International Journal of Information Management Data Insights 2 (2022) 100120

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques

Akib Mohi Ud Din Khanday [a,*], Syed Tanzeel Rabani [a], Qamar Rayees Khan [a], Showkat Hassan Malik [b]

[a] Department of Computer Sciences Baba Ghulam Shah University, Rajouri, Jammu & Kashmir 185234, India
[b] Department of Computer Sciences University of Kashmir, Srinagar, Jammu & Kashmir 190006, India

ARTICLE INFO

Keywords:
COVID-19
Social networks
Misinformation
Hate speech
Machine leaning
Ensemble learning

ABSTRACT

The COVID-19 pandemic has impacted every nation, and social isolation is the major protective method for the coronavirus. People express themselves via Facebook and Twitter. People disseminate disinformation and hate speech on Twitter. This research seeks to detect hate speech using machine learning and ensemble learning techniques during COVID-19. Twitter data was extracted from using its API with the help of trending hashtags during the COVID-19 pandemic. Tweets were manually annotated into two categories based on different factors. Features are extracted using TF/IDF, Bag of Words and Tweet Length. The study found the Decision Tree classifier to be effective. Compared to other typical ML classifiers, it has 98% precision, 97% recall, 97% F1-Score, and 97% accuracy. The Stochastic Gradient Boosting classifier outperforms all others with 99 percent precision, 97 percent recall, 98 percent F1-Score, and 98.04 percent accuracy.

1. Introduction

Information and communication technology (ICT) advancements have altered the artistic method of conveying and receiving information. Regardless of their diversity in behaviour, everyone in our universe wants to be kept up to date. People share views on Online Social Networking sites,various clients use this platform for spreading dubious/false information (Joseph, Kar, & Ilavarasan, 2021). To secure information, we have to secure all links in a chain comprising PPT(People, Process, Technology). In the chain of links, people are usually weakest in any communication. Nowadays, adversarial use of social media is ubiquitous, and it is frequently used to distribute fake or misleading statements, posing a social, economic, and political risk (Spohr, 2017; World Economic Forum, 2017). As the COVID-19 pandemic expands, more and more people practice physically distancing themselves from one another. The coronavirus consists of Severe Acute Respiratory Syndrome (SARS), Middle East Respiratory Syndrome (MERS) and Acute Respiratory Distress Syndrome (ARDS) Viruses. According to "World Health Organisation", symptoms of this virus are Mild Fever, Sore throat, Dry Cough and running nose (Khanday, Rabani, Khan, Rouf, & Mohi ud Din, 2020d). Until 6th July 2020, no vaccine/drug is approved to cure this deadly virus. The COVID 19 pandemic had an undecorated political, economic and social effect. Social media and communication systems are also affected in extraordinary ways. As Classical media has tried to adjust to the quickly evolving situation, Alternate news media on the internet gave coronavirus its ideology spin. These media have been criticized for promoting social confusion and spreading theoretically hazardous "Fake News" or Conspiracy philosophies via social media and other available platforms (Bail et al., 2018; Kar & Aswani, 2021). Facebook a social networking place which also owns WhatsApp and Instagram, published a report in which it was revealed that the messaging has been doubled since the rise in a pandemic. In certain countries, hate speech comes under umbrella of free speech. There are prohibitions against encouraging violence or societal disruption in the United States, Canada, France, the United Kingdom, and Germany (Hua et al., 2013; Gillani, Yuan, Saveski, Vosoughi, & Roy, 2018). Facebook and Twitter have been criticised for not doing enough to prevent their services from being used to assault persons of a certain race, ethnicity, or gender (Opinion | Twitter Must Do More to Block ISIS - The New York Times 2021) . They've stated that they'll work to eliminate prejudice and intolerance (Facebook's Mark Zuckerberg 'Understands Need to Stamp out Hate Speech', Germany says | Daily Mail Online 2021). Meanwhile, therapeutic approaches, such as those used by Facebook and Twitter, have relied on users to report improper remarks, which has been a manual effort (Facebook, Google, and Twitter agree German hate speech deal - BBC News 2021; Grover, Kar, Dwivedi, and Janssen, 2019). This not only necessitates a lot of effort on the part of human experts, but this also increases the risk of bias in judgements. Furthermore, a computer-based solution can accomplish this activity significantly faster than humans, a non-automated process performed by human annotators would have a significant impact on system reaction time. The need to automate the process of detecting online hate speech is highlighted by the tremendous

* Corresponding author.
E-mail address: akibkhanday@bgsbu.ac.in (A.M.U.D. Khanday).

https://doi.org/10.1016/j.jjimei.2022.100120
Received 21 September 2021; Received in revised form 9 September 2022; Accepted 9 September 2022

increase in user-generated content on prior social networking sites, as well as the inability to scale manual screening (Kushwaha, Kar, & Vigneswara Ilavarasan, 2020; Grover, Kar, & Ilavarasan, 2019; Wu & Gerber, 2018).

COVID-19 has created a social crisis by increasing inequality, exclusion and discrimination. Various rumours, philosophies, and propaganda regarding coronavirus were shared massively on various social network platforms like (Facebook, Twitter, WhatsApp, etc.). Logically unfound theories on potential causes and medicines made the rounds, triggering misperception and unsafe behaviour among people who followed these distorted and false recommendations. Hoaxes and propaganda are also being shared enormously through online social networks (Khanday, Khan, & Rabani, 2020b; Aswani, Kar, & Ilavarasan, 2019). With the advent of this pandemic in India, propaganda has blown out many fabrications. Hatemongers spread hate by criticizing a specific community (Khanday, Khan, & Rabani, 2020a). Due to the Tableegi Jaamat event held at Nizamuddin Markaz Delhi, India, various hate speech is being used to target a particular community. According to the latest Twitter data, various trending hashtags are being used to criticize a specific community. Many hate words are being tweeted every day, which can lead to a hazardous situation (Neubaum & Krämer, 2017). It is still a challenge in social networking to detect hate speeches in real-time and attracted many researchers to develop scalable, automated methods for detecting hate speech using semantic content analysis based on Machine Learning (ML) and Natural Language Processing (NLP) (Burnap & Williams, 2015; Ji Ho Park & Pascale, 2017). We extracted data from twitter using trending hashtags, labelling them manually into two Normal and Hate. Using the Twitter API and keywords, about 11K tweets were retrieved. #CoronaJihad, #CoronaTerrorisma, #COVID-19 and #MuslimCorona. The main aim of this study is to create a classifier that can classify tweets into hate and non-hate categories using several Machine Learning techniques that have been fine-tuned.

The noteworthy contributions of this paper are:

- The hybrid features engineering is being performed by merging TF/IDF, Bag of Words and Tweet Length..
- Thirty thousand tweets are extracted from Twitter to form a dataset, out of which 11000 are related to hate speech and are accordingly labeled to a particular class.
- Traditional machine and ensemble learning algorithms are trained and tested based on the proposed hybrid feature selection to classify the hate content shared through Twitter in COVID-19 Era.

The paper consists of VI sections, a brief background of hate speech and machine learning is being given in Section II. Section III provides detail of the proposed methodology. Experimental results are discussed in section IV, section V discusses about implications and limitations of proposed work and section VI concludes our work.

2. Related work

As social media has grown in popularity, research on automated hate speech detection has become a subject of public interest. When used to prevent text posting or blocklisting people, simple word-based algorithms fail to uncover subtle offending content and jeopardize right to free speech and expression (Khanday, 2022). The issue of word ambiguity originates from the fact that a single word can have multiple meanings in different situations, and it is the fundamental reason of these approaches' high false-positive rate. Some conventional NLP techniques are unsuccessful at detecting uncommon spelling in user-generated comment content (Kar & Dwivedi, 2020). This is also known as the "spelling variety problem", and it occurs when single characters in a token are intentionally or unintentionally replaced in order to obfuscate the detection. By and large, the intricacy of natural language constructions makes the process reasonably difficult. But due to the less availability of datasets, the researchers are not able to find the solution using the latest technology. Hate speech detection using supervised learning classification algorithms is not a new concept. Del Vigna, Cimino, Dell'Orletta, Petrocchi, and Tesconi (2017) found that a simple LSTM classifier did not outperform a standard SVM. Another method used a supervised model to detect the objectionable language in tweets (Davidson, Warmsley, Macy, & Weber, 2017).A binary classifier is used for classifying a tweet into abusive language and hate speech. Nobata, Tetreault, Thomas, Mehdad, & Chang (2016) made an attempt to a supervised algorithm using various linguistic features to detect abusive material and grammatical aspects in the text, evaluated at the character unigram and bigram levels, and validated using Amazon data In general, we can highlight the most critical elements. The non-language agnostic feature of NLP-based models is one of their major flaws—the low detection scores.

For classification, machine learning methods can be used, but these algorithms need a huge amount of data for training. Hate speech term was used by Burnap and Williams (2015), Gitari, Zuping, Damien, and Long (2015), Silva, Mondal, Correa, Benevenuto, and Weber (2012). Hate speech detection is typically cast by state of the art as a supervised text classification task (Schmidt & Wiegand, 2017; Dubois & Blank, 2018). Various classical machine learning algorithms that rely on manual feature selection can perform this binary classification (Warner & Hirschberg, 2012; Kwok & Wang, 2013; Waseem, 2016) showed how annotation is important in the classification task. The comparison was performed between expert and amateur annotations. About 6909 tweets are annotated using Crowd Flower. The annotators are chosen based on their knowledge of hate speech. The results showed that expert annotation showed better accuracy in classifying hate speech. Davidson et al. (2017) Hate speech is defined as rhetoric that is used to show hatred toward a certain group or is projected to denigrate, embarrass, or abuse the followers of that group. Crowd-sourcing is used to collect tweets that comprises keywords regarding hate speech. The tweets were labelled into multi-class tweets regarding hate speech, tweets with offensive words, and those that don't contain hate or offensive words. For labelling, Crowd-sourcing was used. The overall Precision, Recall and F1-score of best model are 91%, 90% and 90%, respectively. About 40% of the tweets are misclassified. The Precision of 44% and Recall of 61% are the classification report of hate class. Around 5 per cent of offensive tweets and 2% of inoffensive tweets have been wrongly classified as hate speech.

