INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 2)
Available online at: www.ijariit.com

Fake news analyzer

Ruchika Mahajan
ruchika.mahajan@cumminscollege.edu.in
Cummins College of Engineering for Women, Nagpur, Maharashtra

Sanskruti Godbole
sanskruti.godbole@cumminscollege.edu.in
Cummins College of Engineering for Women, Nagpur, Maharashtra

Vedanti Manjule
vedanti.manjule@cumminscollege.edu.in
Cummins College of Engineering for Women, Nagpur, Maharashtra

Prasanna Lohe
prasanna.lohe@cumminscollege.edu.in
Cummins College of Engineering for Women, Nagpur, Maharashtra

ABSTRACT

Social media is a platform which acts as a chain for spreading of diversified news. Its cheap cost, fast access, and quick circulation and broadcasting of information help people to look out and consume news from social media. In today's era, online social media plays a crucial role during real world practical events, especially climacteric events or events that gain a huge social attention of the population of the world. Apart from legitimate news information, malicious content, mean-spirited, vindictive and misleading information is also put online during these events, which can result in harm, chaos and monetary loss in practical world. In our paper, we draw attention to the role of Twitter in analysing a major ongoing event of India: NRC and Citizenship Amendment Act in spreading fake content about these events. We elaborate the gathering, elucidation, and validation process in detail and perform many exploratory analysis on the recognition of linguistic differences in fake and real or legitimate news content. Also, we aim to find out the source or origin of the fake news and the sources who help spreading the fake content. Secondly, we aim to perform a set of learning experiments to find out accurate fake news and their sources. Besides, we provide relative analysis of the automatic and manual identification of fake news with the legitimate informants of the news content. This paper proposes a system that identifies unreliable and false news after analysing and computing the set of data. This system aims to use various NLP algorithms and classification algorithms or techniques to help achieve maximum accuracy in finding the fake news.

Keywords— *Twitter API, Random Forest, Natural Language Processing Algorithms, Naive Bayes, Bagging*

1. INTRODUCTION

Fake news has been around for years and isn't a concept that we are unaware of. However, the start of the social media can be well judged by the start of the 20th century has provoked the generation and the circulation of false news to a huge extent. Fake news can be fundamentally explained as a chunk of information which is normally written for economic, personal or political benefit.

Fake news in India refers to misinformation in the country which propagate through word of mouth and conventional media and more freshly through digital ways of communication such as manipulated videos, memes, untested advertisements and social media disseminated rumours. Fake news spreading through social media in our country has become a censorious problem, with the passiveness of it resulting in riot as in mob violence, as was the case where at least 20 people were killed in 2018 as a result of circulation of misinformation on social media. Detection of such bogus news articles is possible and can be done by using different NLP techniques, Machine Learning, and Artificial Intelligence. In 2016, the Mr. Narendra Modi, Prime Minister of India announced that large amount of the cash that people have had become of no use and in the duration of one month all this old currency had to be retained in the banks. This led to a chain reaction of a series of fake content being published used mainly for creating un-stability among the citizens of the India, resulting in political and socio-economic gains. This is just a small example of how the spread of false news can impact a greater audience than it may seem. This paper gives an insight into the formulation of detecting fake news, it's implementation and its conclusion. In order to reach a conclusion on the authenticity of the news article we first take the text article that is in the form of tweets, analyse that data and then use various classification algorithms to classify the information as fake or legitimate.

2. LITERATURE SURVEY

- Studies and research papers are being published on this issue to resolve the problem of fake news of any format.
- Existing or work under process solutions are facing various challenges to resolve this problem such as managing different

sources of the same information which makes it quite difficult to find out the source.
- Current processes for this issue are only providing solutions to particular social media platforms, not all platforms.
- Thus, the biggest challenge we will face will be to build a solution that will ensure that this system consider all aspects of this issue and try to provide common solution by overcoming previously occurred or occurring limitations.

3. PARAMETERS FOR DETECTING FAKE NEWS

(a) Date of Account creation.
(b) Comparing authenticated tweets.
(c) Retweet analysis.
(d) Tweet Timing.
(e) Status.
(f) Connections.

