Arabic Tweets Sentimental Analysis Using Machine Learning

Khaled Mohammad Alomari[1(✉)], Hatem M. ElSherif[2], and Khaled Shaalan[2]

[1] Faculty of Arts and Sciences, Abu Dhabi University, Abu Dhabi, UAE
khaled.alomari@adu.ac.ae

[2] Faculty of Engineering and IT, The British University in Dubai, Dubai, UAE
hatem.m.elsherif@gmail.com, khaled.shaalan@buid.ac.ae

Abstract. The continuous rapid growth of electronic Arabic contents in social media channels and in Twitter particularly poses an opportunity for opinion mining research. Nevertheless, it is hindered by either the lack of sentimental analysis resources or Arabic language text analysis challenges. This study introduces an Arabic Jordanian twitter corpus where Tweets are annotated as either positive or negative. It investigates different supervised machine learning sentiment analysis approaches when applied to Arabic user's social media of general subjects that are found in either Modern Standard Arabic (MSA) or Jordanian dialect. Experiments are conducted to evaluate the use of different weight schemes, stemming and N-grams terms techniques and scenarios. The experimental results provide the best scenario for each classifier and indicate that SVM classifier using term frequency–inverse document frequency (TF-IDF) weighting scheme with stemming through Bigrams feature outperforms the Naïve Bayesian classifier best scenario performance results. Furthermore, this study results outperformed other results from comparable related work.

Keywords: Sentiment analysis · Machine learning · Arabic natural language processing

1 Introduction

There has been limited sentiment analysis research addressing Arabic language, where this limitation attributed to the difficulty of analyzing the composition of Arabic Language as well as its complexity and the multiplicity. In 2016, there were 168,426,690 Arabic speaking people using the internet, representing 4.7% of the global internet population (much greater than contribution of Russian [2.9%] and German [2.3%]). Out of the estimated 388,332,877 persons in the digital world, Arabic speaks only 43.4% across the internet. The number of Arabic speaking internet users has been growing to 6602.5% in the last sixteen years (2000–2016) [1]. There is a rapidly growing rate of comments, feedbacks, ratings and reviews added online by Arabic users, which are gaining greater importance and attention. This trend driven by a rapid increase in the number of social media applications channels and users. MSA is the single formal Arabic script mostly based on Classic Arabic and is different from regional Arabic dialects.

S. Benferhat et al. (Eds.): IEA/AIE 2017, Part I, LNAI 10350, pp. 602–610, 2017.
DOI: 10.1007/978-3-319-60042-0_66

Colloquial versions of Arabic differ widely, which can be categorized by regional forms such as (Khaliji, Mesopotamian, Syro-Palestinian, Egyptian, Maghrebi) [2].

In this paper, we introduce a new Arabic Jordanian annotated twitter corpus that are suitable to sentimental analysis. It also investigates several preprocessing and ML strategies for evaluating their performance aiming at finding best ML strategy for sentimental analysis in multiple domain Arabic corpus. Motivated by superior performance achieved by Support Vector Machines and Naïve Bayes classifiers in English [3] and Arabic sentimental analysis research, this study analyzes both selected algorithms performance using different preprocessing strategies such as (stop words, stemming), N-grams and different weighting schemes. The rest of the paper is organized as follows. Section 2, outlines challenges of Arabic sentiment analysis and summarizes the related work. Section 3, describes our approach in building the corpus. Section 4, presents the used methods and experimentations. Section 5, discusses the obtained results. Finally, the main conclusions are presented.

2 Related Work

Many challenges are associated with Arabic Natural Language Processing (ANLP). In [4], a comprehensive survey is provided which discusses several ANLP challenges, such as Arabic diglossia and regional dialect phenomena. Meanwhile, Arabic sentiment analysis inherits the ANLP challenges, mainly due to the nature of the combinations of syntax, morphology and lexical entities from MSA and classical Arabic [5]. In addition, the absence of Arabic language standardization outside of the media and academia is considered a major challenge for ANLP research especially the use of dialects in social media [6–8].

