Informatics Bulletin, Faculty of Computers and Artificial Intelligence, Helwan University
Published Online ***Vol 5 Issue 1, January 2023***
(https://fcihib.journals.ekb.eg)

Aspects Detection Model for Users' Reviews Using Machine Learning Techniques

Saad Khalifa[1,*], Mohamed Marie[1], Mai Eldefrawi[1]
[1]Information System , Computers and Artificial Intelligence , Helwan , Cairo ,Egypt
khalifasaad43@yahoo.com, khalifasaad43@yahoo.com, mai.eldefrawi@gmail.com

***Abstract*—sentiment analysis (SA) is an important field that is concerned with capturing user emotions, modes, and interests through a written text. sentiment analysis (SA) and opinion mining has become a trendy topic today it helps us to catch the emotions of the speaker whether sad or happy etc. As well as some governments may use this field to catch the emotions of their people whether they are satisfied with the way they use to control the country or not. The number of people using social media is rapidly increasing because many people use it to share their opinions, feeling, and emotions so social media platforms like Twitter or Facebook possess a large volume of data for example lots of tweets are posted on Twitter every day, that can be analyzed to capturing the current state or mode of the author. Currently, in the age of coronavirus, many companies obliged their employees to work from home millions of tweets are posted people use this platform to share their opinion, feeling, and their new lifestyle during quarantine. The volume of data become larger, and it became a fertile material for analysis to know how people feel and how they adapt to the new situation. The research objective is to provide a review of using aspect-based sentiment analysis to detect user satisfaction and apply this technique to such a trendy topic, which is working from home during a covid-19 pandemic. In addition to detecting problems, they face.**

***Index Terms*—Sentiment Analysis (SA), Opinion Mining, Social Media, Aspect-Based Sentiment Analysis, Machine Learning (ML).**

I. INTRODUCTION

SOCIAL media make it easy for people to share personal, and opinions. the number of people using social media platforms is increasing rapidly. Social media platforms are used for communication. People always have a point of view (positive or negative) about any event or situation, and they used to express it through social media [1-3]. Sentiment Analysis (SA) and Opinion Mining field are interested in analyzing humans' opinions, modes, situations, and emotions from the text. It is used by organizations, decision-makers, and governments to receive feedback to improve their products and services. It is also used to predict stock market movements and elections. Sentiment analysis or opinion mining attracted attention during the last years because of the great opinionated web content on social media platforms like Facebook, Twitter, Instagram, LinkedIn, etc... are considered a generator of data. Sentiment analysis is a complex task used to identify whether text holds a positive or negative opinion. Because of the number of people using social media increasing significantly business leaders and decision-makers are now making studies and analyses to capture the global opinion about their products and collecting feedback to improve their products and services and attract new customers to keep competing. The number of Arabic speakers using the web is about 185 million around 4.8 % of all users using the internet all over the world. Internet World Stats said that the Arabic language occupies the fourth place in the order of the most used language on the internet [4],[5]. Social platforms like Twitter and Facebook have become popular ways of communication among people. Twitter was established in 2006 it helped the ability to people freely express and share opinions and feelings. Twitter is one of the largest platforms that contain a variety of sentiments. which contains tweets; each tweet has 140 characters There are a lot of tweets posted every day on Twitter Which makes it an ideal example of "big data". Social Networks make it possible for people to share personal information, interests, and opinions. The number of people using social networks is increasing rapidly. In the age of covid-19 social media became the only haven for people to share personal information, interests, and opinions during quarantine.