4. DATA GATHERING

People use different social media platforms to express their views through posts towards all the crises events as well as in general activities. We have gathered the data from Twitter using the Streaming API for the subject Citizenship Amendment Act (CAA) and related subjects such as Citizenship Amendment Bill, National register for of Citizens. We accumulated Twitter Trends API after every day for the current trending topics, and collected tweets corresponding to these topics as query search words for the Streaming API. Also, hourly the data we have downloaded changes according to the live tweets that are being tweeted by twitter account holders. So, we got more than 2 lakh tweets on the topic CAA and tweets with the reference to this event for period of one month. We have categorized the accounts which belong to different states and cities.

Event	Total Tweets	Classification
Citizenship Amendment Act	>2 lakh tweets	1) Authenticated Users 2) Unauthenticated Users

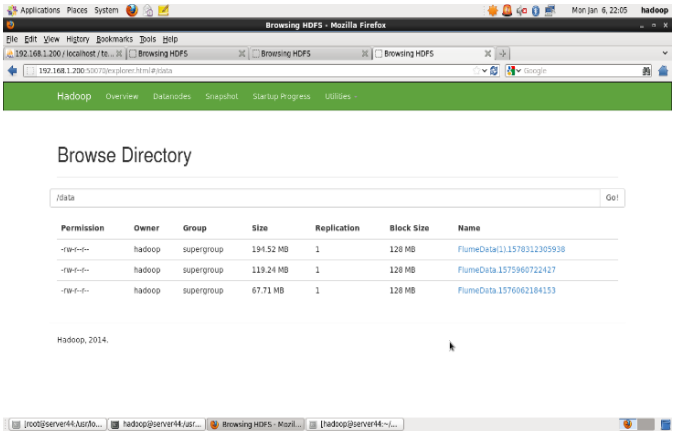

Fig. 1: Twitter Dataset

5. TECHNOLOGIES

Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

Apache Spark is an open source distributed general-purpose cluster-computing framework. Spark provides an interface for programming the entire clusters with implicit data parallelism and fault tolerance. Spark converts unstructured data into structured data.

```
+-----------+--------------------+--------------------+--------------------+--------+
|screen_name|                text|          created_at|          created_at|verified|
+-----------+--------------------+--------------------+--------------------+--------+
|   BBCHindi|CAA: जब सड़क से ल...|Mon Jan 06 12:07:...|Tue Sep 13 10:54:...|    true|
|indiatvnews|Assam NRC: Childr...|Mon Jan 06 12:30:...|Wed Apr 22 11:26:...|    true|
|       ndtv|Internet services...|Tue Dec 24 02:30:...|Fri May 01 20:34:...|    true|
|   TimesNow|.@RahulGandhi tak...|Tue Dec 24 02:39:...|Thu Jan 20 12:17:...|    true|
|   TimesNow|.@RahulGandhi tak...|Tue Dec 24 02:40:...|Thu Jan 20 12:17:...|    true|
|  CNNnews18|#NewsUpdate - Ben...|Tue Dec 24 02:42:...|Fri Jun 01 20:31:...|    true|
|   TimesNow|#Breaking | 2 peo...|Tue Dec 24 02:43:...|Thu Jan 20 12:17:...|    true|
|  CNNnews18|#NewsUpdate - CNN...|Tue Dec 24 02:50:...|Fri Jun 01 20:31:...|    true|
|    ABPNews|CAA कावधिशेः पुड...|Tue Dec 24 02:54:...|Mon May 11 12:25:...|    true|
|   TimesNow|#Breaking | 'No N...|Tue Dec 24 02:57:...|Thu Jan 20 12:17:...|    true|
|   TimesNow|PM Modi taught ho...|Tue Dec 24 03:00:...|Thu Jan 20 12:17:...|    true|
|  CNNnews18|#NewsAlert - Sita...|Tue Dec 24 03:03:...|Fri Jun 01 20:31:...|    true|
|  CNNnews18|#NewsAlert - Inte...|Tue Dec 24 03:04:...|Fri Jun 01 20:31:...|    true|
|   TimesNow|#Breaking | Actor...|Tue Dec 24 03:11:...|Thu Jan 20 12:17:...|    true|
|       ndtv|Assam BJP Backtra...|Tue Dec 24 03:11:...|Fri May 01 20:34:...|    true|
|    ABPNews|Amid #CAA &amp;amp; #...|Tue Dec 24 03:13:...|Mon May 11 12:25:...|    true|
|   TimesNow|#Breaking | A Ger...|Tue Dec 24 03:15:...|Thu Jan 20 12:17:...|    true|
|     aajtak|CAA पर रार के बधै...|Tue Dec 24 03:15:...|Tue May 26 11:31:...|    true|
|   TimesNow|#Breaking | A day...|Tue Dec 24 03:18:...|Thu Jan 20 12:17:...|    true|
|    ABPNews|CAA पर हंगमे के ...|Tue Dec 24 03:20:...|Mon May 11 12:25:...|    true|
+-----------+--------------------+--------------------+--------------------+--------+
only showing top 20 rows
```