Sentiment analysis systems are usually designed for languages with a single form, such as English. Meanwhile, international sentiment requires text analysis in multiple different local language forms. In addition, dialectical diversity and richness of the language challenges hinder the text analysis task [9]. In the literature, studies acknowledged nine principal reasons for Arabic language being particularly problematic for sentiment analyses such as complicated morphology and use of Arabic dialects, with poor spelling. Those concerns can be summarized as either a lack of language resources or problems associated with regional dialects [10, 11]. Table 1 presents a summary of major researches in Arabic sentiment analysis adopting Machine Learning approach.

Arabic Sentiment Analysis researchers have built Sentiment Analysis corpus from diverse electronic online data. As far as Arabic Sentiment Analysis researches is concerned [8, 12–14], publicly available Arabic twitter datasets are still limited in size and coverage of Arabic dialects. Meanwhile, available Arabic Jordanian twitter datasets are very limited such as the work of [12].

Table 1. A summary of major Arabic sentiments analyses using ML Approach.

Author	Dataset name	Size	Multi domain	Available	Source	Type[a]
[13]	Twitter Dataset	2591	Yes	Yes	Tweet/Facebook	MSA/DA
[8]	Twitter Corpus	8868	N/A	Yes	Tweet	MSA/DA
[12]	Tweet as string vector	2000	No	Yes	Tweet	MSA/DA
[15]	LABR	63257	No	Yes	Book	MSA/DA
[16]	AWATIF	<10 k	Yes	No	Wikipedia/Talk Pages/Forums/ News wire	MSA/DA
[17]	OCA	500	No	Yes	Movie	MSA
[18]	AOC	1.4 M	No	Yes	Newspapers Arabic Online	MSA/DA
[14]	Tweets	1000	No	Yes	Twitter	MSA/DA

[a]MSA stands for Modern Standard Arabic and DA stands for Dialectal Arabic

3 Arabic Jordanian General Tweets (AJGT) Corpus

This study introduces an Arabic tweets corpus written in Jordanian dialect and MSA annotated for the purpose of sentiment analyses. In the following paragraphs, we describe the method of collecting Arabic tweets through Twitter application programming interface (API), as well the data preprocessing and generation of the AJGT corpus.

Data Collection: In May 2016, a collection of tweets is retrieved using multiple keywords by narrowing down the search domain to Jordanian related general topics. RapidMiner utilized to retrieve and filter tweets. Manual selection process conducted to identify Jordanian dialect or MSA tweets from different topics based on content and expression of feelings [12].

Data Preparation: collected tweets included a blend of English and Arabic characters. Some characters are repeated (e.g. "حلوووووو "rather than the original "حلو"). We used ASAP Utilities[1] (Excel plugin) for data cleaning by filtering the text and removing links, duplicate tweets, hashtags and Foreign characters.

Corpus Generation: tweets manually annotated as either positive or negative by two human experts, a third expert was consulted. Finally, the generated AJGT corpus consists of 1,800 tweets classified as (900 positives, 900 negatives). The AJGT corpus is publicly available as Github project[2].

[1] http://www.asap-utilities.com.

[2] https://github.com/komari6/Arabic-twitter-corpus-AJGT.

4 Arabic Sentiment Analysis Experimental Framework

The ML model is developed using RapidMiner software platform because it has powerful text processing tools that support Arabic Language and provide tokenization, filtration of tokens and Arabic stemming, among others. Figure 1 shows the process model applied for each tweet. In our experiments, we applied the SVM and NB classifier to nine scenarios: for each of the three N-gram method (Unigram, Bigrams and Trigrams), we used a full stemming, light stemming, and without stemming. In addition, for each analysis we applied either the term frequency–inverse document frequency (TF-IDF) or term frequency (TF) weighting schemes; thus, generating eighteen distinct analysis processes for each weighting scheme per classifier. In order to test different scenarios, we adopted the 10-fold cross-validation approach with shuffled sampling where it builds random subsets of the Test Set and measures the performance of each scenario by computing the following performance metrics (Accuracy, Precision, Recall and F-Score). The proposed models did not include an Arabic stop word removal process, as the preliminary results indicated that including it reduces all scenarios performance.

