Enhancing the Arabic Sentiment Analysis Using Different Preprocessing Operators

Ghadeer AL-Sukkar, Ibrahim Aljarah, Hamad Alsawalqah
The University of Jordan
Amman, 11942, Jordan
ghadeeralsukkar@gmail.com, i.aljarah@ju.edu.jo, h.sawalqah@ju.edu

Abstract

The last decade has seen a huge rise in social media usage and influence. Twitter is one of the highest growing social media web sites in the world where people write tweets to represent their ideas, feedback about some services or products, and their opinions about something. Many organizations and governments need to analyze these tweets to obtain some knowledge for improving their decision-making process. Sentiment analysis is one of the most famous techniques for analyzing data generated by users on social media sites to extract useful information to help in the decision-making process. Most of the sentiment analysis researches focus on the standard English language, and neglected other languages. Moreover, the number of Arabic users on Twitter who write their comments or tweets in Arabic language has increased. Accordingly, this paper focuses on sentiment analysis for Arabic tweets. We analyze collection of Arabic tweets to classify the polarity if the tweet has positive or negative sentiments. To do that we have applied different supervised machine learning techniques to classify these tweets as positive or negative polarities such as Support Vector Machine (SVM), Naïve Bayes (NB), and Decision Tree (DT). Furthermore, we applied different combinations of preprocessing techniques to enhance the quality of the sentiment analysis results. The experimental results show the impact of the preprocessing in achieving better results.

Keywords: Sentiment Analysis, Twitter, Arabic Tweets, Machine Learning

1 INTRODUCTION

The Social media has become the most widely used for sharing news, ideas, information, and opinions. The number of Arabic users in social media has increased, they write their comments and posts in Arabic language [1]. Many organization and governments need to analyze these posts and comment in order to obtain some knowledge about user's feeling and opinions. Although these opinions are meant to be helpful, the massive availability of such opinions and their unstructured nature make it difficult for companies to benefit from them. To solve this issue, a number of techniques for analyzing data generated by users on social media sites have been developed. Sentiment analysis which is known as opinion mining is one such recent techniques where the feeling and opinions are defined as sentiment. Sentiment analysis is used to determine if the user's post or tweet have a positive or a negative polarity. Analyzing these sentiments can provide useful insights to help organizations in creating a competitive advantage over their competitors and to improve the quality of their decisions. Different sentiment analysis techniques for classifying a set of tweet or posts as positive or negative sentiments have been proposed [1]. Machine learning algorithms help in classifying these tweets or posts as positive or negative opinions.

Duwairi et al. worked on sentiment analysis for Arabic tweets [1]. They have used crowd sourcing to collect a large dataset of tweets. They have developed a framework to analyze Twitter comments as positive, negative or neutral sentiments. They also applied three classification techniques: Naïve Bayes, k-nearest classifier, and Support Vector Machines (SVM). Their results showed that the Naïve Bayes had achieved the best results. The authors in [4] built a framework for mining Arabic tweets to measure customer satisfaction toward telecom companies in Saudi Arabia. They have used different preprocessing stages such as normalization, n-grams, and contextual rules in order to improve the classifier accuracy. Al-Ayyoub et al. proposed Multi Way Sentiment Analysis (MWSA) method for Arabic reviews [6]. They presented two different hierarchical structures and then compared these methods with the flat structure based methods using different datasets. Their results showed that the hierarchical classifiers gave enhancement over flat classifiers. In [8], the authors proposed a new model for sentiment analysis of Saudi Arabic-standard and Arabian Gulf dialect tweets to get a feedback from Mubasher products. They proposed a hybrid approach of machine learning techniques and natural language processing to build models for classifying tweet polarity. The authors applied many preprocessing operations such as Term Frequency-Inverse Document Frequency (TF-IDF) and Binary-Term Occurrence (BTO) then they have applied n-grams to enhance the quality of the sentiment. The authors in [9] applied combination of semantic approach and the Arabic linguistic

features. They preprocessed the tweets, and then applied a semantic approach for classification. Moreover, they proposed a new technique which combines lexicons for Arabic and English language in order to classify the Arabic tweets polarity.

In this paper, we will analyze Arabic Twitter users and their tweets using three different machine learning algorithms: Support Vector Machine (SVM), Naïve Bayes (NB), and Decision Tree (DT). In addition, four different preprocessing techniques are applied to enhance the results. We evaluate the results using different measures such as accuracy, precision, recall, and F-measure.

The remainder of this paper is organized as follows: Section 2 describes the proposed methodology. Section 3 demonstrates the experiments and results. Finally, we present the conclusion and future directions in Section 4.

2 PROPOSED METHODOLOGY

The proposed sentiment analysis methodology consists of three steps, as shown in Figure 1 using UML activity diagram.

2.1 Arabic Tweets Dataset

In this paper, we have downloaded a dataset from UCI[1]. The data set contains 2000 Arabic tweets, 1000 positive tweets and 1000 negative ones, and was collected using a tweet crawler in 2014. These tweets include feelings written in Modern Standard Arabic (MSA) and the Jordanian dialect.