Fig. 2: Data in Structured Form

Scala is an acronym for "Scalable Language". It is a general-purpose programming language designed for the programmers who want to write programs in a concise, elegant, and type-safe way. **Scala** enables programmers to be more productive. **Scala** is developed as an object-oriented and functional programming language.

HBase is an open-source non-relational distributed database modelled after Google's Bigtable and written in Java. Apache HBase is used to have random, real-time read/write access to Big Data. On top of clusters of commodity hardware, it hosts very large tables.

6. METHODOLOGY

Firstly, we have gathered Twitter Data of the event using Twitter API, keeping track of every days distinguished content within the data. We have used Apache Flume framework to store the collected data in HDFS into different datasets. The data gathered was downloaded using Apache Spark mentioning current trending hashtags which are as follows:
#CAA,#CAA2019, #CitizenshipAmendmentAct, #NRC ,#CAAJanJagran, #CAA_NRC_Protests, #NRC_CAA, NationalRegisterofCitizens, #CAAshowdown, Citizenship Amendment Bill, #CAAProtest.

The data then gathered was stored into the Database. We then studied the various fields of data, that can help us to get to the results. Some of them were "date_of_creation", "location", "verified", "retweets", "date_of_tweet", "created_at", "text". All of this came under data gathering. Data Manipulation started with removing the retweets and analysed only on the original tweets the account holders tweeted. Several Machine Learning and NLP algorithms were implemented to achieve accurate results out of which Random Forest proved to be the most beneficial.

Natural Language Processing algorithms are based on Machine Learning algorithms, they summarise the main points in a given text or a document. NLP algorithms are also used to distinguish the text according to predefined classes and their classifications and is also used to assemble the data. NLP algorithms are used to examine the text and make it in analysable form. NLP is used in spam filtering as well as in email routing. NLP helps to improve efficiency and accuracy of the documentation. It is the easiest way for sentiment analysis.

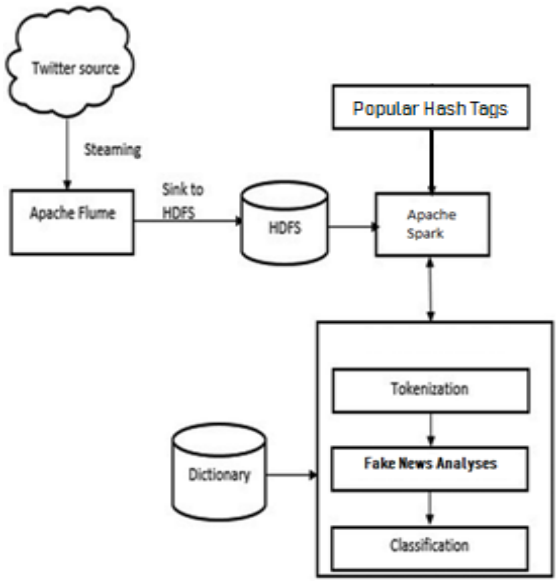

Fig. 3: Flow chart

Random forest is an organized and optimized learning algorithm. The term forest is nothing but group of decision trees usually trained with the "bagging" method. The general concept of the bagging method is that a combination of learning models increases the overall result and reduces the complexity of models. If put into simple words, random forest builds many decision trees and combines them together to get a more accurate, precise and stable prediction of the desired result.

