TWITTER HASHTAG PREDICTION AND ANALYSIS WITH TWITTER DATASET WITH MACHINE LEARNING

*Note: Sub-titles are not captured in Xplore and should not be used

1st Jethin Sai Chilukuri
dept. Computer Science
University of Central Missouri
700746555
jxc65550@ucmo.edu

2nd Sai Leela Otikundala
dept. Computer Science
University of Central Missouri
700747149
sxo71490@ucmo.edu

3rd Navya Sri Bonthu
dept. Computer Science)
University of Central Missouri
700747945
nxb79450@ucmo.edu

4th Joshi Paul Bachala
dept. Computer Science
University of Central Missouri
700745481
jxb54810@ucmo.edu

***Abstract*—Twitter has become a crucial platform for sharing information and opinions, making it an ideal source for social media analysis. The objective of this project is to develop a Twitter hashtag prediction and analysis application to detect tweets in real-world applications, specifically fake news. The project aims to pre-process the dataset, remove noisy and null value data, analyse, and visualize the data for further processing. Predictions will be made using all machine learning methods, specifically the Random Forest algorithm, logistic regression, and linear support vector algorithms etc. The dataset will be divided into two parts, with 70 percent taken to provide training to the machine learning algorithm and the remaining 30 percent taken to the testing part. The Twitter hashtag prediction application will be developed using Python programming language. This application will be helpful in all the web applications where the tweets are spread, and applying the concepts to different types of sources can help to avoid the spreading of bad tweets in real-world applications. The machine learning algorithm will be trained on the pre-processed dataset and will predict the sentiment of the tweets as positive, negative, or neutral. The accuracy of the algorithm will be measured to evaluate the results. The accuracy score of the algorithm in fake news detection will help to evaluate the dataset. The project's significance is to provide a tool for analysing social media data and identifying potentially harmful tweets, specifically fake news. The proposed application will contribute to detecting fake news and reduce its spread on social media. Additionally, the Twitter hashtag prediction and analysis application can be extended to other domains such as product reviews, political opinions, and customer feedback. In conclusion, the proposed Twitter hashtag prediction and analysis application aims to detect tweets in real-world applications, specifically fake news, by pre-processing the dataset, training the machine learning algorithm, and evaluating the accuracy of the results. This project's significance is to provide a tool for analysing social media data and identifying potentially harmful tweets and can be extended to other domains as well.**

***Index Terms*—Hashtag,Prediction,Sentiment Analysis,Tokenization,stemming,vectorization.**

https://github.com/JethinSai/ML$_{Project}$
$https://github.com/NavyaBonthu/Final-Project$
$https://github.com/joshipaul13/Final-paper$
$https://github.com/SaiLeelaOtikundala/final-project$

I. Introduction

In recent years, social media has become a powerful tool for sharing information and shaping public opinion on a variety of topics. The use of social media platforms such as Twitter has become a popular way for people to express their opinions on various topics, The Users often use hashtags to categorize their tweets and make them more discover able to others including environmental change. In this data-set, we have collected around 43000 tweets related to environmental change, which have been reviewed by three commentators, and only tweets that received a consensus among the three commentators have been included. The data-set contains tweets on a wide range of topics related to environmental change, including weather conditions, climate change, and the impact of environmental change on human activities. The analysis of Twitter hashtags related to environmental change reveals that weather conditions are a prevalent topic, with users experiencing this around four to five times every day. The data-set is significant as it provides a valuable resource for studying people's opinions and attitudes towards climate change. The tweets cover a wide range of environmental topics, including the effects of weather conditions on human activities, environmental protection, agriculture, and the impact of climate change on economic and social factors. Each tweet has been labelled as one of the following categories: Anti, Neutral, Pro, or News. The Anti class represents tweets that do not believe in man-made climate change, while the Neutral class indicates tweets that neither support nor reject this idea. The Pro class signifies tweets that support the belief of man-made climate change, and the News class is associated with tweets that report actual news about climate change. Environmental protection is also a significant topic that has changed people's lives and activities for a significant period. The evaluation of environmental information is a continuous decision issue that requires careful consideration. Precipitation is an essential part of environmental change, and its accurate prediction is challenging due to

the complex physical structures involved. The use of climate data also has a significant impact on the profitability of human activities, with implications for economic and social outcomes. Overall, this data-set provides valuable insights into how people perceive and react to environmental change on social media platforms such as Twitter. The analysis of this data-set can help us understand the public's opinions and attitudes towards environmental change and inform policy decisions and communication strategies. The Twitter data-set on environmental change contains tweets that have been labelled as one of the following categories: Anti, Neutral, Pro, or News. Each tweet is associated with a unique tweet ID and contains two columns: the message section shows the content of the tweet, and the opinion section shows the sentiment label not entirely settled by three commentators.