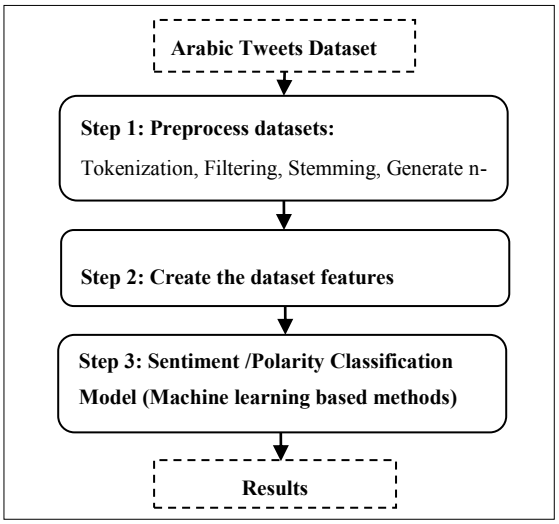

Fig. 1: The workflow of the proposed method in a UML Activity diagram

[1] http://archive.ics.uci.edu/ml/ ,

2.2 Preprocessing

Each tweet is preprocessed in order to apply the supervised machine learning techniques to classify tweets as positive or negative sentiments. We have applied different cases for text preprocessing in order to specify, which case have highest accuracy. Text preprocessing is very important and essential for sentiment analysis. The following are the main steps of the preprocessing stage:

- Tokenization: this step is performed for each tweet in order to divide the tweet into multiple tokens/words based on whitespaces characters.
- Filtering: this step done by filtering the token which is less than length a threshold such as to remove the tokens with length less than 3 characters. Furthermore, we remove the Arabic stop words.
- Stemming: light stemming is used to reduce the feature space.
- Generate n-grams (Terms): creates n-Grams of the terms in a document such as 2-grams, 3-grams, etc.

2.3 Create Dataset Features

The tweets dataset is converted into a matrix where rows represent the tweets and columns represent the features/words. The matrix values are created based on different schemas: TF-IDF, Term Frequency, and Term occurrence.

2.4 Sentiment Classification Model Construction

We have applied Naïve Bayes (NB), Support Vector machine (SVM), and Decision Tree (DT) to construct the model. Furthermore, we have validated the model using the 10-cross validation technique that can be easily implemented. For each machine learning technique, different combinations of the preprocessing have been applied with the following cases:

- With stemming and stop words filtering.
- Without stemming but with stop words filtering.
- With n-gram, stemming and stop words filtering.
- With n-gram, stemming, and without stop words filtering.
- With n-gram but without stemming and without stop words filtering.
- With n gram but without stemming and stop words filtering.

3 EXPERIMENTAL RESULTS

3.1 Environment

We used 32-bit operating system and Rapiedminer[2] 6.5 software, which helps in processing Arabic text and classifying each tweet sentiments. Rapiedminer has different operators which help in text processing, especially it has operators for Arabic text preprocessing such as stemming (Arabic, light), tokenization, and filter (tokenize length), filter (Arabic Stop words), and n-gram for token.

3.2 Evaluation Criteria

We used different evaluation criteria for each case in order to specify which technique has the highest results of sentiment classification. The evaluation criteria in this research are as following:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

$$Precision = TP/(TP + FP) \quad (2)$$

$$Recall = TP/(TP + FN) \quad (3)$$

$$F - Measure = (2 * Precision * Recall)/(Precision + Recall) \quad (4)$$

Where *TP* is true positive; *TN*: true negative, *FP*: false positive, and *FN*: false negative.

3.3 Results

For each case that we have applied, there are different results was observed. Tables 1, 2, and 3 show the experimental results using different classifiers and different preprocessing combination.

Table 1 shows the result of the decision tree classifier. As noted in the table, the best results can be achieved with stemming, unigram and without stop words filter. Table 2 shows the results of Naïve Bayes, which obtained the best results with stemming, bigram, and stop words filtering. Table 3 shows the sentiment analysis results using support vector machines method. The best results are obtained without stop words filtering, unigram, and stemming. Table 4 shows the summary of all cases and shows that the best results are achieved with support vector machines.

TABLE 1: DECISION TREE RESULTS

n gram	Filter Stop words	Stemming	Accuracy	Recall	Precision	F-Measure
Unigram	√	√	67.69%	0.6812	0.7775	0.726171248
Bigram	√	√	66.88%	0.67215	0.71765	0.694155199
Trigram	√	√	66.98%	0.67315	0.71895	0.695296591
Unigram	√	__	65.97%	0.66425	0.7722	0.714168749
Bigram	√	__	66.17%	0.66625	0.77315	0.715730426
Trigram	√	__	66.02%	0.66475	0.7724	0.714543228
Unigram	__	√	**68.91%**	0.6932	0.78365	0.735655
Bigram	__	√	68.46%	0.68765	0.7292	0.707816
Trigram	__	√	68.40%	0.68715	0.72835	0.70715
Unigram	__	__	66.78%	0.67225	0.77505	0.719999
Bigram	__	__	67.14%	0.67575	0.77685	0.722782
Trigram	__	__	67.14%	0.67575	0.77685	0.722782

