Feature-based Sentiment Analysis for Slang Arabic Text

Emad E. Abdallah[1], Sarah A. Abo-Suaileek[2]
Department of Computer Information System, Faculty of Prince Al-Hussein Bin Abdallah II for Information Technology
Hashemite University, Zarqa 13115, Jordan

***Abstract*—The increased number of Arab users on microblogging services who use Arabic language to write and read has triggered several researchers to study the posted data and discover the user's opinion and feelings to support decision making. In this paper, a sentiment analysis framework is presented for slang Arabic text. A new dataset with Jordanian dialect is presented. Numerous specific Arabic features are shown with their impact on slang Arabic Tweets. The new set of features consists of lexicon, writing style, grammatical and emotional features. Several experiments are conducted to test the performance of the proposed scheme. The new proposed scheme produces better results in comparison with others. The experiments show that the system performs well without translating the tweets to English or standard Arabic.**

Keywords*—Sentiment analysis; Arabic features; opinion mining; emotional features; social media

I. Introduction

Social media has become a powerful source of information and part of our daily life. Social media with huge volume of data attract more people every day from different cultures, societies and languages. These data could be analyzed to capture valuable information about various topics. Sentiment analysis (SA) has become gradually popular and turns into an excellent source of information for companies, designers and sales representative. Twitter with 500 million users has turn into a great source to discover the user's opinion, emotions and feelings about services, products, political problems, or any other issues and possibly to create a framework to deal with it in future. Twitter gives users the ability to share their opinion in a short-term message with maximum 140 characters [1].

Sentiment analysis for English text has been researched heavily and several public datasets have been created and made publicly available. For example Stanford twitter sentiment corpus (STS), health care reform HCR [2]. Different types of features like lexicon [3, 4], emotional [5], n-grams [6], part of speech tags, semantic features [7] are used. Most of the sentiment analysis datasets have used positive and negative labels, but some datasets study the neutral and mixed labels.

Sentiment analysis for foreign languages received very little attention in comparison with English language. SA for Arabic language has not been researched seriously due to the language nature. It has many difficulties including the complexity of Arabic grammars, multi-meaning of a single word (ex: "عين المي" The word " عين" means eye and the right meaning in this clause is water source), multi-accents different meaning (ex: "مناظر بتجنن" in Jordanian accent the word "بتجنن" means very beautiful, the same word in Saudi accent is " وايد حلوة", in Egyptian accent "جميلة ئوي"). Other primary problem is the standard public list. There is no standard list for negative, or positive words. Moreover, the slang text make it harder to analyse or even categorized. Previous research in this area translate the Arabic text to English then apply on of the English SA [8-11] without building any special features for the Arabic language.

The reminder of this article is ordered as follows. In Section II, the related work of sentiment analysis and the motivation of the research is provided. In Section III, the methodology of the proposed approach is presented and the feature extraction process is described. Essential steps for feature extraction and classification process are described in Section IV and Section V, respectively. In Section VI, the experimental results are shown to validate the performance of the proposed approach.

II. Related work

In this section, several sentiment analysis techniques for Twitter messages in different languages are presented. Numerous features and tools are described. Existing approaches in SA can be gathered into three main categories: knowledge-based, statistical, and hybrid approaches [12]. Knowledge-based techniques categorize text by classes based on the presence of explicit words [13]. Statistical techniques influence on elements from machine learning such as SVM [14]. Hybrid techniques influence on both machine learning and elements from knowledge representation such as semantic networks [15]. One of the early approaches in the field is presented in [16]. The emotions in the tweets was the focus for their automatic classification algorithm. Supervised learning machine and distant supervision is used. The noisy labels are used in training process. The accuracy over different classifiers like Nivea Bayes, Maximum Entropy and SVM was around 80%.

Another approach presented in [17] studied how twitter can be used for opinion mining purposes. Linguistic analyses are performed over automatically collected corpus to discover singularities and to train sentiment classifier. The operators use syntactic structures to define emotions. In [18] the parts of speech (POS) for specific prior polarity features are introduced. The approach uses unigram features as baseline and creates two types of models, tree kernel and feature based models. The best performance achieved with the specific prior polarity and their tags feature. Ensembles classifier with lexicon is proposed

in [19]. The experiments show that the best feature come from bag of words.

