TEXT ANALYSIS ON TWITTER DATA USING LSA AND LDA

T.Karthik Venkat Sai , K.Lohith , M.Padma Sai , K.Tejaswi , P.M.Ashok Kumar , Karthikeyan C
Department of computer science and Engineering
Koneru Lakshmaiah Education Foundation
Vijayawada, Andhra Pradesh ,India
karthikvenkatsai36@gmail.com , klohith2003@gmail.com , padmasai.1216@gmail.com , tejaswi3003@gmail.com ,
pmashokk@kluniversity.in , ckarthik2k@gmail.com

***Abstract*— The news ecosystem has changed in the modern years from outdated print media to social media sites. Because social media platforms enable us to absorb news much more quickly and with less restrictive editing, fake news is disseminated at an astonishing rate and scale. More and increasingly individuals are using social media as the world becomes more digital since it makes connecting with others relatively simple. But false information is misguiding people. Although fake news is simple to propagate, its effects can be disastrous. Bogus news has frequently resulted in uncontrolled circumstances that killed numerous people. People with limited education might quickly become acclimated to bogus news. Instead of checking the accuracy of the information, they accept what is provided to them. This can be overcome by using Text mining, Statics concepts which can detect fake news more precisely when compared to several machine learning methods. This proposed system is on the analysis of fake news and detection of hatred news using t-distributed Stochastic Neighbor Embedding(t-SNE) to check dimensionality reduction and topic modeling using Latent Dirichlet Allocation(LDA) and data pre-processing using Latent Semantic Analysis (LSA).**

***Keywords*—Fake News, Twitter, social media, t-sne, LSA, LDA**

I. INTRODUCTION

Social media is a Platform of technology which uses network to facilitate social interaction among a large group of people. Social media is rapidly growing and becoming an unavoidable part of daily life as a result of the most recent technological revolution. The increased use of smart phones, Androids, and iPhones can be attributed to this astounding growth. These smart phones enable you to access virtually any social media platform from anywhere. These social media sites are very user friendly because their mobile versions are so simple to use. Furthermore, Map services on mobile devices were widely used to find directions and directions. Because of its low cost, easy access, and rapid information dissemination, the Internet has changed the way people interact and communicate. As a result, many people now prefer social media and online portals to traditional newspapers for news searches and reading. Despite being a powerful source of information social media can provide, it can still affect the outcome of major events. The terms fake news and rumors are closely related. Disinformation is intentionally spread to create a false impression. On the other hand, rumors are unsubstantiated and are not designed to deceive. Spreaders' intentions on social media sites are not always clear. Misinformation is then labeled as such by the Internet. It is tough to tell the difference between accurate and incorrect information, however, have been taken to address this issue. In the case of knowledge verification, various Text Analysis methods are used. Text Analysis can answer questions about various products and brands. A rising number of approaches and models need to be created in order to process texts effectively and extract information that is relevant. The method of quantifying the emotional value in a string of words or text, also known as opinion mining, is used to comprehend the attitudes, views, and feelings expressed about a given object. Text classification and analysis is one kind of information extraction that aids in decision-making in many areas of our lives. Social media is one of these fields that is expanding. People today are very active in obtaining and disseminating various information updates via social media on a 24/7 basis. Here Text categorization is also a practice of categorizing written materials into predetermined groups according to their content. The automatic placement of natural language texts into predetermined groups is known as text classification. Text comprehension systems, which change text in some way, such as providing summaries, answering questions, making judgments, or extracting data, rely heavily on text classification. These systems retrieve texts in response to user queries in order to extract some information.

979-8-3503-4821-7/23/$31.00 ©2023 IEEE

Without a question, social media has developed into a reliable resource for data analysis study. Daily, Digital media users contribute a vast quantity of organized, semi-structured, and un raw data in the form of texts, photos, audio files, videos, and other material. These data, produced through the involvement of end users, garner 236 Businesses, academic institutions, government agencies, and even schools have turned to social data analysis. It is a good deal of analytics for large records analysis, however now no longer all of them may be used to investigate records on social media. There is some false news that circulates more and false advertisements that may change the user's opinions and make them confused. There is a dearth of research covering all facets of text analysis and including several implementation techniques and algorithms, which is quite concerning

II. Literature

This phase covers various fake information detection algorithms. Jing [1] has proposed creating model-based hidden representationsthat size adjustments contextual data ina pplicable posts over time. They ran experiments on five million posts collected from Twitter's hate database. Twitter dataset turned into generated by retrieving tweets labeled as blind hatred, insulting, or neither the use of phrases from the Hate base lexicon [2]. Word-degree unit, bi, and tri-grams with TF-IDF weights and component labels, excellent message analyzing goals, sentiment score, and social media community features such as # count, acknowledges, retweets and URLs had been applied for this supervised class task.

