SindhUniv. Res. Jour. (Sci. Ser.) Vol.48 (4) 707-710 (2016)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Sentiment Analysis for Emiriti Dialects in Twitter

H. AL SUWAIDI, T. R. SOOMRO[++]*, K. SHAALAN

Faculty of Engineering and IT, Birtish University in Dubai, Dubai, United Arab Emirates

Received 13th October 2015 and Revised 12th March 2016

Abstract: Sentiment analysis is not a new field of study and in research there has been focus on word polarity to distinguish between positive, negative, or neutral sentiments of words. However, the Emirati Dialect (and by extension the Arabian Gulf dialects) had received little attention. This paper aims to provide a feasibility study of the possibility of assigning a sentiment value to a word when it comes to the UAE's Arabic Dialect (Emirati). In this study, ten keyword phrases were selected from the Emirati Dialect and were examined to assign their polarity from Twitter collections. During the course of this study and based on the statistical results of the examination of the test sets, it became apparent that it was indeed possible to assign a sentiment value to certain Emirati words.

Keywords: Sentiment Analysis, Emiriti Dialects, Twitter values

1. INTRODUCTION

In recent years, researchers focused gradually more on the Sentiment Analysis of the Arabic text. The purpose of this study is to investigate the possibility of assigning sentiment to words in the UAE's Arabic Dialects (Emirati) using Twitter feed samples. One of the biggest challenges, when one tries to analyze words written in a certain dialect, is the innumerability of the recognized or made up words of certain dialects. According to (Ryding 2005) "Vernacular speech is much more flexible and mutable than the written language; it easily coins words, adapts and adopts foreign expressions, incorporates the latest cultural concepts and trends, and propagates slang, thus producing and reflecting a rich, creative, and constantly changing range of innovation". Today these social media networks hold a treasure trove for analysts of numerous disciplines, from linguists and social scientists to psychologists. Twitter is an online service, provides send and read messages (twits) facilities to its registered users and read messages (twits) facility to unregistered users since (2006).

The popularity of Twitter attests to the growing attachment of the Internet generation, to be connected with rest of the world using these tools. People often misconstrue the many uses of an online social networking service, such as Twitter. As the tweets people frequently publish could be used as a way to converse with others, instead of it only being a one-way communication avenue, as demonstrated by (Honeycutt and Herring 2009). Socrates once said 'The beginning of wisdom is the definition of terms'. For one to better understand what Sentiment Analysis is, they would need to understand what a sentiment is. According to the Cambridge Online Dictionaries[1], Sentiment is a "thought, opinion, or idea based on a feeling about a situation, or a way of thinking about something". The Oxford Online Dictionaries[2] defines Sentiment Analysis as the "process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, is positive, negative, or neutral". In the scientific community, (Pang and Lee 2008) explained the term Sentiment Analysis as follows:

a. sizeable number of papers mentioning “sentiment analysis” focus on the specific application of classifying reviews as to their polarity (either positive or negative), a fact that appears to have caused some authors to suggest that the phrase refers specifically to this narrowly defined task. However, nowadays many construe the term more broadly to mean the computational treatment of opinion, sentiment, and subjectivity in text. Thus, when broad interpretations are applied, “sentiment analysis” and “opinion mining” denote the same field of study (which itself can be considered a sub-area of subjectivity analysis).

This research aims to analyze the sentiments of some of the UAE's Arabic Dialects (Emirati). Twitter users often write their texts using informal language(s) and more often than not would use certain dialects, adapted expressions, or a combination of more than one of those to communicate certain information. The main focus of this study is on the Emirati Dialect. Many studies had been done on Sentiment Analysis over the

[1]http://dictionary.cambridge.org/dictionary/british/sentiment. Retrieved 31.01.2015

[2] http://www.oxforddictionaries.com/definition/english/sentiment-analysis. Retrieved 31.01.2015

[++]Corresponding authors T. R.SOOMRO email: tariq@szabist.ac.ae,

*Faculty of Computing, SZABIST Dubai Campus, Dubai, United Arab Emirates

2014246002@student.buid.ac.ae, khaled.shaalan@buid.ac.ae

years, such as (Pang, Lee and Vaithyana than (2002) Dini and Mazzini (2002) Dave, Lawrence and Pennock (2003) Nasukawa and (2003) Wiebe and Mihalcea 2006), and on Linguistic Analysis of the Arabic text, for example (Rafea and Shaalan 1993; (Shaalan *et al.* 2006). Additionally, (Pang and Lee 2008) attributed the surge of research interest in the areas of Sentiment Analysis and opinion mining to several factors, which include: the rise of machine learning methods in natural language processing and information retrieval; the availability of datasets for machine learning algorithms to be trained on, due to the blossoming of the World Wide Web and, specifically, the development of review-aggregation web-sites; and, of course realization of the fascinating intellectual challenges and commercial and intelligence applications that the area offers.

