RESEARCH ARTICLE OPEN ACCESS

Applying Sentiment Analysis on Arabic comments in Sudanese Dialect

Islam Saif Aldin Mukhtar Heamida [1], EL Samani Abd Elmutalib Ahmed [2], Mamdoh Noureldin Elsayed Mohamed [3], Abd Alhameed Adam Ahmed Salih [4]
Faculty of Computer Science and information technology, University of Alneelain
Faculty of Mathematical Sciences, Khartoum University
Sudan

ABSTRACT

This paper addresses the problem of sentiment classification for the Sudanese text on Facebook using machine learning methods. One of the features of the Sudanese language is the use of a variety of endings depending on deviation, times, and grammatical gender .The Sudanese colloquial have no grammatical or morphological rules, for example, in their slang, they enter the present tense on the word, another common problem in sentiment classification for different languages is that a single word could refer to different meanings, for example, the word "Salem "can refer to a name and also an adjective. Our task was to evaluate how the processing steps and lemmatization libraries that we used affected a Sudanese colloquial text. Two different classifiers were applied; SVM & Naïve Bayes NB to classify comments based on their polarity, whether positive, negative, or neutral. The work was evaluated with four different measures. The results revealed that the use of SVM with lemmatization libraries improves the accuracy of sentiment classification; SVM achieved the best measurement accuracy of 68.6%, while NB achieved 63.1%.

Keywords :- Opinion Mining، Sentiment Analysis, Sudanese Colloquial Dialect.

I. INTRODUCTION

Sentiment analysis is a NLP method [1] that is implemented on a text to determine whether the author's intentions toward a particular topic or product or so are positive, negative, or neutral [2], the Arabic language is divided into three categories: Classical Arabic, Modern Standard Arabic, and Dialect Arabic, (Soliman et, al 2014), Sentiment analysis has been extensively studied in English language literature [3] [13] and many NLP tools are available for this. A significant contribution has been made to the development of sentiment analysis for text messages by researchers from Cornell University (B. Pang and L. Lee) [4][5][6] in 2008. Bang and Lee published the book (Opinion mining and sentiment analysis) [4] which is dedicated to modern methods and techniques for analyzing feelings in text messages [5]. Compared to Arabic texts, relatively few works have been dedicated to study sentiment analysis [7]. Most of the web content is written in a Dialect that has not been studied as much, as far as we know. On the Internet, the content of official websites may be in classical Arabic, but most of the content of social networks is in local dialects, as we know, there are more than 65 local dialects distributed over 22 Arab countries, studies have shown that the software tools developed to address classical Arabic yield poor results with these dialects [8].

the Sudanese are definitely part of the Arab world. Official statistics indicate that the number of users of social media in Sudan reached 15 million people, those open platforms allowed young people to reflect their opinions in their circulating dialect, which increased the need of stakeholders and researchers in applying opinion analysis to the Arabic texts to study those opinions in different areas whether it was political, Economic, or areas of services, etc., where this study aimed at analyzing feelings, analyzing the text written in the Sudanese vernacular based on machine learning methods where the classifier is trained according to available training data. After that, it is possible to classify the Sentiment as positive, negative, and neutral, which was made on 1050 Facebook comments on Internet service in Sudan. As there is no dictionary for Sentiment analyzing about the Sudanese dialect, we have created a dictionary containing 1000 words categorized into positive and negative words.

II. RELATED WORK

Nafissa Yussupova, Diana Bogdanova, Maxim Boyko, emotion analysis to text in Russian based on machine learning approaches [9]. Description of the researchers the problem of sentiment classification in text messages in Russian using machine learning methods - Naive Bayes translator and Support Vector. One of the features of the Russian language is the use of a variety of endings depending on deviation, times, and grammatical gender. Another common problem in sentiment classification for different languages is that different words can have the same meaning (synonyms), thus they may give equal emotional value. Therefore, their task was, how the reduction affects the accuracy of sentiment classification (or other, with or without endings), and to compare the results to Russian and English languages. To assess the effect of synonyms, they use the method of merging

words that refer to the same meaning into a single term. To solve these problems, they used lemmatization libraries and synonyms. The results revealed that the use of the lemmatization of texts in Russian improves the accuracy of sentiment classification. On the contrary, sentiment classification for texts in English without the use of lemmatization gives a better result. The results also showed that the use of synonyms in the form has a positive effect on accuracy.