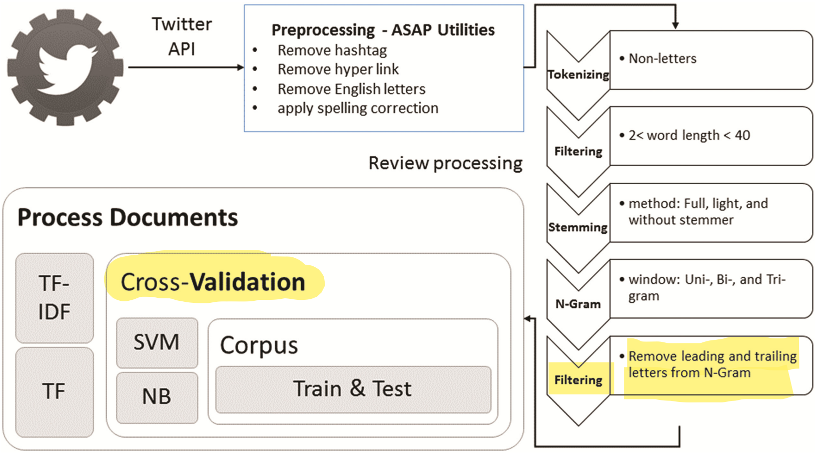

Fig. 1. Steps followed in the generation and validation of the AJGT corpus

5 Results and Discussion

Experiment evaluation results summary shown in Table 2 concluded as follows: (1) SVM classifier using stemmer with TF-IDF weighting scheme through Bigrams achieves the highest Accuracy (88.72%) and F-score (88.27%), (2) NB classifier best scenario results achieved using light stemmer with TF weighting scheme through Trigrams feature (Accuracy: 83.61%) and (F-score: 84.73%), (3) SVM classifier using stemming through Trigrams feature achieved best Precision results (TF-IDF: 92.10%, TF: 91.29%) confirming work of [13], Nevertheless due to the slight lower Precision results in Bigrams feature using TF-IDF weighting scheme (92.08%) and the better Recall results (84.89%) leading to highest F-score (88.27%). Generally, the Trigrams feature achieved the best Precision results in both SVM and NB classifiers, with either TF-IDF or TF weighting schemes, and (4) NB classifier without using stemming through Unigram feature with TF weighting scheme achieved the best Recall result (93.11%) outperforming SVM classifier result.

Table 2. 10-fold cross-validation results using the TF-IDF and TF as weighting schemes.

	N-gram	Stem	Accuracy		Precision		Recall		F	
			NB	SVM	NB	SVM	NB	SVM	NB	SVM
TF-IDF	Unigram	Yes	0.7822	0.8717	0.7546	0.8945	0.8378	0.8444	0.7936	0.8682
		Li	0.8144	0.8783	0.7863	0.9024	0.8644	**0.8489**	0.8233	0.8745
		No	0.8133	0.8678	0.7671	0.8912	0.9011	0.8389	0.8284	0.8638
	Bigrams	Yes	0.8144	**0.8872**	0.7847	0.9208	0.8678	**0.8489**	0.8239	**0.8827**
		Li	0.8200	0.8806	0.7885	0.9175	0.8756	0.8367	0.8295	0.8750
		No	0.8172	0.8606	0.7721	0.9069	0.9011	0.8044	0.8313	0.8523
	Trigrams	Yes	0.8178	0.8861	0.7878	**0.9210**	0.8711	0.8456	0.8271	0.8812
		Li	**0.8228**	0.8789	**0.7916**	0.9195	0.8778	0.8311	0.8321	0.8727
		No	0.8206	0.8583	0.7762	0.9065	**0.9022**	0.8000	**0.8341**	0.8496
TF	Unigram	Yes	0.7917	0.8672	0.7593	0.8885	0.8578	0.8411	0.8047	0.8638
		Li	0.8239	0.8628	0.7839	0.8814	0.8956	0.8389	0.8358	0.8594
		No	0.8222	0.8478	0.7651	0.8790	**0.9311**	0.8078	0.8397	0.8414
	Bigrams	Yes	0.8200	0.8811	0.7798	0.9087	0.8933	**0.8489**	0.8325	0.8771
		Li	0.8317	0.8667	0.7870	0.8954	0.9100	0.8311	0.8439	0.8618
		No	0.8233	0.8467	0.7698	0.8917	0.9233	0.7900	0.8394	0.8374
	Trigrams	Yes	0.8217	**0.8822**	0.7811	**0.9129**	0.8956	0.8467	0.8341	**0.8779**
		Li	**0.8361**	0.8622	**0.7938**	0.8969	0.9089	0.8200	**0.8473**	0.8562
		No	0.8261	0.8411	0.7746	0.8914	0.9211	0.7778	0.8413	0.8302

