2015 6th International Conference on Information and Communication Systems (ICICS)

Sentiment Analysis for Dialectical Arabic

Rehab M. Duwairi
Department of Computer Information Systems
Jordan University of Science and Technology
Irbid 22110, Jordan
rehab@just.edu.jo

***Abstract*— This article investigates sentiment analysis in Arabic tweets with the presence of dialectical words. Sentiment analysis deals with extracting opinionated phrases from reviews, comments or tweets. i.e. to decide whether a given review or comment is positive, negative or neutral. Sentiment analysis has many applications and is very vital for many organizations. In this article, we utilize machine learning techniques to determine the polarity of tweets written in Arabic with the presence of dialects. Dialectical Arabic is abundantly present in social media and micro blogging channels. Dialectical Arabic presents challenges for topical classifications and for sentiment analysis. One example of such challenges is that stemming algorithms do not perform well with dialectical words. Another example is that dialectical Arabic uses an extended set of stopwords. In this research we introduce a framework that is capable of performing sentiment analysis on tweets written using either Modern Standard Arabic or Jordanian dialectical Arabic. The core of this framework is a dialect lexicon which maps dialectical words into their corresponding Modern Standard Arabic words. The experimentation reveals that the dialect lexicon improves the accuracies of the classifiers.**

Keywords*— *Sentiment Analysis, Opinion Mining, Modern Standard Arabic, Dialectical Arabic, Text Mining

I. Introduction

The Semantic Web and social medial channels provided tools and resources for the users to freely comment on various aspects of their lives: sites they visit, products they buy and restaurants they eat at to list few examples. These comments are vital for business owners and organizations as they provide an indication of the satisfaction or acceptance of products or services. For this reason, several businesses and organizations have embarked on developing tools to mine and analyze these comments. Manual analysis is not feasible due to the large amounts of comments and due to time constraints on businesses.

Sentiment Analysis is a field of science that attempts to automatically or semi-automatically determine attitude polarity of phrases embedded in comments. Generally speaking there are two approaches for sentiment analysis. Firstly, unsupervised learning approaches that rely on sentiment lexicons such as the work reported in [7, 15, 16, 18, 19, 21, 22, 27, 28, 29, 33]. Secondly, supervised learning approaches that rely on classification such as the work reported in [1, 3, 4, 5, 6, 10, 11]. Sentiment Analysis can be performed at the whole review or comment [22, 29] or at the aspect or feature level such as the screen resolution of a phone [8, 14, 23, 24]. This latter type of sentiment analysis provides fine grained analysis and has wider applications.

Twitter is a micro-blogging tool that allows users to send and read short messages called tweets [30]. A truly large number of Arabic tweets are generated on daily basis which makes Twitter an ideal choice for the research reported in this paper. Sentiment is language and culture dependent. Any successful sentiment analysis tool should take culture and language aspects into consideration. Arabic language comes in three varieties [12]: Traditional Arabic found in religious scripts, Modern Standard Arabic (MSA) used in formal events and dialectical or colloquial Arabic which is typically spoken not written and is region dependent. A close examination of the comments or reviews posted on social media channels will reveal that dialectical Arabic is present in these written comments.

The reported research utilizes machine learning, classification to be specific, to detect polarity of tweets written in Arabic. To this end, two classifiers, namely: the Naïve Bayes (NB) and the Support Vector Machine (SVM) classifiers were used. Also, as dialect is present in tweets, we decided to handle dialect by translating dialectical words into their corresponding MSA words. The translation utilizes a dialect lexicon that was created for this purpose. We extracted this lexicon from a dataset of 22550 tweets written in Arabic. Every tweet was tokenized into words; these words were examined by two annotators who decided whether a given word is written using MSA or dialect Arabic. For dialectical words, the annotators provided their corresponding MSA words. The experimentation focused on determining the value of this lexicon by first executing the classification tasks without using the dialect lexicon and second by executing the classification tasks with converting the dialectical words to MSA words. The results reveal that the dialect lexicon has a positive impact on the Macro-Precision, Macro-Recall and F-Measure. The results also reveal that the F-measure of the Positive and Negative classes greatly benefited from the dialect lexicon in contrary to the Neutral class.

