Automated System for Fake News Detection using NLP and Machine Learning Approach

 Manish M, Gurukiran, Manoj, Nikil S Kumar, Sandesh Kumar

***Abstract: The large use of social media has tremendous impact on our society, culture, business with potentially positive and negative effects. Now a days, due to the increase in use of online social networks, the fake news for various commercial and political purposes has been emerging in large numbers and widely spread in the online world. Fake news detection topic has gained a great deal of interest from researchers around the world. When some event has occurred, many people discuss it on the web through the social networking. They search or retrieve and discuss the news events as the routine of daily life. Some type of news such as various bad events from natural phenomenal or climate are unpredictable.***

I.INTRODUCTION

Twitter is a micro-blogging service, which has gained popularity as one of the prominent news source and information dissemination agent over the last few years. Each post on Twitter is characterized by two main components: the tweet (content and associated metadata) and the user (source) who posted the tweet. Rumors / fake news posted on twitter during real world events can result in damage, chaos and monetary loss. Today, online social media plays a vital role during real world events such as earthquakes, hurricanes, elections and social movements.

**Manish M,** Electronics and Communication, SMVITM, Udupi (email: manishmani725@gmail.com)

**Gurukiran,** Electronics and Communication, SMVITM, Udupi (email: gurukiranb92@gmail.com)

**Manoj,** Electronics and Communication, SMVITM, Udupi(email:manojdevadiga57@gmail.com)

**Nikil S Kumar,** Electronics and Communication, SMVITM, Udupi (email: nikhilkumar26459@gmail.com)

**Sandesh Kumar,** Electronics and Communication, Assistance Professor ( Senior), SMVITM, Udupi (email: sandesh.ec@sode-edu.in)

Modern life has become quite suitable and the people of the world have to thank the vast contribution of the internet technology for transmission and information sharing. There is no doubt that internet has made our lives easier and access to surplus information viable. This is an evolution in human history, but at the same time it unfocussed the line between true media and maliciously forged media. Today anyone can publish content – credible or not – that can be consumed by the world wide web. Sadly, fake news accumulates a great deal of attention over the internet, especially on social media. People get deceived and don’t think twice before circulating such mis- informative pieces to the far end of the world. This kind of news vanishes but not without doing the harm it intended to cause.

Fake news is news which are created intentionally to misguide the readers. It is a type of propaganda which is published in the form of genuine news. Fake news is spread through traditional news media and social media . Fake news has been a problem from a long time. With the introduction of social media, the spread of fake news is increased and it became diﬃcult to diﬀerentiate between true news and fake news. The spread of fake news is a matter of concern as it manipulates the public opinions. During the American Presidential elections of 2016, it was estimated that over 1 million tweets are related to fake news “Pizzagate” by the end of the elections. The wide spread of fake news can have a huge negative impact on individuals and society as a whole**.**

The given social media sites that play a major role in supplying counterfeit news include Facebook, Twitter, Whatsapp etc. Many scientists believe that counterfeited news issue may be addressed by means of machine learning and artificial intelligence. This is because recently artificial intelligence algorithms have begun to improve work on lots of classification problems (image recognition, voice detection and so on) because hardware is cheaper and bigger datasets are available.

Detecting fake news on social media present unique challenges. First, fake news pieces are intentionally written to mislead consumers, which makes it not satisfactory to spot fake news from news content itself. Thus, we need to explore information in addition to news content, such as social engagements and social behaviours of users on social media. For example, a credible user’s comment that “This is bull shit” is a strong signal that the user doesn’t believe the news and thus the news may be fake. Second, the research community lacks data sets which contain dynamic information to understand how fake news propagates, how users react to fake news, and how we can extract useful temporal patterns for (early) fake news detection and intervention. Thus, it is necessary to have comprehensive datasets that have news content, social context and dynamic information to facilitate fake news research. However, to the best of our know.

