Machine Learning Based Detection of Deceptive Tweets on Covid-19

Amisha Sinha, Mohnish Raval, S Sindhu

Abstract: Social media plays a vital role in connecting people around the world and developing relationships. Social media has a huge potential audience and the circulation of any information does impact a huge population. With the surge of Covid-19, we can see a lot of fake news and tweets circulating about remedies, medicines, and general information related to pandemics. In this paper, we set out to machine learning-based detection of deceptive information around Covid-19. With this paper, we have described our project which could detect whether a tweet is fake or real automatically. The labeled dataset is used in the process which is extracted from the arXiv repository. Data extraction methods are applied for cleaning, training, and testing. Pre-processing, classification, tokenization, and stemming/removal of stop words are performed to extract the most relevant information from the dataset. To achieve better accuracy in comparison with the existing system, we have used two classification techniques: TF-IDF and Bags of words. To achieve better accuracy, we have used two other methodology-SVM and Random Forest and have achieved an F1-score of 0.94 using SVM.

Keywords: Artificial Intelligence, Fake News, Social Media, SVM

I. INTRODUCTION

With the commencement of the pandemic, social media has become a vital source for the spread of information. The large and epiphanic flow of data, the.certitude of the latter is indiscernible. With the pandemic extent, it is the need for an hour to stratify the large data into bumph and gospel. Our project uses various machine learning techniques to analyze the tweets into “Real” and “Fake”. The focus of the project is to help in analyzing the data and for the instances of spreading false information. The model created can be used in a real-time for tracking the source of false information. This will help the government of India to take appropriate actions against the culprit and draconianically deplete wrong tweets circulating on the internet.

Manuscript received on June 14, 2021.
Revised Manuscript received on June 19, 2021.
Manuscript published on June 30, 2021.

* Correspondence Author
Amisha Sinha*, Department of Information Technology, SRM Institute of Science and Technology, Chennai (Tamil Nadu), India. Email: amishasinha84@gmail.com

Mohnish Raval, Department of Information Technology, SRM Institute of Science and Technology, Chennai (Tamil Nadu), India. Email: mohnishraval40@gmail.com

S Sindhu, Assistant Professor, Department of Information Technology, SRM Institute of Science and Technology, Chennai (Tamil Nadu), India. Email: sindhus2@srmist.edu.in

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

The dataset for identifying fake tweets for Covid-19 was extracted from arXiv which is an open repository for data and scientific papers related to diverse fields and can be accessed online.

II. LITERATURE REVIEW

The previous researcher’s machine learning techniques and approaches in fake news detection have been quite successful. Sourya Dipta Das at IIT Madras [1] has used a Heuristic Ensemble framework for detecting fake news about Covid-19. They conducted a prolific study, earmarking assorted levels of prorogativo every level of feature obtaining an F-Score of 0.98. [2] Mohamad K. Elhadad, Kin Fun Li and Fayeze Gebali, constructed a voting ensemble machine learning classifier for detecting inaccurate information for Covid-19 using ten machine learning algorithms and seven feature extraction methods in their survey. They have used TF-idf and bow classifier techniques which we have also implemented in our project. [3] Cody Buntain and Jennifer Golbeck at the University of Maryland research work on automatic detection of fake news in popular news on Twitter include the usage of three datasets to identify fake tweets around CREDBANK. [4] IRJET paper on “Fake News Detection Using ML” uses TF-idf, Naive Bayes Support Vector Classifier (SVC) for their fake news detection model. Various researchers have used their own technique and algorithm, for fake news detection using Artificial Intelligence. Fake news around Covid-19 is mostly negative (63.2%) can lead to threats to public health as stated in Article by Md. Sayeed Al- Zaman on “COVID-19-Related Social Media Fake News in India”[5].

In summary, there are many existing works around fake news detection and for detection, the majority of them rely on manually labelled data. Some existing works are around the detection of fake news around the present Covid-19 situation. As the second wave of Covid-19 has hit India, and we are daily encountering a lot of tweets around this with home remedies and other information with respect to oxygen availability and medicines and there is a lot of fake and misleading news which could be a threat to public and the identification of disinformation instead of the hour.