This paper is organized as follows. Section 1 has provided an introduction to this paper. Section 2, by comparison, describes related work. Section 3 explains, in details, our suggested framework. Section 4, summarizes the experiments and analyzes the results. Finally, Section 5 summarizes the conclusions of this paper and highlights future work.

978-1-4799-7348-4/15/$31.00 ©2015 IEEE

II. Background and Related Work

This section describes related research which addressed sentiment analysis in Arabic reviews. Abbasi et al [1] analyzed sentiment embedded in blogs which are written either in Arabic or English on web forums. The aim of the study was to detect hostility. The authors represented each review using a set of syntactic and stylistic features.

El-Halees [10], by comparison, collected 1143 posts that cover three topics, namely: Education, politics and sports. These posts consist of 8793 sentences. The author uses a sequence of three classifiers to classify the documents in order to increase the accuracy. The final reported accuracy was 80%.

Farra et al [13] proposed a method which targets sentences in Arabic documents. This approach relies on a set of features to represent each sentence such as frequency of positive, negative and neutral words in every sentence, frequency of negations, use of special characters, and frequency of contradiction words. Subsequently, the feature vectors of the documents were fed to the J48 Decision Tree classifier and the accuracy of classification was 62%.

Rushdi-Saleh et al. [25] used classification for sentiment analysis, in particular, the SVM and NB classifiers were used. Their dataset consists of 500 movie reviews written in Arabic. The accuracy of the SVM classifier was 90% and the accuracy of the NB classifier was 84%

SAMAR [2] a two-stage classifier first distinguishes subjective sentences from objective ones. Subsequently, it classifies subjective sentences as positive or negative. The SVM light classifier was used for both stages. The dataset that they have experimented with consists of 8940 sentences.

Mourad and Darwish [20] used the NB classifier to extract sentiment embedded in a dataset of 2300 tweets. In their experimental setup, several alternatives of features were used such as stemming the words of tweets, part-of-speech tags, bi-grams and so on. The accuracy of subjectivity detection, i.e. deciding whether a tweet is subjective or objective, was equal to 76.6%. The accuracy of polarity detection, i.e. deciding whether a tweet is positive or negative, was equal to 80.5.

Shoukry and Refae in [26] used a dataset that consists of a 1000 tweets (500 tweets are positives and 500 tweets are negative). They targeted sentence-level sentiment analysis since a tweet length is restricted to 140 characters. The Neutral class was not addressed by their research. Also, the corpus they used is small. The authors appended some Egyptian words alongside the Modern Standard Arabic to investigate the effects of dialectical Arabic on accuracy. The SVM and NB classifiers were used for polarity classification. The results show that SVM outperformed NB in sentiment analysis with accuracy equals 72.6%.

The reported research uses machine learning, classification to be specific, for sentiment analysis of tweets with the presence of Jordanian dialect. Thus it deviates from the above listed work in that it utilizes a data-driven dialect lexicon; and it employs a large dataset compared with existing work.

III. The Fraework of Sentiment Analysis

A. Overview

The described framework consists of the following phases:

1. *Data collection and annotation*: here 22550 tweets were collected using Twitter API [31] and annotated using the Crowdsourcing Tool described in [9].

2. *Tweet Preprocessing*: during this phase, every tweet is tokenized into words; this is followed by removing stopwords except negation. Also, during this phase emoticons are converted to their corresponding words by using a specialized mapping table that maps common emoticons to their respective words. Table 1 shows an example of emoticons and their corresponding words and polarity labels. Finally, tokens are stemmed using the Khoja stemmer [17].