II. Motivation

The emergence of Twitter as a platform for sharing information and engaging in public conversations has given rise to a plethora of Twitter accounts across the world. The platform has been utilized in various projects for classification tasks, including weather prediction. The primary goal of this project is to extract tweets and analyse the sentiment of positive, negative, and neutral tweets to predict weather conditions. To achieve this goal, pre-processing techniques such as stop word removal and stemming are applied to the data-set containing weather information. Machine learning algorithms are then applied, and the accuracy levels are compared to predict better results. The K-nearest neighbour algorithm has been used to identify potential employees based on their characters and the Myers-Briggs type marker classes, as well as to classify weather data. Deep learning techniques have also been employed to enhance performance in various fields, including image and text data analysis. The bidirectional encoder representations from transformers algorithm, although challenging to use due to its time and cost requirements, has been employed to classify large datasets due to its low training costs. The support vector machine technique has also been used to develop an AI classification method to analyse sentiments in texts. The significance of this project lies in its ability to extract tweets and analyse positive, negative, and neutral sentiments to predict weather conditions accurately. The use of social media platforms such as Twitter for weather prediction can provide valuable insights and improve the accuracy of weather forecasts. The methodology employed in this project can also be used in various other fields such as employee identification and personality assessment.

III. Objectives

- The project focuses on data analysis using a tweet data-set containing weather data.
- Pre-processing techniques, such as stop word removal and stemming, are applied to the data to improve accuracy..
- To apply pre-processing techniques to improve the accuracy of the machine learning algorithms.
- To explore the use of deep learning for various data types, including text.
- Different machine learning algorithms are applied to the data-set to predict weather results more accurately.
- To develop an AI classification method for sentiment analysis in texts using an SVM approach.

IV. Related Work

The author describes a research project that aims to classify tweets into four categories: politics, sports, crime, and natural. The researchers constructed the categories and implemented a preprocessing model on the raw Twitter dataset before using various machine learning techniques, including Random Forest, K-Nearest Neighbors, Naive Bayes, Logistic Regression, Decision Tree, and Support Vector Machine, to classify the Twitter data. The researchers examined the outcomes with and without preprocessing in terms of sensitivity, specificity, and accuracy and found that their proposed preprocessing model enhanced the performance of all the machine learning classifiers. The study suggests that preprocessing can improve the performance of machine learning algorithms for tweet classification into categories.[1] The authors use machine learning methods, specifically the Logistic Regression algorithm, to classify the tweets. They train the classifier using 1800 labelled tweets for each topic and apply several processes in the pre-processing phase, such as removing URLs, punctuation, and stop words, tokenization, and stemming. The pre-processed tweets are then converted into a set of features vectors using the Bag of Words technique. The set of features vectors is then applied to the Logistic Regression algorithm for the classification task. The trained classifier is evaluated using 1800 tweets with 450 for each topic, and the results are analysed using Confusion Matrix. The authors report that the accuracy of tweets classification into the selected topics is 92 percentage, which is considered very high. The proposed web-based application appears to have potential for classifying tweets into topics of interest, which could be useful for information retrieval and analysis purposes. [4] The authors describe a research project that uses Twitter trend analysis to determine the most popular topics being talked about on Twitter. The objectives of the project are to analyze the relative popularity of different hashtags, determine which field has the maximum share of voice, and identify common interests of the community. The authors use machine learning algorithms, including latent Dirichlet allocation, cosine similarity, K means clustering, and Jaccard similarity techniques, to perform feature extraction and trend detection. The project compares the results with Big Data Apache SPARK tool implementation and reports that LDA resulted in an accuracy of 74 percentage, while Jaccard had an accuracy of 83 percentage for static data. The proposed work has implications for various fields, including business, marketing, politics, sports, and entertainment activities, where Twitter trend analysis can play an important role.[10] The authors describe a study that aims to classify sentiments in tweets using a data-set from the Kaggle website based on KFC's challenge and McDonald's of AI.