Bagging: The training algorithm for random forests applies the general technique of bagging. Given a training set $X = \{x_1, ..., x_n\}$ with responses given as $Y = \{y_1, ..., y_n\}$, bagging repeatedly picks a arbitrary sample with replacement of the set and place trees.

After the training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x': or by taking into consideration the greater number of vote in classification trees.

This bootstrapping procedure leads to finer model presentation because it diminishes the dissimilarity of the model, without rising the bias. The predictions are extremely sensitive to hinderance, as long as the trees are not correlated. Only training those many trees on one training set would estimate strongly correlated and suitable trees. Bootstrap sampling is a method to sparse and differentiate the trees by displaying the various training sets.

Naive Bayes classifiers are a group of distribution algorithms depending upon **Bayes' Theorem**. It is a group of algorithms where all of the algorithms share a common principle, i.e. every pair of characteristics being classified is not dependent on each other. Bayes' Theorem notes the probability of an event occurring, given the probability of another different event that has already happened. Bayes' theorem is mathematically expressed as below,

$$P(A|B) = (P(B|A) * P(A)) / P(B)$$

Where A, B are events and P(B) != 0.

- We are trying to note the probability of event A, whereas the event B is true. Event B is also called as **evidence** or **prior probability**.
- The **prior** of A is P(A) (Probability of event before evidence is seen). The evidence is an attribute value of an unspecified instance.
- P(A|B) is a posterior probability of B, which is probability of event after prior probability is seen.

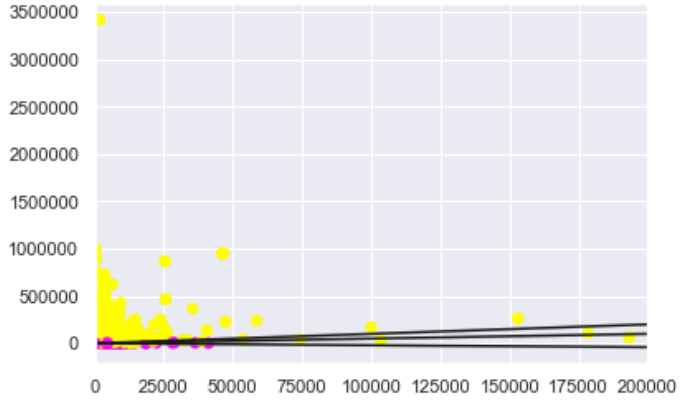

Fig. 4: Analysed Twitted Data

7. CONCLUSION

Our research work provides insights into the behavioural pattern of spread of fake news on the Twitter platform. We have used the NPL algorithms for the analysis of Twitter data from different datasets. We have developed an analysis model to predict that the twitted news is fake or not using comparison technique. With the predictive analysis our results provided a proof of concept that, automated techniques can be used in identifying fake content posted on Twitter. With the help of this technique the data can be analysed and it will be easy to detect the false news which is spread through the social media platforms.

8. REFERENCES

[1] Ruchika Ganar and Shrikant B. Ardhapurkar, *"Exploring social network for forensic analysis to predict civil unrest",* Volume– 04, Issue-04, E-ISSN: 2347-2693.

[2] Kai Shu, Amy Slivaz, Suhang Wangy, Jiliang Tang, and Huan Liu, *"Fake News Detection on Social Media: A Data Mining Perspective",* August 2017

[3] Alessandro Bondielli, Francesco Marcelloni, "*A Survey on fake news and rumour detection techniques",* Published in Inf. Sci. 2019, journal homepage: http://www.elsevier.com/locate/ins

[4] Anisha P. Rodrigues & Niranjan N. Chiplunkar | Marko Robnik-Šikonja (Reviewing editor), *"Real-time Twitter data analysis using Hadoop ecosystem",* Article: 1534519, Published online: 19 Oct 2018.

[5] H. Allcott, M. Gentzkow, *"Social media and fake news in the 2016 election",* Technical Report, National Bureau of Economic Research, 2017.

[6] Carlos Castillo, Mohammed El-Haddad, Jurgen Pfeffer, and Matt Stempeck. *"Characterizing the life cycle of online news stories using social media reactions",* In CSCW'14.

[7] https://twitter.com