Linguistic features in sentiment analysis for English language are studied in [6]. The idea is to present a comparison between POS tags and different linguistic features like lexicons and microblogging features. The experiments show that the POS tags features alone is not powerful enough in comparison with a combination of linguistic features. The impact of semantic features on sentiment analysis area is studied in [8]. Three different sematic features are used for the analysis. The replacement, augmentation and the interpolation. The best results are achieved when interpolating the semantic concept into the unigram language model. The semantic features with the Unigrams and POS sequence are compared and the results show that the semantic feature model outperforms the Unigram and POS.

Two different languages in one application is presented in [20]. The Chinese and English tweets are studied in this approach. The IK Analyzer is employed as a tool for Chinese words segmentation, then the words are trimmed down and select some features using chi-square to modify the model. Comment reviews are collected from very popular movies. English comments are collected from Facebook and tweeter. The accuracy achieved by SVM is higher than N-gram classifier. The accuracy for English comments higher than Chinese comments. The problem of above is that the tweets are not collected based on specific trend hashtag in china or English languages.

Another approach that is built for two languages is presented in [21]. SA application is shown for English and Spanish languages using multilingual hybrid features and machine translated data. One of the interesting results is that the linguistic features is very helpful when moving to other language. However, some linguistic tasks like tokenization and remove stop words may have bad influence on performance. Additional results show that a list of expression that captor strength of polarity in the tweets can reached using unigram and bigram. They showed experimentally that combining the two languages using joint classifiers can assist to increase the performance by removing noisy features.

A language-independent framework is introduced to serve as classifier without giving much care to emotions in text [3]. Semi supervised heuristic labeling and content based features is used. The data set contain tweets in English, German, French and Portuguese languages. The approach achieved good results and it is possible to be applied on new languages.

Sentiment analysis from machine learning perspective for Arabic language is introduced in [10]. The collected dataset consist of 2591 tweet/comment from Twitter and Facebook. Bigram feature is the main feature for their research without creating special features designed for Arabic text. Attention for precision and recall are given special attention. The best results are achieved by SVM classifier.

Motivated by the need for Sentiment analysis approach designed for Arabic Language. We presents the first technique that deals with slang Arabic text in twitter with specific features designed for this purpose. The experiments show that the system performs well without translating the tweets to English or standard Arabic.

III. Methodology

The novelty of our work concentrated on building a framework that deals with Arabic slang text that is wildly used on social media. Several specific Arabic features (SAF) categories is presented. Political Jordanian Arabic dataset (PJAD) is collected to be used for training and for simulation experiments. The dataset is collected using a specially developed tool called tweet collection tool (TCT). The performance of the proposed SFA is evaluated using several machine learning algorithms. Fig. 1 shows the main steps of the Arabic sentiment analysis process.

The Arabic special features are based on four categories: lexicon, writing style, grammatical, and emotional features. The PJAD is collected using a political trend hashtags in Jordan. The Dataset consist of 2000 randomly selected tweets with 1000 positive labels and 1000 negative labels.

A. Tweet Collector Tool (TCT)

In this section, the proposed TCT java tool is presented. The goal is create a complete tool that can be used by any language for sentiment analysis purposes. The tool consist of two main phases: tweets collection and features extraction. Tweet collection phase, the application enables the user to search for any specific hashtag or word and retrieve the collected tweets of any size (ex of hashtags: "2016_انتخابات#","قمة_عمان#","رفع_الاسعار# "). The twitter API is used and it is optimized to use 4 secret keys, consumer key (API key), consumer secret (API secret), access token and access token secret. Fig. 2 shows a simple data flow of the twitter API. Second phase consist of extracting SAF from the collected dataset.