European Commission Joint Research Centre [10]. Presented a method for sentiment analysis specifically designed to work with Twitter data (tweets), taking in to account their structure, length, and specific language. The approach employed makes it easily extendable to other languages and makes it able to process tweets in near real-time. The main contributions of this work are: a)the pre-processing of tweets to normalize the language and generalize the vocabulary employed to express sentiment; b) the use minimal linguistic processing, which makes the approach easily portable to other languages; c) the inclusion of higher-order n-grams to spot modifications in the polarity of the sentiment expressed; d)the use of simple heuristics for features to be employed; e) applying supervised learning using a simple Support Vector Machines linear classifier on a set of realistic data. We show that using the training models generated with the method described we can improve the sentiment classification performance.

A student introduced the Ph.D. Alaa Fadl Al-Hasan, September/2016 a polarity lexicon for Palestinian dialect to improve the accuracy of sentiment analysis for short text in Palestinian dialect. Suggest supervised learning models: Naïve Base (NB) and Support Vector Machine (SVM) for sentiment analysis in Palestinian dialect. The results have shown that the use of the polarity lexicon improves the accuracy of sentiment classification of text in Palestinian dialects. The best result obtained is at 91.78% accuracy for an SVM-based binary sentiment classification model [11].

III. THE PROPOSED METHOD

This paragraph is the same as the researcher in the following set of steps

A. Data label

as a first step, we collected about 1050 comments from Facebook users in the Sudanese dialect and classify it manually by language experts to positive, negative, and neutral. 512 of them were negative, 328 positive, and 210 neutrals.

B. Text pre-processing

in this step, we load data into (Jupyter) to implement several processing techniques for proposed data (NLP) using Python libraries.

C. Data cleaning

in the first step of processing, the researcher removed the spelling inconsistencies, i.e. the standardization of Arabic letters as well as useless data, such as URLs, addresses, names of people, tags, English characters, and numbers, for example (@, &). The researcher also noticed that the data contains some repeated sentences or repeated letters by commentators so to confirm their feelings, as it was removed and set it in the normal context, for example “"سيئئئئئئئئئة" and "غاااالية" changed to"سيئة" "غالية" and "الشبكة كعبة كعبة كعبة"changed to "الشبكة كعبة".

D. Removing Arabic stop- words

There are so many words that are frequently used, and thus they lose their value in meaning, like pronouns"هي"and conjunctions"و" , it is usually a form of linguistic noise and removing it helps in learning to focus on the rest of the sentence. Therefore, the researcher removed some phrases and words that we do not need by filtering the Sudanese stop-words, for example, stop-words in (صباح الخير خدمات الشبكة كعبة " وأسعار الخدمات زايدنها من غير أي مبرر ولا إعتذار حقيقة كمقترح شركة جدا سوداني لو حا تستمر في عدم إستقرار الشبكة قسم بالله حا تفقد العملاء بمجرد ما (تنافسها شركه تانية …. are “لو”,”في “,”بس“,”لقد”.

E. Text Tokenization

Dividing text into words, for example, (وضع الانترنت في السودان لا يطاق) after the Tokenization process " وضع" ، " الانترنت " ، "في" ، "السودان" ، " لا يطاق".

F. Normalization

The text used in the data collected presents many challenges when compared to the formally structured text as it turns out from Nabil and others [12]. It contains unorganized language, many spelling errors, colloquial words, colloquial expressions, acronyms, and idiomatic expressions, and most of them are contradictions. Moreover, some words contain more than one model, which is an issue that highlights the need for normalization, i.e. standardization of Arabic letters applied by the researcher to all data.

TABLE I
EXAMPLE OF PRE-PROCESSING A COMMENT

Pre-processing Steps	comments After Preprocessing
The original comment	شبكه الانترنت سيئة سيئة خالص sudani@ وتكلفتها عاااالية! كمااان!!
Data cleaning	شبكه الانترنت سيئة سيئة خالص وتكلفتها عاااالية كمااان
Normalization	شبكه الانترنت سيئه سيئه خالص وتكلفتها عاااليه كمااان
Removing duplicated word or characters	شبكه الانترنت سيئه خالص وتكلفتها عاليه كمان
Tokenization	شبكه ,الانترنت, سيئه , خالص ,وتكلفتها ,عاليه, كمان
Stop remove word	شبكه الانترنت سيئه خالص وتكلفتها عاليه
Lemmatization	شبكه الانترنت سيئ خالص التكلفة عالي

The researcher also applied different variations to other pre-processing steps in order to study their effects on the accuracy of classification results, its implementation showed clear

results in classification accuracy, these steps can be one of the following:

G. Lemmatization

The purpose of this step is to return speech [13] to its origin, which helps the computer to understand because it reduces the number of vocabulary that it must learn, furthermore the difference in the shape of words might be difficult for the computer to recognize words that have similar meanings. There are two ways to return speech to its normal context. The researcher used the method of lemmatization libraries. In this method, the words are returned to the closest common word in the meaning even if they differ in the source, for example, (جمل => إبل جمال, ناقة) sometimes Sudanese people give the word extra letters like (و, قد, إلخ). All of this was handled by the Lexicon Polarity dictionary which was prepared manually by the researcher. The dictionary contains the source of negative and positive words that express feelings [14].