III. DATASET DESCRIPTION

The dataset for identifying fake tweets for Covid-19 was extracted from arXiv which is an open repository for data and scientific papers related to diverse fields and can be accessed online.
This dataset consists of labeled data in two columns with one column having the tweets and the other column marked as Label with a tag of either ‘0’ or ‘1’. Here, ‘0’ implies a real tweet and ‘1’ implies a potential fake tweet. The dataset consists of 2140 tweets with 1120 real and 1020 fake tweets. It has a vocabulary size of 63032 with a number of URLs and unique usernames.

IV. METHODOLOGY

The Methodology used in our project is based on basic natural language processing. The methodology consists of several components like:

- Collection of labeled data
- Processing of dataset
- Applying classification techniques such as tf-IDF and bow
- Visualization of the dataset
- Tokenizing the tweets in the dataset
- Finding the stems in the tokenized words
- Removal of stopwords

A. Pre-processing

In this work, we mainly focus on the polarity of the tweet and the tweet itself. So, we have pre-processed the tweets by removing unwanted details, stemming, tokenization, and removal of stop words. This involves the removal of mentions, retweets, URLs, and non-alphanumeric characters. The main purpose of data cleaning is to refine the quality of the dataset which is critical for modeling processes. The quality of the dataset determines how the model will perform in a specific circumstance. Mentions consist of the tagged tweets with the person’s username and retweets are a kind of reply to a tweet that plays no role in determining the nature of the tweet. Similarly, the removal of URLs is crucial as they are links to other websites and are types of unstructured data with no meaning. Non-alphanumeric characters provide no meaning to the context and can be removed. These steps of data cleaning ensure structured data which is a salient step before proceeding with the dataset.

B. Tf-idf Technique

TF-IDF is a method used to fetch information supporting the frequency of a term (TF) and its inverse document frequency. It acknowledges a keyword in any given content and based on the number of times it appears in a document, it earmarks its importance to it. Taking a hypothetical situation like we have a group of words that says “Crocin is the answer to COVID-19” and we are keeping that as the reference want to rank the order of priority of other terms in the document, can be started simply by eliminating the sentences without this keyword and often, the maximum frequency definition of the word in the document is used to measure the importance, and the count of the word in a given document is referred to as term frequency. The bag of words model creates a frequency matrix of the words in the document.

C. Bag of words

Bag of words is used for text which converts a particular text into a bag of words i.e., it keeps track of which words appear in a document. The bag of words model creates a frequency matrix of the words in the given document. It converts the text to a vector of fixed size. Major steps involved here are Vocabulary determination (1) where all the words in the document are identified, Counting (2) where all the words along with their frequency are determined in vector form. The bag of words is useful and widely used due to its simplicity and also the fact that its computation is inexpensive.
“Fig. 3.” shows the snapshot of the code, where we have achieved an F-Score of 0.80 using this technique.

D. Tokenization

Tokenization involves splitting sequences of words (Sentences) into keywords, symbols, and other components referred to as tokens. These tokens can be used as input for the processes. Tokens are mostly separated by whitespace or any line breaks. Only words can be included as tokens (any character with a continuous stream of words). To implement this in any dataset, a Natural language toolkit is used which has both word and sentence tokenization functions. Word_tokenizer() is used to split a sentence into tokens. Python offers various tokenization functions which include (1) Sentence Tokenization which is imported from nltk library. (2) PunktSentenceTokenizer which is used when there are large chunks of data and tokenization is not easy using regular sentence tokenizer. (3) Tokenizes sentences of different languages. (4) Word_tokenizer() (5) TreebankWordTokenizer which separates the sentence using punctuation and white spaces. (6) PunktWordTokenizer (7) Using Regular expression which uses regexp_tokenizer() which uses regular expressions to tokenize based on the expression. Tokenization leads to the creation of tokens from a series of words which is useful in finding patterns that are used in further steps of stemming and lemmatization.