N-gram features have a different effect in classifiers average performance for the three used stemming approaches. In case of NB classifier, as shown in Fig. 2, increasing the N-gram improves Accuracy and F-score, and TF weighting scheme (Fig. 2b) outperforms TF-IDF (Fig. 2a) achieving best results using Trigram.

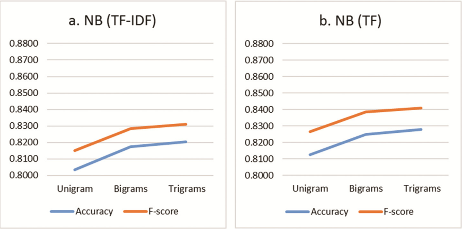

Fig. 2. NB classifier average performance

On the other hand, in case of SVM classifier performance measures (Accuracy and F-score) improvement caused by the increase of N-gram dropped when using the

Trigrams feature as shown in Fig. 3. Although N-gram nearly has the same effect on both weighting schemes. It is clear that SVM classifier performs better using TF-IDF in general and more specifically with the Bigrams feature. Furthermore, results show major differences between both classifiers regarding Accuracy and F-score, while Accuracy average performance is higher than F-score in SVM classifier. As shown in Fig. 3, the NB classifier is the opposite as shown in Fig. 2.

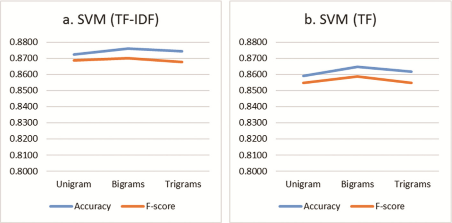

Fig. 3. SVM classifier average performance

As shown in Fig. 4, stemming approaches also have different effect in classifiers average measures performance for the implemented N-grams and weighting schemes. Although the NB results using different stemming approaches approximate, using light stemmer nearly achieved the best result. On the other hand, SVM results of using regular stemmer outperform other approaches.

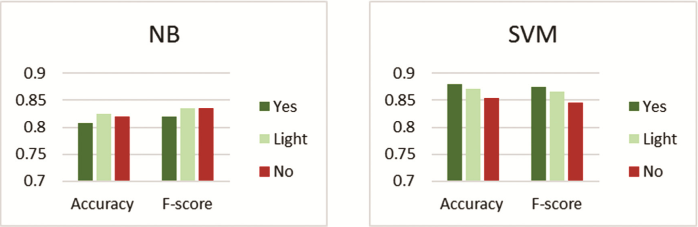

Fig. 4. Classifiers average performance results (Accuracy, F-score) by stemming approaches

Comparing study results with other related work, we observe the following: The work of [14] combining N-grams (Unigrams, Bigrams) on Arabic twitter dataset concludes that SVM classifier performs better than NB. Furthermore, in their extended work [19], results improved using SVM classifier combining N-grams (Unigrams, Bigrams, Trigrams) while investigating different normalization, stop words and stemming approaches. Another enhancement study conducted by same authors [20] confirms that the SVM classifier achieves superior results and using hybrid approaches combining ML and SO further improved the results. This study result shows that our approach through SVM classifier using stemmer with TF-IDF weighting scheme and Bigrams outperforms, in terms of Accuracy and F-score, the best results of [14, 19, 20].