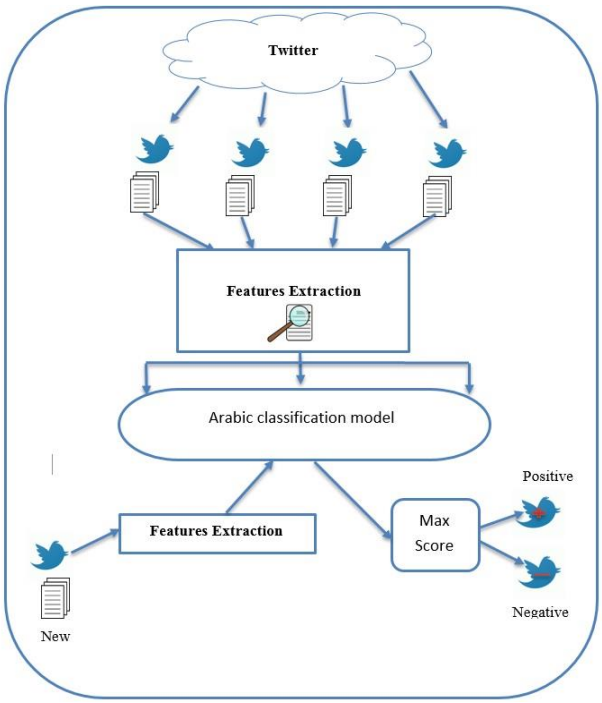


Fig. 1. Arabic Sentiment Analysis Process.

B. Political Jordanian Arabic Dataset (PJAD)

In this section, the PJAD is described; the new dataset topic is the parliamentary elections in Jordan. The tweets contains slang and standard Arabic text. The tweets have political theme and the tweeters use the Jordanian accent in their tweets. Labelling the extracted tweets is done throw several volunteers. Table I shows examples of the collected tweets with their labels. The Dataset consist of 2000 randomly selected tweets with 1000 positive labels and 1000 negative labels.

IV. Features Extraction

Features extraction is the process where important properties (features) are mined from the collected tweets. The features are used later for training purposes. Clearly, using the whole tweets is confusing and misleading to be used for training. The extracted features are mainly, a mix set of four categories: lexicon, writing style, grammatical, and emotional features. Combine all the four types of features are proved through experiments that can increase the classification performance.

A. Lexicon Features

Arabic linguistic features are studied by digging deeply in the word-character structure and how it affects the results. The words state in each individual tweet is used [22] to extract five features: word count [23] character count, word length more than 5 characters, word length more than 6 characters and word length more than 9 characters.

B. Writing Style Features

A set of features are used to extract the user mood, user style and may extract a frequent characteristic of negative or positive text. The writing style feature consists of three groups of features: special characters, occurrences of punctuations and occurrences of digits.

C. Emotional Features

Emotion side in tweets is very important factor to classify the tweets whether is it positive or negative. For Arabic language there is no standard lists for negative and positive emotions. Hence, we create our own lists. A special emotional feature is created to describe the main negative and positive words lists. It contains four features, positive words, negative words, combination of positive words, and combination of negative words. Table II shows an example of emotional features.

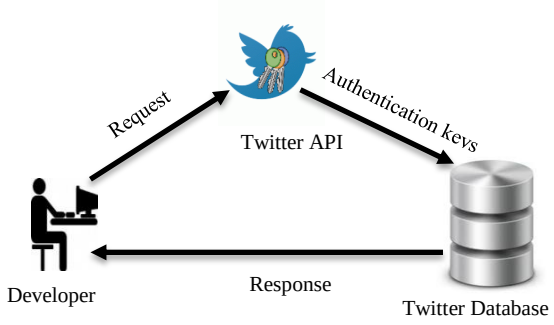


Fig. 2. Twitter API Data Flow.

TABLE I. Examples of Negative and Positive Tweets in PJAD

Polarity	Tweet
Negative	وخلي الحكومة ترفع لحد ما ربنا يرفعهم عنده
Negative	الله يرفع ضغطهم و يديق عليهم يارب
Positive	خير قدوة و خير مثال للعمل الذي يسعى الى رفع مكانه الفرد في مجتمعه
Positive	اطال الله عمرك وابقاك ذخرا للعلم وللاردن مولاي
Positive	أعجز عن وصف يليق بحق هذا الرجل
Negative	مابيجي الفساد غير من اشكالك المواطن الفاسد بنحبس وبنعرف بس انتو مش مبينين

D. Grammatical Features

Grammatical features describe the grammars that are used in Arabic tweets. It consists of seven grammar roles. The features are Kan and sisters (واخواتها كان), Enna and sisters (ان واخواتها), preposition, Plural Words, preposition, Question words and the exception word feature. In English sentiment analysis, the preposition stop words are removed. On Arabic Language it is not easy to remove preposition stop or deals with them as stop words because it is a primary part of Arabic statement.