Arabic content on the internet has increased in volume, especially after the evolution of social networks. On Twitter, there are over 6 million Arabic speakers who write over 10 million tweets per day [1],[4]. Sentiment Analysis (SA) is one of the NLP concepts [1]. This field is used to capture the sentiment of the text. Despite the amount of data is very much now sentiment analysis still in its infancy and a very rich area for development especially in the Arabic language contrary English language because dealing with the Arabic language is considered a great challenge. in this paper we are going to talk about sentiment analysis in the age of corona especially about working from home and its impact on the sentiment of the employee whether they are satisfied or not? many companies obliged their employees to work from home to protect themselves from the fast spread of this virus. There are different

factors in which the employees consider expressing their opinion about working from home. the main objective is to measure the global sentiment about this issue using the help of Twitter data that can be collected and accessed by making Twitter developer accounts to access Twitter data and using the tweepy library [4]. As mentioned above the objective is to capture the sentiment of employees about working from home during the pandemic which can be done with the help of machine learning. Machine learning has rapidly grown and evolved in several applications in a wide variety of research works. Machine learning algorithms play a vital role in building predictive models. These algorithms focus on predicting a higher accuracy. there are many automatic classification techniques, such as Support vector Machine (SVM), Naïve Bayes, and Decision Tree [4]. Here is some Arabic sentiment analysis.

1) ***Spam****: Briefly, this activity consists of writing not real reviews to promote to damage the reputation of a service or product.*
2) ***Sarcasm****: expressing a positive opinion but referring to making fun of something.*
3) ***Arabic morphology****: adding suffixes, prefixes and affixes same word may have important information.*
شعر – شاعر – شعور
4) ***Lack of Data set****: it is difficult to find your corpora.*
5) ***Using Arabic dialects****: most people use modern dialects, not Modern Standard Arabic (MSA).*

II. LITERATURE REVIEW

Work on Arabic language sentiment is still in its infancy and a very rich area for development. Detecting the satisfaction(Satisfied, Neural, Not Satisfied) of a group of users, Applying it on such a trendy topic which is working from home during covid-19 pandemic not covered in the Arabic language even in the English language, it is not covered in such a good way which didn't make use of such trendy ways such machine learning or deep learning, it was done by some libraries such "TextBlob" [1] so that it didn't a achieve the intended Accuracy. Consideration of the text size (opinion) and its impact on the model accuracy is not covered in Arabic or English, besides that their main goal whether in Arabic or English was to detect user sentiment only. Detecting the problems that they during working from home is not addressed whether in Arabic or English.

Victor Ivan Lopez Rodriguez et.al. [1] conducted a study to answer a serious question which is "How do people feel about working from home?" during the coronavirus pandemic and used Twitter data as a dataset and the way used to get this twitter data first created a Twitter developer account from which got the credentials to access Twitter data then used tweepy library to extract data, finally worked on a maximum amount of 2000 tweets and retweets were not included in this research work

To work with data and clean it, it is important to apply Natural Language Processing (NLP) techniques to have clean data some common NLP techniques: Tokenization, Stop Words Removal, Stemming, and Lemmatization, then he becomes ready to work on data which become ready to detect the sentiment of that data going through multiple libraries like "TextBlob" then used another library "seaborn and matplotlib" for plotting the processed information but he didn't achieve good accuracy in addition to not considering the size of opinions.

Akash D Dubey et. Al. (2020) conducted a study in [5] to measure the sentiments of Work-from-Home among the employee using Twitter data as their dataset and they used Twitter API for the collection of the tweets, they collected 100,000 tweets worldwide.
Their search terms were #WorkFromHome or #WFH like what was done in [5] RTweet package was used in R. To avoid the duplication of the tweets, retweets were excluded.

After the data collection process, they went through a data preprocessing phase which contain tokenization, and stop words removal after that they used NRC Emotion Lexicon (Mohammad SM, Turney PD, 2013) was used for the sentiment analysis to analyze the Tweets From the analysis of the sentiments regarding Work-from-Home unlike what was done in [1] who used "TextBlob " library to complete the task of sentiment analysis. According to Akash D Dubey et. Al.(2020), it was found that 73.10 % of the tweets had positive sentiments as compared to 26.10% negative sentiments. This signifies that the people had a positive outlook towards the concept of work-from-home but unfortunately, they didn't consider the size of opinion.

Abdulrahman Radaideh et.al. [6] conducted a study to focus on the usage of specific apps in UAE such as zoom, Totok, and botim for conducting meetings over the internet. Twitter data was used as a dataset and collected almost 833 tweets using Twitter API and converted the image to text after data collection went through pre-processing phase after that they built 2 models to detect the real-time sentiment analysis. Naïve Base Classifier (NBC) and Recurrent Neural Network (RNN) were used. The results showed that NB is more accurate with (84%) than (RNN) with (79%). research work is the data set they worked on is too small and they didn't consider the size of opinion.

