2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA) | 979-8-3503-0426-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCUBEA58933.2023.10392242

Analyzing the Performance of Naive Bayes, Logistic Regression, SVM and Random Forest for Identifying hate speech from Twitter Social Media

Disha S. Wankhede[1]
Dept. of Computer Engg,
Vishwakarma Institute of Information Technology
Pune, India
disha.wankhede@viit.ac.in
https://orcid.org/0000-0001-6245-3097

Gaikwad Vidya S.[2]
Dept. of Computer Engg,
Vishwakarma Institute of Information Technology
Pune, India
vidya.gaikwad@viit.ac.in
https://orcid.org/0000-0001-9785-9249

Akshay Manikjade[3]
Dept. of Mechanical Engg,
Vishwakarma Institute of Information Technology
Pune, India
akshay.manikjade@viit.ac.in

Nikita Meher[4]
Dept. of Computer Engg,
Vishwakarma Institute of Information Technology ,
Pune, India

Aishwarya Ghule[5]
Vishwakarma Institute of Information Technology,
Pune, India

Deepika Gujar[6]
Dept. of Computer Engg,
Vishwakarma Institute of Information Technology ,
Pune, India

Tejas Atkale[7]
Dept. of Computer Engg,
Vishwakarma Institute of Information Technology ,
Pune, India

***Abstract*- The spread of hate on social media and other platforms is of great concern because it has the implicit to be a serious detriment to society and the country and is disastrous. The development of mass media has led to lower exposure of hate speech and discrimination. detest speech generally refers to a person or group of people predicated on race, color, race, gender, race, religion, etc It's defined as humiliating communication grounded on certain characteristics. The donation of Structure through named images, type of caption and words used in the textbook of) can explain the causes of virality and what is associated with it. Fake news and hate speech are not the result of the internet age. Fake news and hate have been around since the morning of mortal history, that is, for times- people have fabricated and fabricated all the time. still, the phenomenon of social media has changed how, where and what it's associated with fake news and hate speech. Where lies and culmination formerly appeared on the internet, now fake news and hate speech are taking their place on social media. The thing of the design is to propose results that stoners can use to identify and filter to count hate speech on Twitter. Using colorful bracket styles, we can determine whether tweets are hate speech. We use traditional shadowing algorithms Analogous as Naive Bayes, Logistic Regression, SVM and Random Forest. From data collection, preprocessing, point engineering and type to, we will estimate and give the delicacy and quality of all classifiers and their results in discrimination quests on Twitter..**

Keywords: Fake news, hate speech, social media, classification, Naive Bayes, SVM, Random Forest

I. INTRODUCTION

How important is security that we read about in the media and publish "intellectual" news? It's easy for anyone to post what they want, and while that's okay, there's also the idea of going too far, like creating fear, using lies to control other people's decisions, or posting false information online. basis. Something else that could have lasting effects. Fake news is the deliberate spread of false information by news sources or social media. Fake information spreads very fast. It turned out that when a fake news is removed, another will immediately take its place. Also, fake news is different from real news because it spreads so fast. People can download information from websites, share information, share it with others, and finally, false information is so far removed from its internal principles that it cannot tell the difference between real news.

In the age of easy media and social media. We must take action against this letter, because these letters will lead to political hatred or cooperation, which will cause great harm to people. With the advent of social media and social media, it has become important to detect fake news and hate speech spread on colored paper or posters. These fake and hateful news can have a negative impact on our society by spreading hatred among communities, religions or different groups of people. The complexity and confusion that the Internet offers makes it crucial that we find and dispose of similar articles and publications.

Access to the global internet has allowed information to circulate around the world and has opened up numerous new openings. On the other hand, besides the good, the expansion and growth of addict- generated content on the internet has also excluded poisons and detest speech. Significant work has been done in the direction of speech discovery. still, it has come important not only to exclude the poisonous content, but also to combat it. While some social connections hamper sensitive content, another consequence may be that poisons were discovered in Stoner's book, and Stoner compartmented it in, giving the textbook a false interpretation.