The data-set consists of more than 14,000 tweets, which are cleaned using Term Frequency-Inverse Document Frequency (TF-IDF) technique. Three classification algorithms, Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT), are applied for testing purposes.[6] The author describes a study that analyzes tweets about the National Education Policy using various machine learning algorithms for sentiment analysis, including Random Forest classifier, Logistic Regression, SVM, Decision Tree, XGBoost, and Naive Bayes. The study aims to determine the sentiment within the given text data and understand people's reactions to the policy. Twitter is used as a platform to collect and analyze people's opinions about specific events. The study concludes that the Decision tree algorithm performs best compared to all the other algorithms used in the analysis. This study has implications for marketing and innovation in understanding people's sentiments and reactions to various issues.[7] The passage describes a research work on multilingual Twitter sentiment analysis, specifically for Hindi and Kannada languages. While most studies in this area focus on English, this work aims to classify positive and negative tweets using language models for Hindi and Kannada. Machine learning models such as Naïve Bayes Classifier, KNeighbors' Classifier, Decision Tree Classifier, and Random Forest Classifier are used for classification. The research aims to understand the semantic nature of the text in order to accurately classify the tweets. The results of this study may have implications for sentiment analysis in other languages and for understanding public opinion in multilingual societies.[9] The author describes a research paper that proposes a sentiment analysis system based on extracting many tweets. The authors have used prototyping to develop the system, which is capable of classifying customers' perspectives on products or services as positive, negative, or neutral.[12] The proposed system aims to use sentiment analysis with Twitter hashtags by applying techniques such as tokenization, stemming, and vectorization. The data-set containing tweets is pre-processed to remove noisy and null value data, and the tweets are analysed and visualized for further processing. The system uses machine learning algorithms, such as Naive Bayes, KNN, SVM, Decision Tree, and logistic regression, for prediction. The system aims to classify the hashtags as disliked, somewhat famous, well-known, highly famous or very famous based on the limit user count. The system evaluates the accuracy of each algorithm for both content and context-oriented features, considering micro and macro F1 scores. The data set is divided into two parts, with 70 percentage used for training and the remaining 30 percentage for testing. The system aims to use Twitter data for data mining and opinion analysis. Pre-processing is used to convert unstructured tweet information from word to vector format to avoid misleading mining results due to incomplete, conflicting or noisy data. The system evaluates the precision, recall, and accuracy of the features and uses micro and macro F1 scores to have a single value for the measurements. The proposed system has implications for marketing and opinion analysis based on Twitter data[13].

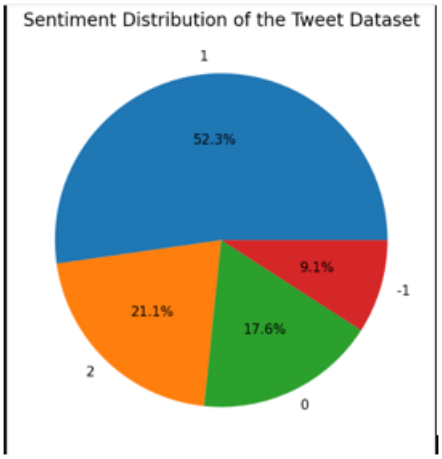


Fig. 1. Sentiment analysis of the tweet dataset

V. Proposed Framework

The proposed framework focuses on analyzing tweets related to environmental change, aiming to classify them into various categories: News, Pro, Neutral, and Anti. The data-set consists of 43,000 tweets, each labeled by three commentators. The framework follows several steps, including data prepossessing, tokenization, stemming, and implementation of different machine learning models to predict the sentiment of each tweet. The accuracy levels of the models are compared to determine the best model for this task. 1. Data Pre-processing: Data pre-processing is an essential step in analysing the noisy textual data obtained from Twitter. This step involves: a. Removing numbers, alphanumeric words, and expanding contractions. b. Tokenization to split the tweet into an array of words. c. Removing stop words to eliminate irrelevant information. d. Stemming to reduce the words to their root forms. 2. Feature Extraction: Vectorization is performed to convert the cleaned and stemmed words into a numerical representation that can be fed into machine learning models. This process includes the creation of a bag-of-words model, representing each tweet as a vector of word frequencies. 3. Model Implementation: The data-set is split into training (75a. Logistic Regression: A supervised learning algorithm used for predicting categorical dependent variables. It provides probabilistic values between 0 and 1. The model achieved an accuracy of 70b. Decision Tree: A tree-structured classifier suitable for both classification and regression tasks. The model uses decision and leaf nodes to represent features, decision rules, and outcomes, respectively. c. Random Forest: An ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and control over-fitting. This model achieved an accuracy of 70d. K-Nearest Neighbours: A classification algorithm that assigns the class of an instance based on the majority vote of its nearest neighbours. The model achieved an accuracy of 43e. Support Vector Machine: A classification algorithm that finds the optimal hyperplane