Due to the nature of social networking sites, such as Facebook, users can create various text content, such as wall posts, comments, and blogs. The vast amount of information that these platforms collect has become available through the Web. Text mining techniques can be used to analyze the data related to individual interactions. In addition, text mining methods combined with social networks can be used to find widespread consensus on any given topic, human patterns of thought, and collective efficacy in complex systems. According to the study's findings, homophobic and racist tweets have high chance to be classified as hateful speech, while misogynistic Twitter posts have greater chance to be categorized as aggressive. It is based solely on observation because there is no official or agreed-upon definition that explicitly differentiates hateful speech from inappropriate words. There is an acceptance that "hate speech is any affirmation aimed at marginalized communities that possibly incite violence or social dysfunction [3]". The authors of the research [5] trained a CNN model on the dataset from [4] to classify tweets as sexist, racist, or none. They trained classifiers using Word2vec, character 4-grams, random sentence vectors, and combinations. employs another interpretable and cutting-edge multi-view SVM approach to classify hate or no hate across 4 distinct sets of data (Hate base Twitter: [10], Stormfront: [7], TRAC Facebook: [6], Hate Val: [8]). Character level unigrams to 5 grams and Word level unigrams to 5 grams the experiments made use of TF-IDF features. Concerning the detection of hate speech, the majority of classifiers' effectiveness is clearly impacted by the improper and reduced dataset. This leads to its consideration in the analysis, as well as all of the challenges in Section I are taken into account and settled through a set of data-building projects. Notwithstanding all of the data frame issues, there are a few major considerations that are not addressed. As a result, a huge number of erroneous positive aspects are produced. Long-term relationships are popular in the input text and can occur in quasi-words. People could not be managed to capture using comparable features such as n-grams, m-skip-n-grams, and so on. These foundational requirements could be utilized to retrieve substantial addiction tuples. Some other problem is that text mining-based functionalities all generate very high dimensions. Word n-grams, m-skip-n-grams, character n-grams, and dependency tuples bi-grams, for illustration, have exceptionally large measurements. High dimensionality data is changed to low dimensional data by acquiring Dimensionality reduction algorithm also recognized as latent semantic analysis (LSA), by trying to identify phonetic similarity and morphemic. In addition to lowering dimensionality, semantic information categorizes the data, which improves the performance of the classifier. Integrating relevant attributes, on the other hand, diminishes classification issues. Enhanced Named Entity Recognition, Dependency Tuples, and other similar functionalities are illustrations. All of these flaws, therefore, are taken into consideration in our way to solving for enhanced quality.

III. PROPOSED METHODOLOGY

A. Data pre-processing:

One of the essential phases in text analysis and data mining is data preprocessing Data preprocessing is the process of putting raw data into a format that is comprehensible. We cannot work with raw data; it is also a crucial stage in data mining. Before using machine learning or data mining methods, the data's quality should be examined.

- **Lower casing:** Lowercasing, which guarantees correlation inside the function set and addresses the sparsity issue, is the maximum efficient textual content instruction method.
- **Removal of Stop Word:** Stop words are tiny words that are used to structure language grammar but are useless for text mining. Prepositions, conjunctions, articles some pronouns, and common terms like a, the, an, by, from, to, about, and so on have all been filtered out.
- **Tokenization:** Tokenization in preprocessing is the technique of dividing prolonged textual content series into tokens (i.e., smaller pieces).

B. Data Splitting:

Although the training data is used to modify and develop such techniques, the unseen dataset is used in to analyses machine learning and deep learning approaches.

1.1 T-distributed stochastic neighbor embedding (t-SNE):

With the use of a two- or three-dimensional map, the statistical technique known as t-SNE makes it possible to see high-dimensional data. The T-SNE method can be used to enhance the visualization of high-dimensional data by embedding it in a low-dimensional area. This method is ideal for analyzing and visualizing complex data in a two or three-dimensional space.

