# Classifying Suspicious Content on Social Media Networks

Noor Alwan Ghanem
*Computer Science Department.*
*College of Science for Women*
*University of Babylon*
Babylon, Iraq
noor.ghanim@student.uobabylon.edu.iq

Haider M. Habeeb
*Information Networks Department.*
*College of Information Technology*
*University of Babylon*
Babylon, Iraq
haiderhabeeb@uobabylon.edu.iq

2021 International Conference on Advanced Computer Applications (ACA) | 978-1-6654-3503-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ACA52198.2021.9626788

***Abstract*— With the emergence and expansion of web technologies, the web now contains a huge amount of data and information for Internet users, in addition to a large amount of new data being generated at every moment. The Internet, with its various platforms, has evolved to be a forum for learning through these platforms, exchanging information, and exchanging opinions and ideas. For example, Twitter and other social networks are rapidly gaining popularity because they encourage users to communicate and convey their opinions on a variety of topics, engage in discussions with different cultures, and send messages around the world. In the areas of sentiment about Twitter information, a lot of studies have been done. This study focuses on the use of sentiment analysis technology on tweets generated by Twitter which is useful for analyzing details in tweets where opinions are highly disorganized, heterogeneous, and either negative, positive, or neutral in some cases. In this paper, the mentioned technique (sentiment analysis) was used to elicit opinions on two suspicious or non-suspicious measures as well as categorize these tweets using machine learning and a lexicon-based approach, along with rating scales. Using different machine learning algorithms such as Naive Bayes and Random Forest Classifier (RFC), we achieved a classification accuracy of 88.07 and 92.61 for NB and RFC. The accuracy of NB and RFC is 88.07 and 92.61, respectively, which is very good when compared to the recent research mentioned below.**

***Index Terms*— Twitter, Naive Bayes (NB), Random Forest (RFC), social media, Sentiment analysis, text mining, Machine Learning**

## I. Introduction

Through the widespread usage of social networking sites such as Facebook and Twitter, the online population is sharing knowledge in the form of views, feelings, emotions, and motivations, all of which represent their aptitude and affiliations for a particular person, case, or policy [1].

Owing to their vast scope and the pace at which knowledge can be spread, popular social networking sites such as Twitter and Facebook have proved to be powerful outlets for disseminating false information, unverified statements, and fake attention-grabbing posts. There has been a spike in the number of alarming cases of false news spreading across social media and having a serious effect on real-world affairs recently [2].

1. Machine learning has been increasingly expanded to include the study of materials and emotions with the advancement of technology. Ferrara et al. [3] Machine learning methods were used to identify user interactions on social networking texts. The proposed scheme was tested on a group of more than 20,000 tweets sent by suspicious accounts that were later suspended by Twitter. The focus was primarily on three tasks: (1) detecting extremist users, (2) identifying users who have been exposed to suspicious posts, and (3) predicting users' reactions to suspicious posts. Aziz [4] proposes a machine learning-based approach to classify suspicious affiliations for the same reason. The classic feature set of the Naive Bayes algorithm is used. which focuses on classifying user reviews into positive and negative categories, as compared to Ferrara et al. [3] Work focusing on terrorist association classification on biased data, their approach uses the NB algorithm on balanced data, yielding more accurate results.

Recently, harassment increased through social media. False news distributed on social networking sites vary according to the intention behind the lie. Suspicious news tends to build a plot rather than conveying facts as in verified news. Also one extreme disinformation is which communications spread false facts to intentionally deceive readers or promote a biased agenda [5].

2. Twitter is one of the most popular online social media and microblogging services that enable its users to send and read text posts of up to 140 characters per message, known as "tweets" [6]. These posts include tweets generated and retweeted from propaganda and so-called ("eye-catching" headlines) of human interest information.

3. Text mining (TM) is the process of extracting interesting and non-trivial patterns or knowledge from text documents. It is also known as text data mining or knowledge discovery from textual databases (KDD). Text mining is an interdisciplinary field that seeks for extracting significant information from unstructured data. It is based on data mining (DM), machine learning, information retrieval, statistics, and computational linguistics [7] [8].