This work can be considered as a flexible work. According to the European Commission, any speech that expresses, encourages, encourages or creates racism, bigotry, antisemitism or other hate speech falls within the scope of the word "discrimination". 4 General statement published by the Committee for the Prevention of Terrorism in the background document for the Council of Europe speech. 5 Delineations aren't only particular, but also include groups and individualities grounded on" race," color, strain, nation or nation, age, disability, language, religion or belief, gender, gender, fornication, and other special characteristics or status. includes the word" demarcation" as a legal term relating to felonious, civil or law enforcement acts analogous to hate, conduct, libel, compulsion, importunity or attacks on public violence.. Other exploration motifs can be set up inanti-Muslim racism, sexism, homophobia and transphobia(borders grounded on sexual or gender identity), antiziganism(borders between Sinti and Roma), capability(to demarcate groups with disabilities), and others. word), classism(bias grounded on social origin), appearance(apparent discrimination).

II. RELATED WORKS

M. S. Al-Zaman (2021) mentioned former literature provides at least eight types of fake news from colorful Tandoc et al. It linked 34 mock studies. linked types of fake news,news lampoon, propaganda, and announcements. Jain, A., Shakya, A., Gupta, & Khatter.(2019) suggested each can be interpreted directly and objectively; some have high delicacy and deception, while others have low delicacy and deception. numerous studies have shown fake news motifs related to the current study.

Drawing on the literal environment of screen ecology, Higdon (2020) explores the four main themes of fake news: race, racism, drama rumors, and fear. While the study tried to give a better understanding of motifs, it didn't accept the complexity of fake news and didn't cover important motifs similar to politics and terrorism.When the media began in the 20th century, it was recorded in 4,444 exploration papers.

1) Accuracy: Accuracy is another name for well-estimated value. It is the estimated quality ratio.

Eventually, Sirafi, H., Rashid, M., & Alwasel, B. N. (2021) suggested more people started using websites and receiving newsletters. To expose bogus news, they employ creative techniques and tools including natural language processing, machine learning, and artificial intelligence. In a press statement, Facebook stated that it is battling fake news in two key areas. M. Granik, V. Mesyura, (2017) mentioned the first step is to eliminate fiscal support because it is the source of the majority of fake news. The alternative is to create newer products to aid in the dissemination of false information, as well as some of the safety precautions adopted by Facebook Advancements. Fake news is less prevalent because of News Feed Ranks.

Direct marketing Outline what is and is not useful. Inconsistent local news reports will be lowered in consumer fare. M. Gahirwal (2018) mentioned WhatsApp has implemented various security mechanisms, such as fake news detection, to counter the spread of false information, although these are only in the beginning stages and are not yet accessible to beta users. The "Unfortunate Call" point is being tested by WhatsApp. By putting red labels on links that it knows will lead to phony websites or other news websites, this point will warn stoners. Additionally, communication will be blocked if it is transferred more than 25 times over the device. The following Table 1 contains various existing classifiers used in various research papers.

K. Bhattacharjee et al (2020), Karnik, M.P et al.,(2023) had proposed Disambiguation Approaches on Unstructured Texts and abstractive text summarization.

TABLE I. SUMMARY OF VARIOUS EXISTING CLASSIFIER

Year of Publication	Classifier Used	Dataset
2023[1]	Hypothesis,Machine Learning,Textual Analysis	Datasets consisting of 691,234 source tweets and ~35.5 million
2023[2]	Machine Learning Deep Learning	A large Arabic hate speech dataset, called arHateDataset
2022[8]	NLP,Text Representation,Hypothesis	Check-worthiness tweets in CheckThat!2022,Fake news spreaders in PAN2020,Hate speech spreaders in 3PAN2021
2022[10]	Machine literacy and Deep literacy	176 fake news events published on Sina Weibo from July 15 to Sept 19
2021[11]	Linear Mixed Models	Telecommunications Ownership and Control Dataset (Freyburg et al., 2021)
2021[12]	Machine Learning,Deep Learning,NLP,Neural Network	Created own dataset by collection of tweets using various hashtags
2020[13],[16]	SVM,Logistic-Regression, Naive Bayes and XGBoost	Hate Speech Dataset from a White Supremacy Forum
2021[14]	Naive Bayes,Machine Learning,NLP	Political News Data(LIAR)dataset
2019[9]	Naïve Bayes,Machine Learning	Social Media,News Website
2019[15]	ICT;Social Media	Dataset of laws from different countries

III. PROPOSED SYSTEM:

Figure 1 depicts the details of the machine learning techniques we have employed in our work to determine whether the given tweet is hateful or not. In this study for binary classification of tweets, a machine learning algorithm was used. Binary categories are bad and bad. In this article, we use traditional tracking methods such as Naive-Bayes, Logistic-Regression, SVM and Random-Forest.The following steps to identify distinctive language are:

i) data collection ii) preprocessing iii) feature engineering iv) classification

A. Data collection:

The information in this study was taken from Twitter. In these extracted data, tweets were collected in two different categories as hate speech and normal speech. The file contains 17000 tweets.

