VOL. 10, NO. 3, FEBRUARY 2015 ISSN 1819-6608

ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ARABIC OPINION TARGET EXTRACTION FROM TWEETS

Marwa Alhazmi and Naomie Salim
Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia
E-Mail: alhazmi_m@hotmail.com

ABSTRACT

Twitter is an ocean of sentiments; users can express their opinion freely on a wide variety of topics. The unique characteristics that twitter holds introduce a different level of challenge in the field of sentiment analysis. Identifying the topic or the target of the expressed opinion is the aim of this study; Opinion target recognition is a task that has not been considered yet in Arabic Language. In this paper we propose a method to extract the opinion target from tweets written in Arabic language. The task is carried out in three phases. Phase 1: preprocess the tweet to delete unnecessary entities like mentions and URLs. Phase 2: construct a feature set from tweet words to be used in the classifying phase; these features are part-of-speech, Named entities, English words, tweet hash tags and part-of-speech pattern. Phase 3: Three classifiers are trained using the extracted features, to assign each word in the tweet to be either an opinion target or not, these classifiers are: Naïve Bayes, Support vector machine and k-nearest neighbor, with an F-Measure result reaching 91%. 500 tweets are used for the experiment, where the opinion target was manually tagged. Finally, a comparison between the results of each model is conducted.

Keywords: opinion aspect, Arabic, Twitter, machine learning.

INTRODUCTION

Twitter is one of the biggest platforms where millions of instant messages are sent by users every day. Users can express their opinion freely on a wide variety of topics, which makes it an ideal source of information to measure user's reactions and opinions; for this reason twitter attracted a great interest of researchers in the field of sentiment analysis. One problem in the context of sentiment analysis and opinion mining is the extraction of information from tweets such as extracting the opinion holder and the target or entity of the opinion. Identifying what the subjective tweets are expressing opinion about is a valuable piece of information towards the construction of a complete opinion summary of the tweet, and which without it the expressed opinion will be meaningless.

Recent researches for opinion target identification are mainly done for English text. Arabic language is one of the fastest growing languages in twitter history, reports shows a growth by 2000% in 12 months[1]. Opinion target recognition is a task that has not been considered yet for Arabic tweets, which will be a base step in the construction of an opinion summary.

In this paper, we propose a method that combines a number of the available resources for Arabic language together with twitter features to identify the opinion target in tweets.

The rest of the paper is organized as follows. In section 2 reviews of the related works, section 3 discuss the preprocessing of Arabic tweets, section 4 and 5 discuss the features and the classifying techniques for extracting the opinion target.

Related work

Many works have been done in the context of opinion target extraction from long review documents in English. In the work of (Ding, Liu et al. 2009) they deal with the two tasks of 'entity discovery', which is the explicit mention of the entity in text, and 'entity assignment' which is the implicit mention of the entity in text. They applied automatic pattern extraction based on POS tags and a starting seed patterns, then assigning entities based on pattern matching. The work provides a solution for entity assignment of implicit opinion that usually appears in comparative sentences. In general, syntactic and statistic patterns yield good results in the domain of long comments in blogs and reviews. (Shang, Wang et al. 2012) argue that these techniques will not score well if it is applied to short comments. They proposed a new method for opinion target extraction from short comments using the representation of a two dimensional vector and back propagation neural network for classification.

In the work of (Hu and Liu, 2004) they explore the problem of identifying opinion target features, and those are the features of the product the review is written about such as picture quality and size of a digital camera review. Features are extracted using POS tags and association rule mining.

Looking specifically at Twitter as the main source of information, the recent researches done are addressing the Named entity recognition problem in general, but not specific to extracting entities that will be assigned as opinion targets in the context of opinion mining. (Liu, Zhang et al. 2011) proposed a semi-supervised learning model for named entity recognition for tweets using K-Nearest Neighbors (KNN) and linear conditional random fields (CRF) as classifiers. (Li, Weng et al. 2012) investigates named entity recognition problem for twitter in a new manner. Their work is different in a way that it does not use the linguistic features of tweets but relies on the relations among segments of tweets.

In the area of Arabic sentiment analysis; a number of researches explored the sentiment classification part of the problem. The studies for opinion target

identification for Arabic has not been widely investigated, one study done by (Hassan, Abu-Jbara et al. 2012) looked into sub grouping expressed opinions in Arabic forums, they identify opinion subgroups by partitioning the signed network representation or by clustering the vector space representation. Their work involved polarity identification and opinion target extraction. The polarity identification is done on two steps: first identifying polarity on a word level using the lexicon Sifaat by (Abdul-Mageed and Diab, 2012) bearing in mind that polarity of some words depends on context the second step involved using a tool called SAMAR developed by (Abdul-Mageed, Kübler et al. 2012) to get context based polarity of the words. For opinion target, they treated noun phrases in the opinionated sentences as opinion targets, but they should appear at least in two posts written by two different participants. To summarize, aspect identification work for English language done using language dependant features that can't be applied directly for the Arabic language. So, in our method we will use an Arabic-Dependant features and machine learning approach to identify opinion aspects in tweets.