Mohammed M.Abdelgwad et.al. [7] conducted a study to focus on ducting aspect-based sentiment analysis in the Arabic language using benchmarked Arabic hotel reviews dataset in addition to using deep learning models for detecting aspect-based sentiment analysis instead of machine learning the results showed that deep learning model is such a good way instead of machine learning with accuracy (83.89%) superior machine learning models by (38.5%) but they not considered the effect of text size (opinion) on the model in addition to the accuracy can be improved.

According to Mohammad Al-Smadi et.al. [8], the focus of this research paper is an aspect-based sentiment analysis of Arabic Hotels' reviews using bidirectional LSTM (Bi-LSTM)

and conditional random field (CRF). The overall achieved accuracy to identify aspect sentiment polarity was 82.7%. but they not considered the effect of text size (opinion) o the model in addition to the accuracy that can be improved.

According to John Pierre D. Delizo1, et. al. [9] This study focused on Philippine Twitter Sentiments during Covid-19 Pandemic using Multinomial Naïve-Bayes. This paper examines the polarity of COVID -19 related opinions on Twitter from January to March 2020 by applying natural language processing. A total of 29,514 tweets were collected and after the data collection phase and data processing phase they built a Naïve Base Classifier (NBC) which 10% of the dataset was used to train the model and the result labeled that 52% of tweets are positive and 48% are negative like what was done in [1].

Fig. 1 shows the proposed methodology adopted in this study.

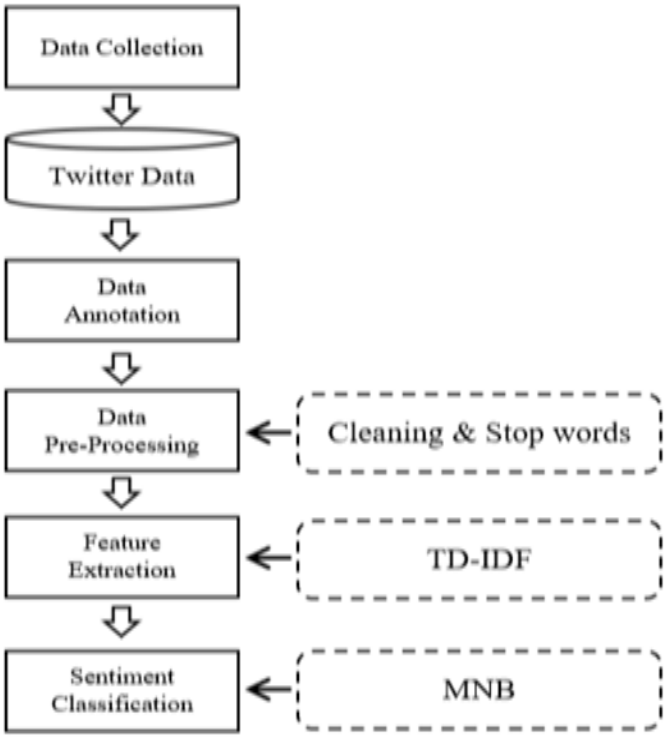

Fig. 1. Methodology

After we have stated the literature review, it is time to mention limitations and shortly comes that were found for example in Home Office due to COVID-19 -Twitter Data Analysis [1]. This research paper didn't achieve good Accuracy in addition to not considering the size of the text (opinion) and its effect on the model**.** According to Worker Perceptions on Working from Home during Covid 19 Pandemic - Empirical Evidence from Bahrain [4] which was based on distributed questionnaires the volume of data they worked on is very small and didn't achieve good accuracy in addition they didn't detect problems that users face during this issue. Arabic aspect-based sentiment analysis using bidirectional GRU-based models [7] didn't achieve such a good Accuracy (83,8%) It needs to improve in addition to not considering the size of text (opinion) and its effect on the model. Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews[8] didn't achieve such a good Accuracy (82,7%) It needs to improve as well as not considering the size of text (opinion) and its effect on the model in addition to not considering problems that users face during working from home. Philippine Twitter Sentiments during Covid-19 Pandemic using Multinomial Naïve-Bayes [9] In this paper it wasn't considered to detect the problems that users face. The accuracy and precision need some more improvement.

