The Use of a Sentiment Analysis/Opinion Mining Technique with Machine Learning Algorithms to Develop Twitter-Related Themes in the Saudi Arabian Context

Salman Dutt*

*Department of Computer Science, Taif University, Taif, Saudi Arabia. E-Mail id: salmanduttcse@rediffmail.com

Received: 17.06.2019 **Revised:** 19.08.2019 **Accepted:** 21.09.2019

***Abstract*---** Twitter is the most popular and valuable resource for sentiment analysis. This observation holds because the platform ensures that individuals express their thoughts and ideas regarding various topics, especially about social life. With users restricted to 140 characters or less, Twitter has evolved as one of the most dominant social networking websites in the world. This research has two main aspects. The first objective is to demonstrate the implementation of sentiment analyzer for Tweeter's comments, which is considered as one of the most critical and freely big data sources. We analyzed Arabic text using Saudi dialect tweets to get the sentiments around a particular topic in Saudi Arabia. After Indeed, we evaluated the algorithms implemented by using confusion matrix. The second objective was to examine the challenges and complexities in Arabic text, especially those that make the Arabic sentiment analysis more complex than other languages.

***Keywords*---** Sentiment Analysis, Twitter Analysis, Data Mining, Machine Learning, Supervised Approach, Weka, Arabic Text.

I. INTRODUCTION

DATA has continually played an essential role in this era, as it is continually increasing in size and value. The available data online is doubling in size – every two years [1-4]. The amount of data online that was created in 2013 was 4.4 Zettabytes (ZB), but it is expected to reach 44 ZB in 2020 [5-7]. Social media such as Twitter and Facebook have become common pathways for cyber communication in society. Since the establishment of Twitter in 2006, users have gained the ability to express, post, and share opinions and feelings freely and easily, especially in public, which can be likened to the SMS style. Indeed, Twitter is deemed the most important site among the various platforms, deemed to be full of sentiment.

From the current literature, Twitter refers to a microblogging site that has a set of tweets, and each tweet contains "140 characters or less". More than 140 million users post over 1 billion Tweets every 72 hours [8]. As a result, Twitter is one of the most commonly used models in big data. On the Internet, Arabic posts have increased, especially after the emergence of social networks. On Twitter, there are over 6.5 Million users in the Middle East [9]. Notably, Saudi Arabia has 2.4 million active users on Twitter [10]. With an increase in the amount of data published online, the use of big data analytic techniques has proved important, especially regarding the need to extract the value and the meaning of the resultant information. Big data analytics seeks

10.9756/SIJCNCE/V7I5/07051011
ISSN: 2321–2403

to analyze huge amounts of data and to establish essential information by using analytical techniques such as Natural Language Processing (NLP) and Machine Learning. Imperatively, Sentiment Analysis (SA) is considered as a critical part of the NLP concept [11].

Sentiment analysis refers to the process of extracting and understanding the emotion from data (known as opinion mining). Data mining techniques help to discover, predict, and explore the hidden information (as well as the opinion) from text data in the World Wide Web [12-14]. Hence, gathering Twitter data is a valuable process in the fields of data analysis and data mining because it gives significant insights into how people think and act on social media platforms. In this study, the main focus is on Twitter as an intense environment for the emotional contents. We chose tweet timelines of the most recent topics in Saudi Arabia as a source of Arabic data. This work has many essential influences and challenges. Firstly, the study uses machine learning algorithms to detect and summarize an overall sentiment. Secondly, the work analyzes the Arabic text – since the Arabic language is more complicated compared to other languages – such as English. The complexity comes from characteristics of the Arabic language in writing. Particularly, the language requires a concentrated stage of preprocessing. Hence, this study seeks to establish or develop a completed vision of social consultation. It is also notable that sentiment analysis helps to track and determine people`s behaviors and reactions to certain subjects. This research focused on specific topics that were issued in Saudi Arabia – such as 'women driving.' The tweets were collected and classified into positive, negative and neutral.

