SENTIMENT ANALYSIS FOR IRAQIS DIALECT IN SOCIAL MEDIA USING MACHINE LEARNING ALGORITHMS

Bayan M. Sabbar[1], Naeem T. Yousir[2], Lamiaa A. Habeeb[3]
College of Information Engineering, Al-Nahrain University, Baghdad, Iraq
dr_b2012@yahoo.co.uk[1], naeemms@yahoo.com[2], abd.lamiaa@yahoo.com[3]
Received:25/3/2018, Accepted:20/5/2018

***Abstract*-In this paper, we designed a system that extract citizens opinion about Iraqis government and Iraqis politicians through analyze their comments from Facebook (social media network). People express their opinion in a random way so extract cleared pattern form their comments need several pre-processing steps, first, we cleaned the dataset by remove special characters, normalize the letters that written in different shape and remove repeated characters. Also, a stemmer specializes for Iraqi dialect was built to stem the vocabulary as much as possible, this is the first attempt to build stemmer for Iraqis dialect. Cleaning and stemming reduce the number of vocabulary in our dataset from 28968 to 17083, these reductions caused reduction in memory size from 382858 bytes to 197102 bytes. Generally, there are two approaches to extract users opinion; namely, lexicon-based approach and machine learning approach. In our work, machine learning approach is applied with three machine learning algorithm which are; NaÃ¯ve base, K-Nearest neighbor and AdaBoost ensemble machine learning algorithm. For NaÃ¯ve base, we apply two models; Bernoulli and Multinomial models. We found that, NaÃ¯ve base with Multinomial models give highest accuracy.**

Keywords: **Iraqis stemmer, Social media analysis, Opinion mining, Text mining**

I. INTRODUCTION

Data mining is a particular data analysis technique that adhered to the process of discovering linkage and patterns in a huge data sets to predict outcomes. Today, social media usage is resurgent increasingly and rapidly growing, but such growing caused a huge unstructured data that belong to a host of domains, including governments, health and business. The increasing importance of social media networks (SMN) recall for data mining techniques that is to simplify reforming the unstructured data and place them within a structured form [1]. Internet based applications, which is called social media, builds on technological and ideological basis of Web 2.0, that allow user to interchange its content. Because of the consistency of social media data, modern data mining methods was called that can effectively treat user generated content with affluent social relations. The development of the new methods is under the field that is called social media mining (SMM). SMM is the job of analyzing, representing, and extracting actionable insights from social media data [2]. Sentiment Analysis (SA) also called opinion mining is an area of study which analyzes people's sentiments, opinions, evaluations, emotions and attitudes from written language. It is concerned with text mining, web mining and data mining. For SA, in most cases, the supervised learning algorithm of this technique are employed. It consists of five Phases: collection of data, pre-processing, training the data, classifying data then plotting results. To train the data, we should provide a collection of labeled corpora. The classifier is offered a set of feature vectors from the training data. A model is building based on the training data set which is employed over the unseen/new text for classification goal [3] Facebook is a social networks service and an online social media its Located in California (Menlo Park). First launched on Feb. 4, 2004 and by Mark Zuckerberg, together with his friends Harvard College students Chris Hughes, Dustin Moskovitz, Eduardo Saverin and Andrew McCollum [4]. FB is the most widely used social networking site in Iraq with the usage about 97.15%,

followed by twitter with 1.67%, 0.78 for Google+ and 0.24% for the remaining social networks using mobile internet [5]. Since the advent of FB, citizens are more joint to politics than ever before, they get the political news directly from a politician's fan page Instead of searching the Internet or watching TV for Fresh information. They can interact with elected officials and candidates for important issues by posting on their walls or sending private messages. Personal connection with politicians gives citizens the power to hold lawmakers accountable for their actions and words and instant access to political information. Citizens can honestly express their opinions about governments or Politicians if they are Satisfied with their governments and Politicians or not [6]. In this work, we built a stemmer to handle Iraqis dialect since there is no previous work in this field. We also expand stop words by adding 340 stop words of Iraqis dialect to modern stander Arabic(MSA).

