Review of Sentiment Analysis for Classification Arabic Tweets

Tagwa Abd Elatif Mohammed
Sudan University of Science and Information Technology

Abstract — The social networking sites are rich environments for the exchange of views and ideas. Therefore, these sites are gaining to grow attention from researchers interested in the knowledge of the prospecting consensus or analysis of the sentiments. The basic goal of the technology of data mining or analysis of feelings is the conclusion of knowledge from massive amounts of data which is known as Sentiment Analysis (SA). SA relies on supported algorithms, which are the basis of the regulations on prospecting and exploration for the data. SA algorithms are derived from many smart sciences, such as statistics, artificial intelligence, expert systems, pattern and machine recognition and other sciences. There are some difficulties related to analyzing Arabic language. Part of the difficulty and complexity of the Arabic language is that it can be found in many forms, Arabic can be found in three forms, namely; classical Arabic, modern standard Arabic (MSA) and colloquial Arabic dialects. So we see few bodies of studies and research handling Arabic sentiment analysis.

In this paper we present some of previous work in sentiment analysis by using two techniques a lexicon based and a Corpus based. Literature review regarding this interest shown difference between the use of Arabic dialects or classical Arabic Language. Techniques and tools used to sentiment analysis in previous work also reviewed.

Keywords— sentiment Analysis (SA); Social media; twitter; Modern Standard Arabic (MSA) ; Support Vector Machines ( SVM); naive bayes (NB);

I. INTRODUCTION

A. Sentiment Analysis

Sentiment Analysis (SA) is an increasingly important area in applied linguistics. Currently SA is very an active area of research. SA has been gaining a lot of attention in the last few years due to the important role it play in several areas such as marketing, education and politics. [1]

Sentiment analysis or opinion mining is used to extract information by identifying data to indicate positive, negative or neutral information in source materials by using text analysis, natural language processing and computational linguistics [2].

Sentiments are very important. Whenever we need to make a decision we need to know others’ sentiment. This is not only true for individuals, but it is also true for organizations and governments. Many tools were built and developed for sentiment analysis [1].

With the increasing use of social media, the researchers have turned to the use of methods of data mining and knowledge to detect people's feelings. Sentiment analysis and exploration of the views aim at finding whether people’s views are positive, negative or neutral. This analysis helps in all areas of life and the policy of education, health and trading [3].

B. Social Media:

The based communication systems, also known as social media or social networking, allow an individual to do many different things. First, one can construct a public, private, and/or semi-public profile. Second, an individual can create a list of friends, contacts, or connections. Finally, the connection connections of others can be viewed and traversed [4].

There are many social networking sites like Twitter, LinkedIn, Facebook, YouTube, Pinterest, and Instagram. The main purpose of social media Social media is social interaction [5]. It consists of popular Web sites and emphasizes the communication and the exchange of opinions and ideas. Some sites have been more popular than others. In Sudan, users use Facebook and Twitter more than other media networks. The social media plays an important role in all aspect of life.

C. Twitter

Twitter which is a micro-blogging service, gives Users the permission to exchange and post short (140-character-long) messages known as tweets [5].

People’s communications via mobile devices on the Internet especially on social media platforms, can be shown by Twitter, IM abbreviations, texting, and chat. Despite grammatical mistakes, misspelling or confusing parts found in these interactional texts.
Texting acronyms are used widely. By shortening a text, the user doesn't have to spell out phrases, sentences or expressions [6]. The use of emoticons is also common in online communications and texting [6].

D. Social Media and emotion

Social networking sites have become one of the most important ports for people to express their feelings and the current mood and the interplay of their joys, sorrows as well as their views concerning what is going on around them.

The interplay of emotions is one of the biggest decisions of users on online activity. Stories on social media spread in different speed and scale, and there is no simple answer on what makes some go viral e.g. sad story, or what enrages users. However, some of studies provide us with details on how the interplay of emotions affects the users of social media [7]. Some studies have shown that reading some of the feelings of happiness or sadness through the means of social communication affects your mood. This is so-called by the infectious moods [8].