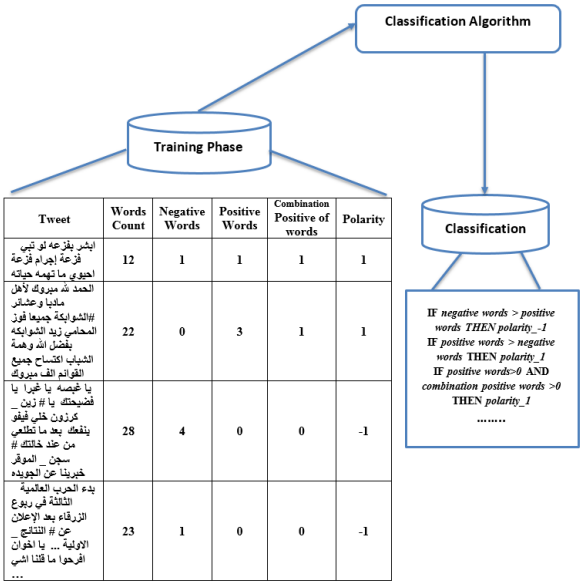
V. Classification

Classification refers to extract models that recognize significant data from classes [24]. The models define data category and classes' labels. Classification helps to determine unseen information which yield to better understanding for the data. Classification has two primary steps: Learning step and classification step. In the learning step, the model is built using several training examples form the dataset. The training examples contains two divided entities with their related features and end with the class label. Fig. 3, depicts a simplified example of the training phase. The classifier model produce general rules to be able to classify new tweet to positive or negative in the future. Testing phase evaluate a new set of data using the extracted rules that are defined from the learning phase. The performance of the classifier is evaluated in terms of classification accuracy.

Several machine learning classifiers are employed including Random forest (RF), Regression (CVR), Dagging, Multi-Class-Classifier (MC), Simple Logistic (SL), Naïve-Bayes (NB) and the MultiBoost.

TABLE II. Emotional Features

Emotional Features	Example
Positive words feature	خير , نشيطة, سعيد , سعيدة , فرح , متفائل , طموح , حلم , مودة , موده , تعاون , امل
Negative words feature	حرام , للاسف , فاتورة , مشاكل اقتصاد , فساد , مش , حرامية , راجعون
Combination of positive words	يحتار عدوك, صادقين مع, معلومه ممتازة, معك حق جميل جدا, الله يبعدنا
Combination of Negative words	يوكل هوى, ارتفاع والله حرام, حرام عليكم المحروقات , ما بتخافوا, جميلة الدعم,



Tweet	Words Count	Negative Words	Positive Words	Combination Positive of words	Polarity
ايدر بفزعه لو تبي فزعة إجرام فزعة احيوي ما تهمه حياته	12	1	1	1	1
الحمد لله مبروك لأهل ماديا وعشائر #الشوايكة جميعا فوز المحامي زيد الشوايكه بفضل الله وهمة الشباب اكتساح جميع القوائم الف مبروك	22	0	3	1	1
يا غيصه يا غبرا يا فضيحتك يا # زين _ كرزون خلي فيفق ينفعك بعد ما تطلعي من عند خالتك # سجن _ الموقر خبرينا عن الجويده	28	4	0	0	-1
بدء الحرب العالمية الثالثة في ربوع الزرقاء بعد الإعلان عن # النتائج _ الاولية ... يا اخوان افرحوا ما قلنا اشي ...	23	1	0	0	-1

Fig. 3. Example of the Training Phase.

The performance of the classifiers is assessed in terms of classification accuracy, average false positive rate, recall, precision, and f − measurements. Classification accuracy is calculated as the number of correctly classified tweets against the total number of tested tweets.

$$Accuracy = \frac{Number\ of\ \text{of correctly classified tweets}}{\text{Total number of tested tweets}}$$

False positive rate (FPR) is calculated as the percentage of all tweets anticipated incorrectly against the sum of the true negatives (TN) and the false positives (FP). The recall or true positive rate (TPR), is the percentage of positives that are correctly identified. It is calculated as the percentage of all tweets anticipated correctly against the sum of the true positive (TP) and the false negative (FN)

$$Recall = TPR = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision is the number of items correctly labeled as to the positive class divided by the total number of elements labeled as belong to the positive class. F-measurements is the average of the recall and precision.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

VI. Experimental Results

Several experiments were conducted to test the performance of the proposed scheme. Numerous feature selections and classification techniques are shown. We demonstrated the potential and the much-improved performance of the proposed technique. The new proposed PJAD dataset is used for the simulation experiments.