3. *Classification*: here the dataset is divided into two subsets. A training subset which is used to build the classification models of NB and SVM and a testing subset which is used to test the accuracy of the classifiers. The weight of every token or word is determined using the Binary Model – where a token is given a weight equals 1 if it is present in the tweet under consideration or is given a weight equals 0 if the token is absent from the tweet. TF-IDF can be used as an alternative to the Binary Model but for sentiment analysis the Binary Model has been used by several researchers. Two versions of the dataset are maintained. The first version consists of the tweets as collected without removal of dialectical words. The second version consists of tweets after replacing dialectical words with their corresponding MSA words by utilizing the dialect lexicon which was built for this purpose. The classification task was executed twice; once using the tweets with dialectical words and once using the tweets after translating the dialectical words.

4. *Results Analysis*: During this phase Precision, Recall and F-Measure were calculated for both classification tasks to assess the effects of translating dialectical words on accuracy.

Table 1: Examples of Emoticons and Their Corresponding Words in Arabic

Emoticon/ Shortcut	Corresponding Arabic Word	Label/ Weight
:'(	بكاء	-1
O:)	وجهه ملائكي(بريئ)	1
3:)	وجهه شيطاني	-1
>:(	وجهه غاضب	-1
^_^	وجهه سعيد جدا	1
o.o	وجهه مرتبك	-1
:)	سعيد	1

B. Dialect Lexicon

We have used a data driven approach to build the dialect lexicon. Firstly, the 22550 tweets were tokenized into words. These words were separated into two subsets by two annotators. The first subset consists of dialectical words and the second subset consists of MSA words. Secondly, the annotators were instructed to translate every dialectical word to its most suitable MSA word. After the second step is over, every dialectical word is associated with its corresponding MSA word. Table 2 displays some statistics about the dialect lexicon. 17.91% of the words used to express positive polarity were dialectical words. 9.12% of words that were used to express negative polarity were dialectical words. 6.03% of neutral words were dialectical words. The percentage of dialect words in the whole dataset for the three classes is 11.5%. The last column in Table 2 expresses the number of distinct dialectical words per class. Table 3 shows a few examples of Jordanian Dialectical words and their corresponding MSA words.

TABLE 2: STATISTICAL PROPERTIES OF THE DIALECT LEXICON

Class Name	Number of all words	Number of Dialectical Words with Repetition	% of Dialectical Words	Number of Distinct Dialectical Words
Positive	87400	15650	17.91%	7987
Negative	73600	6711	9.12%	3561
Neutral	69900	4212	6.03%	4872
Total	230900	26550	11.50%	16420

TABLE 3: EXAMPLES OF DIALECTICAL WORDS AND THEIR CORRESPONDING MSA WORDS

Dialectical Words	Their Corresponding MSA Words
علشان	من اجل
بدها	تريد
رحت	ذهبت
ليش	لماذا
شو	ماذا
بس	فقط
كمك	طرف رداؤك
شوية	قليلا

C. *Dataset*

In this paper, we have collected and annotated a dataset of Arabic tweets which consists of 22550 tweets. This dataset was gathered using Twitter API [31] and was annotated using the Crowdsourcing Tool described in [9]. This dataset is called the ArabicDataset. Table 4 shows the distribution of tweets in the ArabicDataset over the three polarity classes, namely: Positive, Negative and Neutral. As it can be seen from Table 4, 8529 tweets were classified as positive tweets (i.e. they reflect positive sentiment), 7021 tweets were classified to belong to the Negative class and 7000 tweets were classified to belong to the Neutral class. The labels that are assigned to the tweets using the Crowdsourcing Tool are considered the true labels and the predicted labels by the classifiers are compared against these labels to calculate the accuracy of the framework.

TABLE 4: DISTRIBUTION OF TWEETS OVER POLARITY LABELS

Label	ArabicDataset
Positive	8529
Negative	7021
Neutral	7000
Total	22550

D. *Classifiers*

The NB and SVM classifiers, as implemented in the Weka data mining tool [32], were used to classify tweets. These are well known classifiers that have been used frequently by researchers working on topical classification and sentiment analysis.