E. Language used: Arabic Language:

Written Arabic language uses alphabets just like English, but Arabic is written from right to left, and contains 28 letters. Moreover, it is always written in “cursive,” that means the letters are always connected within words.[9] Arabic letters do not contain uppercase and lowercase letters as it is they are in English. Arabic Language belongs to the Semitic languages, which includes Hebrew, Amharic. These languages are found in the Middle East. Arabic can be classified into three basic types Classical Arabic (CA), Modern Standard Arabic (MSA), Colloquial Arabic Dialects [9].

- a) Classical Arabic:
  Classical Arabic, known also as Quran Arabic, is studied by Muslims to read the Holy Quran [29]. Classical Arabic is used in Al-Hijaz region before 1500 years, Written records of the language include Pre-Islamic Poetry that was composed in pre-Islamic times [9].

- b) Modern Standard Arabic (MSA):
  Modern Standard Arabic (MSA) is based on the Classical language [10]. All Arab countries use MSA spoken and written.

- c) Colloquial Arabic:
  Colloquial Arabic is the spoken Arabic used by Arabic speakers in everyday situations. Colloquial Arabic is subject to regional variation, not only between different countries, but also across regions in the same country. Compared with colloquial, MSA is a uniform in all Arab countries [10].
A considerable amount of previous works have been published on sentiment analysis. These studies mostly report analysis of sentiments in messages in English. Fewer studies were done on sentiments in Arabic language.

Methodology proposed by Refae, Eshrag, and Verena Rieser (2014) in [2]; the authors use gold-standard annotated corpus. The corpus contained 8,868 tweets collected manually at different time. Test set (1,365 tweets) development set (7,503 tweets) the result of this analysis showed the top of five frequent words.

Methodology proposed by Al-Kabi et al (2014) in [3]; the researchers used lexicon-based tool to analyze Arabic reviews and comments by using MSA language. This tool can identify the polarity, subjectivity, and strength/intensity of each evaluated Arabic review and comment. The result of this study developed a tool of more accurate results when it is applied on domain-based Arabic.

Methodology proposed by Cheong, France, and Christopher Cheong(2011) in [5]; the work of the authors was done using tweets extracted from Twitter during the Australian 2010-2011 floods. The social network analysis techniques were used to generate and analyze the online networks traffic that emerged at that time. In this study, SNA metrics is used to identify influential members on online communities. However, one limitation of it is the insufficient number of tweets collected for the purpose of generalizing the findings.

Methodology proposed by Kouloumpis et.al. [11] The author investigated the utility of linguistic features for detecting the sentiment of Twitter messages and evaluated the usefulness of existing lexical resources as well as the features that capture information about informal and creative language used in micro blogging. The researcher used three datasets (hashtag, emotion and iSieve) from twitter. In this study, different features are used to classify data, use bigrams, unigrams and sentiment analysis, mainly features representing information from a sentiment lexicon and POS features.

Methodology proposed by Duwairi, Rehab M., et al.(2014)in [12]; the authors used crowd sourcing to collect a large dataset of tweets and developed a framework that made it possible to analyze Tweets to detect the positive, negative or neutral sentiments. However, there was a need to expand the dictionaries.

Methodology proposed by El-Makky et al. (2015) the authors [13] designed a novel subjective and sentiment analysis system for Arabic tweets by building a new Arabic lexicon and merging Modern standard Arabic lexica with two Egyptian Arabic lexica.

The result of novel hybrid approach showed good result compared to the previous work.

Methodology proposed by Amira Shoukry and Ahmed Rafea [14]. They used sentence level sentiment analysis for Arabic. In this study, 1000 tweets are collected from twitter and classified by using two algorithms in Weka program, Support Vector Machine (SVM) and naïve Bayes (NB) algorithms. The result showed that SVM accuracy was higher than NB. However, the study ignored neutral tweets.

Methodology proposed by Laurie BUTGEREIT, (2014) the author [15] aims to monitoring Twitter of weather status for a specific city. It used crowd sourced and collected almost 600 tweets from twitter and used the μ Model to analyze the tweets.

Methodology proposed by Rushdi-Saleh et al (2011) [17] The authors used verity web pages Arabic and blogs to build a small Arabic opinion corpus consists of 500 movie, 250 positive and 250 negative opinion using machine learning support vector machines algorithm and Naïve Bayes algorithm. The study showed good results. The accuracy of the NB classifier was (84%) and the accuracy of SVM classifier was (90%). However, there was a need to expand the corpus.

