World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:16, No:5, 2022

An Enhanced Support Vector Machine-Based Approach for Sentiment Classification of Arabic Tweets of Different Dialects

Gehad S. Kaseb, Mona F. Ahmed

***Abstract*—**Arabic Sentiment Analysis (SA) is one of the most common research fields with many open areas. This paper proposes different pre-processing steps and a modified methodology to improve the accuracy using normal Support Vector Machine (SVM) classification. The paper works on two datasets, Arabic Sentiment Tweets Dataset (ASTD) and Extended Arabic Tweets Sentiment Dataset (Extended-ATSD), which are publicly available for academic use. The results show that the classification accuracy approaches 86%.

***Keywords*—**Arabic, hybrid classification, sentiment analysis, tweets.

I. Introduction

SA is the study of people's comments, and opinions on a selected object. Sentiment Classification (SC) approaches can be divided into three main categories: lexicon-based, machine learning (ML) and hybrid approaches. This paper uses a hybrid approach which aims to incorporate all ML and lexicon-based methods, so that it can take advantage of the benefits of each approach. SVM classifier is also used as it is considered one of the most effective classifiers in the SA field as indicated in the literature surveyed [1].

The proposed approach targets Arabic dialects. However, there is a lack of resources for Modern Standard Arabic (MSA) and even much worse in dialectal Arabic. Considering the importance of an emoji can significantly improve applications that study, analyze, and summarize electronic communications, rather than continuously removing emojis as a preprocessing step. The application of emoji SA is used to boost the sentiment rating.

The remainder of this paper is organized as follows: Section II explores SA related works. Section III illustrates the proposed SA methodology. Section IV shows the results and analysis. Section V presents concluding remarks and discusses the future work.

II. Related Work

Some work done in Arabic SA utilized ML methods; others used a lexicon-based approach. Lexicon-based approaches are unsupervised approaches that depend on external lexica to classify sentiments. ML approaches are mainly supervised approaches that rely on the existence of labeled training documents/phrases using classifiers such as SVM and naïve Bayesian. Hybrid approaches are those that combine lexicon and ML techniques. The interested reader is referred to the survey in [1].

ASTD is a corpus of tweets suggested by [2]. It is composed of 10,006 tweets written in MSA and Egyptian colloquial Arabic. The tweets are classified into 799 positive, 1,684 negative, 832 neutral and 6,691 objective tweets. Different ML approaches were used in this paper. The best recall achieved was 69% using SVM. However, this paper did not mention any pre-processing or cleaning steps. The dataset includes a small number of subjective tweets.

Extended-ATSD: Arabic Tweets Sentiment Dataset is a corpus of tweets proposed by [3] which consists of 7082 tweets written in MSA as well as Egyptian colloquial Arabic. It is composed of 714 positive tweets, 1901 negative tweets, 714 neutral tweets and 3530 objective tweets. The best accuracy achieved was 62.4%.

The SentiStrength software [4] is used to observe and measure the strength of the sentiment expressed within the social media text. It was originally developed for English and later was adapted to different languages including Arabic. Using SentiStrength, the approach proposed by [5] assigned a score to every tweet that indicates the entire sentiment score. They performed a comprehensive analysis of SentiStrength using 11 Arabic datasets consisting of tens of thousands of reviews/comments from different domains and in several dialects. They perform the analysis in terms of positive and negative sentiments. The evaluation results show that overall SentiStrength achieves 62% accuracy, 83.7% precision, 64% recall (positive correct), 68% F1 measure and 55% negative correct.

The authors in [7] presented NileULex, which is an Arabic sentiment lexicon containing approximately 6000 Arabic words and compound phrases. Egyptian or colloquial dialect is about 45% of the lexicon and MSA is about 55%. This lexicon took development period over two years. Whereas the gathering of many of the terms included in the lexicon was done automatically, the particular addition of any term was done manually. One of the vital criteria for adding terms to the lexicon was to disambiguate them as much as possible. The result is a lexicon with a far higher quality than any translated variant or automatically constructed one.