Hate speech can be detected using NLP concepts to use sentences' lexical and syntactic features (Waseem, 2016) and AI-solutions and bag-of-words-based text representations (Dubois & Blank, 2018). Unsupervised learning approaches for detecting offensive remarks in text are extremely widespread. Hatred users employ numerous obfuscation strategies, such as swapping a single character in insulting remarks, making automatic identification more difficult. They were using a binary classifier, for example. It has already been attempted on a paragraph2vec representation of words. Amazon data in the past, however it only worked successfully on an issue of binary classification (Djuric et al., 2015). Another solution based on unsupervised learning, the authors offered a set of criteria for judging whether or not a tweet is offensive (Waseem &Hovy, 2016). They also discovered that changes in user distribution by geography have just a minor influence. The detecting performance is only marginally affected. Another researcher used crowd-sourced strategy for combating hate speech, including constructing a new collection of annotations which supplements the obtainable dataset (Waseem, 2016). The effect of annotators' experience on labelling performance was explored. The authors dealt with tweet classification, but their main focus was on sexism, which they classified as "hostile," "benevolent," or "other" (Jha & Mamidi, 2017).The authors employed Waseem & Hovy (2016) dataset of tweets, They labelled existing 'Sexism' tweets as 'Hostile,' while gathering their own for the 'Benevolent' class, to which they then applied the FastText and SVM (Joulin, Grave, Bojanowski, & Mikolov, 2017). To solve the challenge, a supervised learning model based on a neural network is deployed. The technique

Table 1
Summary of related work.

Refs.	*Features*	*Dataset Used*	*Accuracy*	*Limitations*
Davidson et al. (2017).	TF/IDF and PoS.	StormFront (de Gibert et al., 2019)	73.64%	Only two feature Engeneering Techniques are used.
Davidson et al. (2017).	TF/IDF and PoS.	HatEval [CodaLab - Competition 2021]	73.9%	Only two feature Engeneering Techniques are used.
Davidson et al. (2017).	TF/IDF and PoS.	TRAC (Kumar et al., 2018)	56.04%	Only two feature Engeneering Techniques are used.
Davidson et al. (2017).	TF/IDF and PoS.	HatebaseTwitter (Davidson et al., 2017)	90.07%	Only two feature Engeneering Techniques are used.
Zimmerman et al. (2019).	integrates ten convolutional neural networks with varying initial weights.	StormFront (de Gibert et al., 2019)	80.33%	Words are connected without being adjacent.
Zimmerman et al. (2019).	integrates ten convolutional neural networks with varying initial weights.	HatEval (CodaLab - Competition 2021)	74.70%	Words are connected without being adjacent.
Zimmerman et al. (2019).	integrates ten convolutional neural networks with varying initial weights.	TRAC (Kumar et al., 2018)	53.58%	Words are connected without being adjacent.
Zimmerman et al. (2019).	integrates ten convolutional neural networks with varying initial weights.	HatebaseTwitter (Davidson et al., 2017)	92.13%	Words are connected without being adjacent.
MacAvaney et al. (2019).	TF/IDF.	StormFront (de Gibert et al., 2019)	80.33%	Only TF/IDF is used.
MacAvaney et al. (2019).	TF/IDF.	HatEval (CodaLab - Competition 2021)	75.9%	Only TF/IDF is used.
MacAvaney et al. (2019).	TF/IDF.	TRAC (Kumar et al., 2018)	61.21%	Only TF/IDF is used.
MacAvaney et al. (2019).	TF/IDF.	HatebaseTwitter (Davidson et al., 2017)	91.08%	Only TF/IDF is used.

surpassed any previously known unsupervised learning solution on the same dataset of tweets (Badjatiya, Gupta, Gupta, & Varma, 2017).

Character n-grams extract features, and Gradient Boosted Decision Trees help with the LSTM model. Using character n-grams and word2vec pre-trained vectors, Convolution Neural Networks (CNN) were studied as a potential solution to the hate speech problem in tweets. Ji Ho Park & Pascale, 2017 turned the categorization into a two-step problem, in which abusive language is first differentiated from non-abusive material. The sort of abuse is then determined (Sexism or Racism). Four classes were forecasted using pre-trained CNN vectors, according to the authors (Gambäck & Sikdar, 2017). In terms of F-score, they were marginally better than character n-grams.Even though the success of NLP approaches in hate-speech classification (Schmidt & Wiegand, 2017), we believe Machine learning models can still make a significant contribution to the problem. It's also worth highlighting the challenge's inherent difficulty at this time, as indicated by the fact that no solution has yet managed to attain an F-score higher than 0.93. Table 1 summarizes the related work done in the field of Hate speech on social networks.

It is necessary to do research in order to identify people who use hate speech on social media, focusing on both their features and motivations as well as the social structures in which they are embedded. From the literature review, the following findings can be drawn:

- The majority of the work has been done on the already existing dataset.
- There is more scope for feature engineering, if done properly, the accuracy of the machine learning algorithms may increase.
- The dataset used in the existing work suffers from data Imbalance.

3. Methodology

The proposed methodology which is being used for detecting hate speech using Machine Learning is shown in Fig. 1 depicts a series of steps: (i) Data collection (ii) Preprocessing (iii) Feature Engineering (iv) Machine Learning Classification (v) Ensemble Learning Classification.

3.1. Data collection

Data is being extracted via Twitter, a social media platform mostly used by celebrities and politicians to express their opinions.We used its Application Program Interface (API) (Verma, Khanday, Rabani, Mir, & Jamwal, 2019). Various steps are being followed to extract data using Twitter API. We used hashtag #CoronaJihad, #CoronaTerrorism and #MuslimCorona to extract data from 4th April to 8th April 2020. About 30K tweets were extracted, but only 11K were relevant. The data is saved in the form of a CSV file such that it can be used for future analysis. The extracted dataset consists of about 16 attributes like Created, Text, Id, Screen Name etc.

3.1.1. Tweet length distribution

By this, we get the Length of each tweet in characters such that we can see the size of hate speech and non-hate speech tweets. Fig. 2 gives the tweet length distribution of the whole data set.

3.1.2. Human annotation

Human annotation is a significant step in our research. The labelled tweets are needed for training the supervised machine learning models. Various researchers working in this area were given the task of annotation. They were asked to classify text based on context and words used in the tweet. This task is a binary classification problem, having two classes Hate and Normal. The tweets which contain words like F**k, S**t, hate, worst etc. were put in the class Hate others were put in a Normal class. After annotation, we got a dataset of 11k records, but it was unbalanced so to remove unbalances, we consider 4,093 tweets, as shown in Fig. 3.

3.2. Preprocessing

The data collected from Twitter is in the unstructured form, which contains noise, null values etc. for refining the data, it needed to be preprocessed such that it can be used for classification purposes. Preprocessing is critical for deciphering the meaning of brief texts in classification applications and clustering and anomaly detection. Preprocessing has a large impact on overall system performance, but receives less attention than feature extraction and classification. The preprocessing process includes preparing tweets for tasks such as event recognition, fraudulent information detection, sentiment analysis, and so on. On social media, people frequently adhere to their own set of informal language rules. As a result, each Twitter user has their own writing style, complete with abbreviations, unusual punctuation, and misspelt words. Emoticons and emojis are used in tweets to convey complexity, sentiment, and ideas. Slang and acronyms are common in tweets,

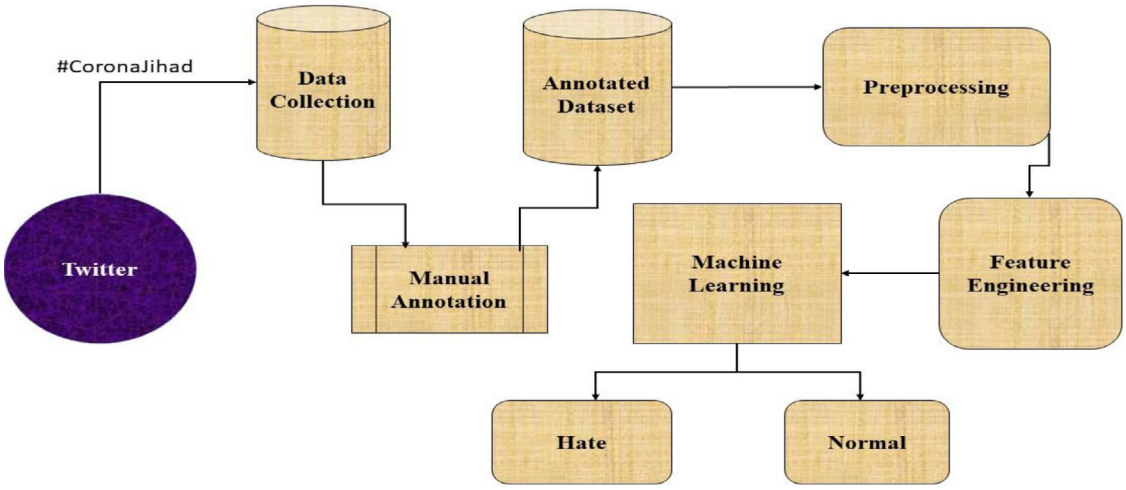


Fig. 1. Proposed methodology.