E. Stemming and Removal of Stop Words

Stemming reduces morphological variants of a word. For instance, it reduces all the words like [ likes, liked, likely] to a root word ‘like’. It is beneficial in reducing akin words. Some algorithms to implement stemming are: Potter’s stemmer, Lovins, Dawson, Krovetz, Xerox, and N-gram. The stemming method is useful to reduce the number of tokens and remove stop words from a domain vocabulary library. It is also very useful to fetch results in an index of a search engine like Google.

```
if gram > 1:
    w = []
    for i in range(len(words) - gram + 1):
        w += [\':'.join([words[i:i + gram]])]
    return w

def stop_words:
    sv = stopwords.words('english')
    words = [word for word in words if word not in sv]
    stemmer = PorterStemmer()
    words = [stemmer.stem(word) for word in words]
    return words
```

Fig. 4. Snapshot - Stemming & Removal of Stop Words

“Fig. 4.” shows the snapshot of our code where we have used Porter Stemmer for stemming.

StopWords syntactically and grammatically supports a document but do not have any significance with the content of the document. Such stop words are removed before indexing. In the index itself, NLTK is used in python to implement it. Here NLTK supports a corpus module that contains a list of stop words, which needs to be imported, then we define all the English stop words to a variable, then for those words in the sentence and if it does contain such stop words, it’s removed from the data frame formed.

Stop Words are basically just common words that are not s significant in the indexing and can be ignored. These stop words can be usual words like “the”, “and”, “an”, “a”, etc. “Fig. 4.” depicts how we have removed stop words from our dataset.

V. Module Description

We have used the Support Vector Machine (SVM) algorithm in our Random Forestinour project.

A. Support Vector Machine (SVM)

SVM is an algorithm that has been facile to embrace but extreme efficient and versatile in implementation. It can be used for classification as well as regression. The framework of this algorithm is illustrated in “Fig. 5”. The space is depicted as a multi-dimensional space of various groups. In “Fig. 5”, the hyperplane must be chosen such that the distance between the support vectors is maximized, referred to as the maximum margin, where support vectors are the points closest to the hyperplane and points of two classes are shown in green and blue.

```
X2
Maximum Margin
Positive Hyperplane
Support Vectors
X1
Negative Hyperplane
```

Fig. 5. Depiction of SVM Model with Maximum Margin between two classes

Using the Hinge loss function, we strive to enhance the margin in the SVM algorithm as shown in “Fig. 6”.

```
SVM Hypothesis:

h_0(x) = \begin{cases} 
1 & \text{if } \theta^T x > = 0 \\
0 & \text{otherwise}
\end{cases}
```

Fig. 6. Hinge Loss Function

On the condition that the value obtained out of prediction and that value of calculation comes on difference at all, then the loss is calculated as zero, else we calculate the loss using the hinge loss function. In order to balance the maximization of margin and loss, we add cost function, as in “Fig. 7”
Machine Learning Based Detection of Deceptive Tweets on Covid-19

For Non-Linear SVM algorithm, the function used is shown in “Fig.8”

For Non-Linear SVM algorithm, the function used is shown in “Fig.8”

Hypothesis: \( h_\theta(x) = \begin{cases} 1 & \text{if } \theta^T f \geq 0 \\ 0 & \text{otherwise} \end{cases} \)

Cost Function: \( J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \phi(\theta^T f^{(i)}) + (1 - \phi(\theta^T f^{(i)})) \)

The only difference between the function used for linear and non-linear SVM algorithms is that “f” is used instead of ‘x’, where f is a function of x. This is called the Kernel function. There are various Kernel functions among which “Gaussian function” is the most popular one as it can be used on the condition where there is no prior understanding of the data.

In our project we have used SVM algorithm, a snapshot of the code and the accuracy is shown in “Fig.9”.

In a nutshell, it comes to the conclusion that the SVM algorithm works well on the clear margin of separation between classes and is effective where the dimensional frequency is greater than that of samples.