IV. EXPERIMENTATION AND RESULTS ANALYSIS

A. *Effects of the Dialect Lexicon on Overall Accuracy*

Table 5 displays the Macro-Precision, Macro-Recall and F-Measure for the three classes under consideration. The second column of Table 5 shows the previous values before utilizing the dialect lexicon. The third column of Table 5 lists the accuracies after using the dialect lexicon. As it can be seen from Table 5, Macro-Precision, Macro-Recall and F-Measure of the NB classifier improved when the dialect lexicon was used. Column 4 of Table 5 shows the value of improvement. Macro-Recall greatly benefited from the lexicon with an improvement equals 0.159. This means that the dialect lexicon aided the classification of true examples of the classes. i.e. reducing the false negative rates.

TABLE 5: THE OVERALL PERFORMANCE OF THE NB CLASSIFIER WITH AND WITHOUT DIALECT LEXICON

	No Dialect Lexicon	With Dialect Lexicon	Improvement
Macro-Precision	0.885	0.905	0.02
Macro-Recall	0.838	0.997	0.159
F-Measure	0.84	0.876	0.036

Table 6, by comparison, shows the Macro-Precision, Macro-Recall and F-Measure for the three classes when the SVM classifier was used. The results shown in Table 6 are consistent with the results shown in Table 5. That is the dialect lexicon helped in improving the accuracy of the SVM classifier as well. The improvement is better in the case of NB.

TABLE 6: THE OVERALL PERFORMANCE OF THE SVM CLASSIFIER WITH AND WITHOUT DIALECT LEXICON

	No Dialect Lexicon	With Dialect Lexicon	Improvement
Macro-Precision	0.871	0.878	0.007
Macro-Recall	0.835	0.868	0.033
F-Measure	0.837	0.867	0.03

B. Effects of Dialect Lexicon on Class Accuracy

Table 7 shows the values of Precision for the Positive, Negative and Neutral classes. As it can be seen from Table 7, the dialect lexicon helped in improving the Precision of the Positive and Negative classes but not the Neutral class. In fact the precision of the Neutral class was adversely affected by the dialect lexicon. The precision of the Negative and Neutral classes are higher than the precision of the Positive class for the NB classifier. This means that the NB classifier did a great job predicting true examples and eliminating false examples of the Negative and Neutral classes.

TABLE 7: CLASS PRECISION FOR THE NB CLASSIFIER

	No Dialect Lexicon	With Dialect Lexicon	Improvement
Positive	0.701	0.758	0.057
Negative	0.997	1	0.003
Neutral	0.998	0.989	-0.009

Table 8 displays the values of Precision for the SVM classifier. Its behavior is consistent with the behavior of the NB classifier. Only the Positive class Precision was improved when the dialect lexicon was used.

TABLE 8: CLASS PRECISION FOR THE SVM CLASSIFIER

	No Dialect Lexicon	With Dialect Lexicon	Improvement
Positive	0.709	0.784	0.075
Negative	0.979	0.966	-0.013
Neutral	0.96	0.903	-0.057

Table 9 illustrates the Recall values when the NB classifier was used. As it can be seen from Table 9, only the Negative class Recall was improved with the dialect lexicon. Table 10 lists the Recall values when the SVM classifier was used. The SVM behavior is in harmony with the NB behavior. i.e. Only the Negative class Recall was improved.

TABLE 9: CLASS RECALL FOR THE NB CLASSIFIER

	No Dialect Lexicon	With Dialect Lexicon	Improvement
Positive	0.998	0.995	-0.003
Negative	0.697	0.927	0.23
Neutral	0.784	0.684	-0.1

TABLE 10: CLASS RECALL FOR THE SVM CLASSIFIER

	No Dialect Lexicon	With Dialect Lexicon	Improvement
Positive	0.964	0.935	-0.029
Negative	0.697	0.932	0.235
Neutral	0.817	0.723	-0.094

C. Results – Revisited

As it can be seen from the values displayed in Table 4 through Table 10, the Precision of the class Positive varies from 0.701 (NB without dialect lexicon) to 0.784 (SVM with dialect lexicon). By comparison, the Recall values of the Positive class vary from 0.934 (SVM with Dialect) to 0.998 (NB without dialect lexicon). We can conclude here that both NB and SVM classifiers did great job in correctly classifying true examples or instances of the Positive class (High Recall) but did less well in eliminating the false examples or instances (Relatively low Precision). Also, the dialect lexicon improved the Precision of the Positive class. However, it did not improve the Recall of the Positive class.