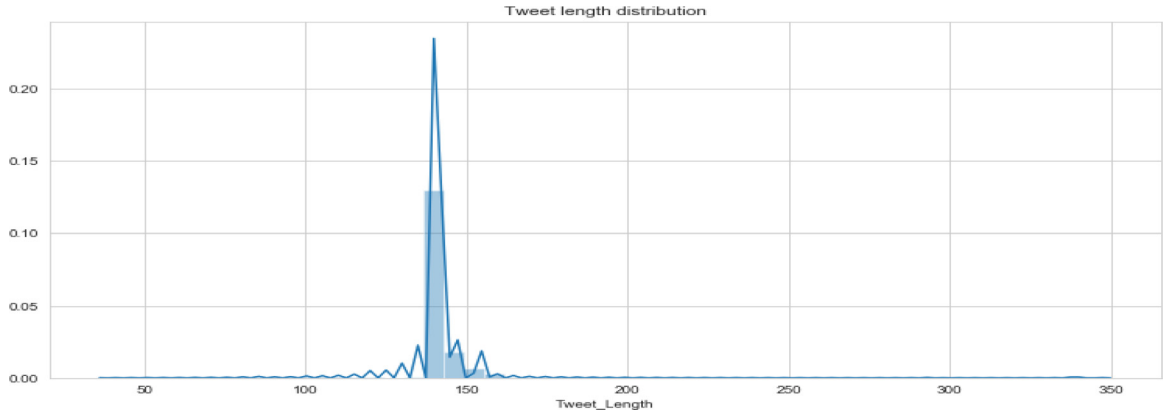


Fig. 2. Length of tweets extracted.

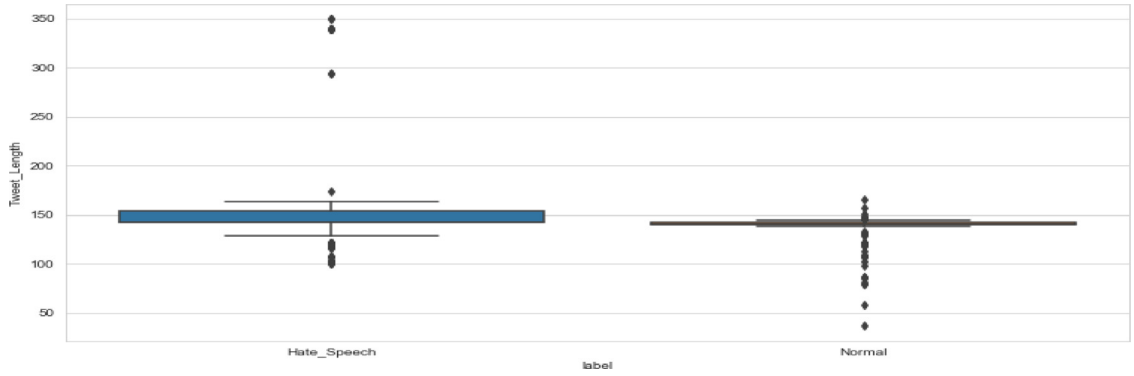
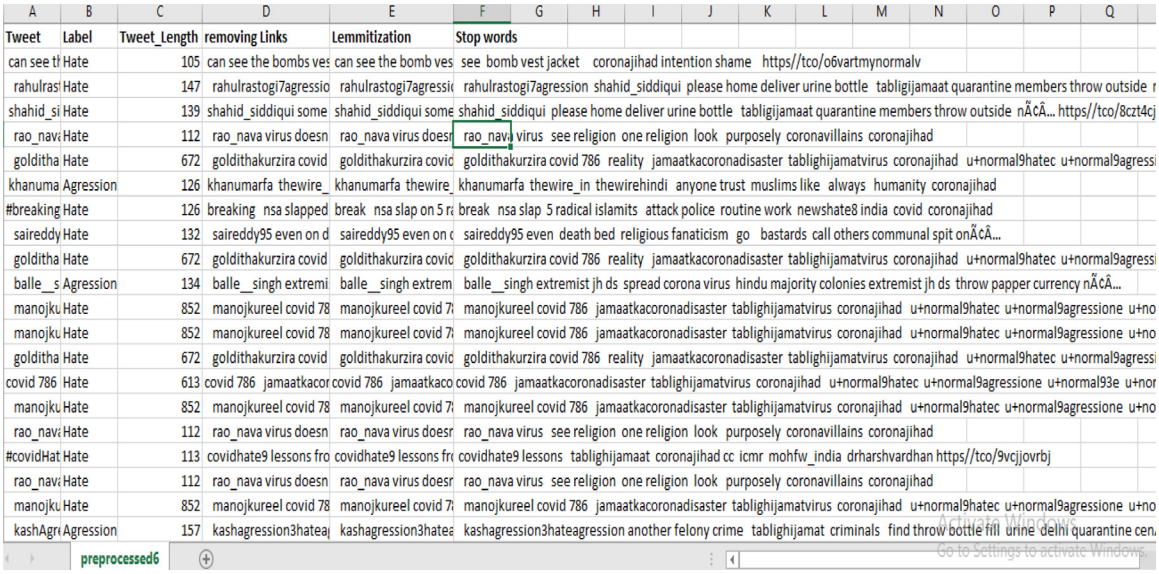


Fig. 3. Balanced data of each class and their corresponding length.

"URLs", "hashtags", and "user mentions". Data noise is caused by unwanted strings and Unicode, which are a result of the crawling process. In addition, practically all user-posted tweets include URLs that link to extra information, user mentions (@username), and the hashtag symbol (#coronaterrorism) to connect their message to a specific subject, and these hashtags can also express mood. These indications provide vital supplementary information to people, but they supply no information to machines and can be considered noise that must be dealt with. Researchers have proposed a number of techniques for dealing with this additional data offered by users, including replacing URLs with tags in one study (Agarwal et al., 2011), and removing user mentions (@username) in another study (Khan, Bashir, & Qamar, 2014).

To communicate sentiment and opinion, Twitter users utilise emoticons and emojis such as:-),;-),:-(, and others. This vital information must be recorded to effectively classify tweets. Words were used to replace emojis and expressions (Gimpel et al., 2011). Twitter's character constraints discourage natural language usage, prompting users to adopt acronyms, abbreviations, and slang. Abbreviations include MIA (missing in action), gr8 (great), and ofc (of course). Slang is a casual way of expressing thoughts or meaning that is sometimes limited to specific individuals or settings. twitter, and OMG usually alludes to a surprise



Tweet	Label	Tweet_Length	removing Links	Lemmitization	Stop words
can see th	Hate	105	can see the bombs ves	can see the bomb ves	see bomb vest jacket coronajihad intention shame https//tco/o6vartmynormalv
rahulrast	Hate	147	rahulrastogi7agressio	rahulrastogi7agressi	rahulrastogi7agression shahid_siddiqui please home deliver urine bottle tablighijamaat quarantine members throw outside r
shahid_si	Hate	139	shahid_siddiqui some	shahid_siddiqui some	shahid_siddiqui please home deliver urine bottle tablighijamaat quarantine members throw outside nÃcÂ... https//tco/8czt4cj
rao_nava	Hate	112	rao_nava virus doesn	rao_nava virus doesr	rao_nava virus see religion one religion look purposely coronavillains coronajihad
golditha	Hate	672	golditha kurzira covid	golditha kurzira covic	golditha kurzira covid 786 reality jamaatkacoronadisaster tablighijamatvirus coronajihad u+normal9hatec u+normal9agressi
khanuma	Agression	126	khanumarfa thewire_	khanumarfa thewire_	khanumarfa thewire_in thewirehindi anyone trust muslims like always humanity coronajihad
#breaking	Hate	126	breaking nsa slapped	break nsa slap on 5 ra	break nsa slap 5 radical islamits attack police routine work newshate8 india covid coronajihad
saireddy	Hate	132	saireddy95 even on d	saireddy95 even on c	saireddy95 even death bed religious fanaticism go bastards call others communal spit onÃcÂ...
golditha	Hate	672	golditha kurzira covid	golditha kurzira covic	golditha kurzira covid 786 reality jamaatkacoronadisaster tablighijamatvirus coronajihad u+normal9hatec u+normal9agressi
balle__s	Agression	134	balle__singh extremi	balle__singh extrem	balle__singh extremist jh ds spread corona virus hindu majority colonies extremist jh ds throw papper currency nÃcÂ...
manojku	Hate	852	manojkureel covid 78	manojkureel covid 7	manojkureel covid 786 jamaatkacoronadisaster tablighijamatvirus coronajihad u+normal9hatec u+normal9agressione u+no
manojku	Hate	852	manojkureel covid 78	manojkureel covid 7	manojkureel covid 786 jamaatkacoronadisaster tablighijamatvirus coronajihad u+normal9hatec u+normal9agressione u+no
golditha	Hate	672	golditha kurzira covid	golditha kurzira covic	golditha kurzira covid 786 reality jamaatkacoronadisaster tablighijamatvirus coronajihad u+normal9hatec u+normal9agressi
covid 786	Hate	613	covid 786 jamaatkacor	covid 786 jamaatkaco	covid 786 jamaatkacoronadisaster tablighijamatvirus coronajihad u+normal9hatec u+normal9agressione u+normal93e u+nor
manojku	Hate	852	manojkureel covid 78	manojkureel covid 7	manojkureel covid 786 jamaatkacoronadisaster tablighijamatvirus coronajihad u+normal9hatec u+normal9agressione u+no
rao_nava	Hate	112	rao_nava virus doesn	rao_nava virus doesr	rao_nava virus see religion one religion look purposely coronavillains coronajihad
#covidHat	Hate	113	covidhate9 lessons fro	covidhate9 lessons fr	covidhate9 lessons tablighijamaat coronajihad cc icmr mohfw_india drharshvardhan https//tco/9vcjjovrbj
rao_nava	Hate	112	rao_nava virus doesn	rao_nava virus doesr	rao_nava virus see religion one religion look purposely coronavillains coronajihad
manojku	Hate	852	manojkureel covid 78	manojkureel covid 7	manojkureel covid 786 jamaatkacoronadisaster tablighijamatvirus coronajihad u+normal9hatec u+normal9agressione u+no
kashAgr	Agression	157	kashagression3hatea	kashagression3hatea	kashagression3hateagression another felony crime tablighijamat criminals find throw bottle fill urine delhi quarantine cen

preprocessed6

Fig. 4. Preprocessed dataset.