B. Random Forest

This supervised algorithm due to its simplicity and its dual purpose to serve classification and regression because of which it has been the most prevailing technique.

In our project we have used Random Forest algorithm, snapshot of the code and the accuracy is shown in “Fig.11”
VI. SYSTEM ARCHITECTURE

![Flowchart - System Architecture](Image)

**Fig.12. Flowchart - System Architecture**

VII. PERFORMANCE OF FINAL METHOD

In our project, we have used two classifiers - tf-idf and bag of words and two algorithms - SVM (Support Vector Machine) and Random Forest.

Table 1: Shows the summary of the accuracies obtained from all the algorithms used:

<table>
<thead>
<tr>
<th>S.No</th>
<th>Algorithms Implemented</th>
<th>Accuracy (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Support Vector Machine (SVM)</td>
<td>94%</td>
</tr>
<tr>
<td>2</td>
<td>Random Forest</td>
<td>92%</td>
</tr>
<tr>
<td>3</td>
<td>Bayes Theorem (TF-IDF)</td>
<td>88.88%</td>
</tr>
<tr>
<td>4</td>
<td>Bayes Theorem (BOW)</td>
<td>77.77%</td>
</tr>
</tbody>
</table>

In our model, we achieved an accuracy of 94 percent using the SVM method from the various classification methods as shown in table 1. Perforce, we conclude that the SVM approach has been found to be the most rational. We bring forward this substructure to assist in the identification of misleading information on any probable public health like the 3rd and 4th waves of coronavirus, which are anticipated analogous to writing this survey. This same framework can be used even in the detection of any misleading news, other than global health as well. For our future work, we would like to look at how other pre-trained models work with the same dataset and also our framework performance with other datasets. It’ll be riveting to see how our system executes against supplementary assimilated Fake News dataset.

VIII. CONCLUSION AND FUTURE WORK

In our model, we achieved an accuracy of 94 percent using the SVM method from the various classification methods as shown in table 1. Perforce, we conclude that the SVM approach has been found to be the most rational. We bring forward this substructure to assist in the identification of misleading info on any probable public health like the 3rd and 4th waves of coronavirus, which are anticipated analogous to writing this survey. This same framework can be used even in the detection of any misleading news, other than global health as well. For our future work, we would like to look at how other pre-trained models work with the same dataset and also our framework performance with other datasets. It’ll be riveting to see how our system executes against supplementary assimilated Fake News dataset.

REFERENCES

1. Ayan Basak, Saikat Dutta and Sourya Dipta Das (2020) - “Heuristic-driven ensemble framework for covid-19 fake news detection” (rozathinking, USA): IIT madras, India
2. Fayeze Gebali, (Life Senior Member, IEEE) , Kinfu Li, (Senior Member, IEEE) and Mohamed K. Elhadi (2020) - “Detecting Misleading Information on COVID-19” IEEE-Department of Electrical and Computer Engineering, University of Victoria, Victoria, Canada
3. Cody Buntain and Jennifer Golbeck - “Automatically Identifying Fake News in Popular Twitter Threads”: University of Maryland
5. Md. Sayeed Al-Zaman - Article on “COVID-19-Related Social Media Fake News in India”: Department of Journalism and Media Studies, Jahangirnagar University, Savar, Dhaka, 1342, Bangladesh
6. Anjali Jain, harsh khater and avinashshakya -(2019) - “A smart system for fake news detection using machine learning”: Dr. Ajabdul kalam university, lunknow, india
AUTHORS PROFILE

Amisha Sinha, B. Tech with specialization in Information Technology from SRM Institute of Science & Technology, Kattankulathur, 2020 graduate. Software Engineer. Core Interest: Machine Learning, Data Science, Algorithm development


S Sindhu, Assistant Professor (OG) at SRM Institute of Science & Technology, Kattankulathur. M.E with specialization in Computer Science from Anna University, 2011. Research Interest: Data Mining, Big data.