1.2 Latent semantic analysis (LSA):

LSA is a method for examining the connections various documents and their terms. It can be used in the processing of natural language. This method is commonly utilized in the distributional semantics area.

1.3 Latent Dirichlet Allocation (LDA):

LDA is a method for supervised classification that is used to build effective machine learning models. Image recognition and marketing predictive analysis are two applications of this type of dimensionality reduction. LDA is used to retrieve the corpus. It is mostly used in the method of topic modelling. LDA is used to setup a generative model. It mainly focuses on observations through the groups of unobserved ones.

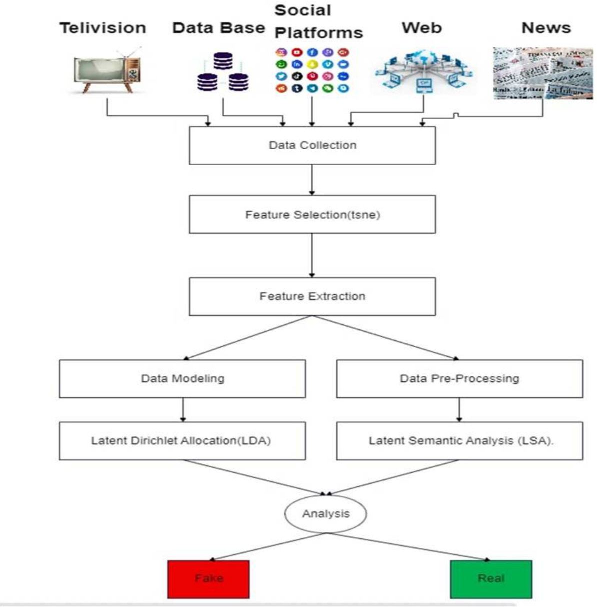


Fig-1 - System Architecture for Fake News Text Analysis

IV. RESULT AND DISCUSSION

A. Data Set 1:

We have gathered the false news tweets and details from Kaggle. The data collection includes Fields for the text, location, and target. Additionally, there is a class label that has the values "0" for false news and "1" for true news tweets. Only the headline and body of the article may be used as models. The remaining 4342 tweets are fraudulent, whereas the first 3272 contain actual news.

B. Data set 2:

The dataset also includes the fields for the ID, keyword, location, and text.
The keywords are of various types, such as "ablaze," "accident," "aftershock," and "ambulance," which indicate the nature of the tweets or the category to which they belong. The tweet's text and location together make up the location of the tweets.

The experiments of the paper had been performed on Jupiter notebook. Ram:25gb Python three and GPU. Four benchmark faux information datasets had been breaking up into 80% of education datasets used to optimize the fashions and signal in cross-validation results, and the 20% attempting out datasets (unseen data) to assess the fashions and signal in the attempting out results.

TABLE 1 TRAINING DATASET

	Id	*target*
count	*7613.000000*	*7613.00000*
mean	*5441.934848*	*0.42966*
std	*3137.116090*	*0.49506*
min	*1.000000*	*0.00000*
25%	*2734.000000*	*0.00000*
50%	*5408.000000*	*0.00000*
75%	*8146.000000*	*1.00000*
max	*10873.000000*	*1.00000*

The Table-1 Shows the statistics of the training dataset and this itself shows the count of cells in the dataset and also the other parameters like mean, Standard deviation, minimum value

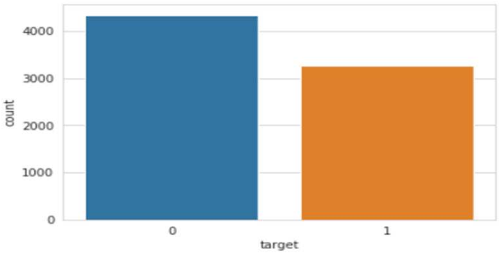


Fig-2-Fake news vs Real news

This Fig 2 lets us to know the variance between fake and try news which indicates the values 1 and 0. There are a total of

979-8-3503-4821-7/23/$31.00 ©2023 IEEE

4342 tweets with target attribute 0 and 3272 tweets with target attribute 1.