TABLE 2
LEMMATIZATION OF WORDS

the words	Lemmatization
بانتظام , منتظمة, و منتظمة , انتظام	منتظم
نافعه مفيدة نافعانا	نافع
تندم , وتندم , ندمت , وندمت , تندمك, وتندمك, ندمان وندمان , ندمانه , وندمانه, تندم , ندما	ندم
اسواء, واسواء , سيئ, سوء, وسيءه , سيئه	سيئ

Note: Lemmatization It is the origin of each word as found in the lexicon of polarity

H. Automated Learning Algorithms as a Tool

The data is divided into training and testing data set (data is divided into two parts, 70% of the data is for training and 30% is for testing). The set of training data used for classification based on SVM and NB classifiers and the data were classified based on their polarity into positive, negative and Neutral categories, while the test dataset is used to predict the polarity of comments. Vector Support Machine (SVM) is a supervised learning model [15] that can be used for either classification or regression challenges. However, it is one of the common methods of classification technology that analyzes data and identifies patterns based on a discriminatory work defined by an overly differentiated class .Which is categorized based on drawing separation lines to distinguish between organisms that have a different organic in separation. Naïve Bayes is a simple model, the classifier model based on Bayesian theory [16] with assumptions of independence between features. We used the Naïve Base (NB) classification model to verify the lateral division used in training data, applying model, and performance process in data testing.

I. Building a Matrix of Numbers

In this step, an array will be created which contains a set of rows "data information", which represent the fundamental comments file, and the "header information" columns represent all words/terms in all comments files after implementing all processing operations. The last columns are Class or Label, which are positive, negative, or neutral according to the file that was manually categorized in advance. The value of each row intersection with a column in the array is determined by typing the categorized word. The researcher compared the comments file with the polarity glossary, in order to filter the comments from words that have no meaning or feeling and only keep the words that have feelings. The comment file was programmatically loaded and each word in the comment file is compared with each word in the positive and negative glossary file, if the word is found in the glossary file we encode it to (1, -1, 2) according to its classification, otherwise, it will not be added.

Documents/term	سوداني	الشبكة	زفت	دسيس	شديد	سينه	Class
سوداني الشبكة زفت شديد	0	0	زفت/ Negative	0	0	0	Negative
دسيس شديد	0	0	0	دسيس/ positive	0	0	Positive
شبكة مرة سيئة	0	0	0	0	0	سيئ / Negative	Natural

TABLE 3
MATRIX OF NUMBERS

IV. EXPERIMENTATION

1) *NLP with Python:* In this study, the researcher uses a natural language processing kit with Python. Natural language processing tools with Python NLTK is one of the leading platforms for working with human language data and Python, the NLTK module is used to process natural languages. NLTK is an abbreviation stand for Natural Language Toolkit [17].

Classification Techniques The researcher used two types of classification methods: SVM and Naïve Bayes NB.

2) *Data-collection:* The data was collected from Facebook, which is one of the most popular social media in Sudan. The researcher focused on collecting comments written in the Sudanese dialect, where the data collected was (1050) comments expressing the opinions of the subscribers on the posts published by the telecommunications company pages about the internet service provided by them.

3) *Operational environment:* The suitable environment for this work was CPU / Intel Core Pentium i5 processor, 8 GB RAM, 64-bit operating system, Windows 10. In addition, we

used Anaconda platform to prepare our data by implementing the learning steps of the classification model using the polarity glossary for our approach and MS Excel 2016 to set up the data file and convert it to CSV files and pass the result to Jupyter as a classification application tool.