 Various models are used to provide an accuracy range of 60-75%. Which comprises of Naïve Bayes’ classifier, Linguistic features based, Bounded decision tree model, SVM etc. The parameters that are taken in consideration do not yield high accuracy. The motive of this paper is to increase the accuracy of detecting fake news more than the present results that are available. By fabricating this new model which will judge the counterfeit news articles on the basis of certain criteria which are as follows: spelling mistake, jumbled sentences, punctuation errors etc.

 The extensive spread of fake news has the potential for extremely negative impacts on individuals and society. Fake news is intentionally written to mislead readers to believe false information, which makes it difficult and nontrivial to detect based on news content. Therefore, the issue of fake new detection is both challenging and crucial. Hence, we decide to take up this challenge and find solution in an efficient way. Refer Fig 1 for the flowchart

II.METHODOLOGY

##  Fig 1: Fake News Detection Cycle

## 2.1 Content extraction from Twitter:

Twitter is a popular social network where users share messages called tweets. Twitter allows us to mine the data of any user using Twitter API or Tweepy. The data will be tweets extracted from the user. The first thing to do is get the consumer key, consumer secret, access key and access secret from twitter developer available easily for each user. These keys will help the API for authentication.

## 2.2 Data pre-processing:

Data pre processing is done to convert the raw data into a required format. Data pre processing can be done by various methods like data cleaning, data reduction, data integration etc. In this , the data sets are collected from diﬀerent resources which have diﬀerent formats and attributes. Hence, the data can be duplicate and they may contain some attributes which are not useful. So, the data is converted into our required format with required attributes which are used to train our model.

### **2.2.1 Data Reduction:**

In data reduction, the cluster representation of the data are used to replace the actual data. It also helps to detect outliers in data. Sampling: Sampling can be used for data reduction because it allows a large data set to be represented by a much smaller random data sample (or subset).

### **2.2.2 Data Cleaning:**

Quality of your data is critical in getting to final analysis. Any data which tend to be incomplete, noisy and inconsistent can effect your result. Data cleaning in data mining is the process of detecting and removing corrupt or inaccurate records from a record set, table or database.

### **2.2.3 Data Transformation:**

Data transformation involves the following: In Normalisation, where the attribute data are scaled to fall within a small specified range, such as -1.0 to 1.0, or 0 to 1.0. Smoothing works to remove the noise from the data. Such techniques include binning, clustering, and regression. It is a fundamental aspect of most data integration and data management tasks such as data wrangling, data warehousing, data integration and application integration.

### **2.2.4 Data Integration:**

 Data Integration is a data pre-processing technique that involves combining data from multiple heterogeneous data sources into a coherent data store and provide a unified view of the data. These sources may include multiple data cubes, databases or flat files.

**2.3 Generating News Feature Vector:**

The most important part of detecting if a given news is fake or not is to convert the news article into a news vector which contains the important features which are used to determine the nature of the news.

There are several ways to generate this feature vector. We tried diﬀerent approaches for the same to determine which method gives the best accuracy. Some of the methods are:

### **2.3.1 Bag of Words:**

 Bag of words is a way of representing text in a format which can be easily processed by the machine learning algorithms. BoW is one of the ways of extracting features from text.

### **2.3.2 TF-IDF:**

 TF-IDF stands for term frequency-inverse document frequency.TF-IDF is a method used to represent text in a format which can be easily processed by the machine learning algorithms. It is a numerical statistic that shows how important a word is to a document in a word corpus. The importance of a word is proportional to the number of times the word appears in the document but inversely proportional to the number of times the word appears in the corpus.

### **2.3.3 Shallow and Deep syntactical Analysis**:

 We generated POS(part-of-speech) tags using the Spacy library. Our POS features will be encoded as tf-idf values for each for these tags. Even though POS tags are eﬀective in detecting fake product reviews, they are not as eﬀective as words. Therefore, we strengthen POS features with unigram/bigram features. For deep syntactical analysis we used the Stanford/Berkeley parser to generate CFG rules for the sentences and we encoded these rules with tf-idf values for each production rule.