III. Sentiment Classification Process

In this section, we will focus on the way that most researchers adapted through working with sentiment classification especially Arabic language sentiment analysis.

A. Data Source

Maintaining a large dataset of various opinions in Arabic (Twitter Data). Social Networks make it possible for people to share personal information, emotions, and current state [10-12]. The number of people using social networks is increasing rapidly. Social media platforms are used for faster communication. Most researchers used the tweepy library to access Twitter data [1].

B. Natural Language Processing (NLP) techniques

Using Natural Language Processing (NLP) techniques to clean the data set noise are some common NLP techniques[1],[13],[14],[15]:

1) Removing duplicate tweets from the dataset.
2) Removing weak relevant information.
3) Removing redundant letters.
Such as "حرر////////م ", "//////////جد" and "خطيييير"
4) Removing of Punctuation: full stops, commas, and punctuation marks.
5) Text Tokenization: dividing a sentence into multiple words.
6) Removing Stop words: stop word is the term that doesn't have any influence on sentiment such as (في,من).
7) Letter replacement: some letters have different forms. Such as أ, إ

Fig. 2 shows an example of preprocessing steps that are related to sentiment analysis while working with the Arabic language which has multiple challenges while working with.

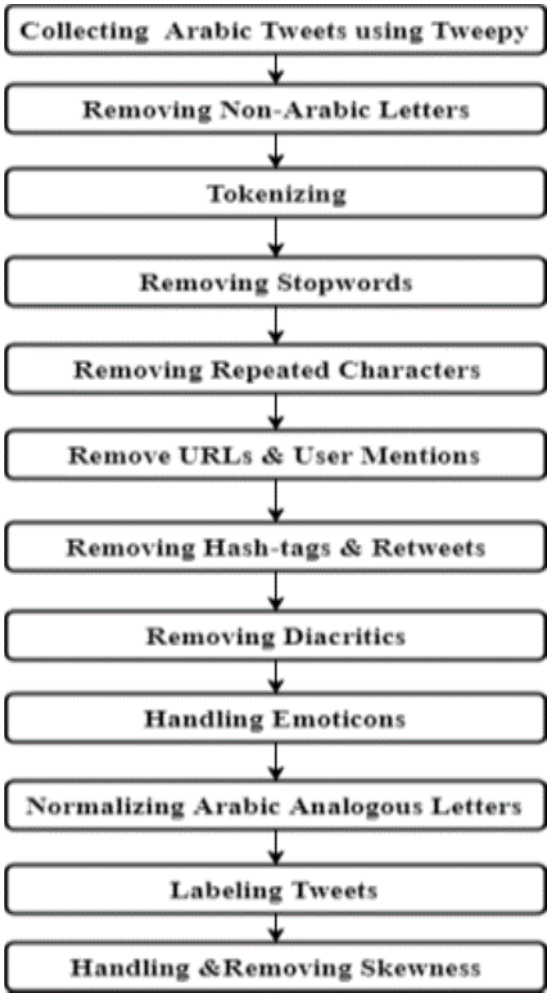

Fig. 2. Preprocessing

C. Feature extraction

Feature extraction is a critical task in sentiment analysis. Selecting the right features correctly will help in gaining high model accuracy and increasing the overall performance of sentiment detection. There are many used features for sentiment analysis such as n-grams, POS tags, stylistic features, and word vectors, syntactic features, lexicon features, semantic features [16].

D. Model Construction

Building a model and training this model dataset to detect sentiment. As mentioned above machine learning has had the lion's share of use in most research on sentiment analysis such as Support vector Machine (SVM), Naïve Bayes, and Decision Tree [3].