or emphasis rather than the literal expansion of oh my God. As a consequence, replacing casual insertions in tweets with their genuine word meaning improves automatic classifier performance without information loss. Abbreviations and slang were translated into word meanings in a study, which were then easily understood using standard text analysis tools (Scheuer et al., 2011). Humans understand punctuation well, but it is less useful for automatic text classification. As a result, eliminate punctuation while preparing text for tasks like sentiment analysis. Some punctuation characters, including! and?, can communicate emotion eliminated punctuation (Lin & He, 2009). However, replacing a question mark or exclamation mark with appropriate tags, such as!, can often express astonishment (2013). Like stemming, lemmatization simplifies a word. In lemmatization, linguistic knowledge is used to turn a word into its base form. Only tweets written in English are considered and converted in lower. Stopwords like a, an, the, etc., are removed using the stopword lexicon. Punctuation is also being performed, and the text is being divided into tokens called tokenization. Stemming is also used to get the root word example understanding will be converted to understand. Links, URLs etc are removed, and lemmatization is done. Fig. 4 shows the visual representation of the preprocessed data set.

3.3. Feature engineering

Feature engineering decides whether a machine learning classifier will perform well or not. In this step, various features are extracted using multiple techniques like TF-IDF, a bag of words, sentence length also emphatic features are taken into consideration. The following Eq. (1). calculates the TF/IDF in context to our corpus.

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

Where t stands for the word as a feature, w stands for each tweet in the corpus, and D stands for the total number of tweets in the corpus (Document space).

Bag of Words features: It is composed of words and lemma.We used bigrams and trigram terms to extract more information from the text. Some of the selected features are corona jihad, COVID, dangerous Muslim, India, India come, dangerous, Muslim Coronavirus, report, coronavirus,coronajihad,billyperrigo etc.

3.4. Classification using traditional machine learning

For performing binary classification of tweets, Machine Learning algorithms are used. The binary classes are Hate and Normal. In this paper, traditional supervised machine learning algorithms are used for performing binary classification tasks. Logistic Regression (LR), Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM) (LR), and Decision Tree algorithms were applied.

3.4.1. Logistic regression

Logistic Regression (LR) forecasts the arithmetic variable of a class constructed on its correlation with labels (Khanday, Khan, & Rabani, 2020c). The input is in the form of a Table with various values. About 50 features are chosen using TF/IDF and Bag of Words during feature engineering. LR usually computes the class relationship possibility, and here are two classes. $y \in \{0,1\}$.The subsequent options can be computed using Eq. (2).

$$P(y=r|x) = \frac{exp^{\varphi^T\theta_r}}{1+\sum_{r=1}^{3} exp^{\varphi^T\theta_k}}\ \forall r=1\ P(y=0|x) = \frac{exp^{\varphi^T\theta_r}}{1+\sum_{r=1}^{3} exp^{\varphi^T\theta_r}} \quad (2)$$

3.4.2. Multinomial naïve bayes

This Machine Learning algorithm uses the Bayes rule for calculating class probabilities of tweets (Rabani, Khan, & Khanday, 2020). Assuming d as the set of classes(Hate, Normal) and N denote the total number of features. In our Problem d= 0,1 and N=50. Multinomial Naïve Bayes allocates test tweet ti to highest probability class $P(d|ti)$ using Bayes rule shown in Eq. (3):

$$P(d \mid t_i) = \frac{P(d)P(t_i \mid c)}{P(t_i)}, d \in d \quad (3)$$

The division of the number of labelled hate speech tweets d to the total number of hate speech tweets gives the value of P(d) value. $P(t_i|d)$ Is the possibility of finding a hate speech tweet like $t_i$in-class c and is computed by:

$$P(t_i|d) = \left(\sum_n f_{ni}\right)! \prod_n \frac{P(w_n|d)^{f_{ni}}}{f_{ni}!}$$

Where f_{ni} = count of word 'n' in tweet 'ti' & $P(w_n|d)$ the probability of word 'n' given class d. the latter probability is calculated from the training data as:

$$P(w_n|d) = \frac{1 + F_{nd}}{N + \sum_{x=1}^{N} F_{xd}}$$

Where F_{xc} = count of word 'x' in training documents having class d.

3.4.3. Support vector machine

It is a type of supervised machine learning algorithm used to classify text into various classes is Support Vector Machine (SVM) (Khanday, Khan, & Rabani, 2021). Assume that 'n' be the number of features of a specific tweet corresponding with a label. We have about 50 features which are unigram and bigrams. Training set data points are $(y_k, x_k)_1^n$ Where n is the number of features chosen. It takes 50 features as input in the form of a table. The motive of the SVM is to build a classifier in the form of Eq. (4).

$$y(x) = sign\left[\sum_{k=1}^{n} \alpha_k y_k \psi(x, x_k) + b\right] \tag{4}$$

Where: α_k = *positive real constant.*
b = *real constant.*

$$\psi(x, x_k) = \begin{cases} x_k^T x : \ Linear\ SVM \\ (x_k^T x + 1)d : Polynomial\ SVM\ with\ Degree\ d \\ exp(-||x - x_k||_2^2/\sigma^2 : \ RBF\ SVM \end{cases}$$

Where: k, σ are constants.

Assuming the following equations classifier can be constructed in which +1 shows the hated class, and -1 shows normal class:

$$\omega^T \varphi(x_k) + b \geq 1, \ if \ y_k = +1$$

$$\omega^T \varphi(x_k) + b \leq -1, \ if \ y_k = -1$$

Which is equivalent to Eq. (5):

$$y_k[\omega^T \varphi(x_k) + b \leq -1, if\, y_k = -1, k = 1, \ldots, n \tag{5}$$

Where $\varphi(.)$ = Nonlinear Mapping function used to map input into higher dimensional space.

Classification is being performed using hyperplane, which distinguishes the two classes (Hate and Normal). For constructing a hyperplane new variable ξ_k is introduced. The Eq. (6) is for the hyperplane and is shown below:

$$y_k \left[\omega^T \varphi(x_k) + b\right] \geq 1 - \xi_k, \ k = 1, \ldots .., n \ \xi_k \geq 0, \ k = 1, \ldots \ldots ., n \tag{6}$$

3.4.4. Decision trees

Decision trees are an alternative method for performing binary classification (Khanday, Khan, & Rabani, 2020a). Decision trees partition the input space into regions and classify each region autonomously. It takes 50 features as input in the form of a table. Space is recursively splitted according to the input. It classifies the Tweets at the bottom of the tree. Leaf nodes do binary classification. An important function needs to be considered while building a Decision tree known "splitting criterion". Splitting Criterion describes how data must be splited to maximize the performance of a decision tree. Information gain ratio is being used in our work, information gain to the intrinsic information gives us the value of information gain ratio shown in Eq. (7).

$$IGR(EX, a) = IG/IV \tag{7}$$

Where IG= Information Gain.
IV= Intrinsic information.
IG can be computed by having the value of Entropy:

$$IG(Ex, a) = H(Ex) - \sum_{v \in values(a)} \left(\frac{|\{x \in Ex | value(x, a) = v\}|}{|Ex|} . H(\{x \in Ex | value(x, a) = v\})\right)$$

Where Ex = Training Set and x$x \in Ex$ which describes the value of a particular training instance 'x' having features 'a'.

H= Entropy and a=features.

IV can be Computed by:

$$IV(Ex, a) = - \sum_{v \in values(a)} \frac{|\{x \in Ex | value(x, a) = v\}|}{|Ex|} . log_2 \left(\frac{(\{x \in Ex | value(x, a) = v\})}{|Ex|}\right)$$

3.5. Ensemble learning techniques

Ensemble classifiers are also used for performing binary(Hate and Normal) classification. Ensemble machine learning classifiers are used to improve accuracy. In our work, we used Bagging, Adaboost, Random Forest and Gradient Stochastic Boosting Ensemble learning techniques for performing binary classification.

3.5.1. Bagging

To increase the efficiency of classification and regression tasks, ensemble learning techniques are applied. The Bagging Technique causes us in abstaining from overfitting. Given a preparation set X having 'n' size, by examining consistently, it produces 'm' preparing sets 'Xi' each of size 'n' with substitutions. The information is presented in the form of a table with various values for the 50 attributes that were picked. Because of substitutions, a few perceptions could rehash in every Xi. If m'=n, at that point for enormous n Xi is relied upon a division (1 - 1/e) to one of a kind instances of X, the rest are copies. The example is recognized as the bootstrap sample. By utilizing 'm' bootstrap samples, 'm' models are fitted and are consolidated by voting.

3.5.2. Adaboost

Adaboost algorithm uses weighting occasions of the dataset (Zimmerman, Fox, & Kruschwitz, 2019). The information is given as a Table by having various values for 50 features that have been selected. Adaboost starts with equivalent weight to every perception and trains a weak algorithm by utilizing the weight information. By playing out this, the weak algorithm is delivered. Contingent upon exhibition of the weak classifier, it picks a coefficient 'α', which is misclassified. It focuses on improving weights and lessening weights effectively. A weak learning algorithm is used to generate a weak classifier using newly weighted data. Reiterating the process results in the development of an AdaBoost Classifier..