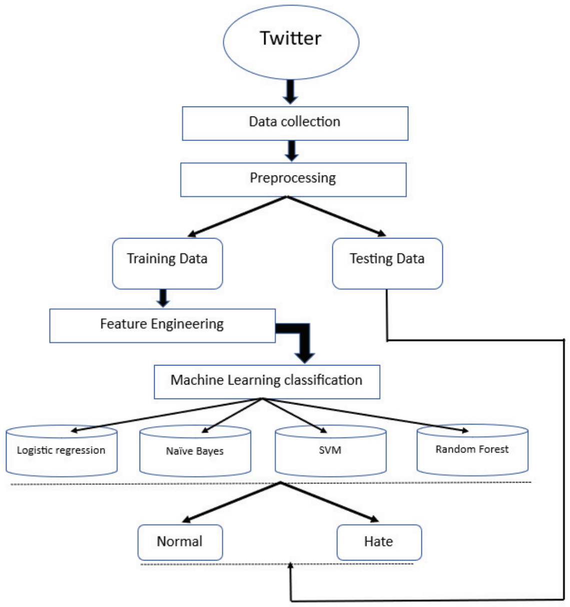


Fig. 1. System Architecture

B. Preprocessing:

Unstructured information pertaining to noise, value, etc. was collected from Twitter. Data organization requires preprocessing. As a result, in this instance, we filter away noise and uninformative characteristics using several prioritization techniques. tweets. In the past, we substituted tweets with a smaller number. In addition, we eliminate all hashtags, symbols, stop words, URLs, users, locations, and matching pattern data from tweets.

A) Expand Contractions

B) Lower Case

data['twitter_text'] = data['twitter_text'].lower()

C) Remove punctuations

data["twitter_text"] = data['twitter_text"].apply(lambda x: re.sub('[%s]' % re.escape(string.punctuation), '' , d))

D) Remove words and digits containing

digitsdata['twitter_text'] = data['twitter_text"].apply(lambda d: re.sub('W*dw*','',d))

E) Remove Stopwords

data['twitter_text'] = data[['twitter_text'].apply(lambda d: remove_stopwords(d))

F) Rephrase text

#email-id

df['twitter_text'] = df['twitter_text'].apply(lambda d: re.sub('b[w-.]+?@w+?.w{2,4}b', 'emailadd',d))

#url

df['twitter_text'] = df['twitter_text'].apply(lambda d:re.sub('(http[s]?S+)|(w+.[A-Za-z]{2,4}S*)', 'urladd', d))

G) Stemming and Lemmatization

df['twitter_text'] = df['twitter_text'].apply(lambda d: stem_words(d))

df['twitter_text'] = df['twitter_text'] .apply(lambda text: lemmatize_words('twitter_text'))

H) Remove Extra Spaces

df['twitter_text'] = df[['twitter_text'].apply(lambda text: re.sub(' +', ' ', d)

C. Feature Engineering:

Feature engineering determines how well a machine learning system performs. In this study, TF-IDF video extraction technique was used. The following equation calculates in TF/IDF.

$$TFIDF(t,w,D) = TF(t,w) * IDF(t,D)\ IDF(t,D) = log\frac{|D|}{1+|\{w \in D : t \in w\}|}$$

Where , t represents message-based features, w represents each tweet, and and D represents all tweets.

D. Classification using Machine Learning Algorithm:

Machine learning algorithms are used for binary classification of tweets. The dual classes are Hate and Normal. In this article, we use traditional methods such as Naive Bayes, Logistic Regression, SVM and Random Forest.