The overall goal was to provide a feasibility study of the possibility of assigning a sentiment value of being positive, negative, or neutral to a specific Emirati phrase. The test set for this research is Twitter, because Twitter limits the user's input to 140 characters per message. In March 2012, (Twitter 2012) the company announced that the service was made available in Arabic, Farsi, Hebrew, and Urdu among others. The experiments of this research used a small corpus of a newly collected twitter feed. Word polarity will be used to try and determine the viability of sentiment classification in Emirati Dialect by manually assigning values to words. Research in the Arabic Dialect Sentiment Analysis faces many problems. Its main hindrance is the lack of publicly available annotated corpora that could be used to test new techniques.

Communication is an integral part of life. Most living things communicate with one another using different means. (Russell and Norvig 2010, p.888) explained that communication "is the intentional exchange of information brought about by the production and perception of signs drawn from a shared system of conventional signs". Language is one of those means for humans.

A language is basically a system that uses a collection of characters to form words and a combination of words to form sentences. (Anderson 2012) explained that a particular language can be identified when people speak in the same way and form a cluster of similar systems. The Arabic language is an example of such a system. All members of the Arab League, which span from the Arabian Gulf to Northern Africa, use Arabic as their official language. And while the diversity of the regional dialects could be attributed to the vast geography and land barriers between the different regions, their cultural background, religion, and shared history worked to unify the Arab society. (Ryding 2005).The Arabic language can be categorized into the Modern Standard Arabic (MSA), Classical Arabic (CA), and Dialectal Arabic (DA). The first two are known to be similar save for their styles and vocabulary. MSA is used for both Arabic public speaking (i.e. Television and Radio) and written media (i.e. Newspapers). On the other hand, the Dialectal Arabic is used by people for informal exchanges. Moreover, most Arabic people know at least one dialect (which they could speak fluently), and they might also understand more one dialect, i.e. from other regions.

2. MATERIALS AND METHODS

This section discusses the techniques, methodology, collection source, and type of data that was used during the analytical process of this research. The main aim of this research was to analyze the sentiments of some of the UAE's Arabic Dialects (Emirati) phrases. In order to conduct this study, two Twitter sets were collected at two different dates. The assignment of a sentiment to a word is more complicated than one might think. The purpose of this analysis was to mainly see if a word can be classified as being positive, neutral, or negative. Consequently, the first collection was examined and a value was assigned for each tweet. The second collection was also examined and values were assigned. At the end of the examination the overall sentiment assignment was the combined averages of the two examinations, which provided the possible sentiment for each phrase.

In order to collect the Twitter test set for the analysis of sentiment, 10 phrases were selected that are commonly used in everyday conversation using the Emirati Dialect **(Table 1).** Each phrase can be said to lean more towards a specific sentiment more than the other based on its intended use. The two collections of Twitter were collected on February 7th, 2015 and on February 14th, (2015). The source of these collections was from random

Table 1: 10 Phrases of Emirati Dialect

#	Phrase	How It's Read	Synonym in English
1	اسولف	Asolef	Talking
2	اندوكم	Endokm	Here you have it
3	جزاك	Jazak	Praise
4	عبالهم	Abalhum	They think
5	متفيج	Metfayej	Has time
6	يغربل	Yegarble	Disorganize or Garble
7	بهالعمر	Behal'emr	In this Age
8	تحقر	Tahger	Despise
9	ما تستحي	Ma'Testehi	Not Ashamed
10	وهق	Wahag	Get Someone in Trouble

3. RESULTS AND FINDINGS

As stated previously, due to the time limitations on this study, the final count of used tweets in this research was limited to 1000 tweets per collection. Thus, each

phrase from the selection was evaluated and at the end of the examination the final tally was counted; to provide a possible sentiment for each phrase based on the results of the examination. For instance, for the phrase "اسولف"; the first collection of this word contained over 1000 tweets, of which only 100 tweets were randomly evaluated. The second collection of this word also contained over 1000 tweets, and again only 100 tweets were randomly evaluated. At the end of the evaluation period, the overall rated sentiment was the combined averages which provided the possible sentiment for each phrase.