[2] www.rapidminer.com

TABLE 2: NAÏVE BAYSE RESULTS

n-gram	Filter Stop words	Stemming	Accuracy	Recall	Precision	F-Measure
Unigram	√	√	**81.14%**	0.81	0.82245	0.816178
Bigram	√	√	**81.39%**	0.8126	0.8247	0.818605
Trigram	√	√	**81.39%**	0.8125	0.82575	0.819071
Unigram	√	__	78.25%	0.78065	0.80105	0.790718
Bigram	√	__	78.75%	0.7857	0.8075	0.796451
Trigram	√	__	78.70%	0.7851	0.80855	0.796652
Bigram	__	√	81.19%	0.81055	0.82235	0.816407
Trigram	__	√	81.14%	0.81	0.8227	0.816301
Unigram	__	__	78.09%	0.77915	0.7996	0.789243
Bigram	__	__	78.60%	0.78415	0.80605	0.794949
Trigram	__	__	78.60%	0.7841	0.8078	0.795774

TABLE 3: SUPPORT VECTOR MACHINE RESULTS

n gram	Filter Stop words	Stemming	Accuracy	Recall	Precision	F-Measure
Unigram	√	√	83.01%	0.82845	0.8489	0.83855
Bigram	√	√	70.28%	0.69875	0.79955	0.745759
Trigram	√	√	60.29%	0.5972	0.78045	0.676637
Unigram	√	__	77.69%	0.7742	0.82085	0.796843
Bigram	√	__	77.69%	0.7742	0.82085	0.796843
Trigram	√	__	56.85%	0.56225	0.7661	0.648533
Unigram	__	√	**83.16%**	0.8301	0.84765	0.838783
Bigram	__	√	71.50%	0.7112	0.80065	0.753279
Trigram	__	√	60.90%	0.6034	0.7799	0.68039
Unigram	__	__	78.14%	0.77895	0.8194	0.798663
Bigram	__	__	64.50%	0.64005	0.77755	0.702132
Trigram	__	__	57.15%	0.56535	0.76715	0.650969

TABLE 4: SUMMARY RESULTS

Classifier	n gram	Filter Stop words	Stemming	Accuracy	Recall	Precision	F-Measure
SVM	**Unigram**	**__**	**√**	**83.16%**	**0.8301**	**0.84765**	**0.838783**
NB	Bigram/ Trigram	√	√	81.39%	0.8126	0.8247	0.818605
DT	Unigram	__	√	68.91%	0.6932	0.78365	0.735655

4 CONCLUSION

Sentiment analysis is one of the fastest growing research areas which uses the natural language processing, text mining and computational linguistic to extract useful information to help in the decision-making process. In this paper, sentiment analysis is applied on Arabic users on Twitter to analyze their Arabic tweets. The proposed sentiment analysis method combines different classification and preprocessing techniques to classify tweets. The results show that when we used Decision Tree the highest accuracy was without filter Stop words and with unigram without filter Stop words. For Naïve Bayes result show that highest accuracy was with filter Stop words, Stemming and with bigram and trigram. Support vector machine results show that highest accuracy was without filter Stop words and with unigram without filter stop words (*Best One*). For future work, we plan to use Neural Network and compare its results with the previous results. Moreover, we plan to apply the same methodology

5 REFERENCES

[1] R. M. Duwairi, Raed Marji, Narmeen Sha'ban, Sally Rushaidat" Sentiment Analysis in Arabic Tweets", 2014 5th IEEE.

[2] Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117-126.

[3] Duwairi, R. M., Alfaqeh, M., Wardat, M., & Alrabadi, A. (2016, April). Sentiment analysis for Arabizi text. In *Information and Communication Systems (ICICS), 2016 7th International Conference on* (pp. 127-132). IEEE.

[4] Almuqren, L., & Cristea, A. I. (2016, July). Framework for sentiment analysis of Arabic text. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media* (pp. 315-317). ACM.

[5] Al-Horaibi, L., & Khan, M. B. (2016, July). Sentiment analysis of Arabic tweets using text mining techniques. In *First International Workshop on Pattern Recognition* (pp. 100111F-100111F). International Society for Optics and Photonics.

[6] Al-Ayyoub, M., Nuseir, A., Kanaan, G., & Al-Shalabi, R. (2016). Hierar-chical classifiers for multi-way sentiment analysis of arabic reviews. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *7*(2), 531-539.

[7] Al-Kabi, M., Gigieh, A., Alsmadi, I., Wahsheh, H., & Haidar, M. (2013). An opinion analysis tool for colloquial and standard Arabic. In *The Fourth International Conference on Information and Communication Systems (ICICS 2013)* (pp. 23-25).

[8] Al-Rubaiee, H., Qiu, R., & Li, D. (2016, March). Identifying Mubasher software products through sentiment analysis of Arabic tweets. In *Industrial Informatics and Computer Systems (CIICS), 2016 International Conference on* (pp. 1-6). IEEE..

[9] Al-Horaibi, L., & Khan, M. B. (2016). Sentiment Analysis of Arabic Tweets Using Semantic Resources. *International Journal of Computing & Information Sciences*, *12*(2), 149.