IV. IMPLEMENTATION

By comparing products, we can determine whether a newspaper is fake or not. After collecting important information, it is necessary to make some progress in the old newspaper. Download research papers and use the most established distributions. To train data classification, we must first extract many features. The next step is to determine if the news is hate speech.The classifiers we use are Random Forest, Logistic Regression, Support Vector Machine, Naive Bayes, and Random Forest, which is currently using in our work, is a machine learning algorithm commercialized by Leo Breiman and Adele Cutler, combining the results of various decisions. wood is combined to achieve a result. It is easy to use and adopt as it can solve the problem of deployment and recovery. That's why we used random forests. Logistic retrogression is a machine literacy system used to break distribution problems. It's a system of assessing prognostications grounded on the conception of probability. The bracket algorithm logistic retrogression was used to estimate the probability of the categorical dependent variable. One of the quick and straightforward machine learning techniques for predicting data classes is naive Bayes. It can be utilized for multiclass and double brackets. In comparison to other algorithms, it performs well in many vaccination classes. most common solution to bracket problems in textbooks You can use Naive Bayes to determine whether a tweet is typical. It is a textbook classification algorithm. Token operation is connected to tweets that could be normal or spiteful, and Bayes' theorem is employed to determine how delicate the news.

A supervised machine learning approach for bracket and retrogression called random timber, often referred to as arbitrary decision tree, uses trees to identify the most important classes. By combining these results, the major goal is to improve the delicate vaticination process. A data analysis method called logistic retrogression is used to forecast a double overgrowth similar to a yes-or-no (double distribution) outcome based on previous data set compliances.

It's important to note that the variable in logistic retrogression is a double variable, meaning it can only take a value of 1(yes, the result is positive) or 0. Different performance measures are used to estimate the performance of the design. Some performance criteria in the textbook are compactly explained below. The performance of the product is estimated by calculating true negative(TN), negative(FP), negative(FN), and positive[TP].

1) Precision : Precision is another name for well-estimated value. It is the proportion between estimated and actual quality.

Precision is equal to TP/(TP + FP).

2) Recall : The proportion of actual to anticipated quality.

Recall is TP / (TP + FN).

3) F-Measure: The harmonic approach to recall and precision (shown in Equation 3). Precision and recall are equally weighted in the conventional F measure (F1).

Measurement factor F is equal to 2 * (precision * recall) / (precision + recall).

4) Accuracy: This is the total number of correctly positive and incorrectly negative events classified.

Accuracy is equal to (TP+TN)/(TP+FP+TN+FN)

5) Confusion Matrix:

TABLE II. CONFUSION MATRIX

	Projected No	Projected Yes
Actual No	TN	FP
Actual Yes	FN	TP

The classifiers have also been tested by varying the dataset size. The accuracy, precision and recall variation with expansion of datasets is depicted in the tables below.Table 2 shows how the accuracy of the classifier varies as the dataset is expanded. Table 3 shows the variation of precision with expansion of the dataset. Table 4 depicts the variation of recall with expanding dataset.

System Configuration

OS Windows 11 Pro , Memory 16 GB DDR4, Microprocessor Intel Core i7 @ 2.7GHz Simulation Time 9.110 seconds

TABLE III. ACCURACY VARIATION WITH DATASET EXPANSION

Dataset	Naive Bayes	Random Forest	SVM	Logistic Regression
5000	94	99	99	98
10000	95	99	98	98
15000	95	99	98	97
20000	94	99	98	97
31962	95	98	98	97