Phase 1: Preprocessing

Three forms of preprocessed tweets are generated to be tested for the best presentation for entity classification:

Form 1 consists of deleting the retweet token 'RT' [which indicate that the tweet is actually a resend of a tweet "retweet"], and deleting mentions and URLs from the tweet.

Form 2 consists of form 1 preprocessing steps plus the deletion of leading and trailing spaces, line breaks and leading and trailing non alphabetic characters such as: " : \ | * -,

Form 3 consists of keeping the Arabic and English alphabets, numbers and punctuation marks only. This phase is done by checking the Unicode of each character in the tweet and only keeps the allowed Unicode. Plus replacing the occurrence of some Arabic alphabets by the basic form of that alphabet, as an example, the letters أ and ¦ will both be replaced by the basic form of the letter which is ا .Table-1 shows all the letters that are replaced and their replacement. And the final step was to delete the repetition in characters.

Table-1. Letter replacement.

Letter	Replacement
أ	ا
إ	ا
آ	ا
ٱ	ا
ھ	ه
ة	ه
ؤ	و
ئ	ى

Here is an example of one tweet and the three forms of preprocessing:

Table-2. Preprocessing examples.

Tweet	RT @Oman_Falcon: ★ أبل تعلن أنها تستبدل مجانا آيفون 5 الذي اشتكى مستخدموه من وجود عيب تقني في زر التشغيل بعد أن توقف الجهاز عن العمل…
Form 1	أبل تعلن أنها تستبدل مجانا آيفون 5 الذي اشتكى ★ مستخدموه من وجود عيب تقني في زر التشغيل بعد أن توقف الجهاز عن العمل…
Form 2	أبل تعلن أنها تستبدل مجانا آيفون 5 الذي اشتكى مستخدموه من وجود عيب تقني في زر التشغيل بعد أن توقف الجهاز عن العمل
Form 3	ابل تعلن انها تستبدل مجانا ايفون 5 الذي اشتكى مستخدموه من وجود عيب تقني في زر التشغيل بعد ان توقف الجهاز عن العمل

Table-2 shows how the tweet went through the three phases of preprocessing. In the original tweet we see that it contains a mention (@Oman_Falcon) and (RT). Form 1 of preprocessing deletes the retweet token and the mention. Form 2 of preprocessing deleted the non-alphabet leading star (★) and the trailing dots (…). Form 3 of preprocessing replaces the alphabets أ with ا as in the words (أبل) to (ابل).

Features

In features extraction, we took advantage of some of the existent tools built for Arabic language. We used named entity recognition tool and part-of-speech tagger, plus we included features that are twitter specific. The features used for classification are:

Part of speech tags (POS)

We used Stanford Arabic part of speech tagger[1] to tag tweet's words. In addition to the POS tags for the word, we constructed a POS pattern for each word by getting the POS tag of the word before and the POS of the word after. POS tags were obtained from all three forms of preprocessed tweets; an initial test was carried out to check the most accurate results, which was obtained from preprocessing form number 3.

Entities

To identify entities, we used the LinqPipe[2] tool to extract named entities from Arabic tweets. For each word we create a binary feature to indicate if the word is marked as entity by the tool or not. An initial test was carried out against a dataset of 500 tweets which are tagged manually for entities in tweet. By comparing the entities obtained from LinqPipe to the manually tagged entities, the best results obtained by the tool are when using form 1 of preprocessed tweets.

[1] http://nlp.stanford.edu/software/tagger.shtml
[2] http://alias-i.com/lingpipe/index.html

Twitter specific

We created two binary features that are twitter specific; the first feature is the hash tag, each word is marked if it is either a hash tag or not. The second feature is spotting English words in Arabic tweets, each word is marked as English word or not, this step is done by using Unicode code of characters in the word. Form 1 of the preprocessed tweets is used in this step.

Phase 3: Classification and results

For testing our method we used twitter API to collect 500 Arabic tweets in general topics, collected between the dates 26/4/2014 and 1/6/2014. The opinion target in the tweets was manually tagged.

For classification training, we used Weka, three classifiers were trained; these are Support vector machine, Naive bayes and K-nearest neighbor. Each word of the tweet is treated as a training entity with its own set of features. Giving that we are treating each word as a training entity, the number of words that are opinion aspects compared to the non opinion aspect words is very low. 6019 words are not entities compared to 562 words that are entities. So, before training the classifier, balancing the dataset is performed by applying the Synthetic Minority Oversampling Technique [SMOTE] in which the minority class is over-sampled by creating synthetic examples to get equal number of classes in the dataset. Resulting in total number of instances of 12038.