4. Sentiment analysis is a broad and growing field that is being treated as an activity in NLP. Several algorithms and technological databases have been introduced to process the problem statement, ranging from text-level classification to word-level classification. Opinion mining can be done in two solutions to these problems: using a lexicon-based approach that uses pre-built sentiment dictionaries, and using a machine

learning-based approach that trains the machine to use available data. Retrieval of data from Twitter, language recognition, and interpretation of sentiment on the document using machine learning algorithms make up the entire technology. Ibrahim's article [9] proposes a way to predict election outcomes with Twitter data collected throughout the campaign season, and then perform a sentiment analysis by assigning the polarity of each tweet and measuring the sentiment of all tweets based on this polarity. In contrast to using autonomous machine learning methods, Rincy Jose [10] suggested an ensemble method with voting to increase precision. The obtained data is analyzed utilizing a lexicon-based technique.

Bak and Paroubeek [11] proposed a model for classifying tweets into objective, optimistic, and negative categories. They created a Twitter group by manually grouping tweets and commenting on them with tokens using the Twitter API. They used the combination to create a sentiment classifier based on the Naive Bayes polynomial technology, as well as features including Stamps and Ingram. Since the training dataset they used only included tweets about tokens, it didn't work. Parikh and Movassate [12] used two methods to distinguish between tweets: the Bigram Naive Bayes model and the Maximum Entropy model. They found that Naive Bayes classifiers outperformed the Maximum Entropy model. With their training data consisting of tweets with codes that act as noisy labels.

Tripathi et al. [13] Supervised learning algorithms including Naive Bayes, SVM, random forest, and linear discriminant analysis were used to describe the reactions based on their polarity. The proposed method contained four measures to do so. The first step was to delete stop terms, binary codes, and special characters from the data. The text labels are then translated into a numeric array. Finally, the produced vectors were fed into four separate classifiers as inputs. Then, the results were obtained by classifying two data sets. The efficacy of the proposed solution was then evaluated using various parameters including classification accuracy, f-measurement, recall, and accuracy. The random forest classifier outperformed other classifiers such as Naive Bayes In addition, general issues, and Twitter sentiment implementations were addressed.

Ullah, Mohammad Aman, et al [14] Learning techniques only involve categorizing text, symbols, or images only as symbols with text are always neglected, thus many emotions are ignored. This research suggested an algorithm and method for analyzing feelings, using both text and emoji. In this work, this data was analyzed with both Naive Bayes and Random Forest machine learning algorithms using Twitter-based airline data where several features such as TF - IDF, Bag of Words, N-gram, and emoticon lexicons were used. This research demonstrates that whenever emoji's are used, the associated emotions dominate the emotions conveyed by the analysis of textual data. Also, the accuracy of the machine learning algorithms was obtained from the NaivyBase algorithm 52% and random forest algorithm 76%. [22] [23]

Veny Amilia et al [15], The data used in this study are Twitter comments tweets from Indonesians captured using Python and archived in CSV files. Comments search keywords using hashtags and query searches are associated with anti-LGBT case studies. 3,744 comments / tweets collected. 75% of the data will be used as training data and 25% of the data will be used as test data, the algorithm used to perform sentiment analysis is Naïve Bayes because it is highly accurate in classifying sentiment analysis. The stages of conducting sentiment analysis in this study are pre-processing data, data processing, classification, and evaluation. Sentiment analysis using Naïve Bayes, Decision Tree and random forest algorithm where 86.43% accuracy was obtained from testing data using Naïve Bayes, where the accuracy is higher than other algorithms such as decision tree and random forest which is 82.91%.

## II. SENTIMENT ANALYSIS (SA)

also known as Opinion Analysis/Mining (OM), is a computational method for studying the opinions and feelings of people towards a particular subject to discern opinions in text into positive, negative, and/or neutral through NLP. SA had become one of the most attractive research fields in computer science. The importance of SA has grown with the growth of social media such as Twitter, YouTube, and Facebook [16].