TABLE IV. PRECISION VARIATION WITH DATASET EXPANSION

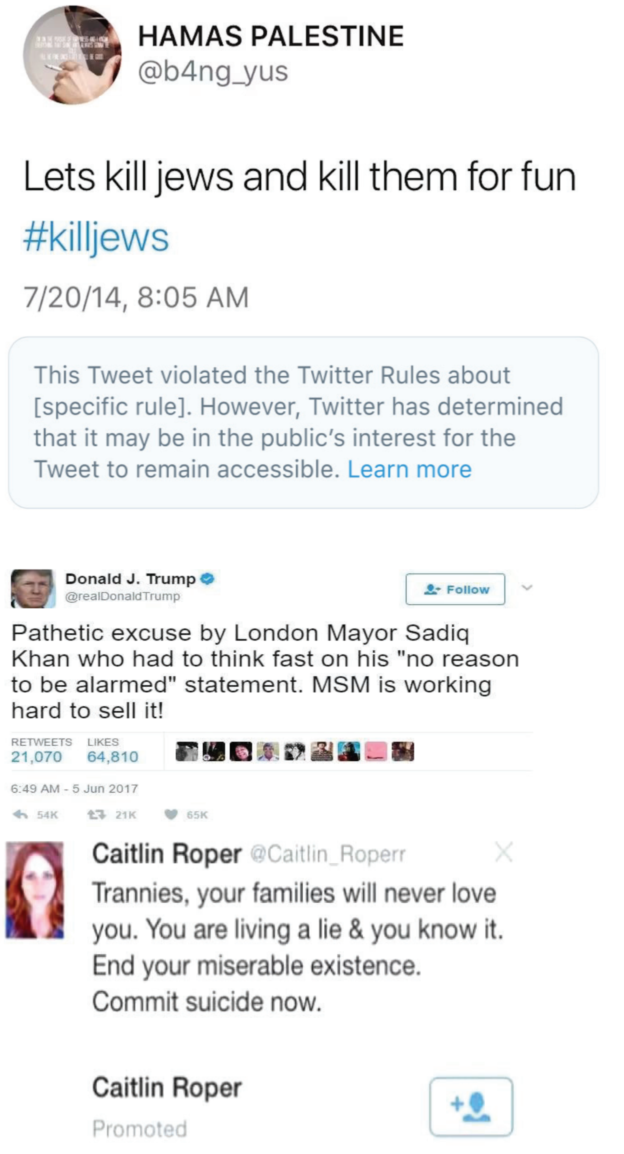
Dataset	Naive Bayes	Random Forest	SVM	Logistic Regression
5000	92	98	98	97
10000	92	98	97	96
15000	91	98	97	96
20000	92	98	97	96
31962	91	96	97	96

TABLE V. RECALL VARIATION WITH DATASET EXPANSION

Dataset	Naive Bayes	Random Forest	SVM	Logistic Regression
5000	100	100	100	100
10000	99	100	100	99
15000	99	99	99	99
20000	98	99	99	99
31962	96	100	99	99

For example

There’s few examples where people have said bad and a lot of negative things on twitter and spreadings lies all around. Here are some examples



HAMAS PALESTINE
@b4ng_yus

Lets kill jews and kill them for fun
#killjews

7/20/14, 8:05 AM

This Tweet violated the Twitter Rules about [specific rule]. However, Twitter has determined that it may be in the public's interest for the Tweet to remain accessible. Learn more

Donald J. Trump
@realDonaldTrump

Follow

Pathetic excuse by London Mayor Sadiq Khan who had to think fast on his "no reason to be alarmed" statement. MSM is working hard to sell it!

RETWEETS 21,070 LIKES 64,810

6:49 AM - 5 Jun 2017

54K 21K 65K

Caitlin Roper @Caitlin_Roperr

Trannies, your families will never love you. You are living a lie & you know it. End your miserable existence. Commit suicide now.

Caitlin Roper
Promoted

Fig. 2.

The tweet made by people about killing Jews is violating twitter rules and causing lots of problems among the people. We identify it as negative and hate speech. We suggest blocking them and making it not visible to the audience.

V. RESULTS

We use Twitter hate speech data labeled as hate speech or normal speech. The dataset is analyzed and figure 3 shows the confusion matrix of the dataset analysis using the four detection algorithms. The search algorithms are as follows:

- Naive Bayes Classifier
- Support Vector Machine
- Random Forest Classifier
- Logistic Regression