The study results outperformed [12] results and confirms their conclusion, in case of Arabic twitter dataset SVM classifier with light stemmer (without using N-gram) gives better Accuracy. Nevertheless, using stemmer with TF-IDF weighting scheme through Bigrams gives much better results. Another study [7] used the [12] corpus achieved better results using the rule-based (lexicon-based) approach. However, although rule-based approach achieves superior results comparing to ML in single domain corpus, ML as a supervised approach is domain-independent [14] and more suitable for twitter multi-subject corpus.

The study results summary shown in Fig. 5 illustrates the best scenario. First, SVM classifier using stemmer with TF-IDF weighting scheme through Bigrams achieves the best performance which aligns with the work of [21] in English sentimental analysis. On the other hand, NB classifier achieves best performance results using light stemmer with TF weighting scheme through Trigrams feature. Comparing both classifiers best scenario, we conclude that SVM classifier proposed scenario that outperforms NB classifier results in terms of Accuracy and F-score.

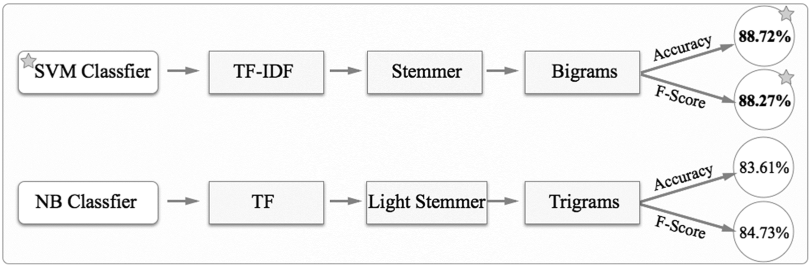

Fig. 5. Summary of best scenario results

6 Conclusions

This study main goal is to investigate the ML supervised approach and best scenario for Arabic sentimental analysis in multiple domain corpora. In order to achieve this goal: First, we introduced a new publicly available Arabic tweets corpus (1,800 annotated tweets) written in Jordanian dialect and MSA from different general topics, namely: AJGT corpus, contributing to the Arabic sentiment analyses research domain limited resources. Second, experiments are conducted comparing two machine learning algorithms (SVM and NB) utilizing different features and preprocessing strategies. We have used several N-grams (Unigram, Bigrams, Trigrams) with different weighting schemes (TF, TF-IDF) and applied alternative stemming techniques (no stemmer, stemmer, light stemmer).

This study concluded that adapting machine learning supervised approaches in Arabic sentimental analysis give promising results. The experiment concludes that SVM classifier using stemmer with (TF-IDF) weighting scheme through Bigrams is the best scenario achieving highest performance results outperforming NB classifier best scenario. Moreover, the proposed system following the best performing scenario using SVM classifier (Accuracy: 88.72% and F-score: 88.27%) outperforms other Arabic sentiment analyses research results. For our further work, we plan to increase the dataset, add third classification (neutral) and consider hashtags. In addition, investigate the proposed approach using other related Arabic regional dialect such as Palestinian Arabic dialect.