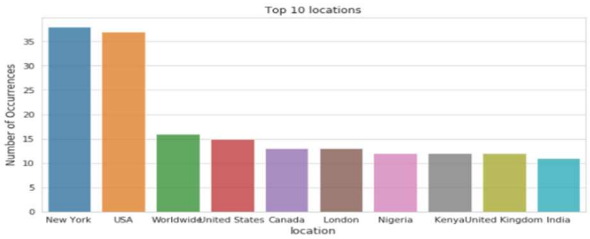


Fig-3-Occurrences of top 10 locations

The above Fig -3 shows the top occurrences of the 10 locations in test dataset with more no of tweets

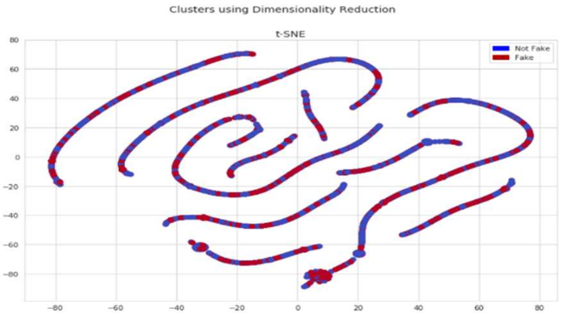


Fig-4-Clustering Dimension

At first in the t-SNE method we get the above results as the Fake and Not Fake news are Represented as the Blue and Red colors in the results. We find the results are too clumsy that are nearer to each other and it is hard to recognize which is fake news and not for the end user. Now let's use the keywords analysis to make clearer in identification of the fake and not fake news.

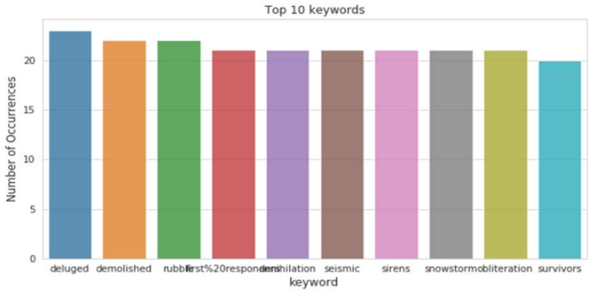


Fig-5-Top-10 keywords in datasets with different occurrences

This Fig 5 shows top 10 keywords in the datasets. Regarding the clumsiness, this can be overcome by intersection of both training and test datasets. This merges duplicate keywords which were found in the dataset. the intersection between the two datasets gives us the same number of unique keywords, so the keywords used in both datasets with different occurrence of count 222.

Now we will perform the t-SNE (t-Distributed Stochastic Neighbor Embedding) once again so that the

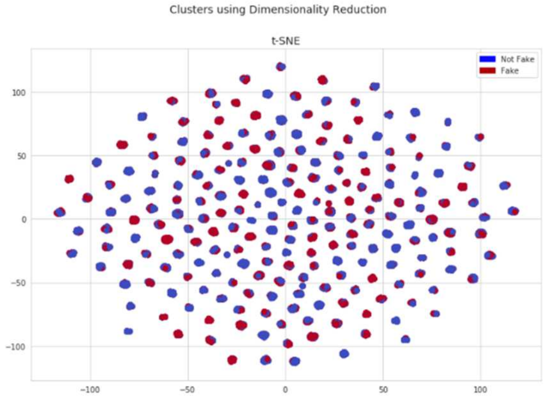


Fig-6-Segregation of fake vs real tweets

Now the Text analysis process begins, In the first step we remove the duplicated tweets, it would be a retweet. If we know that a tweet is fake or not so the other duplicated tweets will get the same class.

We observe that there are 110 tweets which are duplicated in training dataset and 20 duplicated tweets in the test data. The second step of this process is topic modelling. Now here we are going to apply LDA (Linear Discriminant Analysis) to the tweets and split them into topics. In this process remove lower casing, URL's, special symbols, stop word and also perform tokenization and stemming.