Two sets of experiments are carried out with different features to evaluate the effectiveness of POS patterns as a feature in the classification process. The first set of experiments used the POS tags, Entities and twitter specific features. Figure-1 shows the F-measure obtained from each classifier.

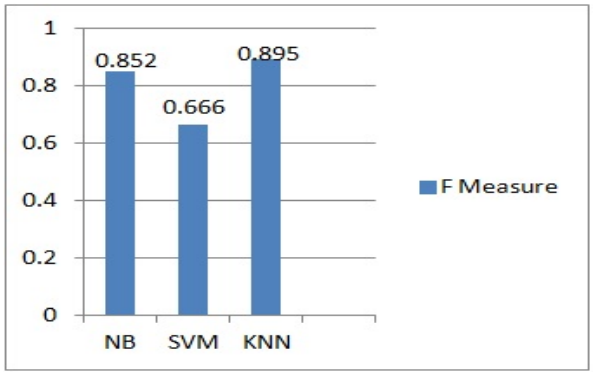

Figure-1. Experiment-1.

The second set of experiments, add the POS pattern to the set of features. Then the three classifiers are trained with this added feature. Figure-2 shows the F-measure obtained from each classifier.

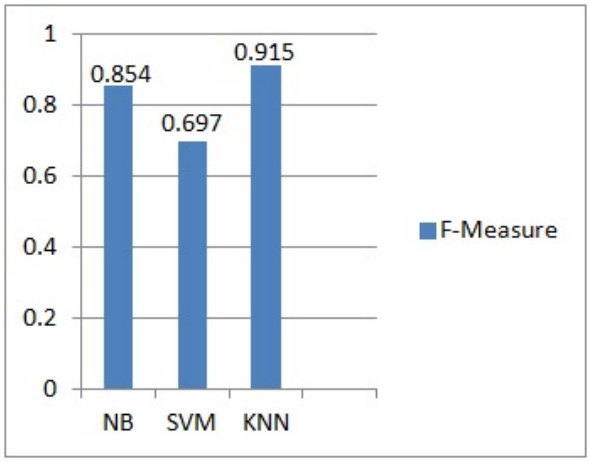

Figure-2. Experiment-2.

From the figures, we see that improvements in the results in general are very small. But SVM scored the highest improvement by 0.031 by using POS patterns although the F-measure is low compared to other classifiers. KNN scored the highest F-Measure among the three classifiers in both experiments. Table-3 shows the results of the precision, recall and F-Measure of the two experiments.

Table-3. Evaluation results.

	Without POS pattern			With POS pattern		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
NB	0.861	0.853	0.852	0.854	0.854	0.854
SVM	0.784	0.693	0.666	0.744	0.708	0.697
KNN	**0.904**	**0.896**	**0.895**	**0.922**	**0.915**	**0.915**

Conclusion and future work

Our experiments shows that the improvements achieved by POS pattern is not so significant, and that might comes to the fact that in twitter it is popular to usually use the dialectal Arabic, which is an informal form of the Arabic language. So, given that the POS tagger is trained on a Modern Standard Arabic Dataset, the results from the POS patterns are not of high accuracy.

More research is needed to exploit Arabic language specific feature to improve the results of the classification and also look into the two different types of Arabic: Modern Standard Arabic language and Dialectal Arabic which is widely used in twitter.

ACKNOWLEDGEMENTS

This work is supported by the ministry of higher education (MOHE) and research management centre (RMC) at Universiti Teknologi Malaysia (UTM) under

research university grant category (Vot: Q.J130000.2528.07H89).

REFERENCES

Abdul-Mageed, M. and M. Diab (2012). Toward building a large-scale Arabic sentiment lexicon. Proceedings of the 6th International Global WordNet Conference.

Abdul-Mageed M., S. Kübler et al. (2012). Samar: A system for subjectivity and sentiment analysis of arabic social media. Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, Association for Computational Linguistics.

Ding X., B. Liu et al. (2009). Entity discovery and assignment for opinion mining applications. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Hassan A., A. Abu-Jbara et al. (2012). Detecting subgroups in online discussions by modeling positive and negative relations among participants. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics.

Hu M. and B. Liu (2004). Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Li C., J. Weng et al. (2012). Twiner: named entity recognition in targeted twitter stream. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM.

Liu X., S. Zhang et al. (2011). Recognizing named entities in tweets. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics.

Shang L., H. Wang et al. (2012). Opinion target extraction for short comments. PRICAI 2012: Trends in Artificial Intelligence, Springer. pp. 528-539.