3.5.3. Random forest classifier

Random Forest is used for classification tasks having similar functions as the decision tree. Bootstrap amassing strategy is utilized for training this ensemble classifier. Averaging forecasts make the expectation of all single regression trees. For classification trees, the more significant part vote is taken. Random Forest utilizes an altered tree knowledge which chooses and split every learning procedure by irregular features subset. The data is presented in the form of a table, with varying values for the 50 attributes that were chosen. This algorithm makes a forest by utilizing a lot of decision trees from a subset of information that is arbitrarily chosen and summarises the decisions in favour of the choice tree to choose the last class of the article.

3.5.4. Stochastic gradient boosting

The Stochastic Gradient Boosting permits trees, which are eagerly made from the training dataset. The data is presented in the form of a table, with varying values for the 50 attributes that were chosen. It is utilized for decreasing the connection among the trees in inclination boosting. Every cycle, a subsample of preparation information is drawn

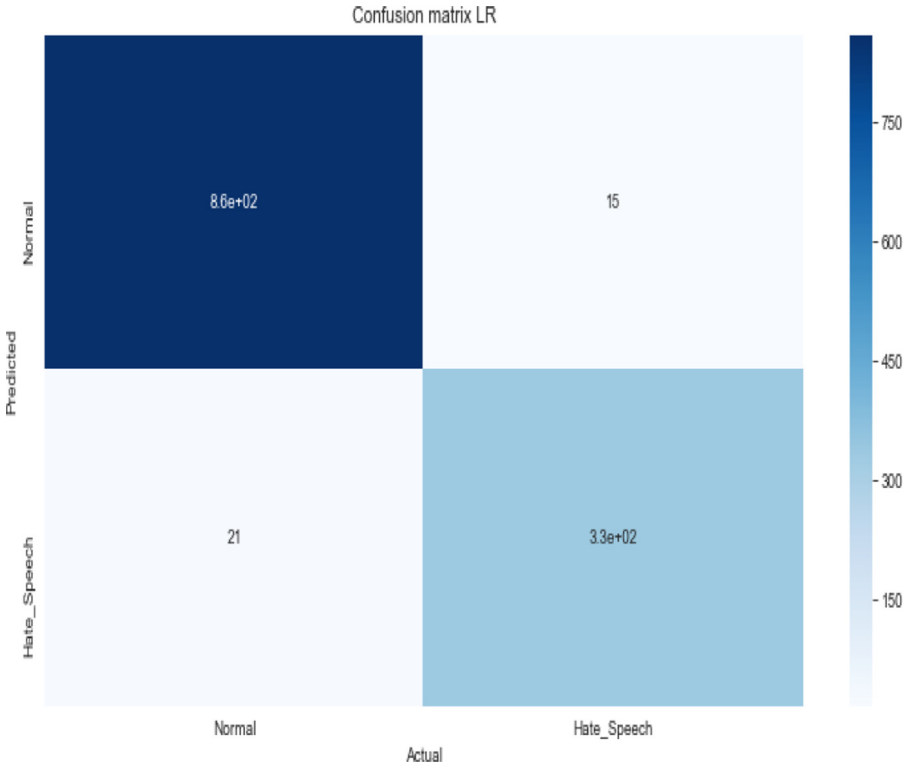


Fig. 5. Logistic regression.

Table 2
Classification report of proposed methodology with ML classifiers.

Classifiers.	Precision.	Recall.	F1-Score.	Accuracy.
Logistic Regression	97%	96%	96%	97.06%
Multinomial Naïve Bayesian	97%	96%	97%	97.39%
Support Vector Machine	98%	96%	97%	97.71%
Decision Tree	98%	97%	97%	97.96%

aimlessly lacking substitutions from the complete preparing dataset. The haphazardly chosen subsample is utilized rather than the full example to adequately the base learner.

4. Results

The experimentation is performed on a workstation having 4 GB Ram and 6 parallel 2.3GHz processors. Machine Learning is being performed using SCIKIT Learn toolkit. Other Libraries like Natural Language tool kits (NLTK) are also used for performing tasks like Tokenisation, Lemmatization, StopWord removal etc. In the wake of performing arithmetical figuring, further knowledge about the information was accomplished. We have used a 70:30 ratio for performing this task, where 70% of tweets are taken for training the ML models, and 30% are used for testing. We extracted 30K tweets, out of which 11K were essential. After annotation, 4,093 tweets were considered to adjust the dataset. They were marked into two classes, Hate and Normal. Hybrid features are selected by merging standard feature engineering techniques (TF/IDF,Bag of Words and Length).The classification was performed with the help of various Machine and Ensemble Learning algorithms by providing them features. Fivefold cross-validation was done as we don't have some other information by which the model can be validated. Table 2 shows the classification report of all machine learning classifiers.

The results showed that the Decision Tree Classifier outperformed all other traditional machine learning algorithms. The Precision of 98%, Recall of 97% and Accuracy of 97.96% is achieved. Figs. 5–8 show traditional machine learning algorithms' actual and predicted tweets by visualizing them by Confusion Matrix. Table 3: shows the classification report of all Machine and Ensemble learning classifiers. The results showed that the Stochastic Gradient Boosting classifier outperforms all other algorithms.

Table 3
Classification report of machine and ensemble learning techniques using proposed methodology.

Technique.	Precision.	Recall.	F1-Score.	Accuracy.
Logistic Regression.	97%	96%	96%	97.06%
Multinomial Naïve Bayesian.	97%	96%	97%	97.39%
Support Vector Machine.	98%	96%	97%	97.71%
Decision Tree.	98%	97%	97%	97.96%
Bagging.	98%	97%	97%	97.96%
Adaboost.	98%	97%	97%	97.96%
Random Forest.	99%	97%	97%	97.96%
Stochastic Gradient Boosting.	99%	97%	98%	98.04%

Fig. 9–12 shows the confusion matrices of the corresponding ensemble learning techniques. The results showed that the Decision Tree gives better results, 98% precision, 97% recall, 97% F1 score and 97.9% overall accuracy, indicating the algorithm outclasses all other traditional algorithms. Stochastic Gradient Boosting classifier shows the highest performance among all ensemble and machine learning classifiers. It gives 99% precision, 97% recall, 98% F1 Score and 98.04% of overall accuracy. Other ensemble learning classifiers like Random Forest, Boosting and Adaboost also show promising results. The accuracy of these models can be improved by supplying more data. Fig. 13 depicts a comparative study of all algorithms employed in our research.

5. Discussion

Hate Speech detection on social media is a pressing issue, and in this paper, we used Machine Learning Algorithms to detect hate speech in COVID-19 era. As the pandemic rose, Online Social Networks saw a drastic change in the behaviour, as users shared information regarding COVID-19 at an enormous pace. Hatemongers find the Pandemic

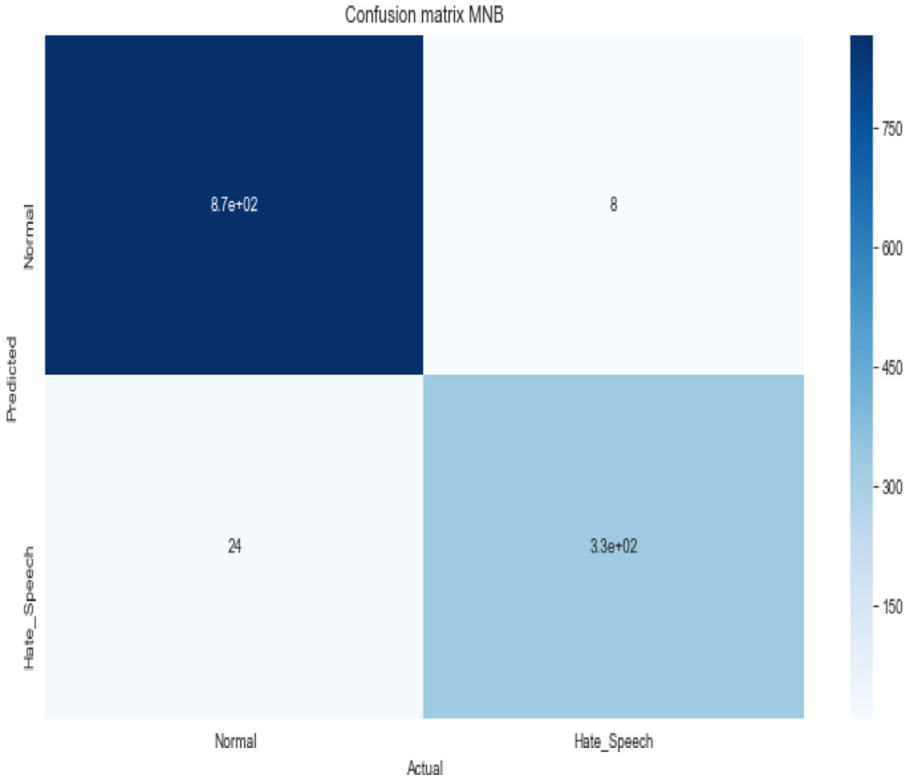


Fig. 6. Multinomial naïve bayes.