## III. RANDOM FOREST

Random forest is indeed a classification algorithm for classifying and forecasting data. It is, however, mostly used to resolve classification problems. A forest is composed of plants since we all know, and more trees imply a more stable forest. The algorithm of random forest, on the other hand, creates decision trees from sets of data, derives forecasts from each, and afterward votes on the optimum response. It's an aggregation solution, not a single decision tree because it averages the outcomes to reduce overfitting.

The following measures will help us grasp how the Random Forest algorithm works. −

- Stage 1 − Begin by selecting random specimens from a specified dataset.
- Stage 2 − Following that, this algorithm would build a decision tree for each sample. The forecast findings from each decision tree will then be obtained.
- Stage 3 − Voting will take place in this stage with each expected outcome.
- Stage 4 − Finally, choose the forecast outcome with the most votes as the actual forecasting finding.

The illustration that follows will demonstrate how it works.

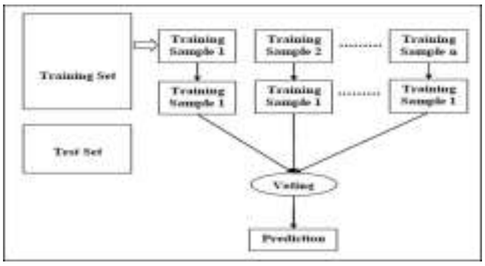


Fig. 1. The diagram for illustration Random Forest working

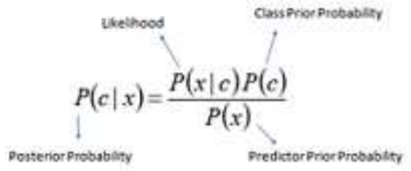
## IV. NAIVE BAYES

It's a categorization method focused on Bayes' Theorem and the principle of prediction freedom. A Naive Bayes classifier, in basic words, implies that the existence of one function in a class is irrelevant to the existence of some other feature.

For instance, if the fruit is red, circular, and around 3 inches in diameter, it is called an apple. Even if these characteristics are contingent on one another or the presence of other characteristics, they both add to the likelihood that this fruit is an apple, which is why it is called "Naive."

The Naive Bayes model is simple to construct and is particularly useful for broad data sets. Naive Bayes is considered to outperform even the most advanced categorization approaches due to its simplicity.

The Bayes theorem allows you to calculate posterior likelihood P(c |x) from P(c), P(x), and P(x| c) using P(c), P(x), and P(x |c). Have a look at the following formula:



$$P(c\mid X) = P(x_1\mid c)\times P(x_2\mid c)\times\cdots\times P(x_n\mid c)\times P(c) \quad (1)$$

Above,

- The possibility function of class (c, target) provided a prediction (x, attributes) is P(c| x).
- P(c) is the prior possibility of class.
- P(x| c) is the probability that is the likelihood of the predicted class.
- P(x) is the prior possibility of the predictor. [1]
-

## V. PROPOSED SYSTEM

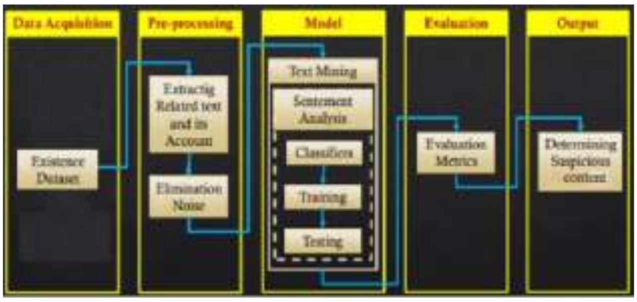


Fig.2. Diagram showing the proposed system

- *Data Acquisition*:

Due to the many problems and the difficulty of getting permission from Twitter API and the problems of the Internet, the dataset for our proposed system was obtained from the Kaggle website, which is a CSV file called (Sentiment140), it contains (1048377) tweets with the following fields:

1 - ids: The id of the tweet.
2 - date: the date of the tweet.
3 - user: the user that tweeted
4 - text: the text of the tweet.