G. S. Kaseb is a postgraduate researcher and M. F. Ahmed is an associate professor in Computer Engineering Cairo University, Giza, Egypt (e-mail: gehad.kaseb@gmail.com, mona_farouk@eng.cu.edu.eg).

III. Proposed SA System

This section outlines the methodology and resources employed in the proposed work, the used datasets, preprocessing, classification architectures, and the accuracy measurements used to evaluate them. The system components are described in detail in the following sections.

A. Used Datasets

For all the conducted experiments, two datasets are used which are available for academic use; ASTD [2] and Extended-ATSD [3].

ASTD and Extended-ATSD both have four labels: positive, negative, neutral and objective. Due to the highly skewed distribution of the classes, and since our focus is to perform opinion classification rather than subjectivity classification, we excluded the objective and neutral tweets. So, we focused on positive and negative tweets only. Then, the datasets will be preprocessed and cleaned by a series of proposed steps to improve the classification accuracy. Table I shows the used datasets statistics.

TABLE I
ASTD and Extended-ATSD Used Dataset Polarities

	ASTD	Extended-ATSD
Positive	798	714
Negative	1680	1901
Total	2478	2847

B. Preprocessing

In this phase, the data is prepared before being fed to the classifiers either in the training phase or in the testing phase. Preprocessing phase includes those sequential steps:

Step 1: Normalization

- Remove Taskeel: taskeel is collected in this list (Fathatan, Dammatan, Kasratan, Fatha, Damma, Kasra, Shadda, Sukun, Maddah above, Hamza above, Hamza below, Subscript Alef, Inverted Damma, Mark Noon Ghunna, Zwarakay, Vowel Sign Small V above, Vowel Sign Inverted Small V Above, Vowel Sign Dot below, Reversed Damma, Fatha with two dots, Wavy Hamza below, Letter Superscript Alef)
- Remove Tatweel: for example انـــــــــــــــــا will be انا.
- Remove honorific sign (Arabic Sign Sallallahou Alayhe Wa Sallam "صلي الله عليه وسلم", Arabic Sign Alayhe Assallam "عليه السلام", Arabic Sign Rahmatullah Alayhe "رحمه الله", Arabic Sign Radi Allahou Anhu "رضي الله عنه", Arabic Sign Takhallus)
- Remove koranic annotation list (Arabic Small High Tah, Arabic Small High Ligature Alef With Lam With Yeh, Arabic Small High Zain, Arabic Small Fatha, Arabic Small Damma, Arabic Small Kasra, Arabic Small High Ligature Sad With Lam With Alef Maksura, Arabic Small High Ligature Qaf With Lam With Alef Maksura, Arabic Small High Meem Initial Form, Arabic Small High Lam Alef, Arabic Small High Jeem, Arabic Small High Three Dots, Arabic Small High Seen, Arabic End Of Ayah, Arabic Start Of Rub El Hizb, Arabic Small High Rounded Zero, Arabic Small High Upright Rectangular Zero, Arabic Small High Dotless Head Of Khah, Arabic Small High Meem Isolated Form, Arabic Small Low Seen, Arabic Small High Madda, Arabic Small Waw, Arabic Small Yeh, Arabic Small High Yeh, Arabic Small High Noon, Arabic Place Of Sajdah, Arabic Empty Centre Low Stop, Arabic Empty Centre High Stop, Arabic Rounded High Stop With Filled Centre, Arabic Small Low Meem)
- Normalize the letters which have more than one form such as Alef (replace the Alef with Hamza above "أ", and Alef with Hamza below "إ" and Alef Madda "آ" to Alef "ا"), Haa (replace the Taa Marbuta "ة" with Haa "ه") and Yaa (replace the Dotless Yaa "ى" with Yaa "ي")

Step 2: Emoji Word Converter

Emoticons and emojis are extracted using the "emoji" java library [6], and then they are replaced with their Aliases using a manually-prepared list of emotion-word converter Table II. These words are then used in the emotion word lexicon.