For example:

1) ***Naïve Bayes classifier*** is frequently used for sentiment classification; it is a probabilistic model based on the theorem of Bayesian. Naïve Bayes classifier particularly produces a high accuracy and performance, and it is used when the number of the dimensions of the inputs is high. It assumes that the features document is independent of one another [4].

2) ***Decision tree*** is one of the most popular techniques used for classification. The decision tree is composed of branches connected with nodes. the final decision is the leaf node. Inner nodes split the classes into subsets of classes according to feature checks that were selected. The most important features are placed at the top of the tree and add the next features recursively checks in the following node levels until we reach the tree's leaves, which label the target text to a category [4].

E. WordCloud Method

Detecting problems that users face for example detecting problems that employees face during working from home during covid-19 that can be done by detecting some words that commonly occur like quarantine, safe, time, good, home, need, space, best, good, and corona can specify an important reason for the tweet sentiment These captured words can notify companies about what can affect employees opinions to detect and solve this problem.

These words can let companies know what influence workers' opinions and work on a strategy to resolve this problem like supporting an employee with for example laptops and headphones etc.[1]. Fig. 3 shows an example of a wordcloud that is associated with working from home during a coronavirus pandemic.

Fig. 3. WordCloud of the tweets related to work from home

Briefly, the main steps are :

1- Data set gathering

2- Tokenization

3- *Pre-processing*
4- *Feature engineering*
5- *Model buildering*
6- *Model evaluation*
7- *Detecting the effect of text size (opinion) on the model*

Fig. 4 shows an overview of the proposed sentiment analysis model.

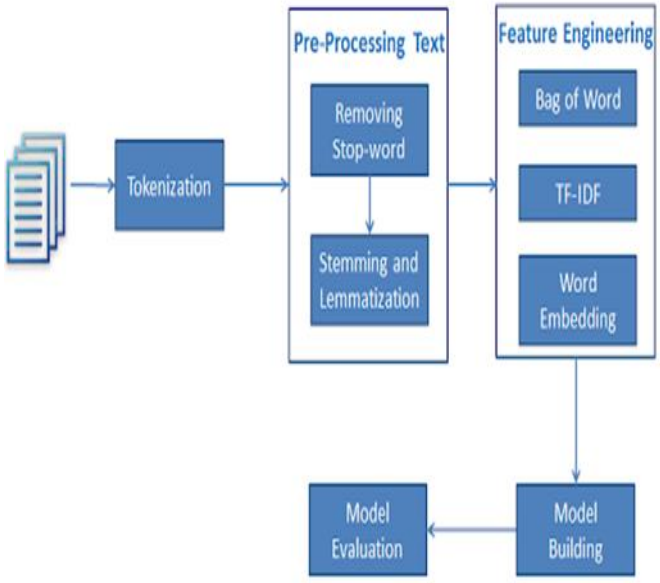

Fig.4. Overview of the proposed sentiment analysis approach

IV. CONCLUSION

Working on Arabic language sentiment is still in its infancy and a very rich area for development and this research will contribute to the sentiment analysis of Arabic text [16-20]. Detecting users' satisfaction in the Arabic language using aspect-based sentiment analysis as well as considering of size of text and its effect on the model accuracy these points have some limitations.

Future work is to construct a proposed model to detect user satisfaction about working from home using Aspect-based Sentiment analysis and detecting problems that they face through creating a large data set from Twitter data using tweepy library using

#المنزل + كورونا#,#المنزل + كوفيد,العمل_في _المنزل#,العمل_من_ المنزل#, العمل # ,العمل _عن_بعد + كوفيد # ,العمل _عن _ بعد#,#العمل + الفيروس _من _ المنزل + كورونا

with adapting Natural language processing (NLP) techniques in addition to measuring the effect of the text size (Opinion) to model accuracy and problems that employee face during working from home.

V. REFERENCES

[1] L. Rodríguez, V. Iván, "Home Office due to COVID-19 - Twitter Data Analysis," 2020.

[2] A. Al-Obaidi and V. Samawi, "Opinion Mining: Analysis of Comments Written in Arabic Colloquial," 2016.