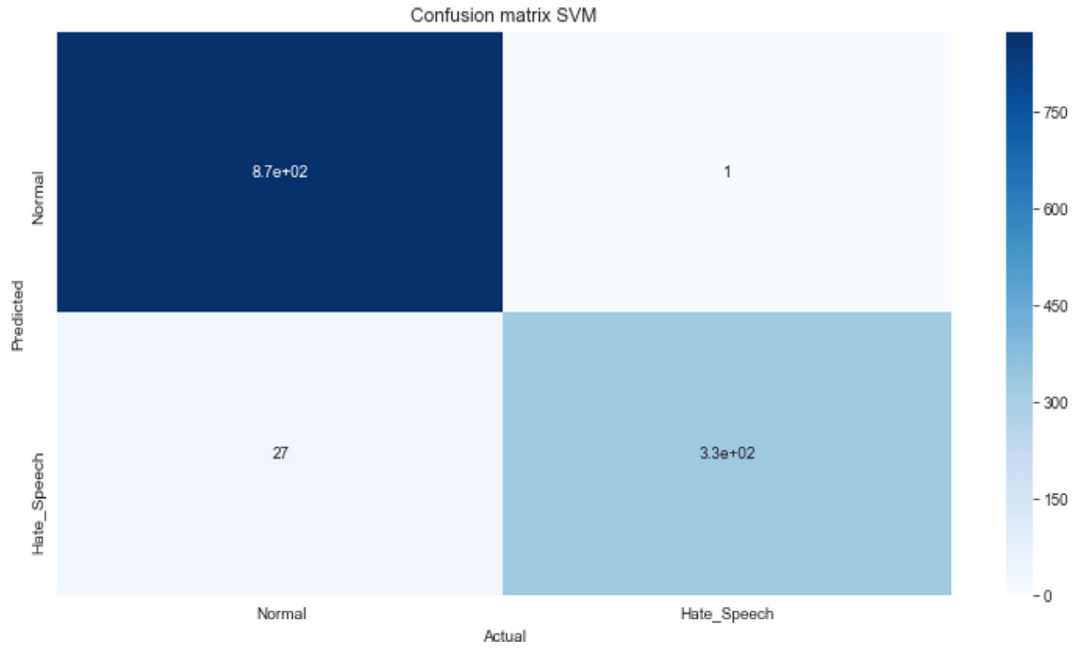


Fig. 7. Support vector machine.

to share hate and panic, triggering mass hysteria. The Twitter API is used to extract data from Twitter using various hate-related terms in this project. For supervised machine learning dataset needs to be labelled, manual annotation is being performed to label the tweets into Hate and Non-Hate Class. Due to tweets' semantic and contextual nature, manual annotation is being preffered by many researchers. Various techniques like Tokenization, Stemming, Normalisation, etc., are used for performing data preprocessing.

Since hate is spread in the form of text, feature selection is one of the important step for detecting hate. Features are selected by techniques, TF/IDF and Bag of Words. After performing data exploration, it was found that the tweet's Length play a vital role in spreading the hate. Due to this critical role, Length was also considered as a feature. After Selecting features supervised Machine Learning classifiers are trained and tested in the ratio of 70:30. When used with our methodology, it was found that the Decision tree showed better accuracy among other algorithms. This work will somehow help government officials tackle hate speech by analyzing the tweets regarding hate speech.

When compared with existing work, dataset HatebaseTwitter (Davidson, Warmsley, Macy, & Weber, 2017) was used with proposed

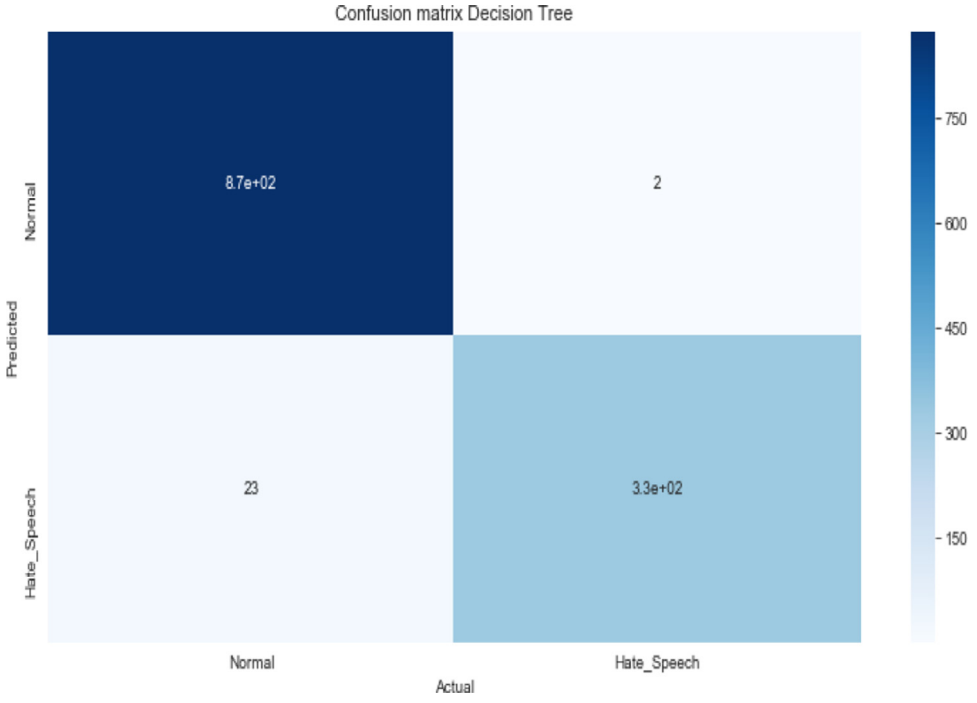


Fig. 8. Decision tree.

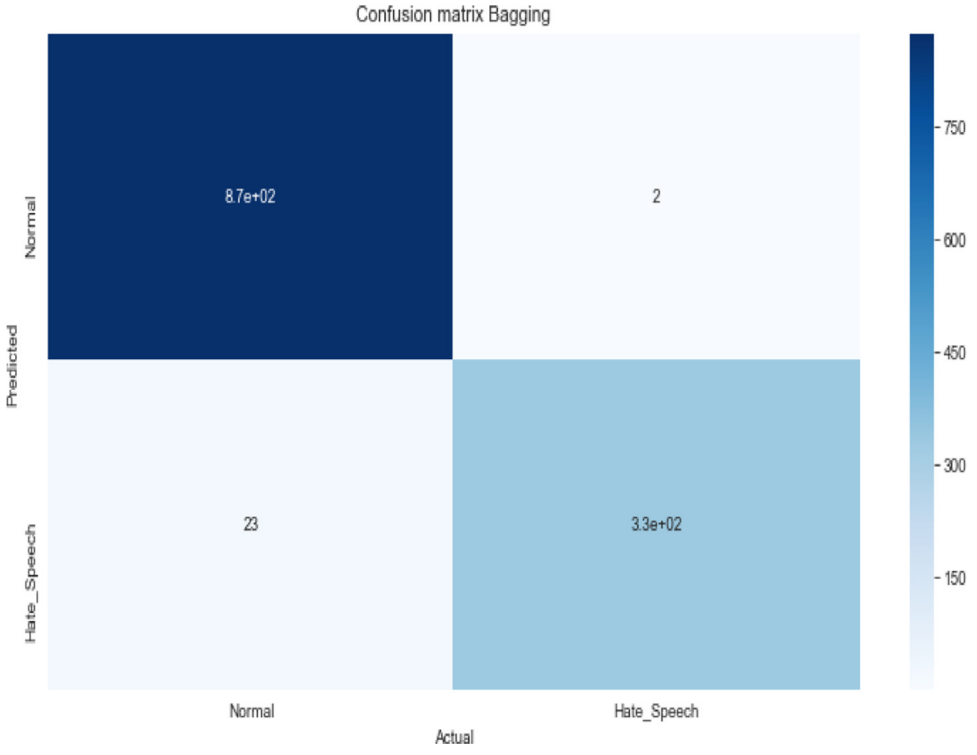


Fig. 9. Bagging.

methodology and the results showed that, Decision Tree and Stochastic Gradient Boosting showed better accuracy then all other algorithms. Table 4 shows the comparative analysis of our best performing algorithms with previous work.

5.1. Contribution to literature

Following the completion of a series of experiments, it was determined that our method performed far better than other methods that had been utilised in earlier investigations concerning hate speech. When

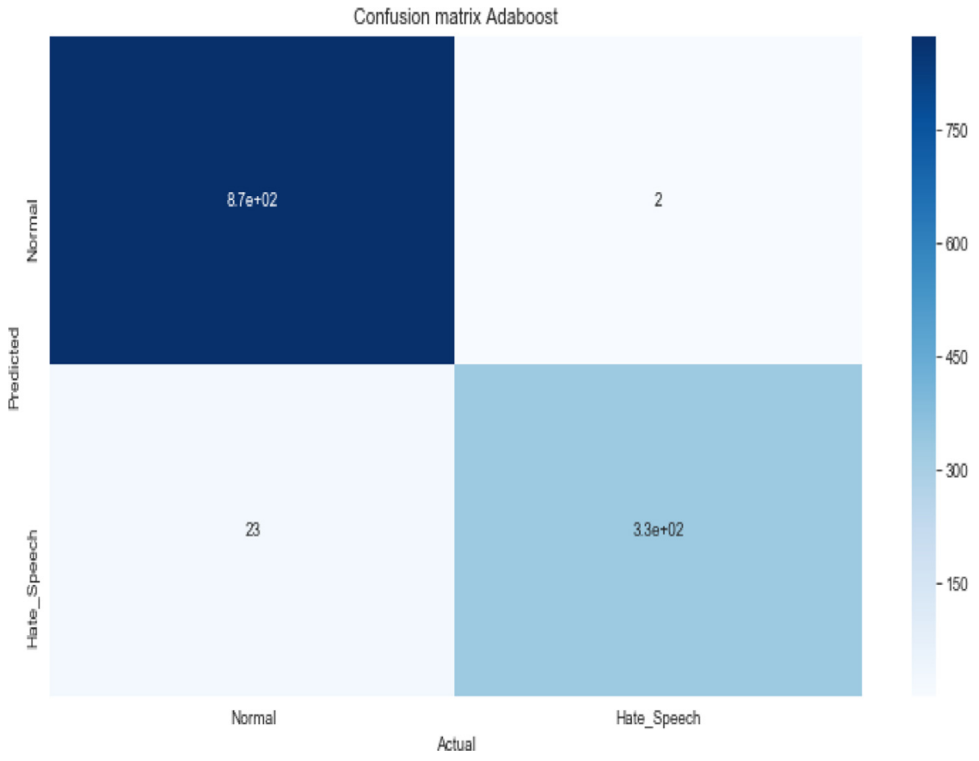


Fig. 10. Adaboost.