TABLE II
Emotion-Word Converter List

Emotion-Word	Emotion-Word
وجهكاء = ':)	وجهسعيد = ^_-
وجهكاء = ":)	وجهسعيد = -_-
وجهزعلان = :->	وجهسعيد = -_^
وجهشيطان = 3:)	وجهسعيد = *_*
وجهشرير = >-)	وجهسعيد = :-)
وجهغاضب = >:)	وجهملاك = O:)
وجهسعيد = ^_^	وجهسعيد = :)
وجهزعلان = :/	وجهغاضب = :-@
وجهمندهش = :-O	وجهحضن = (((H)))
وجهزين = :)	وجهضحك = :D
وجهحزين = :-)	وجهمرتبك = o.o
وجهمرتبك = O.o	وجهقبله = :-*
وجهمرتبك = o.O	وجهمتغاظ = :P
وجهقلب = <3	ان شاء الله = ISA
وجهضحك = LOL	برايي = Imo
اهلا بعودتك = Tyt	جزاك الله خيرا = Jak

Step 3: Arabic Named Entities Recognition

Named Entities Recognition (NER) becomes an important part of SA not only when the task is to identify an opinion holder but additionally for the task of determining semantic orientation. The reason for this is that the majority of Arabic first names, and to lesser extent family names, are derived from Arabic adjectives that can be easily confused for sentiments. Some Arabic male names that demonstrate this point include: Adel, Nabil, Said, and Hakim. The meanings of these names are: Just, Noble, Happy, and Wise. Examples of female names include: Gamila, Latifa, Sara, and Wafia, whose meanings are: Beautiful, Nice, Happy, and Loyal [8].

The Named Entities are removed after recognizing them by using gazetteer lists. We use ANERGazet [9] which is a collection of three Gazetteers. (i) Location Gazetteer: this dictionary consists of 1,950 continents names, countries, cities, rivers and mountains found in the Arabic version of Wikipedia. (ii) Person Gazetteer: this dictionary consists of

2,309 Arabic and non-Arabic names of people found in Wikipedia and other websites. (iii) Organizations Gazetteer: which consists of a listing of 262 names of companies, football teams and other organizations from different web sources as well [10].

Step 4: Stop Words Removal

A manually-prepared stop words list was collected that consists of 5866 words starting from three Arabic stop words lists [11]-[13] then adding days and months names, country and capital names [14], and ANERGazet dictionaries then removing duplicates in these stop words.

Step 5: Misspelling Correction

Users sometimes repeat a character more than once to emphasize and stress their meaning. For example, the word "كثيييير", which implies "moooore" in English, should be written as "كثير"; but the letter "ي" is continual.

The foremost word used is "هههههههه" which implies "hahahaha" for laughing so we detect first the hahaha word with any length then replace it with "ضحك" which implies "laugh"; we do that firstly because the next step will remove any other repetition. Secondly, deleting repeated characters is needed in order to have the base form of the words. However, some words already have repeated characters, such as "written" in English. To handle this matter, a Java program was used to delete repeated characters for words that are not in MSA.

Step 6: Other Cleaning

We remove punctuations (؟ ? ! . : | ( ) - # / @ $ ^ _ % & * + \ } { [ ] " ' ; , ،< >), symbols and other special characters.

C. Lexicons

Sentiment lexicons containing opinion terms, along with their polarity and strength are an essential part of any SA tool. There are currently limited publicly available colloquial Arabic sentiment lexicons. In order to have a good SA, sentiment scores of each tweet are calculated. These sentiment scores consist of positive score and negative score. Different lexicons were used:

1) A large-scale Arabic Sentiment Lexicon (ArSenL) was built by [15] and it is available for academic use. ArSenL is constructed using a combination of English SentiWordnet (ESWN), Arabic WordNet, and the Arabic Morphological Analyzer (AraMorph). This lexicon has sentiment for words in the MSA. ArSenL has more than 28 thousand lemmas with almost 158 thousand synsets, which means that each lemma may have a different part of speech, or different sentiment scores. Each line in this sentiment lexicon represents one word, with the lemma of the word analyzed using Aramorph analyzer, POS, positive score and negative score. The other information is disregarded in our use. It can also be seen that each lemma has two different scores which are positive score and negative score. Moreover, the words in the lexicon are represented using lemmas not Arabic characters. We use buckwalter to Unicode converter [16] to get the corresponding Arabic words. Fig. 1 shows a snapshot of the Arabic words written in English of sentiment lexicon and Fig. 2 shows the corresponding Arabic words.

```
raHomap_1;n;0;0;50;NIL;01071411;NIL
raHomap_1;n;0.625;0;50;NIL;01227495;NIL
raHomap_1;n;0.5;0.125;50;NIL;04829282;NIL
raHomap_1;n;0;0.75;50;NIL;04829550;NIL
raHomap_1;n;0;0.5;50;NIL;07553741;NIL
raHomap_1;n;0.125;0.5;50;NIL;07554500;NIL
raHomap_1;n;0.125;0;50;NIL;14474435;NIL
```

Fig. 1 Snapshot of the English sentiment lexicon

```
1_رَحْمَة;n;0;0;50;NIL;01071411;NIL
1_رَحْمَة;n;0.625;0;50;NIL;01227495;NIL
1_رَحْمَة;n;0.5;0.125;50;NIL;04829282;NIL
1_رَحْمَة;n;0;0.75;50;NIL;04829550;NIL
1_رَحْمَة;n;0;0.5;50;NIL;07553741;NIL
1_رَحْمَة;n;0.125;0.5;50;NIL;07554500;NIL
1_رَحْمَة;n;0.125;0;50;NIL;14474435;NIL
```

Fig. 2 Snapshot of the Arabic sentiment lexicon

More complexities would arise from the presence of dialectal words, idioms and compound phrases. So, dictionaries are needed for all these words categories.

2) Nile University's Arabic sentiment Lexicon NilULex v0.27 was proposed by [7]. It contains about six thousand Egyptian Arabic and MSA sentiment words and their polarities. The class distribution within this lexicon is shown in Fig. 3.

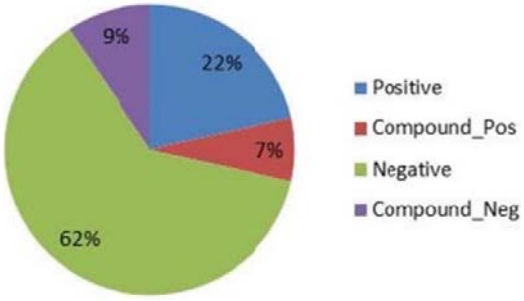

Fig. 3 NilULex Class distribution

3) Large multi-domain lexicons for SA in Arabic were proposed by [17]. The resources are publicly available in [18]. The authors showed that the generated lexicons are effective and reliable for Arabic SA. The combined lexicon (ALL_lex.csv) includes hotel, library, movie, production and restaurant opinion words with their polarity. The lexicons are domain specific lexicons, semi automatically generated from the datasets with total size of two thousand.
4) An Emoji lexicon was proposed by [19]. It contains emoji Unicode with its positive, negative and neutral score as shown in Fig. 4.
5) An Emoticons lexicon was also manually built depending on the emotions words converted in step 2 in the preprocessing phase. Table III shows 16 words corresponding to the converted emoticons with their

positive-negative annotation value; positive annotation takes value equal 1 and negative annotation takes value equal -1.