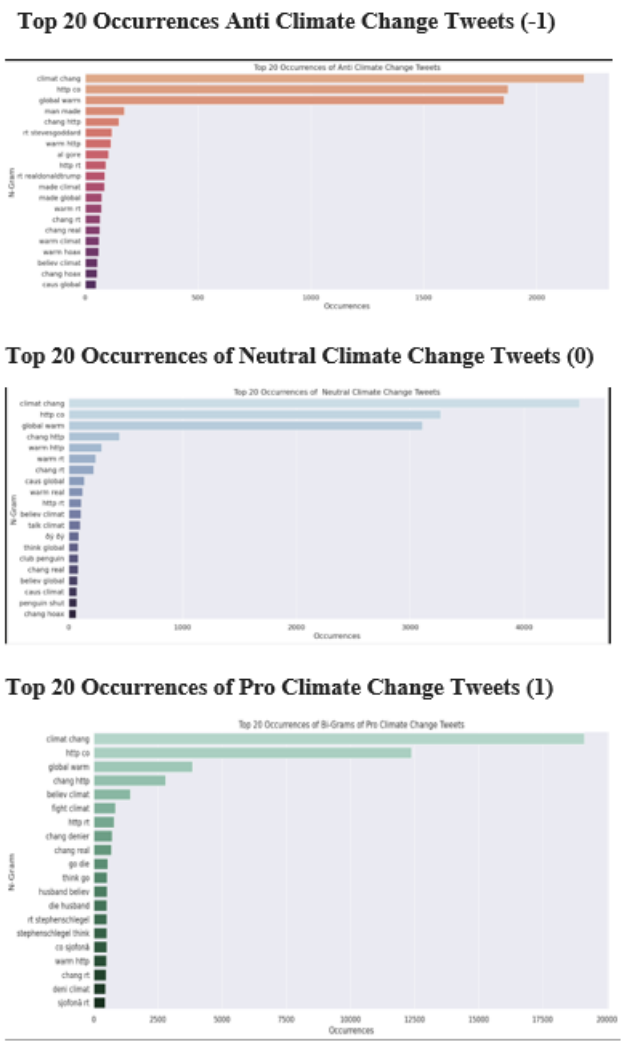


Fig. 2. Climate change tweets of anti ,neutral , pro

to separate classes in a multi-dimensional space. The model achieved an accuracy of 724. Model Evaluation: The accuracy levels of the different machine learning models are compared to determine the best model for this task. In this framework, the Support Vector Machine outperformed the other models, achieving an accuracy of 72

VI. Data Description

We need to import different python library files such as:

- nltk.downloader: This module is used to download additional NLTK resources such as corpora and models.
- RegexpTokenizer: This class is used to tokenize text into words based on regular expressions.
- PorterStemmer: This class implements the Porter stemming algorithm, which is used to reduce words to their root form (e.g., ”running” becomes ”run”).
- CountVectorizer: This class is used to convert a collection of text documents into a matrix of token counts. This is a common prepossessing step for text-based machine learning models.
- Counter: This class is used to count the frequency of elements in a list or other iterable object.
- train testnsplit: This function is used to split data into training and testing sets for machine learning models.
- metrics: This module provides various functions for evaluating the performance of machine learning models, such as accuracy, precision, recall, and F1 score. Install the python dataset by using the pip command.
- Import the ”data” function from the ”pydataset” package, which allows you to load various datasets that are commonly used in statistics and machine learning and then import several machine learning models from the ”sklearn” (scikit-learn) package, which is a popular library for machine learning in Python.
- Import the ”pickle” and ”json” modules, which are used for serializing and deserializing Python objects. This can be useful for saving and loading machine learning models, among other things.
- Import the ”makedirs” and ”path” functions from the ”os” module, which are used for working with directories and file paths.

  We include different process like: Tokenization: Tokenization is the process of splitting text data into smaller units called tokens, which are usually words, but can also be phrases, numbers, or punctuation marks. In the case of tweets, tokenization involves splitting the text of the tweet into an array of individual words.