The main purpose of this research is to examine a recently-issued domain in Saudi Arabia, especially in terms of or regarding women driving. Based on New York Times news and CNN, the subject of allowing Saudi Women to drive is considered as one of the breaking news, dominating in the September and November 2018. Based on this trend, this research aims to investigate the overall opinions around this topic by using a sentiment analysis or Opinion Mining technique, coupled with machine learning algorithms.

II. METHODOLOGY

Sentiment analysis can be achieved through five essential steps. These steps include retrieving the data, preprocessing and filtering, data labeling, the application of algorithms, and sentiment results. Thus, we conducted the study by specifying the keyword of the domain through which to collect the data from Twitter. After collecting the data, the pre-processing and filtering steps were applied to clean up the tweets and make it ready for the next step. Then, the tweets were labeled and tagged. We used three different categories for labeling: positive, negative, and natural. The final stage entailed working with machine learning algorithms to build the classifier – and evaluate the accuracy of the classifier. The resultant data was divided into two groups: one set was used for the training step and named as a training dataset while another set was used during the testing step (to verify our classification method and yield a test dataset).

Sentiment analysis at sentence level	
Text	**Sentiment**
The touch screen is cool	Positive
Sentiment analysis at document level	
Text	**Sentiment**
I bought an iPhone a few days ago. It is such a nice phone, although a little large. The touch screen is cool. The voice quality is clear too. I simply love it!"	Positive

Sentiment analysis at Aspect level		
Text	**Aspect**	**Sentiment**
The iPhone's call quality is good, but its battery life is short.	*call quality*	Positive
	battery life	Negative

Figure 1- Sample Sentiment Analysis

Data Collection

We considered the most recent tweets in Saudi Arabia – based on the use of three different hash-tags. The hash-tags were: *#Women_driving_car*, *#reject_womn_driving* and *#The _King_suppotrs_women_driving_in_SaudiArabia* (see Table 1).

Table 1 - Three Different Hash-tags that Were Used to Collect the Data

Hash-tag 1
Hash-tag 2
Hash-tag 3

The collected data had 14 attributes, including the ID of the user, the user name, the universal time stamp, local time stamps, text, languages, profile images, sources, locations, time zone hash-tags, URLs, user mentions, and the media. In

10.9756/SIJCNCE/V7I5/07051011
ISSN: 2321–2403

this work, we considered texts and locations.

Data Pre-Processing and Filtering

Preprocess is a method that aims to clean up the data from any redundant contents, as well as remove any noisy data. The techniques of preprocessing are an essential and important stage in this work because applying the sentiment and classification of the original textual data could lead to inaccurate results, as the data is likely to be noisy. One of the techniques is *Tweet cleaning* which helps to remove any noisy data such as:

- Remove the spam tweets, which are comments that have a harmful link or advertisement
- Re-tweet tweets, which start with "RT"
- Duplicate tweets, which are repeated more at least twice
- Irrelevant information, which is not related to the textual tweets; for example #, @, URLs, User name and any non-Arabic words
- Opinions that are unrelated to women driving or the subject under investigation
- Remove non-textual contents such as videos and images
- Punctuations from the tweets. This factor has an essential impact on the classification stage in the WEKA.

After the tweet cleaning technique, the output is only the texts. Also, the outcome texts require the implementation of other techniques to filter it further to ensure that it is ready for analysis. This technique is the *Filtering*, which helps to clean each word in the texts – to make it smooth and worth processing. The filtering process involved:

- Correcting the misspelling in some words manually because of grammar or spelling mistakes
- Normalizing the words with repeated letters by removing repeated letters to get better classification results. For example, the word ارامسسكو implies that the repeated letters could be removed and replaced with an alternative word such as ارامكو
- Removing stop words, which have no particular meaning and do not carry any crucial information, including conjunctions and prepositions. This task helped us to identify the most influenced words in the text. Notably, a list of Arabic stop words is available in El-Khair [13].
- Removing non-letters from the words, examples being +, $, and &
- Removing Arabic short diacritics from the words
- Few of Twitter users compress two words by disregarding the space between two words because the number of characters is limited. This issue was normalized by creating spaces between the words manually

The preprocessing and filtering stages were important to this project, especially because the success of the preprocessing part was poised to yield more accurate results, hence better performance.