After the phrases' sentiment evaluation of the first **(Table 2)** and second collections **(Table 3),** the overall averages where calculated to reach the final rating of each phrase **(Table 4).** The overall phrase sentiment was reached at the end based on the final rating **(Table 5).** Some negation words were identified in the course of this study. The negation words were (لا, ما, مب) and they mostly mean (No or Not). And while the Sentiment Analysis of Dialect words holds the spotlight in this study, identifying these small keywords might further future research as they sway the word polarity from one side to the other in some instances. Also, no special attention was made in this study to identify when a word turned negative or positive due to the presence.

A number of studies had focused previously on Twitter as a test bed for Sentiment Analysis, like (Shoukry and Rafea 2012; Ahmed *et al.* 2013); (El-Beltagy and Ali 2013; Abdulla *et al.* 2013) which used a dataset of tweets that ranged in size between 500-2000 tweets. Thus, this study falls within the dataset size of previously examined samples, as shown previously the size of the examined data included a total of 2000 tweets. Additionally, while some work was done on Sentiment Analysis of Arabic Dialects, the only work that's been done that could be found so far was on the Egyptian Dialect like (Shoukry and Rafea 2012); (El-Beltagy 2013) or on the Jordanian Dialect like (Abdulla *et al.* 2013). Therefore, no other work could be related to the results reached in this study as the UAE's Arabic Dialect (Emirati) hasn't received much attention when it comes to Sentiment Analysis.

Table 2: Sentiment Evaluation of the First Set

#	Phrase	Positive	Neutral	Negative
1	اسولف	0.00	1.00	0.00
2	اندوكم	0.00	0.44	0.56
3	جزاك	1.00	0.00	0.00
4	عبالهم	0.00	0.82	0.18
5	متفيج	0.00	0.98	0.02
6	يغربل	0.07	0.17	0.76
7	بهالعمر	0.07	0.80	0.13
8	تحقر	0.12	0.61	0.27
9	ما تستحي	0.02	0.6	0.38
10	وهق	0.13	0.28	0.59

Table 3: Sentiment Evaluation of the Second Set

#	Phrase	Positive	Neutral	Negative
1	اسولف	0.00	1.00	0.00
2	اندوكم	0.00	0.82	0.18
3	جزاك	1.00	0.00	0.00
4	عبالهم	0.00	0.94	0.06
5	متفيج	0.00	0.92	0.08
6	يغربل	0.03	0.29	0.68
7	بهالعمر	0.00	0.92	0.08
8	تحقر	0.28	0.31	0.41
9	ما تستحي	0.00	0.48	0.52
10	وهق	0.00	0.13	0.87

Table 4: Combined Averages

#	Phrase	Positive	Neutral	Negative
1	اسولف	0.00	**1.00**	0.00
2	اندوكم	0.00	**0.63**	0.37
3	جزاك	**1.00**	0.00	0.00
4	عبالهم	0.00	**0.88**	0.12
5	متفيج	0.00	**0.95**	0.05
6	يغربل	0.05	0.23	**0.72**
7	بهالعمر	0.035	**0.86**	0.11
8	تحقر	0.2	**0.46**	0.34
9	ما تستحي	0.01	**0.54**	0.45
10	وهق	0.065	0.205	**0.73**

Table 5: Overall Phrase Sentiment Evaluation

#	Phrase	Sentiment
1	اسولف	Neutral
2	اندوكم	Neutral
3	جزاك	Positive
4	عبالهم	Neutral
5	متفيج	Neutral
6	يغربل	Negative
7	بهالعمر	Neutral
8	تحقر	Neutral
9	ما تستحي	Neutral
10	وهق	Negative

4. DISCUSSION

The overall goal was to provide a feasibility study of the possibility of assigning a sentiment value of being positive, negative, or neutral to a specific Emirati phrase. The test set used for this research was Twitter, as Twitter limits the user's input to 140 characters per message. Ten keyword phrases were selected from the Emirati Dialect. And two Twitter collections were created using these phrases on two different dates. Due to the time limitations, the final count of the used tweets was limited to 1000 tweets per collection during the phrase evaluation stage. Finally, this research concluded that as some phrases from the Emirati Dialect can be shown to indicate a certain sentiment (positive, negative, or neutral) than the others, it is possible to assign sentiments to words in the Emirati Dialect.