A. Experiment 1: Hold-Out Test

In this test, we partitioned the 2000 tweets into two independent datasets. Nearly 70% of the tweets (1400 tweets) are used to train the classifiers and build the classification model. The testing tweets (30%) are then used by the classification model. The anticipated tweets are compared with the right class and the accuracy is calculated. Table III show the number of tweets that are correctly classified, not classified correctly classified, true positive rate (TPR) and false positive rate (FPR) for each classifier. Best results are achieved by Simple Logistic classifier (71.5%), TPR (0.715) and low FPR (0.288) due to the given weights for features like liner function tends (see Fig. 4 for accuracy comparisons). Fig. 5, 6 and 7 depict the detailed analysis, we show the F-measurements, recall, and Precision over four classifiers with 70% - 30% hold-out-test.

TABLE III. Accuracy Results with Percentage Split. (30% of the Tweets are used for Testing)

Classifier	Accuracy	Correctly class.	Miss class.	TPR	FPR
NB	66.5%	400	201	0.665	0.325
AdaBoost M1	69.66%	419	182	0.697	0.303
Simple Logistic	71.5%	430	171	0.715	0.288
SVM	70.66%	425	176	0.707	0.293

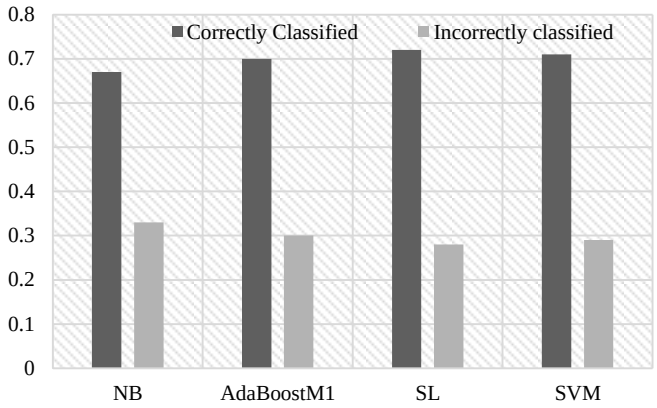


Fig. 4. Correctly and Incorrectly Classified Tweets with 70% SPLIT.

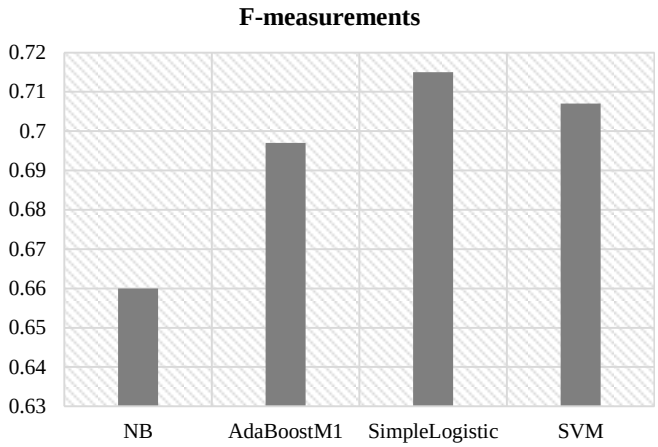


Fig. 5. F-Measurement Results Hold-Out Test (70%-30%).

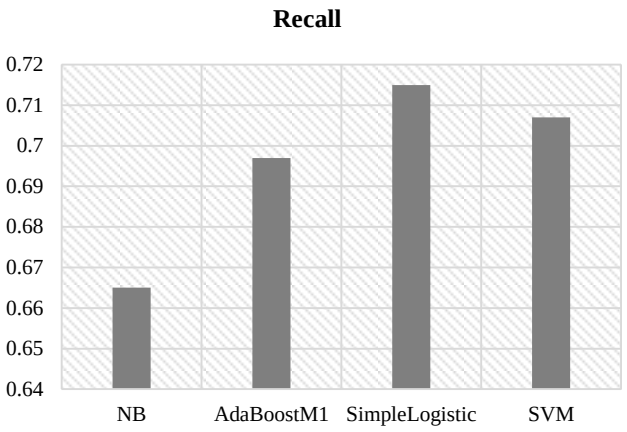


Fig. 6. Recall Results Hold-Out Test (70%-30%).