V. RESULT

Four different measures - Precision, Recall, Accuracy and F-Measure for both SVM and NB classifiers - were used to assess the validity of the rating of test comments as positive or negative or Neutral, the results of the experiments described in Tables

Tables of Basic Classifiers Results for 3 classes in terms of accuracy

TABLE 4
SVM CLASSIFIER

Algorithm	Type Process	Precision	Recall	F-measure	Accuracy
SVM	Stp1=data cleaning	**56.%**	**24.2%**	**49.%**	**55.8%**
	#step2=Stp1+Removing Arabic Sudanese Stop Words	**55%**	**56.1%**	**49.%**	**55.5%**
	#Step3=step2+Normalization	**58.3%**	**59.2%**	**54%**	**59.3%**
	Step4=step3+Lemmatization	**85.3%**	**65.6%**	**66.3%**	**68.6%**

From table 3 above we notice that SVM achieved good result For precision which equal to 85.3%

TABLE 4
NB CLASSIFIER

Algorithm	Type Process	Precision	Recall	F-measure	Accuracy
NB	Stp1=data cleaning	**52%**	**40%**	**35%**	**52.3%**
	#step2=Stp1+Removing Arabic Sudanese Stop Words	**53.2%**	**52%**	**52.5 %**	**52.3%**
	#Step3=step2+Normalization	**61%**	**54%**	**54%**	**53.9%**
	Step4=step3+Lemmatization	**72.5%**	**63%**	**59.%**	**63.1%**

From table 3 above we notice that NB achieved good result For precision which equal to 72.5%

From the table above, we observe that if SVM is used in learning a rating model, the best results we get are 68.6% for a triple classification model with a pre-processing step that includes Lemmatization, Normalization, Removing Arabic Sudanese Stop-words, and data cleaning. On the other hand, for the NB classifier, we got the best result, which was 63.1% when focusing only on pre-processing. Again, as we can see from the results, the number decreases a lot when we learn a classifier without the Lemmatization step. These results generally indicate that SVM performed better than NB, but in all cases, we need to improve three-category rating models.

VI. CONCLUSIONS

The evaluation of the various sentiment classification models presented in this paper revealed that SVM-based classifiers yield results of higher accuracy than that of NB-based classifiers. We also found that using Arabic in the pre-processing step of the text would give better results in using the Arabic language. Moreover, the primary processing steps that we suggested in this research greatly improved the accuracy of the classification of sentiment. Besides, our approach to sentiment analysis in which we expand the area of features with features extracted from the Lexicon Polarity glossary improves sentiment classification results. We obtained 85.78% accuracy for Class III Classifiers with SVM Classifier. For future work, we plan to expand our model and polar glossary to improve the accuracy of classification of sentiment for the text in the Sudanese accent. We also plan to publish the Polar Lexicon for public use to classify sentiment, as well as looking at other features that may be discriminatory, such as negations, symbols, and abbreviations.
From table 9 above we notice that, K-Nearest Neighbor achieved good result for Recall which equal to 94.94%.

REFERENCES

[1] https://dovetailapp.com/blog/2018/sentiment-analysis/

[2] BING LIU 2015, Mining opinions, sentiment and Emotion, Sentiment Analysis

[3] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 2(1-2):1–135, January.

[4] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval", Vol. 2, 2008, pp. 1-135.

[5] B. Pang and L. Lee, "Thumbs up? Sentiment Classification using Machine Learning Techniques", Proceedings of the Conference on Empirical Methods in Natural. Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.

[6] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Mini-mum Cuts", Proceedings of the ACL, 2004, pp. 271-278.

[7] Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

[8] Alaa Fadl Al-Hasan, September/2016.PHD, Building a Sentiment Lexicon for the Palestinian Dialect

[9] Nafissa Yussupova, Diana Bogdanova, Maxim Boyko. Applying of Sentiment Analysis for Texts in Russian Based on Machine Learning Approach, The Second International Conference on Advances in Information Mining and Management, no: 978-1-61208-227-1

[17] Perkins, J., 2010. Python text processing with NLTK 2.0 cookbook. Packt Publishing Ltd.

[10] Sentiment Analysis in Social Media Texts. European Commission Joint Research Centre Vie E. Fermi 2749 21027 Ispra (VA), Italy alexandra.balahur@jrc.ec.europa.eu

[11] Building a Sentiment Lexicon for the Palestinian Dialect. Alaa Fadl Al-Hasan, September/2016.PHD

[12] Nabil, M., Aly, M. and Atiya, A. (2015). Astd: Arabic Sentiment Tweets Dataset. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, September 2015. Association for Computational Linguistics

[13] https://www.tutorialspoint.com/python_data_science/python_stemming_and_lemmatization.htm

[14] Ghadah Alwakid, Taha Osman and Thomas Hughes-Roberts, Challenges in Sentiment Analysis for Arabic Social Networks, 2017, 5-6 November 2017, Dubai, United Arab Emirates

[15] Lin, Keng-Pei, and Ming-Syan Chen. "On the design and analysis of the privacy-preserving SVM classifier." IEEE transactions on knowledge and data engineering 23.11 (2010): 1704-1717.

[16] Rain, Callen. "Sentiment analysis in amazon reviews using probabilistic machine learning." *Swarthmore College* (2013).