Methodology proposed by Siqi zhao (2013) [18]. The work developed Sport Sense to detect sport events targeted for the National Football League (NFL) games. The results showed that major game events can be accurately and effectively extracted by using open access Twitter data. Sport Sense API is accessible for developers for using system in order to create Twitter-enabled applications. The application was tested and used by the National Football League (NFL) games as a targeted domain.

Methodology proposed by Sarah Vieweg et al (2010) in [19]. The authors analyzed the tweet posts in emergency events in North America during the Oklahoma grass fires and Red River Floods in spring 2009 by using Twitter Search API and geo-location. The aim of the study is to improve situational awareness in emergency situations.

Methodology proposed by Godfrey et al (2014) in [20]. The authors used cluster analysis to find topics in the collection of tweets. The researchers clustered the tweets using k-means, a commonly used clustering algorithm, and the Non-Negative Matrix Factorization (NMF). A comparison of the results showed that two algorithms gave similar results, but NMF proved to be faster and provided more easily interpreted results.
Methodology proposed by Duwairi, Rehab M (2015) in [22]. The authors introduced a framework for sentiment analysis on tweets written using either Modern Standard Arabic or Jordanian dialectical Arabic. In this study, the researchers produced dialect lexicon which maps dialectical words into their corresponding Modern Standard Arabic words. The researchers used dataset consists of 22550 tweets, NB and SVM. Classifiers were used to determine the polarity of the tweets [22].

Methodology proposed by Korayem et al (2015) in [23]. The study presents a survey of different techniques for subjectivity and sentiment analysis (SSA) for Arabic. These surveys describe the main existing techniques and test corpora for Arabic SSA that has been introduced in the literature.

Methodology proposed by Rushdi-Saleh, Mohammed, et al. (2011) in [24] They used different machine learning algorithms to classify the polarity of Arabic reviews extracted from specialized Web pages related to movies and films. The authors have translated the Opinion Corpus for Arabic (OCA) corpus into English, generating the EVOCA corpus (English Version of OCA). The result has shown that the accuracy of EVOCA is low because of translation.

Methodology proposed by Mnahel and Naomie (2013) in [25]. In this study, the researchers used 65 studies to determine features and methods used for twitter opinion mining, the results of this study showed n-grams features is commonly used for twitter sentiments analysis and also used for Arabic tweets.

The study also reported that the methods most common used is the Lexical based classification using Support Vector Machines (SVM) and Naive Bayes (NB) [25].

Methodologies proposed by Farra, Noura, et al (2010) in [27]. The authors used sentiment mining of Arabic text at both sentence level and document level. This study work in a grammatical approach and a semantic approach for sentence-level sentiment mining in Arabic text, and a document-level, they construct semantic dictionary Includes Arabic roots, The results showed high accuracy in grammatical approach [27].

Methodologies proposed by El-Beltagy, Samhaa R., and Ahmad Ali (2010) in [28]; the study reviewed the problems and challenges for sentiment Analysis in social media. The researchers investigate the possibility of determining the semantic orientation of Arabic Egyptian tweets and comments by using two data sets. Twitter dataset contained 500 tweets. 310 tweets were classified as negative, 155 as positive, and 35 as Neutral. Dostour dataset consisted of 100 comments randomly collected. 38 comments were negative, 40 classified as positive and 22 as neutral. The outcome of study was an Egyptian dialect sentiment lexicon.

Methodology proposed by Al-Kabi, Mohammed N., Nawaf A. Abdulla, and Mahmoud Al-Ayyoub (2013) in [41] the authors using 4625 reviews/comments collect from Yahoo and Maktoob social media to determine length and likes/dislikes reviews, the dataset contains MSA language and various Arabic dialects. the results shows the accuracy of SVM (68.2%) and the accuracy of NB (61.43%).