Fig. 3. Tokenization

Stemming: Stemming is a process in natural language processing (NLP) that involves reducing words to their base or root form, known as a stem. This is often done to improve the efficiency and effectiveness of text analysis tasks such as search, classification, and clustering.

Vectorization Vectorization is the process of converting text data into a numerical representation that can be used by machine learning algorithms. In natural language processing (NLP), vectorization involves representing each text document or sentence as a vector of numbers that captures the important semantic information contained in the text.

Bi grams and trigrams: Bigrams and trigrams are types of n-grams in natural language processing (NLP) that capture

Fig. 4. Stopwords removal

Fig. 5. Vectorization

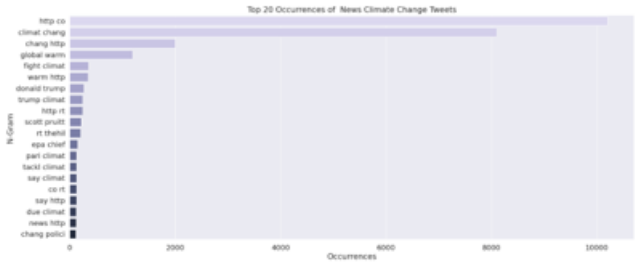
sequences of adjacent words in a text document. A bigram represents a sequence of two adjacent words, while a trigram represents a sequence of three adjacent words.

VII. Results

Implementation of Machine Learning models: The dataset is divided into training and testing data. The spitted dataset is passed into the different machine learning algorithm models and the accuracy levels were found.The models are evaluated by the train and test data. Create training set with 75 percentage of data and test set with 25 percentage of data. Build the model with the raining data. Logistic Regression

Logistic regression is one of the most famous machine learning calculations, which goes under the Directed Learning strategy. It is utilized for foreseeing the downright reliant variable utilizing a given arrangement of free factors. Logistic regression predicts the result of an unmitigated ward variable. Hence the result should be a clear cut or discrete worth. It very well may be either Yes or No, 0 or 1,, and so on yet rather than giving the specific worth as 0 and 1, it gives the probabilistic qualities which lie somewhere in the range of 0 and 1. Decision Tree Decision Tree is a Directed learning strategy that can be utilized for both characterization and Relapse issues, yet for the most part it is liked for taking care of Order issues. It is a tree-organized classifier, where inner hubs address the elements of a dataset, branches address the choice standards, and each leaf hub addresses the result. In a Decision tree, there are two hubs, which are the Decision Hub and Leaf Hub. Choice hubs are utilized to pursue any

Top 20 Occurrences of Bi-Grams of Factual Climate Change Tweets (2)



Top 20 Occurrences of Tri-Grams of Anti Climate Change Tweets (-1)

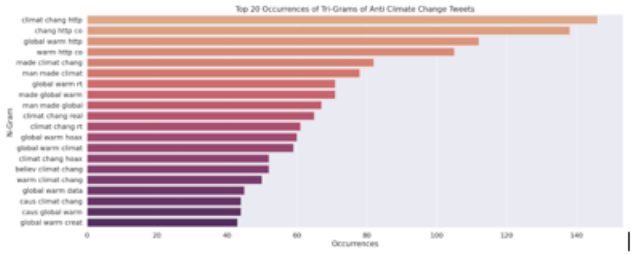
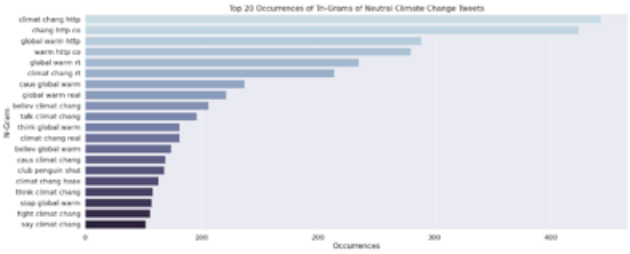
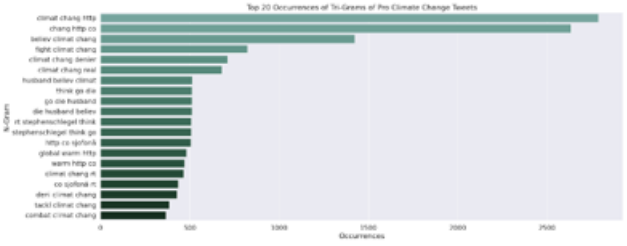


Fig. 6.