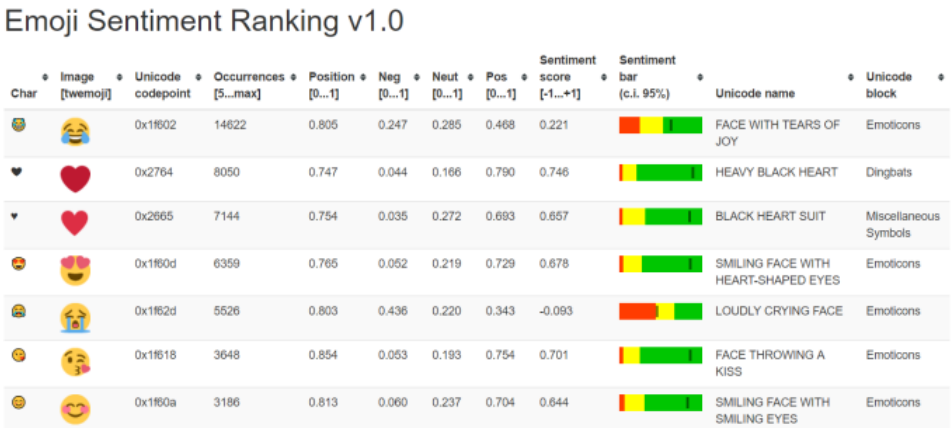

Fig. 4 Emoji lexicon

TABLE III
EMOTION WORD AND POS-NEG VALUE CONVERTER LIST

Word-value	Word-Value
وجهسعيد 1	وجهملاك 1
وجهحزين -1	وجهبكاء -1
وجهشيطان -1	وجهحضن 1
وجهشرير -1	وجهزعلان -1
وجهغاضب -1	وجهضحك 1
وجهمرتبك -1	وجهمندهش -1
وجهمتغاظ -1	وجهقبله 1
وجهقلب 1	ضحك 1

There are Arabic words that do not exist in the lexicons. In this case, these words will be disregarded, which means the sentiment scores are calculated using only the words that exist in the lexicons. The final dataset after adding lexicon features will be represented by six features e.g. (Arabic Tweet, annoted sentiment, 0,0,0,0,0,0). The six features are representing positive and negative scores for ArSenL, NileUnv and Emoticons lexicon, respectively.

D. Classification

From the study, comparison and analysis of the different proposed methodologies for SA, it was observed that SVM yield the best performance [1]. So, SVM with Term frequency-inverse document frequency (TF-IDF) feature vector was used.

E. Evaluation Measures

We report the results of each experiment using accuracy metric to measure the performance. Accuracy reports the ratio of correctly classified tweets to the total number of tweets regardless of their class.

IV. RESULTS AND ANALYSIS

We run the algorithm 100 times using a random seed to shuffle and partition the dataset into training set (80%) and test set (20%). This randomness changes the training and test sets in every iteration, which means that the methodologies are tested using 100 combinations of the training and test sets. The three values, Max., Avg. and Min. shown in Table IV which represents the maximum, average and minimum accuracy achieved over the 100 iterations.

TABLE IV
ACCURACY RESULTS FOR BOTH DATASETS ASTD AND EXTENDED-ATSD

	ASTD	Extended-ATSD
Max	86.0	85.9
Avg.	79.4	79.1
Min	75.6	72.1

The authors in [20] extracted a subset of the ASTD dataset with the positive and negative tweets only. They reached accuracy: 57.1%, Precision: 38.5% Recall: 55.7% and F1: (45.5%).

The authors in [21] extracted a subset of the ASTD dataset with the positive and negative and neutral tweets. These data are split into a training set (70%), a development set (10%) and a test set (20%). The results show that RNTN achieves the best performance (Accuracy = 58.5% and Average F1 = 53.6%) although it was trained on a dataset that is different from that used for testing.

The authors in [22] extracted a subset of the ASTD dataset with the positive and negative tweets only. The resultant accuracy reaches 75.9% when the model has been trained in a balanced form and 79.07% in an unbalanced form.

The authors in [23] extracted a subset of the ASTD dataset with the positive and negative tweets only. They then combined it with ArTwitter [24], and QCRI [25]. They achieved recall: 76.5%, precision: 83.0%, F-measure 79.62% and macro-accuracy: 80.21%.

V. CONCLUSION

This work presented a model for Arabic SA including the preprocessing steps, the methodology and the used lexicons. The model was trained and tested using two datasets, ASTD and Extended-ATSD. The results show improved accuracy

which achieved 86%. The intended future work is to create a large scale lexicon and to propose a deep learning model to enhance Arabic SA. In addition, it is planned to build Arabic SA software which annotates Arabic tweets online.