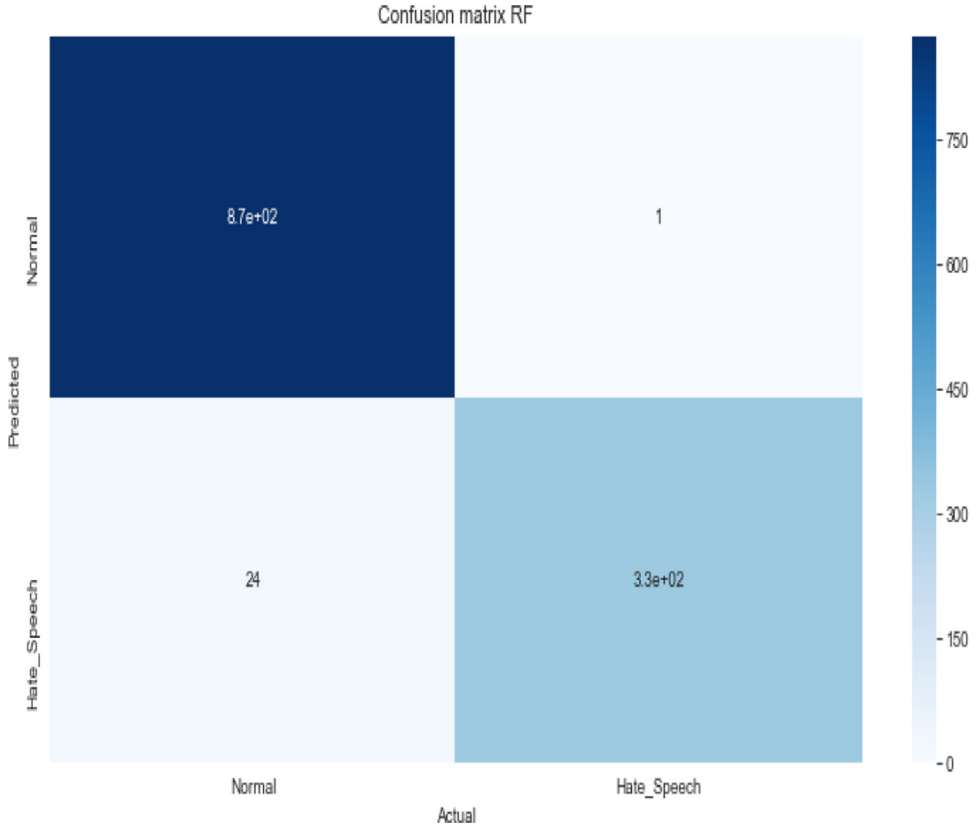


Fig. 11. Random forest.

training the model, the authors Zimmerman et al. (2019) solely employed TF/IDF features, whereas other researchers selected TF/IDF features as well as PoS features. In this study, the hybrid features Bag of Words, TF/IDF, and Tweet Length have been chosen for consideration. During the course of this research, novel data was generated, and tweets were collected without regard to their spatial context.

5.2. Practical implications

There are many different practical implications of this work. One of these implications is that the ability to detect hate speech in real time will help us combat hate speech on social networks. Hateful people take advantage of the fact that social media platforms can be used as

Table 4
Comparative analysis with existing work.

Author/Technique.	*Dataset.*	*Features used*	*Accuracy.*
Davidson et al. (2017)	HatebaseTwitter (Davidson et al., 2017)	TF/IDF and PoS	90.07%
Zimmerman et al. (2019)	HatebaseTwitter (Davidson et al., 2017)	10 CNN with Weights.	92.13%
MacAvaney et al. (2019)	HatebaseTwitter (Davidson et al., 2017)	TF/IDF	91.08%
Decision Tree based upon proposed Approach.	HatebaseTwitter (Davidson et al., 2017)	TF/IDF, Bag of Words and Tweet Length.	96.96%
Stochastic Gradient Boosting based upon proposed Approach.	HatebaseTwitter (Davidson et al., 2017)	TF/IDF, Bag of Words and Tweet Length.	97.04%

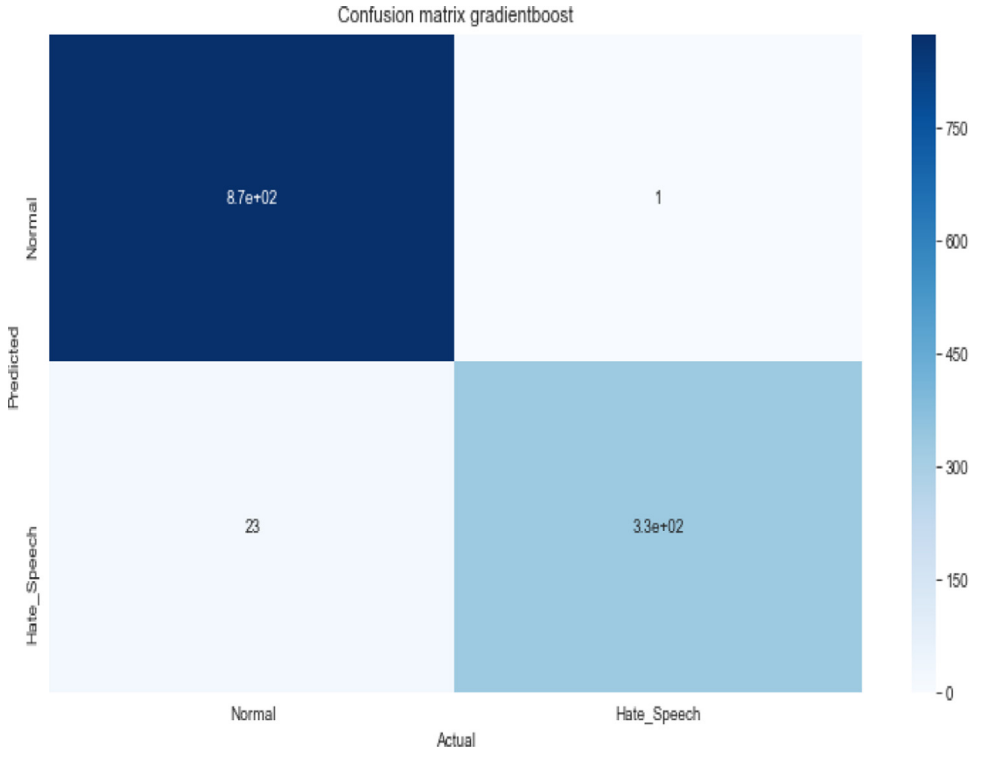


Fig. 12. Stochastic gradient.

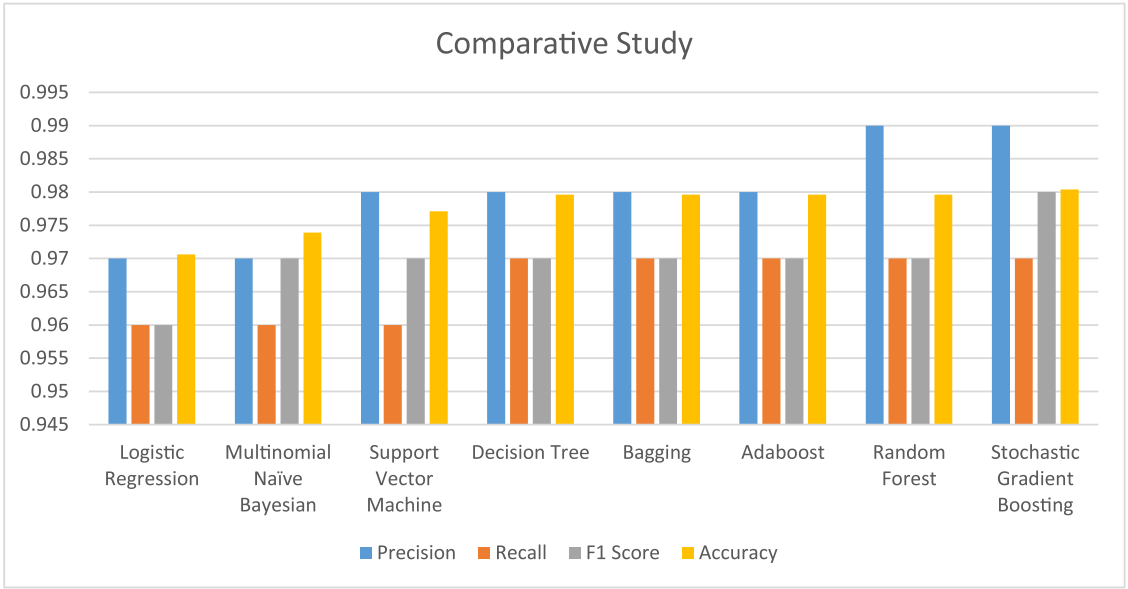


Fig. 13. Comparative study of ML and ensemble learning classifiers.

a medium for communication, and as a result, these platforms are used to spread hatred among users. Checking the credibility of hate speech by manually is a rigorous and time-consuming process. Machine learning can be used to identify those who engage in cybercrime. It was discovered that the length of the tweet containing hate speech about COVID-19, measured in characters, is significantly longer than a typical class tweet. This work has the potential to be expanded to be investigated on other social networking platforms such as Facebook, Linkedin, and Reddit, amongst others. If automatic annotation programmes existed, they would have made a significant contribution to this body of work. The magnitude of the dataset would have allowed for more effective training of machine learning classifiers if automatic annotation had been used. Additionally, the phrases that are used to describe hate speech fluctuate depending on the subject that is being discussed. In the near future, features based on emphasis and semantics may be employed to improve hate speech prediction.