### **2.3.4 Semantic Analysis**:

 A widely used open-source resource for incorporating semantic information is Empath (developed by Stanford). Empath is a lexicon of words grouped into semantic categories relevant to psychological processes. Several research works have relied on semantic analysis to build deception models using machine learning approaches and showed that the use of semantic information is helpful for the automatic identiﬁcation of deceit. Empath has 194 semantic categories, some of these semantic classes are emotional tone(positive or negative), anger, nervousness. We get a score between 0- 100 for each semantic class. The lexicon we get is converted to a TF-IDF vector by taking the score for a semantic class(like nervousness) as its frequency.

**2.3.5 Combining features to form ﬁnal news vector:**

We considered 3 methods for generating feature vectors

1.TF-IDF bigram vector of the news article.

2.Feature Vector generated by Syntax Analysis of the news article.

3.Feature Vector generated by the semantic analysis of the news article.

After generating these features and generating their individual feature vector, we have to combine these features to form the ﬁnal news vector on which classiﬁcation is performed.

The method we approached for combining the feature vectors is

1.Take the most important features for the 3 feature vectors

2.Assign weights to each vector and then take the weighted combination of the 3 feature vectors to generate the ﬁnal feature vector. If x is the weight corresponding to the ﬁrst feature vector, y for the second, and 1-x-y for the third. The ﬁnal feature vector will be the linear combination of these feature vectors multiplied by their corresponding weights.

## 2.4 Classification:

 After generating the news feature vector, now we classify the vector to whether it is fake or real. We aim to use the following classiﬁcation algorithms for the purpose of classiﬁcation.

**2.4.1 Naive Bayes:**

Naive Bayes’ is a supervised learning algorithm which is used for classiﬁcation. It is based on Bayes’ theorem assuming that features are independent of each other. It calculates the probability of every class, the class with maximum probability is chosen as the output. Refer Fig 2 for the equation.

****

Fig 2: Posterior probability using Bayes’ Theorem

**2.4.2 Logistic Regression:**

The logistic regression module performs the simple task of taking the dataset, splitting it into two parts, viz. test and train set. The train dataset which is the BuzzFeed dataset is used to train the regression model for the user input which is the news to be tested in this case. It is a classification algorithm used for machine learning that predicts the likelihood of a categorical dependent variable, whereit will be either fake or authentic henceforth logistic regression will help to describe a relationship between a set of independent variables and categorical dependent variables. The dependent variable in logistic regression is a binary variable that includes data encoded as 1 (the user given news is fake) or 0 (the news is authentic), hence these are the only two classes.

The model gives an authenticity value between 0 and 1 later on converted into percentage and hence can be easily categorized as how much the news is authentic or fake.

### **2.4.3 SGD Classifier:**

 SGD Classifier is a Linear classifier with SGD training. It is a simple and efficient approach for discriminative learning of linear classifiers under convex loss functions such as (linear) Support Vector Machines and Logistic Regression.

The gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule (aka learning rate). SGD allows mini batch (online/out-of-core) learning via the partial fit method. For best results using the default learning rate schedule, the data should have zero mean and unit variance.

This implementation works with data represented as dense or sparse arrays of floating-point values for the features. The model it fits can be controlled with the loss parameter; by default, it fits a linear support vector machine (SVM).

**2.4.4 Linear SVC:**

The objective of a Linear SVC is to fit to the data you provide, returning a "best fit" hyper plane that divides, or categorizes, your data . From there, after getting the hyper plane, you can then feed some features to your classifier to see what the "predicted" class is.

**2.5.1 Languages used:**

**PYTHON 3.6:** Python 3.6 is specifically used because, some of the libraries like scikit-learn, random etc. Works only on this platform.

**NLP:** It is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

**2.5.2 Libraries used:**

**Tweepy**: It is a Python library for accessing the Twitter API. It is great for simple automation and creating twitterbots.