The Negative class scored high precision that varies from 0.966 (SVM with dialect lexicon) to 1 (NB with dialect lexicon). The Recall of the Negative class varies from 0.697(Both NB and SVM without dialect lexicon) to 0.932 (SVM with dialect lexicon). The behavior of the classifiers with the Negative class can be summarized as follows: both classifiers did great task in eliminating false examples (High Precision) and less well in classifying true examples (Low Recall without using the dialect lexicon). The dialect lexicon improved the Recall for the Negative class to reach 0.932 in the case of the SVM classifier.

The behavior of the Neutral class is similar to the behavior of the Negative class in that it secured high Precision and relatively low Recall without using the dialect lexicon. The dialect lexicon helped in improving the Recall of the Neutral class.

V. CONCLUSIONS AND FUTURE WORK

This paper has addressed sentiment analysis in Arabic tweets in the presence of dialectical words. A dataset, which consists of 22550 tweets, was collected and annotated. Dialectical words were translated into their corresponding MSA words by utilizing a dialect lexicon. This lexicon consists of dialectical words alongside their corresponding MSA words. The NB and SVM classifiers were used to determine the polarity of the tweets. These classifiers built their classification models by using two versions of the same dataset. The first version consists of tweets that contain dialectical words and the second version consists of tweets after translating the dialectical words.

The results reveal that replacing dialectical words with their corresponding MSA words improves the overall Precision, Recall and F-Measure. When examining the results at the class level, we conclude that Precision of the Positive class was slightly improved when the dialect lexicon was used with the two classifiers. The Precision of the Negative class was slightly improved when the dialect lexicon was used with the NB classifier. The Precision of the Neutral class was not improved when the dialect lexicon was used. The Recall of the Negative class was greatly improved when the dialect lexicon was used with both classifiers.

ACKNOWLEDGMENT

The Author of this paper would like to thank Mosab Alfaqeh for his help. This Project is supported by Jordan University of Science and Technology, grant No. 20130174.

REFERENCES

[1] A. Abbasi, C. Hsinchun, and S. Arab. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems (TOIS). Vol. 26, no. 3, 2008, pp. 1-34.

[2] M. Abdul-Mageed, S. Kubler, and M. Diab. SAMAR: A system for subjectivity and sentiment analysis of Arabic Social Media. 3rd Workshop on Computational Approaches for Subjectivity and Sentiment Analysis WASSA, Satellite Workshop, Jeju, Koera, June, 2012.

[3] K. Ahmad, D. Cheng, and Y. Almas. Multi-lingual Sentiment Analysis of Financial News Streams. Proceedings of the 1st International Workshop on Grid Technology for Financial Modeling and Simulation, Palermo, Italy, 2006.

[4] K. Ahmad, and Y. Almas. Visualizing Sentiments in Financial Texts. Proceedings of the Ninth International Conference on Information Visualization. 2005, PP. 363 – 368, Washington, USA.

[5] G. Alec, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009, pp. 1-12.

[6] E. Boiy, and M.F. Moens. A Machine Leaning Approach to Sentiment Analysis in Multilingual Web Texts. Information Retrieval. Vol. 12, Issue 5, 2009, pp. 526 – 558.

[7] K. Denecke. Are SentiWordNet Scores Suited for Multi-domain Sentiment Classification? In Fourth International Conference on Digital Information Management (ICDIM), 2009. 2009, pp. 33-38, Nov, Ann Arbor, MI.

[8] X. Ding, B. Liu, P.S. Yu, A holistic lexicon based approach to opinion mining. Proceedings of the International Conference on Web Search and Web Data Mining (WSDM'08), 2008, pp. 231–239.

[9] R. Duwairi, R. Marji, N. Sha'ban, S. Rushaidat. Sentiment Analysis in Arabic Tweets, Proceedings of the 5th International Conference on Information and Communication Systems, Irbid, Jordan, April 1-3, 2014.