II. PREVIOUS WORK

Arabic SA area has been performed by Many studies. Researchers have proposed interesting approaches and developed various systems to transact with this problem, for instance in reference [7], two approaches to SA for the Arabic language which are machine learning approach and lexicon-based approach were addresses. A thousand Tweets for each positive and negative class from twitter have been collected about different topics such as: arts and politics and they deal with Modern Standard Arabic (MSA) and Jordanian dialect. For the first approach, RapidMiner software was used to implement the experimentation, Support Vector Machines (SVM), NaÃ¯ve Bayes (NB), Decision Trees (D-Tree) and K-Nearest Neighbors (KNN) were used, they found that SVM and NB have better accuracy than other classifiers which was 87.2% and 81.3% for SVM and NB respectively, while the accuracy of KNN and D-Tree were about 51.45% and 50% respectively. The results from both approaches shows that lexicon-based approach has much lower accuracy compared with the machine learning approach.

In reference [8] researchers extract useful information about user's sentiments and behaviors from Tunisian user's statuses on "FB" during the "Arabic Spring" era. They collect 260 FB statuses through the Tunisian revolution [01/01/2011 - 01/06/2011]. NB and SVM were used. A sentiment lexicon based on acronyms, interjections and emoticons were constructed from extracted statuses. WEKA which is machine learning toolkit was used for their experiments. SVM's algorithm outperformed NB in all cases, SVM achieves 89.2% F-measure for positive class and 60% for negative class, while with NB, F-measure was 75% and 73.1% for positive and negative class respectively. The accuracy of positive class was better than negative once because of that the model was built with many more positive statuses than negative ones and the testing data contains mostly positive samples.

At [9] researchers describe an enterprise system developed for extracting sentiment from vast volumes of social data in Egyptian and Saudi dialects. They use rule-based approach to perform SA and the data set was Twitter data focus on three cases: Telecommunications in Egyptian Arabic, Government in Egyptian Arabic, and Employment in Saudi Arabic. The data collected based on location from which the tweets originated, time of the tweets posted and language variations in the tweets. They build lexicon contain feathers words with its polarity and also build blockers lexicon. After normalize the data set: polarity words and blocker terms were extracted using polarity lexicons and blockers lexicon respectively.

In [10] researchers achieved SA in Arabic reviews using machine learning approach. KNN, SVM and NB algorithms were applied on data set collected from tweeter and FB social media network, this data set addresses different topic such as political news, sports and education with size of 2591 for positive and negative Tweet/Comment. classification models built used RapidMiner software. They labeled the tweets using the crowdsourcing tool and manually labeled FB comments. The data set processed used four operations Tokenization, Stem(Arabic), Stop words(Arabic) removable and Generate-n-Grams(Terms). The data was divided into training and testing sets used 10-fold cross validation. SVM achieved the best precision equals to 75.25, best recall was accomplished by KNN equals to 69.04.

Reference [11] introduces an Arabic dataset that was collected from Twitter for opinions about health services. NB, SVM and LR beside Deep and Convolutional Neural Networks were used in their experiment, the final dataset size after filtering was 2026 tweets which is annotated manually by three annotators to positive or negative class. The NB algorithm involved Bernoulli Naive Bayes and Multinomial Naive Bayes models and the SVMs involved Linear Support Vector Classification, Support Vector Classification, Nu-Support Vector Classification and Stochastic Gradient Descent. The best Accuracy was 91.87% with Stochastic Gradient Descent, on the other side, the Accuracy for both Deep and Convolutional Neural Networks were 85% and 90% respectively. Reference [12] designed and implemented Arabic text classifier regarded to King Abdul-Aziz University student's opinions. SVM and NB were used. Total dataset size was 1121 tweets grabbed from twitter that were labeled manually into (-1, 0, 1) for negative, neutral and positive classes respectively. Two experiments were performed, firstly, positive and negative classes were used only and second with a neutral class. For positive and negative classes only, the best accuracy was 84.84% and 73.15% when neutral class was added, that was achieved by SVM with n-gram feature.