6. Conclusion

The world is paralyzed due to COVID-19 as no Vaccine or medication is available until 26th July and has affected social life. Online social networks are used enormously in this pandemic for communicating with each other, a vast amount of information is being shared through these platforms. Many misinformation and Hate speech is being shared on this deadly virus. Hatemongers use COVID-19 as a platform for spreading hatred. Tweets were extracted using various hashtags like #CoronaJihad, #CoronaTerrorism, etc. and were labelled into Hate class and Normal Class. Hybrid Feature selection is made using TF/IDF and Bag of Words after training and testing all Machine Learning Models. The Decision Tree classifier shows promising results, 98% precision, 97% recall, 97% f1 score and an accuracy of 97.9%. Ensemble Models are also trained and tested for performing binary classification. Among all Ensemble Learning classifiers, Stochastic Gradient Boosting shows the highest performance, 99% precision, 97% recall, 98% F1 Score and Accuracy of 98.04%. Random Forest, Adaboost and Boosting also showed promising results. The effectiveness of classifiers can be improved by expanding measures of information. In the future, hate speech may be categorized based on gender. Long Short Term Memory (LSTM) and Convolutional Neaural Network (CNN) may also be used soon for performing Multi-class Classification. Algorithms 1

Algorithm 1
Classification of Tweets into Hate and Non-Hate Class

Require: Filtered Tweets (T_{input})**Ensure:** Hate Tweet (H_T) and Non-Hate Tweet (NH_T)

Filtered Tweets → T_{input},
Hate Tweet→H_T, *Non-Hate Tweet*→NH_T
Tokenization→*TK, StopWordRemoval*→*SWR, Stemming*→ *S, Total Number of Tweets*→*T*
Term Frequency/Inverse Document Frequency→*TF/IDF, Bag of Words*→*B*
START
For *i from 1 to n do*
C [i] = T_{input} *[i] + Label*
T [i] = Tweetlength(C [i])
End For
For *i from 1 to n do*
Hate [i] = Tk(C [i])
Hate [i] = SWR(Hate [i])
Hate [i] = S(Hate [i])
End For
For *i from 1 to n do*
F [i] = B(TF/IDF(Hate [i]))
F [i] = F [i] + T [i]
End For
Classifier (SVM/DT/LR/MNB)
END

References

Balahur, Alexandra (2013). *Sentiment Analysis in Social Media Texts* (pp. 120–128). Atlanta, Georgia: Association for Computational Linguistics.

A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," pp. 30–38, 2011.

Aswani, R., Kar, A. K., & Ilavarasan, P. V. (2019). Experience: Managing misinformation in social media – Insights for policymakers from twitter analytics. *Journal of Data and Information Quality, 12*(1) Nov..

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on world wide web confrence 2017* (pp. 759–760). WWW 2017 Companion.

Bail, C. A., et al., (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences of the United States of America, 115*(37), 9216–9221.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet, 7*(2), 223–242.

CodaLab - Competition. [Online]. Available: https://competitions.codalab.org/competitions/19935. [Accessed: 22-Dec2021].

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the eleventh international AAAI conference on web and social media, ICWSM* (pp. 512–515). Icwsm.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated hate speech detection and the problem of offensive language*.

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2019). *Hate speech dataset from a white supremacy forum*, 11–20.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. *MATECWeb of Conferences, 125*, 86–95.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the WWW Companion - 24th International Conference on World Wide Web* (pp. 29–30). May.

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: The moderating effect of political interest and diverse media. *Information, Communication & Society, 21*(5), 729–745.

Facebook, Google and Twitter agree German hate speech deal - BBC News. [Online]. Available: https://www.bbc.com/news/world-europe-35105003. [Accessed: 22-Dec-2021].

Facebook's Mark Zuckerberg 'understands need to stamp out hate speech', Germany says | Daily Mail Online. [Online]. Available: https://www.dailymail.co.uk/news/article-3464501/Mark-Zuckerburg-understands-needs-stamp-hate-speech-Facebook-says-German-minister-meeting-discuss-deleting-neo-Nazi-comments-faster.html. [Accessed: 22-Dec- 2021].

B. Gambäck and U.K. Sikdar, "Using convolutional neural networks to classify hate-speech," no. 7491, pp. 85–90, 2017.

Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., & Roy, D. (2018). Me, my echo chamber, and I: Introspection on social media polarization. In *Proceedings of the world wide web conference on world wide web* (pp. 823–831).

Gimpel, K., et al., (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the ACL-HLT -49th annual meeting of the association for computational linguistics: Human language technologies: 2* (pp. 42–47).

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering, 10*(4), 215–230.

Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2019). Polarization and acculturation in US Election 2016 outcomes – Can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change, 145*, 438–460.

Grover, P., Kar, A. K., & Ilavarasan, P. V. (2019). Impact of corporate social responsibility on reputation – Insights from tweets on sustainable development goals by CEOs. *International Journal of Information Management, 48*, 39–52.

Hua, T., et al., (2013). Analyzing civil unrest through social media. *Computer (Long. Beach. Calif)., 46*(12), 80–84.

Jha, Akshita, & Mamidi, Radhika (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science* (pp. 7–16). Vancouver, Canada: Association for Computational Linguistics.

Ji Ho Park, & Pascale, Fung. (2017). *One-step and Two-step Classification for Abusive Language Detection on Twitter* (pp. 41–45). Vancouver, BC, Canada: Association for Computational Linguistics.

Joseph, N., Kar, A. K., & Ilavarasan, P. V. (2021). How do network attributes impact information virality in social networks? *Information Discovery and Delivery, 49*(2), 162–173.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics EACL: 2* (pp. 427–431).

Kar, A. K., & Aswani, R. (2021). How to differentiate propagators of information and misinformation–Insights from social media analytics based on bio-inspired computing. *Journal of Information and Optimization Sciences, 42*(6), 1307–1335 Aug..

Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research – Moving away from the 'What' towards the 'Why. *International Journal of Information Management, 54*, Article 102205.

Khan, F. H., Bashir, S., & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems, 57*(1), 245–257.

Khanday, A. M. U. D., Khan, Q. R., & Rabani, S. T. (2020a). Identifying propaganda from online social networks during COVID-19 using machine learning techniques. *International Journal of Information Technology*.

Khanday, A. M. U. D., Khan, Q. R., & Rabani, S. T. (2020b). Detecting textual propaganda using machine learning techniques. *Baghdad Science Journal*, 199–209 December.

Khanday, A. M. U. D., Khan, Q. R, & Rabani, S. T. (2020c). Analysing and predicting propaganda on social media using machine learning techniques. In *Proceedings of the 2nd international conference on advances in computing, communication control and networking (ICACCCN)* (pp. 122–127).

Khanday, A. M. U. D., Khan, Q. R., & Rabani, S. T. (2021). SVMBPI: support vector machine-based propaganda identification. In Cognitive Informatics and Soft Computing (pp. 445–455). Singapore: Springer.

Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Mohi ud Din, M. (2020d). Machine learning based approaches for detecting COVID-19 using clinical text data. *International Journal of Information Technology*.

Khanday, Akib Mohi ud Din, et al., (2022). NNPCov19: Artificial Neural Network-Based Propaganda Identification on Social Media in COVID-19 Era. *Mobile Information Systems*, 1–10. In this issue https://www.hindawi.com/journals/misy/2022/3412992/. https://doi.org/10.1155/2022/3412992.

Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking aggression identification in social media. *Trac*, (1), 1–11.

Kushwaha, Amit Kumar, Kar, Arpan Kumar, & Ilavarasan, P. (Apr 2020). Predicting Information Diffusion on Twitter a Deep Learning Neural Network Model Using Custom Weighted Word Features. *19th Conference on e-Business, e-Services and e-Society (I3E)*, 456–468 Skukuza, South Africa. hal-03222872. https://doi.org/10.1007/978-3-030-44999-5_38.

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. *Association for the Advancement of Artificial Intelligence*, 1621–1622.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the ACM international conference on information & knowledge management* (pp. 375–384).

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *Plos One, 14*(8), 1–16.

Neubaum, G., & Krämer, N. C. (2017). Opinion climates in social media: Blending mass and interpersonal communication. *Human Communication Research, 43*(4), 464–476 Oct..

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international world wide web conference WWW* (pp. 145–153).

Opinion | Twitter Must Do More to Block ISIS - The New York Times. [Online]. Available: https://www.nytimes.com/2017/01/13/opinion/twitter-must-do-more-to-block-isis.html. [Accessed: 22-Dec- 2021].

Rabani, S. T., Khan, Q. R., & Khanday, A. M. U. D. (2020). Detection of suicidal ideation on Twitter using machine learning & ensemble approaches. *Baghdad Science Journal, 17*(4), 1328–1339.

Scheuer, C., et al., (2011). Twitter sentiment analysis: The good the bad and the OMG!. *Physical Education and Sport for Children and Youth with Special. Needs: Researches – Best Practices – Situation*, 538–541.

Silva, L., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. In *Proceedings of the 10th international conference on web and social media, ICWSM* (pp. 687–690).

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review, 34*(3), 150–160 Aug..

Verma, P., Khanday, A. M. U. D., Rabani, S. T., Mir, M. H., & Jamwal, S. (2019). Twitter sentiment analysis on Indian government project using R. *International Journal of Recent Technology and Engineering, 8*(3), 8338–8341.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. *Association for Computational Linguistics*, 19–26.

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138–142).

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In *Proceedings of the NAACL HLT 2016 the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies student research workshop* (pp. 88–93).

World Economic Forum. (2017). The global risks report 2017 12th edition. *Global Competitiveness Risks Team*, 103.

Wu, C., & Gerber, M. S. (2018). Forecasting civil unrest using social media and protest participation theory. *IEEE Transactions on Computational Social Systems, 5*(1), 82–94.

Zimmerman, S., Fox, C., & Kruschwitz, U. (2019). Improving hate speech detection with deep learning ensembles. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)* (pp. 2546–2553).

Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1-10).