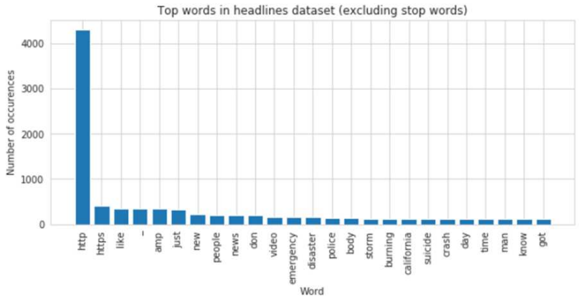


Fig-7-Top words in headlines dataset

Then, to comprehend the different word types utilized across the corpus, we create a histogram of the headline word lengths. To do this, each headline string must first be converted to a Text Blob before using the postages method on each one to get a list of marked keywords for every headline.

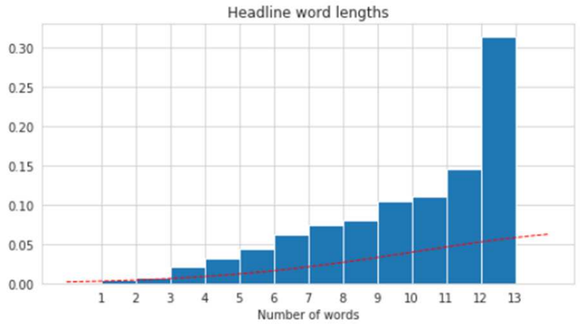
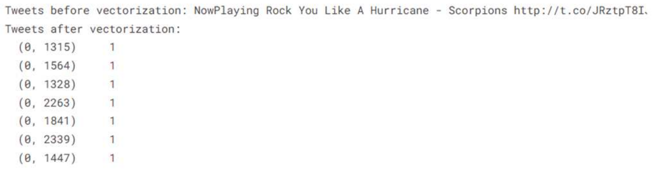


Fig-8-Sorting headline word lengths

Now we will apply a clustering algorithm to the headlines Now we will apply a clustering algorithm to the headlines corpus in order to study the topic, as well as how it has evolved through time. To do so, we first experiment with a small subsample of a dataset in order to detect which of the two potential clustering algorithms is most appropriate – once this has been ascertained, we then scale up to a larger portion of the available data.

The only pre-processing in our situation, a necessary stage is feature building, where we take the sample of text tweets and represent them in some tractable feature space. In practice, this simply means converting each string to a numerical vector. This can be done using the Count Vectorizer object from SK Learn, which yields an n×K document-term matrix where K is the number of distinct with a maximum of max features and less stop words across the n headlines in our sample.



```
Tweets before vectorization: NowPlaying Rock You Like A Hurricane - Scorpions http://t.co/JRztpT8I.
Tweets after vectorization:
  (0, 1315)     1
  (0, 1564)     1
  (0, 1328)     1
  (0, 2263)     1
  (0, 1841)     1
  (0, 2339)     1
  (0, 1447)     1
```

Fig – 9-Vectorization

Now that we have our (very rising and scarce) data for learning, small textual phrase matrix, and cluster technique, we can put it into action. Latent Dirichlet Allocation or Latent Semantic Modeling will be our choice of approach. Both will generate an n x N topic matrix, where N is the number of subject categories, and use our document-term matrix as their input (which we supply as a parameter). We'll assume that this equals five similar categories for the time being.

Let's begin by testing out LSA. This is essentially simply an r= n topic's greatest singular values preserved truncated singular value decomposition of an (extremely high-rank and sparse) document-term matrix. So, using the most common words inside each group, we were able to turn our initial tiny sample of headlines into a list of anticipated topic categories. The use of a bar chart makes it simple to see the relative magnitudes of the various categories.

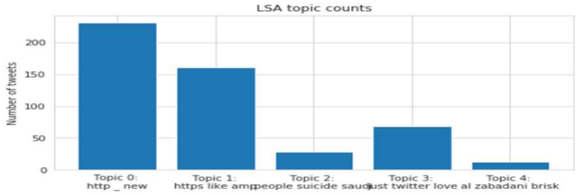


Fig-10-LSA topic Counts

a) Nevertheless , this doesn’t offer a reliable benchmark for comparision with other clustering alogorithms. We employ t-sne, a dimensionality reduction method, to correctly compare LSA with LDA and to futher illuminate the effectiveness of the clustering procedure.

b) We can now trace the groups with bokeh because we have condensed the dimensional vectors of these objects to two-dimensional representations. Before doing so, however, it's helpful to derive the location of each topic's centroid to better contextualize our visualization

c) It only remains to draw the
d) grouped questions.