REFERENCES

[1] Gehad S. Kaseb, Mona F. Ahmed. Arabic Sentiment Analysis approaches: An analytical survey. International Journal of Scientific & Engineering Research, Volume 7, Issue 10, October-2016 712 ISSN 2229-5518.

[2] Mahmoud Nabil, Mohamed Aly and Amir F. Atiya. ASTD: Arabic Sentiment Tweets Dataset. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2515–2519, Lisbon, Portugal, 17-21 September 2015.

[3] Kaseb, Gehad S., and Mona F. Ahmed. "Extended-ATSD: Arabic Tweets Sentiment Dataset" Journal of Engineering and Applied Sciences 14.14 (2019): 4780-4785.

[4] M. Thelwall, Heart and soul: sentiment strength detection in the social web with sentistrength, in: Proceedings of the CyberEmotions, 2013, pp. 1–14.

[5] A. Rabab'ah, M. Al-Ayyoub, Y. Jararweh, M. Al-Kabi, Evaluating sentistrength for Arabic sentiment analysis, in: 2016 7th International Conference on Computer Science and Information Technology (CSIT), 2016, pp. 1–6.

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

[7] El-Beltagy, Samhaa R., 2016. NileULex: A Phrase and Word Level Sentiment Lexicon for Egyptian and Modern Standard Arabic. In proceedings of LREC 2016, Portorož, Slovenia.

[8] Open Issues in the Sentiment Analysis of Arabic Social Media: A Case Study, Samhaa R. El-Beltagy

[9] ANERGazet available at: http://users.dsic.upv.es/grupos/nle/?file=kop4.php

[10] Benajiba, Y., Rosso, P., Bened´ı Ru´ız: ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy. In: Proceeding of CICLing-2007, Mexico. Lecture Notes in Computer Science 4394, Springer-Verlag

[11] Arabic Stop Words list available at: https://www.arabeyes.org/%D9%85%D8%B3%D8%AA%D8%A8%D8%B9%D8%AF%D8%A7%D8%AA_%D8%A7%D9%84%D9%81%D9%87%D8%B1%D8%B3%D8%A9

[12] Arabic Stop Words list available at: https://sites.google.com/site/kevinbouge/stopwords-lists/stopwords_ar.txt?attredirects=0&d=1

[13] Arabic stop words list available at: http://www.ranks.nl/stopwords/arabic

[14] Arabic country and capital names available at: http://www.nationsonline.org/oneworld/countrynames_arabic.htm

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

[16] Buckwalter to Unicode converter available at: http://www.comp.leeds.ac.uk/andyr/software/

[17] ElSahar, Hady, and Samhaa R. El-Beltagy. "Building large arabic multi-domain resources for sentiment analysis." International Conference on Intelligent Text Processing and Computational Linguistics. Springer International Publishing, 2015.

[18] Lexicons github website: https://github.com/hadyelsahar/large-arabicsentiment-analysis-resouces-last

[19] Petra Kralj Novak, Jasmina Smailovic, Borut Sluban, ´ and Igor Mozetic. 2015. Sentiment of emojis. ˇ PloS one, 10(12):e0144296.

[20] A. Rabab'ah, M. Al-Ayyoub, Y. Jararweh, M. Al-Kabi, Evaluating sentistrength for Arabic sentiment analysis, in: 2016 7th International Conference on Computer Science and Information Technology (CSIT), 2016, pp. 1–6.

[21] Baly, Ramy, et al. "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) (2017): 110.

[22] Dahou, Abdelghani, et al. "Word Embeddings and Convolutional Neural Network for Arabic Sentiment Classification."

[23] Altowayan, A. Aziz, and Lixin Tao. "Word embeddings for Arabic sentiment analysis." Big Data (Big Data), 2016 IEEE International Conference on. IEEE, 2016.

[24] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. AlAyyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on, Dec 2013, pp. 1–6.

[25] A. Mourad and K. Darwish, "Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs," in Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, 2013, pp. 55–64.