- preprocessing

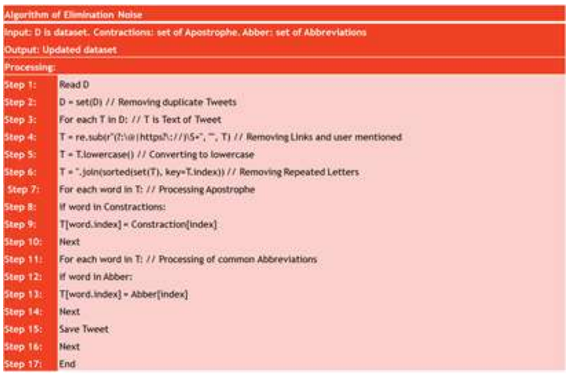
Data is prepared for review using preprocessing procedures. The most important part of data analysis is pre-processing. Since the data is obtained as a result of the experiment, the next stage is data modeling. to derive valuable knowledge in this project, we use pre-processing to exclude any data we don't need, such as tags and hashtags, to create transparent data that we can use to classify, namely texts. for sentiment analysis and classification. we take the dataset line by line, each line representing one tweet, then we check if this tweet has any text after the @ symbol without space and delete it, then we look for hashtag any text after the # symbol without space in this tweet and delete it, and we do the same for retweets and ties.

Remove stop words in English from this tweet as the next stage in preprocessing. Stop words are a list of widely encountered words of every language. Then we exclude all symbols, numbers, or non-alphabetic code. To get consistent details for sentiment analysis and grouping, we replicating these measures on all tweets in the dataset as follows:-

1- Elimination Noise: Data preprocessing prepares data for the methodology of the proposed system. Dataset of the proposed system will be set during the data preprocessing stage. It consists of the following processes:

- Removing the Duplicated Tweets
- Removing Links
- Letter Lowercase
- Removing Repeated Letters
- Processing Apostrophe
- Processing of Common Abbreviations

2- Removing Stopwords: Stop words are words that do not have any significance or useful information in the text such as prepositions, pronouns, and conjunctions [17].



```
Algorithm of Elimination Noise
Input: D is dataset. Contractions: set of Apostrophe. Abber: set of Abbreviations
Output: Updated dataset
Processing:
Step 1:   Read D
Step 2:   D = set(D) // Removing duplicate Tweets
Step 3:   For each T in D: // T is Text of Tweet
Step 4:   T = re.sub(r"(?:\@|https?\://)\S+", "", T) // Removing Links and user mentioned
Step 5:   T = T.lowercase() // Converting to lowercase
Step 6:   T = ''.join(sorted(set(T), key=T.index)) // Removing Repeated Letters
Step 7:   For each word in T: // Processing Apostrophe
Step 8:   If word in Constractions:
Step 9:   T[word.index] = Constraction[index]
Step 10:  Next
Step 11:  For each word in T: // Processing of common Abbreviations
Step 12:  If word in Abber:
Step 13:  T[word.index] = Abber[index]
Step 14:  Next
Step 15:  Save Tweet
Step 16:  Next
Step 17:  End
```

Fig 3. The algorithem for Pre-processing

### *A.SENTIMENT ANALYSIS*

The method of determining whether a text is positive or negative is known as sentiment analysis. Natural language processing (NLP) and machine learning methods are used within the framework of sentiment analysis to analyze text to assign weighted sentiment scores to people, subjects, patterns, and groups within a sentence or word. [3].

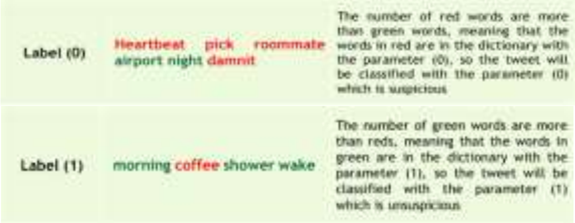
Sentiment analysis (SN) is used to classify tweets into suspicious and non-suspicious. 10,000 tweets with 5,000 suspicious tweets and 5,000 unsuspicious have manually labeled.

At first, from the 5000 suspicious tweets, all the important words would be extracted without repeat then it will be saved as a Dictionary. Applying the same method on the 5000 unsuspicious tweets then creates a new dictionary with the extracted unrepeated words.