III. METHOD/EXPERIMENTAL WORK

Machine Learning, in general, involves automatic computing processing based on binary or logical operations, that learn the function from series of samples. The goal of machine learning is to generate classifier simple enough to be understood easily by the human [13]. Machine learning can be considered as the most famous techniques having interest of researcher because of its accuracy and adaptability. For SA, in most cases, the supervised learning algorithm of this technique are employed. It consists of three phases: collection of Data, Pre-processing, Training the data, classifying data then plotting results [14]. Two ML algorithm used in this work which are NaÃ¯ve base (with MultinomialNB [15]) and KNN [16] [17].

1) Data Collection: Facebook comments regarded to politic topic collected form politic pages that contains the comments of users about Iraqis governments and Politicians. Python scripts used to capture thousands of comments from FB using the FB Graph API, all steps that explain how to create an access token which never expires and how to scrap data from FB is explained in [18].

2) Filtering: After data collection stage, now we have thousands of FB comments, but we found that the collected comments have many problems that effects on accuracy. First problem is that these data contain comments that don't have any opinion because we didn't find enough posts that ask people about their opinion honestly, so, we read these comments one by one and the comments that didn't have any opinion was removed. Also, the comments contain links, mention comments and photo (photo scraps as '[[photo]]'); all these comments were removed.

3) Annotation: After filtering stage and to train the system, we labeled the comments manually into two classes with (1 and 0) labels which represents positive and negative class respectively. We annotated 12k comments, 6800 comments as negative class with 0 label and 4200 comments as positive class with 1 label. Negative comments more than positive because of that; people's opinion about governments and politicians was not good, so we had difficulty with collecting positive comments. We read each comment and assign label to it, assuming that each comment has opinion. Fig.1 shows steps of filtering and annotated comments .

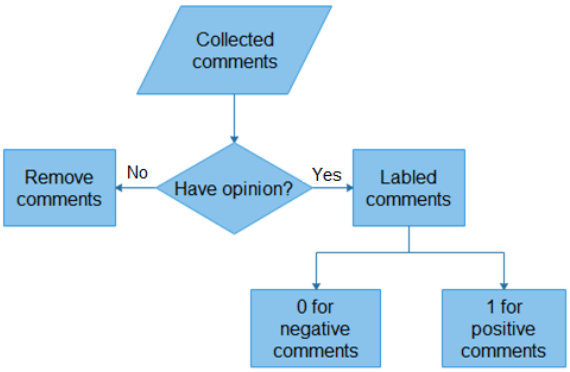

Figure 1: Filtering and annotated comments

4) Data Preprocessing: Since we are deal with the noisy nature of Iraqis dialect on FB social site, the data need to be pre-processed to reduce huge number of vocabulary because these vocabularies used as a features next time.

5) Tokenization: First of all, the text will tokenize into tokens or words to treat with each token separately, the text tokenized depending on the spaces between tokens.

6) Data Cleaning: Cleaning the data include the following steps:

- Remove special char such as "@, $, %, ^, &, *, : < >" and so on.
- Remove numbers and non-Arabic text.
- Remove repeated char form token and excluded the word that have repeated char in its original form.
- Normalize the letter as following:
  - (إ , آ , أ) replaced with (ا)
  - (ة) replaced with (ه) from the end of the word.
  - (ؤ) replaced with (و)
  - (ظ) replaced with (ض)

7) Stemming: The process of reducing words to their original root (base form) called Stemming, in our work this step is difficult because we treat with dialect not modern stander Arabic, it's difficult to get pattern from Iraqis dialect. Increasing difficulty that; we deal with data from social media, users write their comments and posts in different way so, getting rules

helps to stem the words is difficult. We try to reduce words as much as possible but not actually get stem because this task nearly impossible, also in Iraq there is difference in dialect according to the region. Following steps explain our trying to stem words:

- Remove prefix consisting of 3 letter (يال , كال, فال , وال, بال, ولل) for words with length >6 .
- Remove prefix that consist of 2 letters (ال , لل) for words with length >5 .
- Remove " و"from beginning word and exclude words that beginning with "و" in its origin shape, for example (وزير, ولد, وضح, وجه …….etc.) from this process.
- Extract the root from the following pattern according to its length, the pattern may have only suffix or prefix or suffix and prefix, table 1 shows this pre-process.

8) Stop Word Removable: In natural language processing, stop words means word that do not have any meaning such as pronouns, prepositions, conjunctions etc. and are thus eliminated prior to classification. Stop word different according to language and even dialect, so there is no predefined list of stop words, we build list of stop word related to Iraqis dialect and this list add to stop word found in [19] which is related to MSA, total number of stop words for both MSA and Iraqis dialect is 925 words.

TABLE I
EXTRACT THE ROOT FROM VOCABULARY ACCORDING TO ITS LENGTH

Length of pattern	Pattern	Root
9	مفاعلاتكم – مفاعلاتجن – مفاعلاتهم – مفاعلاتهن – مفاعلاتهه – مفاعلاتنه -	مفاعله
8	فعلنالهه – فعلنالهم – فعلنالهن – فعلنالكم – فعلنالجن – فعلتيلهه – فعلتيلهم – فعلتيلهن – فعلتولهه – فعلتولهم – فعلتولهن – فعلتولنه – يفعلولنه – يفعلولهم – يفعلولهن – يفعلولجن – يفعلولكم – تفعلولهه – يفعلولهه – تفعلولهم – تفعلولهن – افعلولهن – افعلولهه – افعلولهم – تفعلولنه – افعلولنه- فاعلناكم	فعل
7	فاعلوهم – فاعلوكم – فاعلوهن – فاعلوجن – فاعلتني – فاعلتهم – فاعلتهن – فاعلتنه – فاعلتكم – فاعلتجن – فاعلتهه – افاعلكم – نفاعلكم – تفاعلكم – يفاعلكم – افاعلجن – نفاعلجن – تفاعلجن - يفاعلجن	فاعل
	افعالكم –افعالجن - افعالهه – افعالهم – افعالهن – افعالنه تفعيلهم -تفعيلنه – تفعيلهن – تفعيلكم – تفعيلجن – تفعيلهه – فعلاتهم – فعلاتهن – فعلاتجن – فعلاتكم – فعلاتنه - فعلاتهه – افعللهه – يفعلوهم – افعللهم – يفعلوهن – افعللهن – يفعلوني – افعللجن – يفعلولك – تفعللهه – يفعلولج – نفعللهم – تفعللهم – يفعللهن – تفعللهن – نفعللهن – نفعللكم – نفعللجن – يفعللهه – يفعللهم – يفعلوله – يفعللكم – تفعللكم – يفعللجن – تفعللجن - يفعلوهه – يفعلونه – يفعلوجن –	فعل