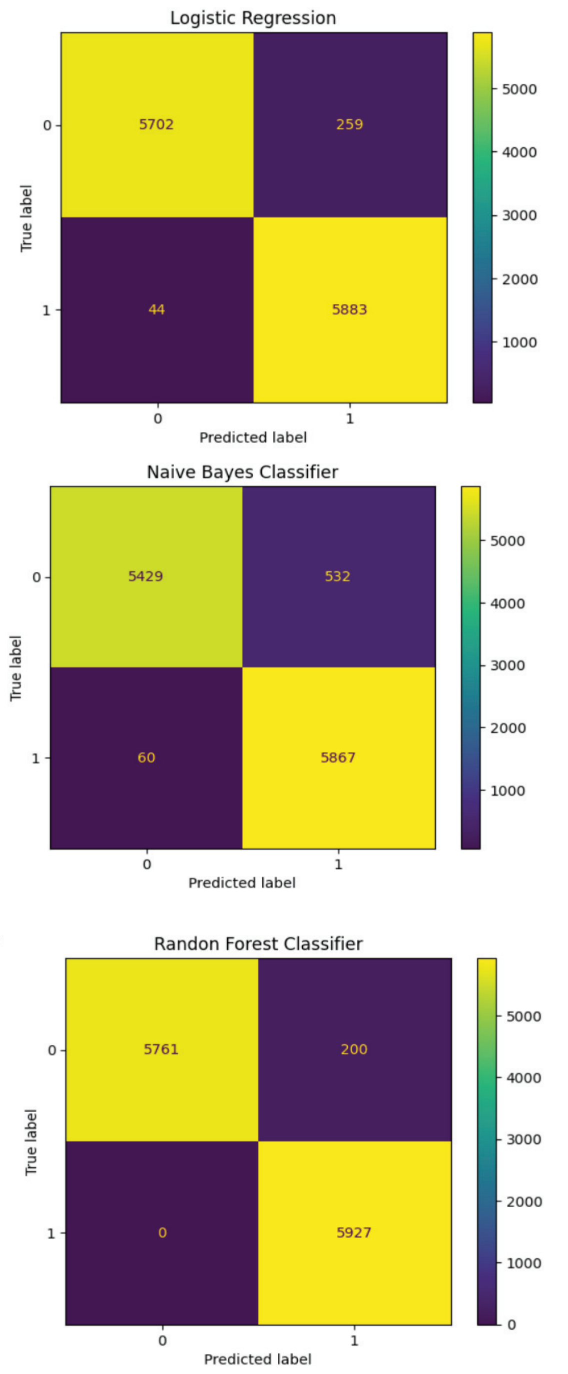


Fig. 3. Confusion Matrix for all algorithms

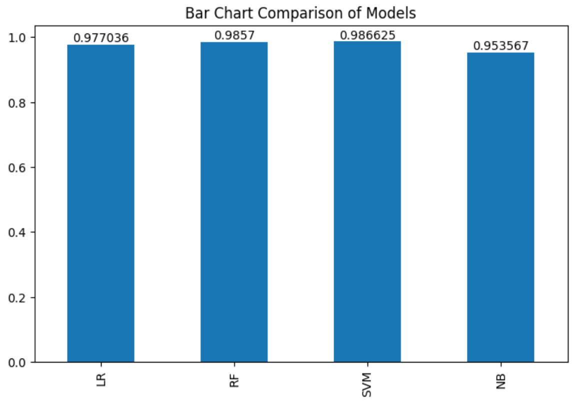


Fig. 4. Accuracy results for all algorithms

The accuracy of these algorithms is shown in figure 4. Random forest and SVM have the highest accuracy at 98%, followed by logistic regression at 97% and naive Bayes at 95%.

TABLE VI. CLASSIFIER COMPARISON BASED ON PRECISION, RECALL AND F1 SCORE.

Classifier	Precision	Recall	F1 Score
Random Forest	96	100	98
Logistic Regression	96	99	97
SVM	97	99	98
Naive Bayes	91	96	93

Table 5 shows the classifier comparison based on: Precision, Recall, F-Measure. SVM has the highest accuracy of 97 and thus the best classifier.The recall of the Random Forest classifier is 100. F1 score combined and inverse, random forest and SVM outperformed the other at 98.

VI. CONCLUSION

Social media has grown exponentially and so has the spread of hate speech. Racism is speech that degrades a person on the basis of race, gender, color, ethnicity, or religion. People suffer from hate speech. We reviewed the literature to identify the best methods for detecting hate speech and concluded that Naive Bayes classifiers, Random Forest classifiers, logistic regression and SVMs are the most useful methods. This study uses text classification techniques to identify hateful tweets. In addition, this study compares four machine learning algorithms to classify hateful tweets. Random Forest and SVM show better results than Logistic-Regression and Naive Bayes. The lowest performance is found in Naive-Bayes. First, the proposed ML model is ineffective as a real-time data estimator. Finally, simply dividing hate speech into two different categories is not enough to analyze the seriousness of the message.