During the course of this study a review of previous similar work was done prior to selecting the technique used for this research. Most of the conducted research in the area of sentiment analysis relied on an automation technique for word annotation as well as a Lexicon. There are many drawbacks to doing manual analysis on a collection including the time limitation, and there is no known and acceptable threshold for the sample size to reach a definite conclusion. One recommendation is to have another unbiased native speaker of the Emirati Dialect go through the collections and compare the end results with the conclusions of this study. Another recommendation is to create a Lexicon for Emirati Dialect words, which could be mapped to both Arabic and English Lexicons to accelerate the process of word annotation and sentiment analysis. Additionally, the creation of a publicly available resource for different Arabic Dialect phrases, adapted expressions, or made up words would help speed up the process for any work in this field that relies on automation techniques.

REFERENCES:

Abdul-Mageed, M., and M. Diab, (2012). AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), May 23-25 2012, Istanbul, Turkey, European Language Resources Association (ELRA).

Abdulla, N., M. Shehab, M. Al-Ayyoub, (2013). Arabic Sentiment Analysis: Lexicon-Based and Corpus-Based. In the Proceedings of the IEEE Conference on Applied Electrical Engineering and Computing Technologies, Dec 3-5 2013, Jordan, IEEE Press.

Abdul-M., M. Diab, and M, Korayem, (2011). Subjectivity and Sentiment Analysis of Modern Standard Arabic. In the 49th Annual Meeting of the Association for Computational Linguistics. Human Language Technology: short papers, Vol. 2, 587-591.

Abdul-M., M. Diab, and M, Korayem, (2012). SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media. In the Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, Jeju, Republic of Korea, 12 July 2012 19-28.

Ahmed, S., M. Pasquier, and G. Qadah, (2013). Key Issues in Conducting Sentiment Analysis on Arabic Social Media Text. In the 9th International Conference on Innovations in Information Technology (IIT), 17-19 March 2013, Abu Dhabi, Al Ain, UAE.

Anderson, S. (2012). Languages: A Very Short Introduction. Oxford: Oxford University Press. ISBN 978-0-19-959059-9.

Andreevskaia, A., (2006). Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. In Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, 209-216.

Dave, K., S. Lawrence, and D. Pennock, (2003). Mining the peanut gallery: Opinion Extraction and Semantic Classification of Product Reviews. In Proceedings of WWW2003, May 20-24, Budapest, Hungary, 519–528.

Dini, L., (2002). Opinion Classification Information Extraction. In Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields (Data Mining), 299–310.

Elarnaoty, M., S. Abdelrahman, A. Fahmy, (2012). A Machine Learning Approach for Opinion Holder Extraction in Arabic Language. In the International Journal of Artificial Intelligence and Applications (IJAIA), Vol.3, No.2, 45-63.

Farra, N., H. Hajj (2010). Sentence-level and Document-level Sentiment Mining for Arabic Texts. In IEEE International Conference on Data Mining Workshops (ICDMW).

Honeycutt, C., and S. C. Herring, (2009). Beyond Microblogging: Conversation and collaboration via Twitter. In 42nd Hawaii International Conference on System Sciences, Los Alamitos, CA, IEEE Press.

Pang, B. and L. Lee, (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, Vol. 2, No. 1-2, 1-135.

Rafea, A., and K. Shaalan, (1993). Lexical Analysis of Inflected Arabic Words Using Exhaustive Search of an Augmented Transition Network. Software Practice and Experience, Vol. 23(6), 765Pp

Russell, S. and P. Norvig, (2010). Artificial Intelligence: A Modern Approach, 3rd Edition. New Jersey: Inc. 888Pp.

Shaalan, K., A AbdelMonem, A Rafea, H. Baraka, (2006) Mapping Interlingua Representations to Feature Structures of Arabic Sentences. The Challenge of Arabic for NLP/MT, International Conference, the British Computer Society, London, 149–159.

Turney, P., and M. Littman, (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Transaction on Information Systems.

Wiebe, J., and R. Mihalcea, (2006). Word Sense and Subjectivity. In Proceedings of the Conference on Computational Linguistics/Association Computational Linguistics (COLING/ACL).