	يفعلوكم – يفعلوله – تفعلوله – تفعلونه – تفعلو هن – تفعلو هم – تفعلو هه – تفعليني – يفعلنهه – تفعلينه – تفعليلي – فعلتلهه – فعلتلهم – فعلتلهن – فعلتلكم – فعلتلجن – فعلنالهه – فعلناهه – فعلناهم – فعلناهن – فعلنالك – فعلنالج – فعلناكم – فعلناجن – فعلولهه – فعلولهم – فعلولهن – فعلولكم – فعلولجن – فعلتيله – فعلتيهم – فعلتيهن – فعلتوله – فعلتو هه – فعلتو هم – فعلتو هن – فعلتنهه – فعلتنهم- فعلتنهن – فعلتولي – فعلتوني – فعلتونه – فعلتلنه – فعلناهه	
6	فعلتله – فعلتهه – فعلتهم – فعلتهن – فعلتلك – فعلتلج – فعلتجن – فعلتكم – فعلناك – فعلناج – فعللهه – فعللهم – فعللهن – فعللكم – فعللجن – فعلوله – فعلو هه – فعلو هم – فعلو هن – فعلولك – فعلولج – فعلوكم – فعلوجن – فعلنهن – فعلنهم – فعلتنه – فعلولي – فعلوني – فعلتني – فعلتلي – فعلونه – افعلله – افعلهه – تفعلله – تفعلهه – افعلهم – افعلهن – يفعلون – تفعلون – افعلكم – تفعلكم – افعلجن – تفعلجن – نفعلله – نفعلهم – تفعلهم – نفعلهن – نفعللك – تفعللك – نفعللج – تفعللج – نفعلكم – تفعلكم – يفعلكم – يفعلله – يفعللك – يفعلجن – يفعلهه – يفعلهم – يفعلهن – يفعللج – يفعلوج – يفعلوك – تفعلنه – يفعلني – تفعلني – يفعلنه – فعلاته – افعاله – تفعيلك - تفعيله	فعل
	فعالنه – فعالهم – فعالكم – فعالجن – فعالهه – فعالهن	فعال
	فاعلتي – فاعلنه – فاعلتم – فاعلتن – فاعلهم – فاعلهن – فاعلتك – فاعلتج	فاعل
5	فعلته – فعلنه – فعلله – فعلهه – فعلهم – فعلهن – فعللك – فعللج – فعلكم – فعلجن – فعلتك – فعلتج – فعلتي – فعلني – افعله – افعلك – افعلج – نفعله – نفعلك – يفعله – يفعلك – يفعلج – تفعله – تفعلج – تفعلك – افعلي – أفعال - افاعل	فعل
	فعاله – فعالي – فعالك – فعالج	فعال
	فاعلت	فاعل

9) Classification : To train the classifiers, the dataset will split into two datasets 70% for training and 30% for testing, to have enough samples for training and depending on previous literature, this splitting has been chosen. 5-fold Cross-Validation [20] used to assess the predictive performance of the models. The accuracy of the models is calculated according to [21].

IV. Results and Discussion

1) Dataset Collection : FB scraper tool used to collect FB comments regarding to Iraqis Governments and Politicians topic. After filtering step, we have 4200 and 6800 comments for positive and negative opinions respectively to be used in our work as showed in Fig.2

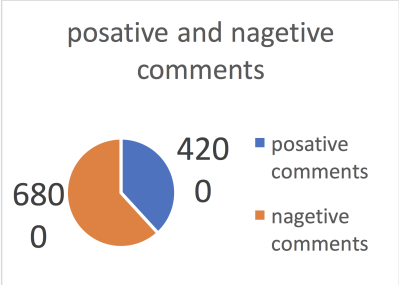

Figure 2: The collected comments

2) The Effective of Preprocessing onto the Data: Preprocessing stage consist of data cleaning and normalization (C&N), stemming and stop word removable. After these stage, the number of vocabulary reduced from 28968 to 16688, also memory size reduced from 382858 bytes to 193564 bytes as in Fig.3 and Fig. 4. The effective of this stage onto the MultinomialNB classifier is as shown in Fig. 5 a, Fig. 5 b, Fig. 5 c and Fig. 5 d.