REFERENCES

[1] Solovev, K., & Pröllochs, N. (2022). Moralized language predicts hate speech on social media. PNAS Nexus. https://doi.org/10.1093/pnasnexus/pgac281

[2] Ramzi Khezzar1 Abdelrahman Moursi1 Zaher Al Aghbari1 (2023)

[3] M. S. Al-Zaman (2021).Indian fake news on social media. 9(1), 25–33, Asian Journal for Public Opinion Research. citation: 10.15206/ajpor.2021.9.1.25.

[4] Jain, A., Shakya, A., Gupta, & Khatter.(2019) a clever machine learning-based fake news detecting system. Paper presented at the ICICT Conference in September 2019 with the following DOI: 10.1109/ICICT46931.2019.8977659.

[5] Sirafi, H., Rashid, M., & Alwasel, B. N. (2021). Kazham, Z. Using Machine Learning Techniques to Spot Fake News. Materials Science and Engineering, IOP Conference Series, 1099, 012040. DOI: 10.1088/1757-899X/1099/1/012040..

[6] M. Granik, V. Mesyura, (2017) Detecting fake news with a naïve Bayes classifier. (pp. 900–903) in 2017 IEEE 1st Ukr. Conf. Electr. Computer Eng. UKRCON 2017 - Proc. Reference: 10.1109/UKRCON.2017.8100449..

[7] M. Gahirwal (2018). International Journal of Advance Research, Ideas and Innovations in Technology, 4(1), 817-819. Fake News Detection.

[8] 8) Bader, L., Bender, J., and Beutel, I. (2022) COUNTER-FAKE: A scientific foundation for a strategy against hate speech and false information. 591-604 in Open Cultural Studies, 3(1). 10.24989/ocg.v.342 is the doi.

[9] Anjali JainAvinash Shakya et al.(2019)A smart System for Fake News Detection Using Machine Learning September 2019 DOI: 10.1109/ICICT46931.2019.8977659

[10] Wang, Xin ; Chao, Fan ; Ma, Ning ; Yu, Guang (2022) ,Exploring the Effect of Spreading Fake News Debunking Based on Social Relationship Networks Publication: Frontiers in Physics, vol. 10, id. 833385 Pub Date: April 2022 DOI: 10.3389/fphy.2022.833385

[11] Lisa Garbe,Icon,Lisa-Marie Selvik,Icon &Pauline Lemairec (2021) How African countries respond to fake news and hate speech Pages 86-103 2021, Published online: 09 Nov 2021,doi.org/10.1080/1369118X.2021.1994623

[12] Perifanos, Konstantinos, and Dionysis Goutsos. 2021. "Multimodal Hate Speech Detection in Greek Social Media" Multimodal Technologies and Interaction 5, no. 7: 34. https://doi.org/10.3390/mti5070034

[13] Samujjwal Goswami, Manoj Hudnurkar, Suhas Ambekar. (2020). FAKE NEWS AND HATE SPEECH DETECTION WITH MACHINE LEARNING AND NLP. PalArch's Journal of Archaeology of Egypt / Egyptology, 17(6), 4309 - 4322. Retrieved from https://archives.palarch.nl/index.php/jae/article/view/1686

[14] Z Khanam1, B N Alwasel1, H Sirafi1 and M Rashid,(2021))Fake News Detection Using Machine Learning Approaches DOI 10.1088/1757-899X/1099/1/012040

[15] Santuraki, S.U. (2019)."Trends in the Regulation of Hate Speech and Fake News: A Threat to Free Speech?," Hasanuddin Law Review, 5(2): 140-158 DOI: 10.20956/halrev.v5i2.1625

[16] Ganesh Udge, et al(2019), "Statistical Analysis for Twitter Spam Detection", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 6 Issue 2, pp. 624-629, March-April 2019. Available at doi : https://doi.org/10.32628/IJSRSET1962170

[17] K. Bhattacharjee et al., "Survey and Gap Analysis of Word Sense Disambiguation Approaches on Unstructured Texts," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 323-327, doi: 10.1109/ICESC48915.2020.9155947.

[18] Karnik, M.P., Kodavade, D.V. (2023). A Survey on Controllable Abstractive Text Summarization. In: Abraham, A., Pllana, S., Casalino, G., Ma, K., Bajaj, A. (eds) Intelligent Systems Design and Applications. ISDA 2022. Lecture Notes in Networks and Systems, vol 715. Springer, Cham. https://doi.org/10.1007/978-3-031-35507-3_30