentiment analysis of colloquial Arabic tweets. In *ASE BigData/SocialInformatics/PASSAT/BioMedCom 2014 Conference, Harvard University*.

Hady ElSahar and Samhaa R El-Beltagy. 2015. Building large arabic multi-domain resources for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pages 23–34.

M M Itani, R N Zantout, L Hamandi, and I Elkabani. 2012. Classifying sentiment in arabic social networks: Naïve search versus Naïve Bayes. In *2012 2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*. IEEE, pages 192–197.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 142–150.

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

Mahmoud Nabil, Mohamed A Aly, and Amir F Atiya. 2015. ASTD: Arabic Sentiment Tweets Dataset. In *Empirical Methods on Natural Language Processing (EMNLP)*. pages 2515–2519.

Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, and Babita Majhi. 2016. Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System*.

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

Stefan Räbigera, Mishal Kazmia, Yücel Saygına, Peter Schüllerb, and Myra Spiliopoulouc. 2016. SteM at SemEval-2016 Task 4: Applying active learning to improve sentiment classification. *Proceedings of SemEval* pages 64–70.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, pages 97–106.

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.

Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. 2016. Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management* 52(1):5–19.

Mohammad Salameh, Saif Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after Translation: A Case-Study on Arabic Social Media Posts. In *2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. The Association for Computational Linguistics, pages 767–777.

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

Duyu Tang, Bing Qin, and Ting Liu. 2015. Deep learning for sentiment analysis: Successful approaches and future challenges. *Wiley Int. Rev. Data Min. and Knowl. Disc.* 5(6):292–303.