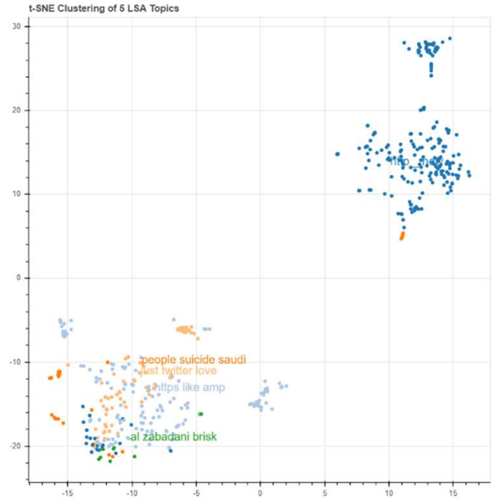


Fig-11Top-3 probabilities of LSA and LDA

The top three words of each group are also included, which will be ranked. We haven't reached our goal yet with the LSA procedure. Now we repeat this procedure with LDA instead of LSA. LDA is instead a generative probabilistic process designed with the specific aim of discovering the structure of latent subject matter in text corpora. However, to compare LDA to LSA, let's take this matrix of topics again and project it in two dimensions.

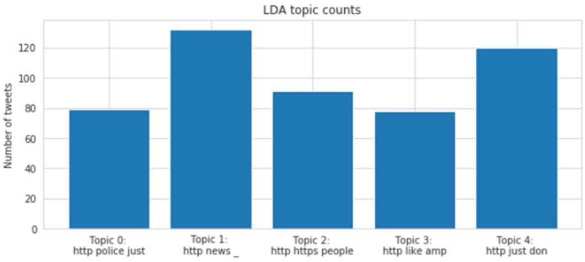


Fig-12
(LDA topic counts)

V. CONCLUSION AND FUTURE WORKS

In this analysis, the largest obstacles are listed among the top. There we have performed the Text analysis Through the Tweet data. At first, we found out top locations where there are maximum tweets located. Next, we performed TSNE (T-distributed Stochastic Neighbor Embedding) to check dimensionality reduction, but the results are too clumsy. Again, we performed the above step by using the top keywords and performed TSNE once again, now the results of Fake and Not Fake news are observed clearly. Then we perform data pre-processing steps stop word removal, Removal of URL's, removing stop word and tokenization process. Then we observe the top words in form of histograms. Then, in order to comprehend the different word types utilized across the corpus, we create a histogram of the headline word lengths. To do this, you must first convert each headline string to a Text Blob and use the post tags method to return a list of each headline's tagged terms. In further we perform LSA and LDA. We get better results in the LDA at the end.

REFERENCES

[1] Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F. and Cha, M., 2016. Detecting rumors from microblogs with recurrent neural networks.

[2] Jokić, D., Stanković, R., Krstev, C. and Šandrih, B., 2021. A Twitter Corpus and lexicon for abusive speech detection in Serbian. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[3] Moran, L.J., 2001. Affairs of the heart: Hate crime and the politics of crime control. *Law and Critique*, *12*(3), pp.331-344.

[4] Waseem, Z. and Hovy, D., 2016, June. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88-93).

[5] Kumar, R., Ojha, A.K., Malmasi, S. and Zampieri, M., 2018, August. Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)* (pp. 1-11).

[6] Gambäck, B. and Sikdar, U.K., 2017, August. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online* (pp. 85-90).

[7] De Gibert, O., Perez, N., García-Pablos, A. and Cuadros, M., 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

[8] Khanday, A.M.U.D., Rabani, S.T., Khan, Q.R. and Malik, S.H., 2022. Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. *International Journal of Information Management Data Insights*, *2*(2), p.100120.

[9] MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N. and Frieder, O., 2019. Hate speech detection: Challenges and solutions. *PloS one*, *14*(8), p.e0221152.

[10] Davidson, T., Warmsley, D., Macy, M. and Weber, I., 2017, May. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media* (Vol. 11, No. 1, pp. 512-515).

[11] Aggarwal, C.C. and Wang, H., 2011. Text mining in social networks. In *Social network data analytics* (pp. 353-378). Springer, Boston, MA.