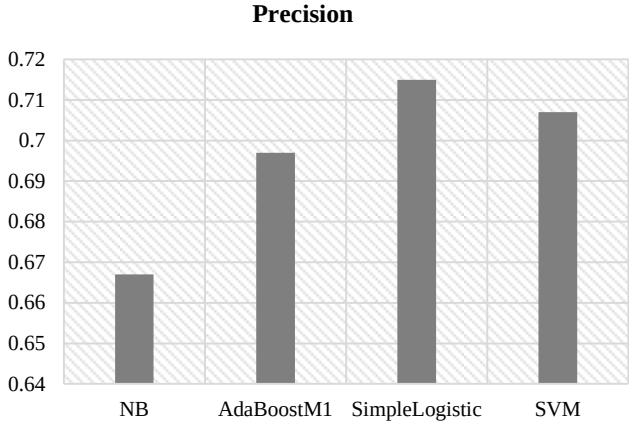


Fig. 7. Precision Results Hold-Out Test (70%-30%).

B. *Experiment 2: Cross-Validation Test*

For the second experiment, all the 2000 tweets of the dataset are used and divided into ten separate sets where nine of them are used for training and the tenth one is used for testing. The algorithm runs for ten times and the average accuracy across all the folds is calculated.

The accuracy of the correctly and incorrectly classified instances is presented in Fig. 8. Clearly, the results are acceptable for all classifiers taking into consideration that some tweets are confusing where the tweets are neutral. Table IV shows the number of tweets that are correctly classified, not classified correctly classified, true positive rate (TPR) and false positive rate (FPR) for each classifier. Best results are achieved by Dagging classifier (72.83%), TPR (0.728) and low FPR (0.272) due to it is training over SVM classifier. Several classifiers are presented to prove that the extracted features are effective and the system perform well with any classifier. Fig. 9, 10 and 11 depict the detailed analysis, we show the F-measurements, recall, and precision over several classifiers with 10 cross validation.

The cross-validation test results are often less than the hold-out test as the procedure used with the cross-validation test is to test the datasets ten times rather than one time for hold-out test. However, the results remain even with small increase. There are no unnatural changes between the two validation tests.

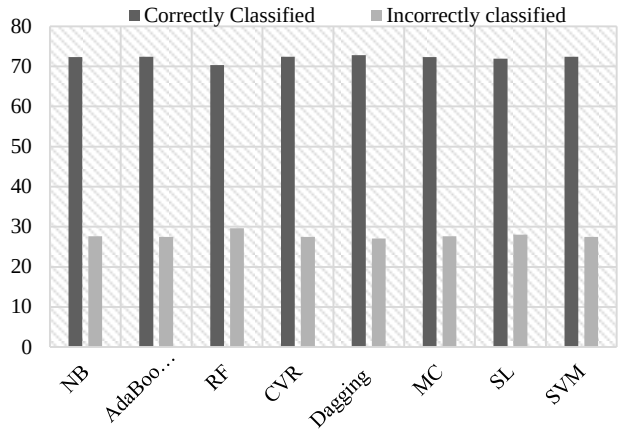


Fig. 8. Correctly and Incorrectly Classified Tweets with Cross Validation (10-Cross)

TABLE IV. ACCURACY RESULTS WITH PERCENTAGE SPLIT. (30% OF THE TWEETS ARE USED FOR TESTING)

Classifier	Accuracy	Correctly classified	Miss classified	TPR	FPR
NB	72.33%	1446	553	0.723	0.277
AdaBoostM1	72.48%	1449	550	0.725	0.275
RF	70.33%	1406	593	70.3	0.297
CVR	72.43%	1448	551	0.724	0.275
Digging	72.83%	14456	553	0.728	0.272
MC	72.33%	1446	553	0.723	0.277
SL	71.93%	1438	553	0.719	0.281
SVM	72.43%	1448	551	0.724	0.276