Data Classifier

Depending on the meaning of the influenced words in the text, we established the three categories and their associated words as shown in the table below.

Arabic Tweet	Translation	Label
اتمني ان هذا القرار كان من زمان	We wish the decision has been released long time ago.	Positive
نحن فعلا نرفض قيادة المرأه في السعوديه	Indeed, we reject women drive in Saudi Arabia.	Negative
لايهمني موضوع قيادة المرأه	I don't care if women drive or not.	Neutral

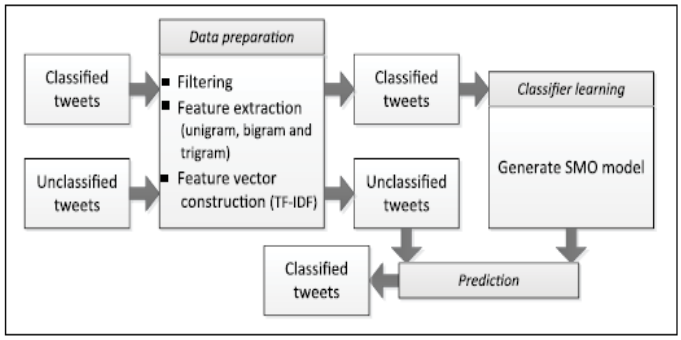

Figure 2- SVM Classifier Implementation Steps

III. Classification and Analysis

Building the Training Dataset

The training dataset was built from the sentiment text and their label. Notably, the dataset needs to show in either

Comma Separated value (CSV) or Attribute-Relation File Format (ARFF). ARFF is the standard default format for WEKA and is an ASCII text file that represents a list of instances sharing a group of attributes [15].

Building Classifier

Next step to the classification was running the supervised machine learning algorithms. Naïve Bayes and Support Vector Machine (SVM) was the machine learning algorithms that were used in this research. These algorithms were selected because they have been affirmed to work well in text classification and that they yield good results. Using specific filters to enhance the objective of the classification process in WEKA, the *Class Assigner* filter was used to assign the category column as the class for the data.

Training Classifier

The third step in the classification was training the classifier. The machine learning in WEKA was not appropriate with the string data type. As a result, The *StringToWordVector* filter was applied to convert the string into words or numeric.

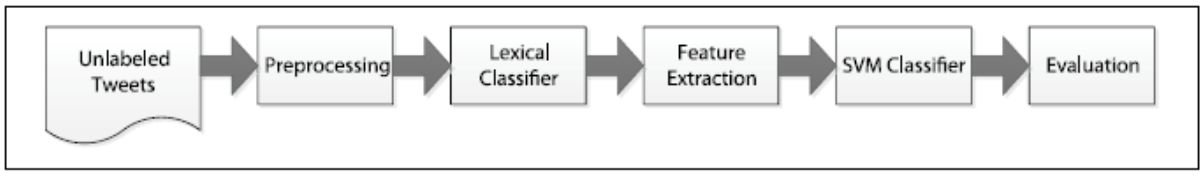

Figure 3 - Summary of the Research Design

IV. Results and Discussion

Monkey Learn entails sentiment analysis software that supports machine learning algorithms. The Software provides trained data for some languages – such as English. However, it does not provide trained data for the Arabic text. Naïve Bayes have shown 88.3117% accuracy of classification while the Support Vector Machine (SVM) has shown an accuracy of 81%.

Monkey Learn software as used in classifying texts by typing the texts and submitting them. The texts were analyzed and classified based on the training dataset that we trained before. The first comment, which is *"Woman is my mother, my sister, my wife and my daughter. If women drive necessary, then we all support it,"* was categorized correctly as a positive.