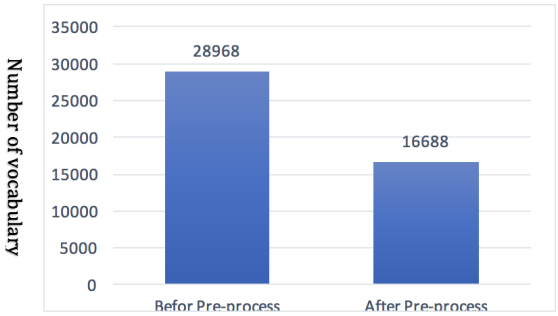

Figure 3: Number of vocabulary after Pre-Processing

Figure 4: Memory size in bytes after Pre-Processing

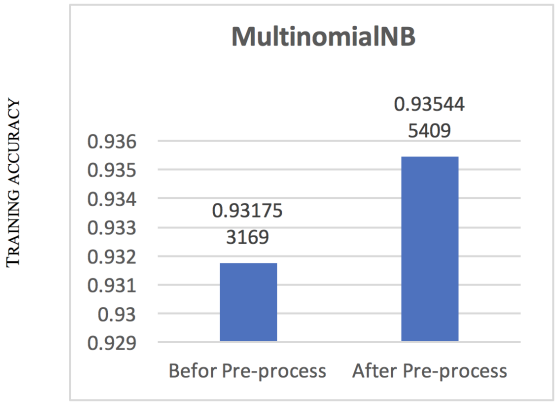

Figure 5 a: Training accuracy with MultinomialNB after Pre-Processing

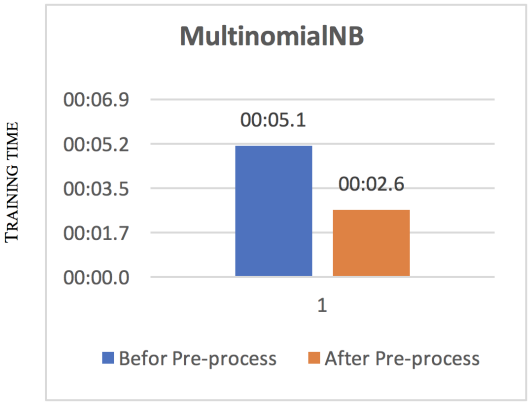

Figure 5 b: Training time with MultinomialNB after Pre-Processing

The accuracy increased in both steps, and there is significant reduction in time, the time of prediction step more than training time because of that; KNN is Instance based learning. So, in prediction step, the reduction in time is significant.

Because of the reduction that happened in the number of vocabulary and since KNN is Instance-based learning, prediction time will reduce.

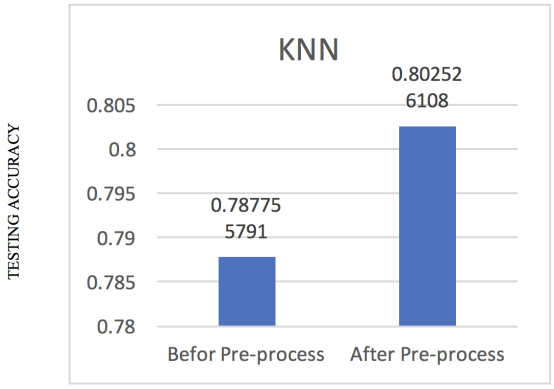

Figure 6 a: Testing accuracy with KNN after Pre-Processing

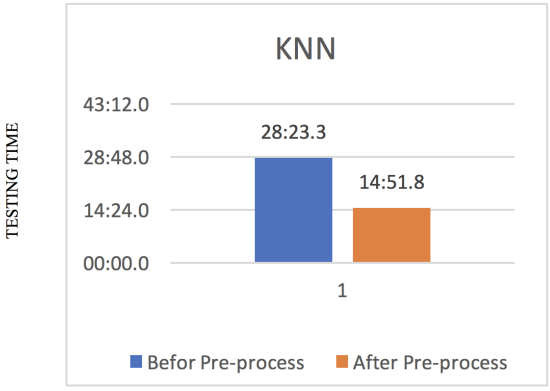

Figure 6 a: Testing time with KNN after Pre-Processing

In general, with MultinomialNB we get higher accuracy and less time, MultinomialNB work good with text classification, also the time of prediction step for MultinomialNB much less than prediction time of KNN. Finally, and in our work, we found that; MultinomialNB much better than KNN in the case of text classification.