[3] D. Najar and S. Mesfar, "Opinion mining and sentiment analysis for Arabic on-line texts: application on the political domain," Int J Speech Technol, vol. 20, no. 1, pp. 575–585, 2017.

[4] B. Almarzooqi and F.A. Alaamer, "Worker perceptions on working from home during covid-19 pandemic empirical evidence from bahrain," 2020. https://ssrn.com/abstract=3643890.

[5] A.D. Dubey and S. Tripathi, "Analysing the sentiments towards work-from-home experience during covid-19 pandemic," Journal of Innovation Management, vol. 8, no. 1, pp.13-19, 2020.

[6] A. Radaideh, F. Dweiri and M. Obaidat, "A novel approach to predict the real time sentimental analysis by naive bayes & rnn algorithm during the covid pandemic in uae," in *International Conference on Communications, Computing, Cybersecurity and Informatics (CCCI)*, pp. 1-5, 2020.

[7] M. Mustafa, T. Soliman, A. Taloba and M. Farghaly, "Arabic aspect based sentiment analysis using bidirectional GRU based models," 2021.

[8] M. AL-Smadi,, B. Talafha, M. Al-Ayyoub and Y. Jararweh, "Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews," *International Journal of Machine Learning and Cybernetics*, 2019.

[9] Mideth. Abisado, Ma. De Los Trinos and John. Delizo, "Philippine Twitter Sentiments during Covid-19 Pandemic using Multinomial Naïve-Bayes", vol. 9, no. 1.3, pp. 408-412, 2020.

[10] A. Ghallab, A. Mohsen and Y. Ali, "Arabic Sentiment Analysis: A Systematic Literature Review," *Applied Computational Intelligence and Soft Computing*, vol. 21, 2020.

[11] M. Hammad and M. Al-awadi," Sentiment Analysis for Arabic Reviews in Social Networks Using Machine Learning," *Advances in Intelligent Systems and Computing*, 2016. DOI: 10.1007/978-3-319-32467-8_13

[12] F. Hemmatian, M. Karim," A survey on classification techniques for opinion mining and sentiment analysis," vol. 52, 2019.

[13] O. Oueslati, E. Cambria, M. Ben HajHmida, H. Ounelli, "A review of sentiment analysis research in Arabic language," *Future Generation Computer Systems*, vol. 112, pp. 408–430, 2020.

[14] G. Alwakid, T. Osman and T. Hughes-Roberts, "Challenges in Sentiment Analysis for Arabic Social Networks," *Procedia Computer Science*, vol. 117, pp. 89-100, 2017, ISSN 1877-0509

[15] S. Alotaibi and M. B. Khan, "Sentiment Analysis Challenges of Informal Arabic Language," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 2, pp. 278–284, 2017.Available from: https://www.researchgate.net/publication/346052809_Comparative_

[16] S. Jasti and T.S. Mahalakshmi, "A Review on Sentiment Analysis of Opinion Mining," vol 768, 2019. Springer, Singapore. https://doi.org/10.1007/978-981-13-0617-4_58

[17] M. Sadegh, I. Roliana and Z. Ali Othman. "Opinion mining and sentiment analysis: A survey," *International Journal of Computers & Technology*, no. 2.3, pp. 171-178, 2012.

[18] M. Al-Kabi, G. Amal, A. Izzat, W. Heider and M. Haidar, "Opinion Mining and Analysis for Arabic Language," *International Journal of Advanced Computer Science and Applications*. vol. 5, pp. 181-195, 2014.

[19] D. Osimo, and F. Mureddu, "Research Challenge on Opinion Mining and Sentiment Analysis," 2010. Available from: https://www.researchgate.net/publication/343685163_SOCIAL_MEDIA_ANALYSIS_WITH_AI_SENTIMENT_ANALYSIS_TECHNIQUES_FOR_THE_ANALYSIS_OF_TWITTER_COVID-19_DATA

[20] A. Alsayat and N. Elmitwally, "A comprehensive study for arabic sentiment analysis (challenges and applications)," Egyptian Informatics Journal, vol. 21, no. 1, pp.7-12, 2020.