table 1

<table>
<thead>
<tr>
<th>Author</th>
<th>Approach</th>
<th>Language</th>
<th>Limitations</th>
<th>Analysis Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>[12]</td>
<td>Collect 350,000 tweets.</td>
<td>Arabic dialects, Arabizi and Emoticons.</td>
<td>Need to expand the dictionaries</td>
<td>SVM+NB+KNN</td>
</tr>
<tr>
<td>[17]</td>
<td>Built a small corpus consists of 500 positive 250 and 250 Negative</td>
<td>MSA</td>
<td>Use Small corpus</td>
<td>SVM ÷ NB</td>
</tr>
<tr>
<td>[22]</td>
<td>Produced dialect lexicon which maps dialectical words into their corresponding Modern Standard Arabic words.</td>
<td>MSA</td>
<td>The result of Neutral class is weak comparing with Negative class</td>
<td>SVM ÷ NB</td>
</tr>
<tr>
<td>[25]</td>
<td>Investigate features and methods used for twitter opinion mining, the authors use 65 papers were used in our synthesis of evidences.</td>
<td>Modern Standard Arabic</td>
<td>Studies undiversified</td>
<td>SVM ÷ NB</td>
</tr>
<tr>
<td>[26]</td>
<td>Investigated the ML approach for sentence level sentiment analysis for Arabic using 1000 tweets from twitter</td>
<td>Egyptian Dialect and MSA</td>
<td>Ignored neutral tweets.</td>
<td>SVM ÷ NB</td>
</tr>
<tr>
<td>[27]</td>
<td>Use a grammatical and a semantic approach for sentence-level sentiment mining in Arabic text, and a document-level,</td>
<td>MSA</td>
<td>To build lexicon’s only use words’ root</td>
<td>SVM ÷ NB</td>
</tr>
<tr>
<td>[28]</td>
<td>Two data sets (twitter dataset 500 tweets) and (Dostour dataset content 100 comments)</td>
<td>Egyptian Dialect and MSA</td>
<td>Small Dataset by used limited resources.</td>
<td>double polarity (DP) method and Sum method for calculating semantic orientation</td>
</tr>
<tr>
<td>[41]</td>
<td>Dataset content 4625 comment/review from Yahoo ÷ Maktoob</td>
<td>MSA language and various Arabic dialects</td>
<td>-</td>
<td>SVM+NB+ KNN</td>
</tr>
</tbody>
</table>

III. OPEN ISSUES

As a result of investigation of a broad comprehensive previous research, we have concluded that there are some important gaps and open issues that need to be covered in future research. The summary below shows those gabs:

- An open area in the Arabic sentiment analysis attributed to the lack of research in this area.
- There is no research on sentiment analysis in Khartoum area; neither there are studies on the Sudanese dialect.

IV. CONCLUSION AND FUTURE WORK

At the conclusion of this paper, the authors have addressed some experiments and studies that deal with sentiment analysis in Arabic. There is a scarcity of irrigation and feelings in Arabic attributed to the difficulty of analyzing the composition of Arabic Language as well as the complexity and the multiplicity of Arabic language. It contains three kinds, Classical Arabic (CA), Modern Standard Arabic (MSA), and Colloquial Arabic Dialects. Most of these studies used NB and SVM for data Classification.
There is no research on sentiment analysis in Khartoum area, and there are no studies on the Sudanese dialect. We want to develop a data collector tool which collects tweets from Khartoum every 30 minutes. We also need to build a Lexicon that contains various words which indicate joy and sorrow to put them in a dictionary called feelings for classifying tweets.

My Projects aims to use sentiment classification for Arabic tweets around Khartoum. In my research, I use different techniques for Arabic sentiment analysis applied in Arabic tweets around Khartoum and decide if sentiment is happiness (positive), sadness (negative) or neutral. The methodology of my research is creating corpus Arabic tweets around Khartoum, and using some applications. Regarding collecting data, I use Meltego program and classification of Arabic tweets using Weka program through two types of classification techniques, a support vector machine and naive Bayes. The final step is to build the opinion of lexicon for Arabic opinion word to my corpus document. The Arabic opinion lexicon contains a total of words divided in two groups; the words indicating happiness (positive) and sadness words (negative) with experts in language.

REFERENCES
[18] Siqi zhao, Detecting events from twitter in Real time, 2013
[29] https://en.wikipedia.org/wiki/Modern_Standard_Arabic access date 6/12/2015