These two dictionaries would be used for analyzing the main dataset to determine if each tweet is suspicious or not. Each word in the tweet from the main dataset would be compared with the words in both dictionaries (suspicious and unsuspicious). This operation would be repeated on each word in the selected tweet. The Comparison would be done by applying classification algorithms.

It has been taking each tweet in the main dataset and checking each word in it with the suspicious and not suspicious dictionaries. It has been replicating this procedure on each word in the message, and if the majority of words are found in the suspect dictionary, the tweet is suspicious. If the majority of words are found in the not suspicious dictionary, the tweet is not suspicious. Repeat this process with and tweet in the dataset to generate a new dataset that can be used to decide whether or not a tweet is suspect.

TABLE I : SHOWING HOW SENTIMENT ANALYSIS WORKS



| | | |
|---|---|---|
| Label (0) | Heartbeat pick roommate airport night damnit | The number of red words are more than green words, meaning that the words in red are in the dictionary with the parameter (0), so the tweet will be classified with the parameter (0) which is suspicious |
| Label (1) | morning coffee shower wake | The number of green words are more than reds, meaning that the words in green are in the dictionary with the parameter (1), so the tweet will be classified with the parameter (1) which is unsuspicious |

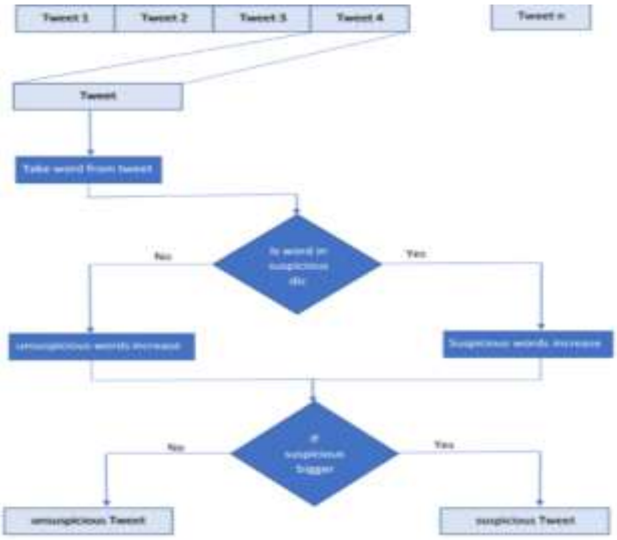


Fig 4. The diagram for illustration Sentiment analysis

### *B. CLASSIFICATION STAGE*:

Classification algorithms are applied. The classification stage involves the use of two machine learning algorithms to classify tweets as suspicious or unsuspicious based on the training set in the classification stage:

Suggested system algorithms:

1- Naive Bayes algorithm

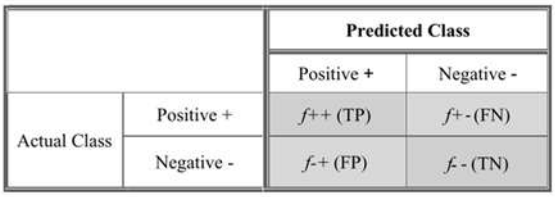
2 - Random Forest algorithm

### *C. EVALUATION:*

A confusion matrix was used to evaluate the model by calculating the accuracy, recall, accuracy, and f1 score.

confusion matrix which is a matrix that summarizes the number of examples properly or wrongly predicted by a classification model as shown in the following table II.

TABLE II : SHOWING THE CONFUCION MATRIX



| | | Predicted Class | |
|---|---|---|---|
| | | Positive + | Negative - |
| Actual Class | Positive + | *f*++ (TP) | *f*+-(FN) |
| | Negative - | *f*-+ (FP) | *f*- - (TN) |

f

The main values of this table are described below:

1 - True Positive (TP): denotes to the positive examples that are properly classified.