[10] A. El-Halees A. Arabic Opinion Mining Using Combined Classification Approach. Proceedings of the International Arab Conference on Information Technology (ACIT), 2011, Riyadh, Saudi Arabia, Dec.

[11] M. Elhawary, and M. Elfeky. Mining Arabic Business Reviews. Proceedings of the IEEE International Conference on Data Mining. 2010, pp 1108 – 1113, Dec, Mountain View, USA.

[12] A. Farghaly, and K. Shaalan. Arabic Natural Language Processing: Challenges and Solutions. ACM Transactions on Asian Languages Information Processing, Vol. 8, No. 4, Article 14, 2009.

[13] N. Farra, E. Challita, R.A. Assi, H. Hajj. Sentence-Level and Document-Level Sentiment Mining for Arabic Texts. Proceedings of the ICDM Workshop, 2010. pp. 1114-1119.

[14] M. Hu, B. Liu, Mining opinion features in customer reviews. Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI), 2004, pp. 755–760.

[15] K. Jaap, M.J. Marx, J. Robert, and M.D. Rijke. Using WordNet to measure semantic orientations of adjectives. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). Vol. IV, 2004, pp. 1115–1118.

[16] J. Jiao, and Y. Zhou. Sentiment polarity analysis based multi-dictionary. Physics Procedia. Vol. 22, 2011, pp. 590 – 596, 2011.

[17] S. Khoja, and R. Garside. Stemming Arabic Text. Computing Department, Lancaster University, Lancaster,UK.1999 http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps

[18] T. Mike, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology 63, No. 1, 2012, pp. 163-173.

[19] A. Moreo, M. Romero, J.L. Castro, and J.M. Zurita. Lexicon-based comment-oriented news sentiment analyzer system. Expert Systems with Applications. Vol. 39, 2012, pp. 9166 – 9180.

[20] A. Mourad, and K. Darwish. Subjectivity and sentiment analysis of modern standard Arabic microblogs. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2013, pp. 55-64, Atlanta, Georgia, USA.

[21] B. Ohana, and B. Tierney. Sentiment Classification of Reviews using SentiWordNet. In 9th. IT & Conference, Dublin, Ireland, 2009.

[22] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002), 2002, pp. 79–86.

[23] A. Popescu, O. Etzioni, Extracting product features and opinions from reviews. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP), 2005, pp. 339–346.

[24] C. Quan, F. Ren, Unsupervised product feature extraction for feature-oriented opinion determination. Information Sciences, 272 (2014) 16–28.

[25] M.R. Saleh, M.T. Martin-Valdivia, L.A. Urena-Lopez, and J.M. Perea-Orteg A. OCA: Opinion Corpus for Arabic. Journal of the American Society for Information Science and Technology. Vol. 62, Issue. 10, 2011, pp 2045-2054.

[26] A. Shoukry, and A. Rafea. Sentence-level Arabic sentiment analysis. Proceedings of Collaboration Technologies and Systems (CTS). 2012, pp. 546-550.

[27] J. Steinberger et al. Creating sentiment dictionaries via triangulations. Decision support Systems. Vol. 53, 2012, pp. 689 – 694.

[28] S. Tan, and Q. Wu. A random walk algorithm for automatic construction of domain-oriented sentiment lexicon. Expert Systems with Applications. Vol. 38, 2011, pp. 12094 -12100.

[29] P. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002), 2002, pp. 417–424.

[30] https://twitter.com, last accessed 27-Nov-2014.

[31] Twitter API, https://dev.twitter.com/docs/api/1/get/search, last accessed 27-10-2014.

[32] http://www.cs.waikato.ac.nz/ml/weka/. Last accessed 16-Dec-2014.

[33] H. Xia, J. Tang, H. Gao, and H. Liu. Unsupervised Sentiment Analysis with Emotional Signals. Rio de Janeiro, Brazil ACM 978-1-4503-2035-1/13/05, 2013.