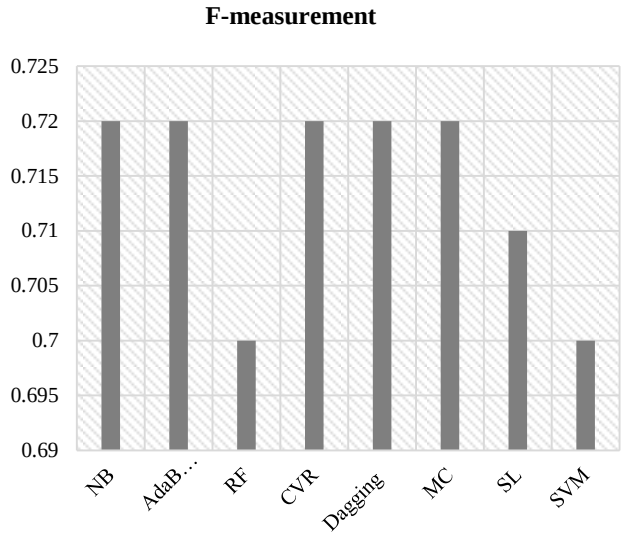


Fig. 9. F-Measurement Performance Over 2000 Tweets (10-Cross).

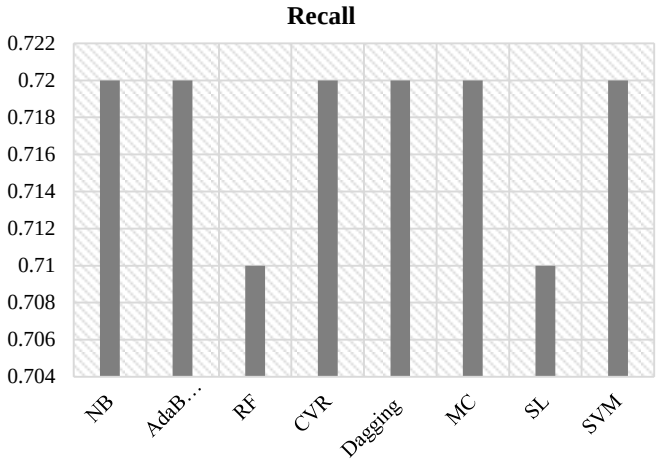


Fig. 10. Recall Performance Over 2000 Tweets (10-Cross).

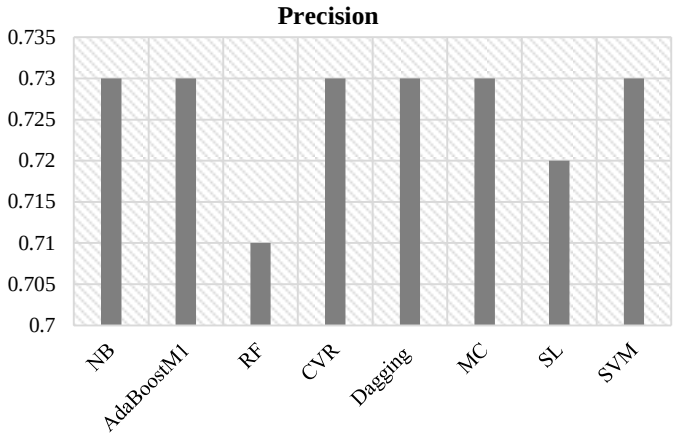


Fig. 11. Precision Performance Over 2000 Tweets (10-Cross).

It is clear from the tables above that the proposed features set is powerful for distinguishing between the positive and negative tweets with a performance accuracy of 72% and low FPR of 2%. In general, the proposed features set is proved to be effective to this problem with acceptable accuracy rate (considered high) taking into consideration, the tweets are randomly chosen and never been filtered. Moreover, the neutral tweets need to be classified as positive or negative tweet and this might confuse any trained classifier. Some Arab tweets authors have positive and negative opinion in the very same tweet, making the tweet classification is quite difficult. Different writing behaviors for each writer and the baffling words in Arabic language makes it a challenging task to recognize whether the tweet is positive or negative.

VII. Conclusion

A new sentiment analysis framework for Arabic language is proposed. The main attractive idea of the new schema is the Arabic features that deals with the complexity of the language. Four general features categories are proposed including lexicon, grammatical, writing style, and emotional features. Several numerical experiments were performed to demonstrate the potential and the much-improved performance of the proposed method. The best results achieved by Dagging classifier for political data set gathered with Jordanian accent. For future work, it would be interesting to analyze the relationship among the number of features used, classification technique and the correctly classified tweets.