V. Conclusion

Stemming was the most step that effect onto our dataset, the reduction of the vocabulary after stemming was significant so that because of only stemming the number of vocabulary reduced by nearly 33%, while reduced by 10 because of cleaning and reduced by 3% because of stop words removable, in spite of the difficulty of stemming Iraqis dialect, our stemmer was good and the results was satisfied. Our stemmer stemmed words with length > 5 almost correctly, but for words with length ≤ 5 this task will be difficult and causes error. MultinomialNB was much better KNN in the case of text classification. There are a lot of comments found in politic pages that don't have any opinion such as appeal comments, we can't obtain accurate result from our classifiers with these comments, so to be close to reality the classifiers must be trained with these comments which can be add as a third class called 'natural class'

References

[1] M.A.Olowe, M. M. Gaber, and F. Stahl, "A survey of data mining techniques for social media analysis", arXiv preprint arXiv:1312.4617, 2013
[2] R. Zafarani, M. A. Abbasi and H. Liu, "Social media mining: an introduction", Cambridge University Press, 2014.
[3] S. B. Hamouda and J. Akaichi, Social Networks Text Mining for Sentiment Classification: The case of Facebook statuses updates in the Arabic Spring Era, ISSN 2319 4847 IJAIEM , 2013.
[4] N. B. Ellison, "Social network sites: Definition, history, and scholar ship", Journal of computer mediated Communication13, no. 1, 210-230, 2007.
[5] https://www.statsmonkey.com/sunburst/21375-iraq-mobile-social-media-usage-statistics-2015.php
[6] "https://www.lifewire.com/facebook-and-politics-1240558"
[7] N. A. Abdulla, N. A. Ahmed, M.A.Shehab and M. Al-Ayyoub, "Arabic Sentiment Analysis: Lexicon-based and Corpus-based", IEEE, 2013.
[8] S. B. Hamouda and J. Akaichi , Social Networks Text Mining for Sentiment Classification: The case of Facebook statuses updates in the Arabic Spring Era, ISSN 2319 4847 IJAIEM, 2013.
[9] H. Wang, V. R. Bommireddipalli, A. Hanafy, M.Bahgat, S. Noeman, O.S. Emam, "A System for Extracting Sentiment from Large-Scale Arabic Social Data".
[10] R.M. Duwairi, I.Qarqaz, "Arabic Sentiment Analysis using Supervised Classification", Irbid 22110, Jordan, 2014.
[11] A. M. Alayba, V.Palade, M. England and R. Iqbal, " Arabic Language Sentiment Analysis on Health Services".
[12] H. AL-Rubaiee, R.Qiu, K. Alomar and D. Li, "Sentiment Analysis of Arabic Tweets in e-Learning", 2016.

[13] D. Michie, D.J. Spiegelhalter, C.C. Taylor," Machine Learning, Neural and Statistical Classification", February 17, 1994.
[14] H. Thakkar and D. Patel. "Approaches for sentiment analysis on twitter: A state-of-art study", arXiv preprint arXiv:1512.01043, 2015.
[15] H. Shimodaira, "Text classification using naive bayes", Learning and Data Note 7, 1-9, 2014.
[16] https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/
[17] Roiss Alhutaish and Nazlia Omar. "Arabic text classification using k-nearest neighbour algorithm." Int. Arab J. Inf. Technol.(IAJIT) 12, 190-195, 2015.
[18] "https://nocodewebscraping.com/facebook-scraper/"
[19] https://raw.githubusercontent.com/stopwords-iso/stopwords- ar/master/stopwords-ar.json
[20] P. Refaeilzadeh, L.Tang, and H. Liu. "Cross-validation", In Encyclopedia of database systems, pp. 532-538. Springer US, 2009.
[21] http://scikit-learn.org/stable/modules/modelevaluation.html/accuracy-score