Top 20 Occurrences of Tri-Grams of Neutral Climate Change Tweets (0)



Top 20 Occurrences of Tri-Grams of Pro Climate Change Tweets (1)



Top 20 Occurrences of Tri-Grams of Factual Climate Change Tweets (2)

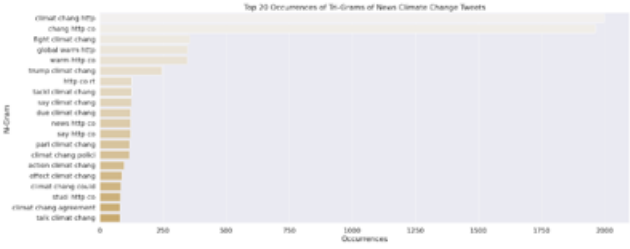


Fig. 7.

Logistic Regression

Fig. 8. Accuracy of logistic regression

Random Forest Classification

The random forest classifier gets the results of 70% of accuracy.

K-Neighbours Classification

K Neighbours classifier provides the accuracy of 43% in the evaluation of the ML model.

Linear Support Vector

Support vector provides the accuracy of 72% in the evaluation of the machine learning model.

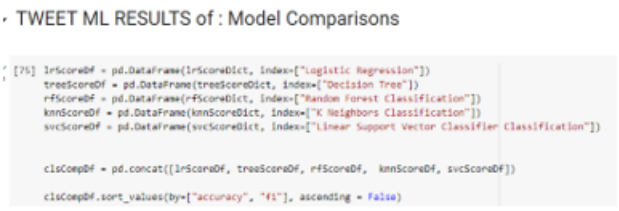


Fig. 9. Caption

	accuracy	recall	precision	f1
Logistic Regression	0.734025	0.734025	0.727631	0.729033
Linear Support Vector Classifier Classification	0.727107	0.727107	0.730806	0.710121
Random Forest Classification	0.702439	0.702439	0.709524	0.680830
Decision Tree	0.607591	0.607591	0.624698	0.551640
K Neighbors Classification	0.433825	0.433825	0.594123	0.448024

Fig. 10. Comparison Table

choice and have numerous branches, while Leaf hubs are the result of those choices and contain no further branches. The choices or the test are performed based on elements of the given dataset. It is a graphical portrayal for getting every one of the potential answers for an issue/choice considering given conditions. It is known as a choice tree on the grounds that, like a tree, it begins with the root hub, which develops further branches and builds a tree-like design

Random Forest Classification:

The accuracy levels of the different type of machine learning algorithms are compared with the graphical report.

The tweets of different messages have been analysed with the machine learning algorithms and the accuracy levels of the algorithms are compared.

REFERENCES

[1] A. Sarker, M. R. Islam and A. Y. Srizon, "A Comprehensive Preprocessing Approach for High-Performance Classification of Twitter Data with several Machine Learning Algorithms," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 630-633, doi: 10.1109/TENSYMP50017.2020.9230590..

[2] Barhate, Sanket, et al. "Twitter bot detection and their influence in hashtag manipulation." 2020 IEEE 17th India Council International Conference (INDICON). IEEE, 2020.

[3] Alzamzami, Fatimah, Mohamad Hoda, and Abdulmotaleb El Saddik. "Light gradient boosting machine for general sentiment classification on short texts: a comparative evaluation." IEEE access 8 (2020): 101840-101858.

[4] S. T. Indra, L. Wikarsa and R. Turang, "Using logistic regression method to classify tweets into the selected topics," 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Malang, Indonesia, 2016, pp. 385-390, doi: 10.1109/ICACSIS.2016.7872727.

[5] Y. Marini and K. T. Setiawan, "Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A Comparative Evaluation," FATIMAH ALZAMZAMI 1, MOHAMAD HODA 2, AND ABDULMOTALEB EL SADDIK 1, (Fellow, IEEE) VOLUME 8, 2020, IEEE Access. Conference Series: Earth and Environmental Science, vol. 149, no. 1, 2018, doi: 10.1088/1755-1315/149/1/012055.

[6] J. Singh and P. Tripathi, "Sentiment analysis of Twitter data by making use of SVM, Random Forest and Decision Tree algorithm," 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2021, pp. 193-198, doi: 10.1109/CSNT51715.2021.9509679.