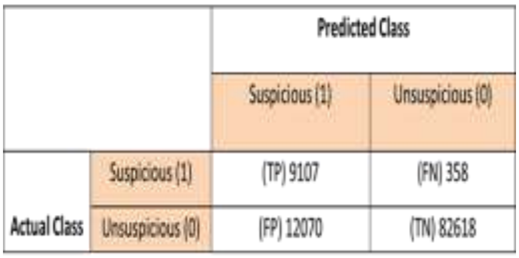
2 - False Negative (FN): denotes to the positive examples that are incorrectly classified.

3 - False Positive (FP): denotes to the negative examples that are incorrectly predicted and classified.

4 - True negative (TN): denotes to the negative examples that are properly predicted by the classification model.

Confusion Matrix has used for evaluating the model by computing the accuracy, recall, precision, and f1-score.

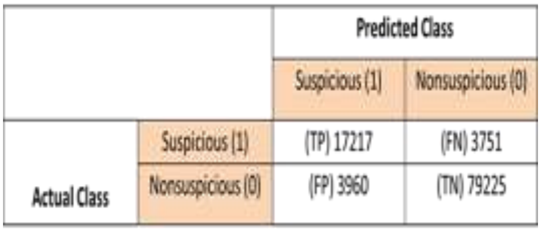
TABLE III : NAIVE BAYES ALGORITHM CLASSIFICATION RESULTS



| | | Predicted Class | |
|---|---|---|---|
| | | Suspicious (1) | Unsuspicious (0) |
| Actual Class | Suspicious (1) | (TP) 9107 | (FN) 358 |
| | Unsuspicious (0) | (FP) 12070 | (TN) 82618 |

- Accuracy score :- 88,06
- Precision score :- 87,25
- Recall score :- 99,56
- F1 score :-93,00

TABLE IV: RANDOM FOREST ALGORITHM CLASSIFICATION RESULTS



| | | Predicted Class | |
|---|---|---|---|
| | | Suspicious (1) | Nonsuspicious (0) |
| Actual Class | Suspicious (1) | (TP) 17217 | (FN) 3751 |
| | Nonsuspicious (0) | (FP) 3960 | (TN) 79225 |

- Accuracy score: 92,59
- Precision score: 95,23
- Recall score: 95,47
- F1 score: 95,350

## V. Result

In our work, we use naïve Bayes algorithm and random forest algorithm and get accuracy

- Naïve Bayes Accuracy Score = 88.07%
- Random Forest Accuracy Score = 92.61%.

TABLE V: REPRESENTING THE RESULTS OF COMPARATIVE RESEARCH

| No | Study | Technique | Accuracy |
|---|---|---|---|
| 1- | [13] | Machine learning with sentiment analysis | obtained from Naive Base algorithm 52% and random forest algorithm 76%. |
| 2- | [14] | sentiment analysis on twitter | Sentiment analysis using Naïve Bayes, Decision Tree and random forest algorithm where 86.43% accuracy was obtained from testing data using Naïve Bayes, where the accuracy is higher than other algorithms such as decision tree and random forest which is 82.91%. |

## VI. Conclusion

It is very important to categorize suspicious content on social media as pre-processing techniques have been implemented as an important step in the proposed system because this can lead to cleaning all tweets of unnecessary information, then use analytical techniques such as sentiment analysis techniques and implement algorithms such as Random Forest Classifier and investigate. Best results with 92.61% accuracy.

Several trends can be highlighted based on the findings of this current thesis. These include, but are not limited to Analysis of Tweets written in other languages. Can show work like this

The spread of suspicious and unnatural thoughts is based on the opinions of Twitter users in different countries and cultures. This classification can avoid users on social networks from fake accounts as well as help their platforms to clear such suspicious accounts as well: -

- Using sentiment analysis technique with machine learning approaches lead to accurate results.
- Random Forest Classifier achieves better results than Naïve Bayes Classifier when using Sentiment Analysis with it.

## References

[1] F. Hao, D.-S. Park, and Z. Pei, 'When social computing meets soft computing: opportunities and insights', Human-centric Comput Inf Sci, vol. 8, no. 1, pp. 1–18, 2018.

[2] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, 'Deep learning for hate speech detection in tweets', in Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.