Finally, we visualized the amount of each category in different locations. The location represented the place of users who had posted the tweet(s). The majorities of the tweets were from Saudi Arabia and while others originated from other countries, including Qatar, Baghdad, England, Kuwait, USA, and Canada. Moreover, positive comments were more than negative. There was missing data regarding some of the users' countries, as some of them had not declared their locations (or countries) in their profiles. To address this dilemma, we replaced the missing values by null; a step accomplished during the pre-processing stage.

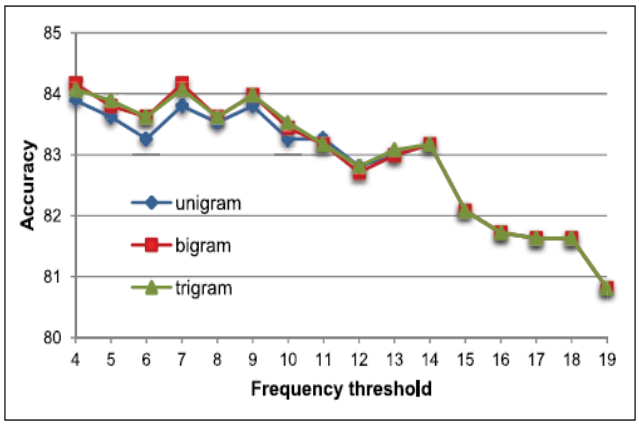

Figure 4 - Accuracy in Low-frequency Feature Removal for the Proposed Hybrid Classifier

V. Conclusions

Social media platforms such as Twitter are very valuable and form platforms that enable individuals to capture opinions surrounding a subject – and applying the sentiment analysis technique. Data mining techniques help to discover, predict, and explore the hidden information and opinion from text data. In this work, we presented the problem of getting the sentiment of Arabic texts that are related to the tweets produced in Saudi Arabia. We encountered challenges about analyzing the Arabic content, especially those that make the Arabic sentiment analysis more complex than other languages. For future work, there is a need to use the unsupervised method to build a lexicon of Arabic words and

10.9756/SIJCNCE/V7I5/07051011
ISSN: 2321–2403

implement the bag of word models, which might help to enhance the goal of sentiment analysis. In addition, there is a need to extend the current work by taking into account different domains – and analyzing more amounts of data. Also, it is important to consider the re-tweet tweets and the site users' emotions, which might suggest interesting insights regarding the role of sentiment analysis.

REFERENCES

[1] Ahlqvist, T., Bäck, A., Halonen, M., & Heinonen, S. (2008). Social media roadmaps: exploring the futures triggered by social media. *VTT Tiedotteita Research Notes 2454, Espoo, Finland*, 2008.

[2] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, *53*(1), 59-68.

[3] Nielsen Social Media Report 2012, http://www.nielsen.com /us/en/newswire/2012/social-media-report-2012-social-media -comesof-age.html

[4] Dubai School of Government. Arab Social Media Report (ASMR), 2010, issue no. 2, http://www.dsg.ae/en /ASMR2/ASMRHome2.aspx

[5] Mart.nez-C.mara E, Mart.n-Valdivia, MT, Ure.a-L.pez LA., & Montejo-R.ez AR. (2013). Sentiment analysis in Twitter. *Natural Language Engineering*, 20(1), 1–28.

[6] Arab News. #Saudi Arabic world's 2nd most Twitter-happy nation. Issue 20 May 2013.

[7] BBC. Saudi join Twitter campaign for higher pay, 13 September 2013.

[8] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135.

[9] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1): 1–167;

[10] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 79-86.

[11] Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010), 627-666.

[12] Somasundaran, S., Wilson, T., Wiebe, J., & Stoyanov, V. (2007). QA with Attitude: Exploiting Opinion Type Analysis for Improving Question Answering in On-line Discussions and the News. In *ICWSM*.

[13] O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth International AAAI Conference on Weblogs and Social Media*, 122–129.

[14] Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, 492-499.

[15] Tumasjan, A., Sprenger, T.O., Sandner, P.G., & Welpe, I.M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media*, 178–185.

10.9756/SIJCNCE/V7I5/07051011
ISSN: 2321–2403