References

[1] Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E. and González-Castaño, F.J., “Unsupervised method for sentiment analysis in online texts,” Expert Systems with Applications, 58, 57-75. (2016).

[2] Saif, H., Fernandez, M., He, Y., & Alani, H. "Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-Gold" (2013).

[3] Xiang, B., Zhou, L., & Reuters, T. "Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training." ACL, 2, 434-439. (2014).

[4] Mohammad, S. M., Kiritchenko, S., & Zhu, X." NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets." arXiv preprint arXiv:1308.6242. (2013).

[5] Kouloumpis, E., Wilson, T., & Moore, J. D. "Twitter sentiment analysis: The good the bad and the omg!" Icwsm, 11, 538-541. (2011).

[6] Pak, A., & Paroubek, P., “Twitter as a corpus for sentiment analysis and opinion mining” In LREc, 10, 1320-1326. (2010).

[7] Saif, H., He, Y., & Alani, H. "Semantic sentiment analysis of twitter." International Semantic Web Conference Springer Berlin Heidelberg. 508-524, (2012).

[8] Nabil, M., Aly, M., & Atiya, A. F. "Astd: Arabic sentiment tweets dataset." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2515-2519. (2015).

[9] Al-Kabi, M. N., Gigieh, A. H., Alsmadi, I. M., Wahsheh, H. A., & Haidar, M. M. "Opinion mining and analysis for arabic language. "International Journal of Advanced Computer Science and Applications (IJACSA), SAI Publisher. 5(5). (2014).

[10] Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., & Perea-Ortega, J. M. "Bilingual experiments with an arabic-english corpus for opinion mining." Recent Advances in Natural Language Processing, RANLP. (2011).

[11] Hamed, A. R., Qiu, R., & Li, D. "Analysis of the relationship between Saudi twitter posts and the Saudi stock market." .Seventh International Conference on Intelligent Computing and Information Systems (ICICIS) . 660-665. (2015).

[12] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. "New avenues in opinion mining and sentiment analysis.” IEEE Intelligent Systems, 28(2), 15-21. (2013).

[13] Ortony, A. C., & Collins, G. A. The cognitive structure of emotions. 1st edition 1988. Cambridge Univ. Press.

[14] Kim, S. M., & Hovy, E. "Identifying and analyzing judgment opinions." Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics , 200-207, (2006).

[15] Cambria, E., & Hussain, A. Sentic computing: Techniques, tools, and applications 2ed edition 2012. Springer Science & Business Media.

[16] Go, A., Bhayani, R., & Huang, L. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford, 1, 12. (2009).

[17] Narr, S., Hulfenhaus, M., & Albayrak, S." Language-independent twitter sentiment analysis." Knowledge Discovery and Machine Learning (KDML), LWA, 12-14. (2012).

[18] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. "Sentiment analysis of twitter data." Proceedings of the workshop on languages in social media Association for Computational Linguistics. 30-38, (2011).

[19] Da Silva, N. F., Hruschka, E. R., & Hruschka, E. R. "Tweet sentiment analysis with classifier ensembles." Decision Support Systems, 66, 170-179. (2014).

[20] Yan, G., He, W., Shen, J., & Tang, C."A bilingual approach for conducting Chinese and English social media sentiment analysis." Computer Networks, 75, 491-503. (2014).

[21] Balahur, A., & Perea-Ortega, J. M. "Sentiment analysis system adaptation for multilingual processing: The case of tweets." Information Processing & Management, 51(4), 547-556. (2015).

[22] Otoom, A. F., Abdallah, E. E., Hammad, M., Bsoul, M., & Abdallah, A. E. "An intelligent system for author attribution based on a hybrid feature set." International Journal of Advanced Intelligence Paradigms, 6(4), 328-345. (2014).

[23] Abdallah, E. E., Abdallah, A. E., Bsoul, M., Otoom, A. F., & Al–Daoud, E. "Simplified features for email authorship identification." International Journal of Security and Networks, 8(2), 72-81. (2013).

[24] Alpaydin, Ethem "Introduction to Machine Learning." MIT Press. p. 9. ISBN 978-0-262-01243-0. (2010).