[3] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, 'Predicting online extremism, content adopters, and interaction reciprocity', in International conference on social informatics, 2016, pp. 22–39.

[4] F. Hao, D.-S. Park, and Z. Pei, 'When social computing meets soft computing: opportunities and insights', Human-centric Comput Inf Sci, vol. 8, no. 1, pp. 1–18, 2018.

[5] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, 'Deep learning for hate speech detection in tweets', in Proceedings of the 26th international conference on World Wide Web companion, 2017, pp. 759–760.

[6] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, 'Predicting online extremism, content adopters, and interaction reciprocity', in International conference on social informatics, 2016, pp. 22–39.

[7] S. A. Azizan and I. A. Aziz, 'Terrorism detection based on sentiment analysis using machine learning , J Eng Appl Sci, vol. 12, no. 3, pp. 691–698, 2017.

[8] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas, "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter," ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 2, no. January 2017, pp. 647–653, 2017, doi: 10.18653/v1/P17-2102.

[9] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas, "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter," ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 2, no. January 2017, pp. 647–653, 2017, doi: 10.18653/v1/P17-2102.

[10] J. Martinez-Romo and L. Araujo, "Detecting malicious tweets in trending topics using a statistical analysis of language," Expert Syst. Appl., vol. 40, no. 8, pp. 2992–3000, 2013, DOI: 10.1016/j.eswa.2012.12.015.

[11] A.-H. Tan, "Text Mining: The state of the art and the challenges," Proc. PAKDD 1999 Work. Knowl. Discovery from Adv. Databases, vol. 8, pp. 65–70, 1999, DOI: 10.1.1.38.7672.

[12] D. W. FREEMAN and W. A. SISTRUNK, "Effects of Post-Harvest Storage on the Quality of Canned Snap Beans," J. Food Sci., vol. 43, no. 1, pp. 211–214, 1978, DOI: 10.1111/j.1365-2621.1978.tb09773.x.

[13] M. Ibrahim, O. Abdillah, A. F. Wicaksono, and M. Adriani, 'Buzzer detection and sentiment analysis for predicting presidential election results in a Twitter nation', in 2015 IEEE international conference on data mining workshop (ICDMW), 2015, pp. 1348–1353.

[14] R. Jose and V. S. Chooralil, 'Prediction of the election result by enhanced sentiment analysis on Twitter data using classifier ensemble Approach', in 2016 international conference on data mining and advanced computing (SAPIENCE), 2016, pp. 64–67.

[15] A. Pak and P. Paroubeek, 'Twitter as a corpus for sentiment analysis and opinion mining.', in LREC, 2010, vol. 10, no. 2010, pp. 1320–1326.

[16] R. Parikh and M. Movassate, 'Sentiment analysis of user-generated twitter updates using various classification techniques', CS224N Final Rep, vol. 118, 2009.

[17] A. Tripathy and S. K. Rath, 'Classification of the sentiment of reviews using supervised machine learning techniques, Int J Rough Sets Data Anal, vol. 4, no. 1, pp. 56–74, 2017.

[18] M.A. Ullah, et al. "An algorithm and method for sentiment analysis using the text and emoticon." ICT Express 6.4 (2020): 357-360.

[19] V. A. Fitri, R. Andreswari, and M. Azani Hasibuan. "Sentiment analysis of social media twitter with case of anti-lgbt campaign in indonesia using naïve bayes, decision tree, and random forest algorithm." Procedia Computer Science 161 (2019): 765-772.

[20] K. Ravi and V. Ravi, A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, vol. 89, no. June. Elsevier B.V., 2015.

[21] R.M. Sallam, H. M. Mousa, and M. Hussein, . "Improving Arabic text categorization using normalization and stemming techniques.", Int. J. Comput. Appl. 135. 2 (2016): 38- 43

[22] SB Sadkhan, "A Developed DS-CDMA Detection based on ICA", 2021 International Conference on Communication & Information Technology (ICICT).

[23] A Borany, SB Sadkhan, "Decision-making approach in Cognitive Radio using Tsukamoto and Mamdani FIS", 2021 1st Babylon International Conference on Information Technology and Science (BICITS).