References

1. INternet World Stats: Internet World Users by Language. Top 10 Languages. http://www.internetworldstats.com/stats7.htm
2. Al-Kabi, M., Al-Qudah, N.M., Alsmadi, I., Dabour, M., Wahsheh, H. (eds.): Arabic/English Sentiment Analysis: An Empirical Study (2013)
3. Agarwal, B., Mittal, N.: Prominent Feature Extraction for Sentiment Analysis. Springer, Cham (2016)
4. Farghaly, A., Shaalan, K.: Arabic natural language processing: challenges and solutions. TALIP **8**, 1–22 (2009)
5. Ray, S.K., Shaalan, K.: A review and future perspectives of arabic question answering systems. IEEE Trans. Knowl. Data Eng. **28**, 3169–3190 (2016)
6. Bani-Khaled, T.A.: Standard Arabic and Diglossia. A problem for language education in the Arab world. Am. Int. J. Contemp. Res. **4**, 180–189 (2014)
7. Siddiqui, S., Monem, A.A., Shaalan, K.: Towards improving sentiment analysis in Arabic. In: Hassanien, A.E., Shaalan, K., Gaber, T., Azar, A.T., Tolba, M.F. (eds.) Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016, pp. 114–123. Springer, Cham (2017)
8. Refaee, E., Rieser, V.: An Arabic Twitter Corpus for subjectivity and sentiment analysis. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, 26–31 May 2014, pp. 2268–2273 (2014)
9. Shaalan, K.: A survey of Arabic named entity recognition and classification. Comput. Linguist. **40**, 469–510 (2014)
10. El-Makky, N., Nagi, K., El-Ebshihy, A., Apady, E., Hafez, O., Mostafa, S., Ibrahim, S.: Sentiment analysis of colloquial Arabic Tweets (2015)
11. Al-Twairesh, N., Al-Khalifa, H., Al-Salman, A.: Subjectivity and sentiment analysis of Arabic: trends and challenges. In: 2014 IEEE, Doha, Qatar, 10–13 November 2014, pp. 148–155. IEEE, Piscataway (2014)
12. Abdulla, N.A., Ahmed, N.A., Shehab, M.A., Al-Ayyoub, M. (eds.): Arabic sentiment analysis: Lexicon-based and corpus-based. In: 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT) (2013)
13. Duwairi, R.M., Qarqaz, I. (eds.) Arabic sentiment analysis using supervised classification. In: 2014 International Conference on Future Internet of Things and Cloud (FiCloud) (2014)
14. Shoukry, A., Rafea, A.: Sentence-level Arabic sentiment analysis. In: International Conference on Collaboration Technologies and Systems (CTS), 21–25 May 2012, Denver, Colorado; Proceedings, pp. 546–550. IEEE, Piscataway (2012)
15. Aly, M., Atiya, A.: LABR: large scale arabic book reviews dataset. In: Meetings of the Association of Computational Linguistics (ACL) (2013)
16. Abdul-Mageed, M., Diab, M.T.: AWATIF: a multi-genre corpus for modern standard arabic subjectivity and sentiment analysis and evaluation. In: Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, 23–25 May 2012, pp. 3907–3914. European Language Resources Association (ELRA) (2012)

17. Rushdi-Saleh, M., Teresa, M.-V.M., Ureña-López, A.L., Perea-Ortega, J.M.: OCA: opinion corpus for Arabic. J. Am. Soc. Inf. Sci. **62**, 2045–2054 (2011)
18. Zaidan, O.F., Callison-Burch, C.: The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, vol. 2, pp. 37–41. Association for Computational Linguistics, Portland, Oregon (2011)
19. Shoukry, A., Rafea, A.: Preprocessing Egyptian Dialect Tweets for sentiment mining. In: Fourth Workshop on Computational Approaches to Arabic, AMTA 2012, pp. 47–59 (2012)
20. Shoukry, A., Rafea, A.: A hybrid approach for sentiment classification of Egyptian Dialect Tweets. In: Gelbukh, A., Shaalan, K. (eds.) Advances in Arabic Computational Linguistics. First International Conference on Arabic Computational Linguistics: ACLing 2015, 17–20 April 2015, Cairo, Egypt: Proceedings, pp. 78–85. IEEE, Piscataway (2015)
21. Rushdi Saleh, M., Saleh, R., Martín-Valdivia, M.T., Montejo-Ráez, A., Ureña-López, L.A.: Experiments with SVM to classify opinions in different domains. Expert Syst. Appl. **38**, 14799–14804 (2011)