[7] K. Agarwal, S. Deepa, R. V. SivaBalan and C. Balakrishnan, "Performance Analysis of Various Machine Learning Classification Models Using Twitter Data: National Education Policy," 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India, 2023, pp. 862-870, doi: 10.1109/IITCEE57236.2023.10091034.

[8] R. Caraka, R. C. Chen, T. Toharudin, M. Tahmid, B. Pardamean, and R. M. Putra, "Evaluation performance of SVR genetic algorithm and hybrid PSO in rainfall forecasting," ICIC Express Letters, Part B: Applications, vol. 11, no. 7 631, p. 639, 2020, doi: 10.24507/icicelb.11.07.631.

[9] algorithm and hybrid PSO in rainfall forecasting," ICIC Express Letters, Part B: Applications, vol. 11, no. 7 631, p. 639, 2020,ICIMTech 2020, pp. 185–190, 2020, doi:10.1109/ICIMTech50083.2020.9211246.

[10] Loyola-González, O., Monroy, R., Rodríguez, J., López-Cuevas, A., Mata-Sánchez, J.I. (2019). Contrast Pattern-Based Classification for Bot Detection on Twitter. IEEE Access, 7, 45800-45817.

[11] L. Yung-Hui, Y. Nai-Ning, K. Purwandari, and L. N. Harfiya, "Clinically applicable deep learning for diagnosis of diabetic retinopathy," Proceedings - 2019 12th International Conference on Ubi-Media Computing, Ubi-Media 2019, pp. 124–129, 2019, doi: 10.1109/Ubi-Media.2019.00032.

[12] Satish, G. Yamini, M. Ashok Kumar, and K. Sudhakar. "Emotion Recognition on Twitter: Comparative Study and Training a Unison Model. "October 2022

[13] E. Dogariu, S. Garg, B. Khadan, A. Potts and M. Scornavacca, "Using Machine Learning to Correlate Twitter Data and Weather Patterns," 2019 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 2019, pp. 1-4, doi: 10.1109/URTC49097.2019.9660487.

[14] Alsaqer, M.; Alelyani, S.; Mohana, M.; Alreemy, K.; Alqahtani, A. Predicting Location of Tweets Using Machine Learning Approaches. Appl. Sci. 2023, 13, 3025. https://doi.org/10.3390/app13053025

[15] Anisha P. Rodrigues, Roshan Fernandes, Adarsh Bhandary, Asha C. Shenoy, Ashwanth Shetty, M. Anisha, "Real-Time Twitter Trend Analysis Using Big Data Analytics and Machine Learning Techniques", Wireless Communications and Mobile Computing, vol. 2021, Article ID 3920325, 13 pages,

[16] Devi, P. Suthanthira, R. Geetha, and S. Karthika. "Trending-tags—Classification and Prediction of Hashtag Popularity Using Twitter Features in Machine Learning Approach Proceedings." Computational Intelligence in Data Mining: Proceedings of the International Conference on ICCIDM 2018. Springer Singapore, 2020.

[17] A. Budiarto, R. Rahutomo, H. N. Putra, T. W. Cenggoro, M. F. Kacamarga, and B. Pardamean, "Unsupervised News Topic Modelling with Doc2Vec and Spherical Clustering," Procedia Computer Science, vol. 179, pp. 40–46, 2021, doi:

[18] Sanket Barhate, Ritika Mangla "Twitter bot detection and their influence in hashtag manipulation," 978-1-7281-6916-3/20/31.00 ©2020 IEEE.

[19] R. E. Caraka, S. A. Bakar, M. Tahmid, H. Yasin, and I. D. Kurniawan, "Neurocomputing fundamental climate analysis," Telkomnika (Telecommunication Computing Electronics and Control), vol. 17, no. 4, pp. 1818–1827, 2019, doi: 10.12928/TELKOMNIKA.v17i4.11788.

[20] R. Rahutomo, A. Budiarto, K. Purwandari, A. S. Perbangsa, T. W. Cenggoro, and B. Pardamean, "Ten-year compilation of savekpk twitter dataset," Proceedings of 2020 International Conference on Information Management and Technology, ICIMTech 2020, pp. 185–190, 2020, doi: 10.1109/ICIMTech50083.2020.9211246.

[21] T. B. Pramono et al., "A Model of Visual Intelligent System for Genus Identification of Fish in the Siluriformes Order," IOP Conference Series: Earth and Environmental Science, vol. 794, no. 1, 2021, doi: 10.1088/1755-1315/794/1/012114.