**Regular expression(re):** "re" module included with Python primarily used for string searching and manipulation. Also used frequently for webpage "Scraping" (extract large amount of data from websites).

**CSV**: A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format. A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number offields.

**TextBlob:** It is a Python library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

**NumPy:** NumPy is a python library used for working with arrays. It also has functions for working in domain of linear algebra, Fourier transform, and matrices.

**Scipy:** The SciPy library depends on NumPy, which provides convenient and fast N-dimensional array manipulation. The main reason for building the SciPy library is that, it should work with NumPy arrays.

**Skicit-learn:** It is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems. This package focuses on bringing machine learning to non- specialists using a general-purpose high-level language.

**Random:** It generate pseudo-random numbers. Almost all module functions depend on the basic function random(), which generates a random float uniformly in the semi-open range [0.0,1.0].

**Pickle:** It is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it “serializes” the object first before writing it to file. Pickling is a way to convert a python object (list, dict ,etc.).

III.RESULTS

A complete, production-quality classiﬁer will in corporate many different features beyond the vectors corresponding to the words in the text. For fake news detection , we can add as features the source of the news, including any associated URLs, the topic (e.g., science, politics, sports, etc.), publishing medium(blog, print, social media), country or geographic region of origin, publication year, as well as linguistic features not exploited in this exercise use of capitalization, fraction of words that are proper nouns (using gazetteers), and others. Besides, we can also aggregate the well- performed classiﬁers to achieve better accuracy.

It mainly consists of four phases:

**Phase 1:** Taking the inputs by pulling the data.

**Phase 2:** Cleaning the obtained raw data.

**Phase 3:** Creating the model on the basis of given inputs and providing the model accuracy.

**Phase 4:** Testing the inputs by entering the Tweets and displaying whether the entered news is REAL or FAKE with the corresponding confidence.



**Fig 9:** Demonstration of first phase

****

**Fig 10:** Demonstration of second phase



**Fig 11:** Demonstration of third phase

****

**Fig 12:** Demonstration of fourth phase

IV.CONCLUSIONS AND FUTURE WORK

We obtained a result with an accuracy of 82% by using the weights for feature vectors derived by bigrams, syntax and semantic analysis. The first algorithm used for classification was Naive Bayes’ (Multinomial and Bernoulli’s), where no hyper-parameter was required. This helped to set a reference point for further analysis. It was followed by SVC model where we selected the normalizing parameter (T) as1 2.

REFERENCES

1. Allcott, H., and Gentzkow, M., “ Social Media and Fake News in the 2016 Election“,https://web.stanford.edu/œgentzkow/research/fakenews.pdf,January, 2017.
2. Datasets,“*Kaggle*”[,https://www.kaggle.com/](http://www.kaggle.com/c/fake-news/data)c[/fake-news/data,](http://www.kaggle.com/c/fake-news/data) February, 2018. Source Code Repository, GitHub, https://github.com/FakeNewsDetection/FakeBuster, April, 2018.
3. Arushi Gupta, Rishabh Kaushal, “*Improving Spam Detection in Online Social Networks*”, 978-1-4799-71718/15/$31.00 ©2017IEEE.
4. Conroy, N., Rubin, V. and Chen, Y. (2015). “*Automatic deception detection: Methods for finding fake news.*”, Proceedings of the Association for Information Science and Technology, 52(1), pp.1-4.
5. S. Maheshwari , “*How fake news goes viral: A case study*”, Nov.2016. [Online]. Available: htt[ps://www.nytimes.com](http://www.nytimes.com/) / 2016 / 11 / 20 / business / media / how- fake - news -spreads.html (visited on11/08/2017).
6. Nikita Munot, Sharvari S. Govilkar, “ *Comparative Study of Text Summarization Methods*”, International Journal of Computer Applications (0975 – 8887) Volume 102– No.12, September2017.
7. Mykhailo Granik, Volodymyr Mesyura, “*Fake News Detection Using Naive Bayes Classifier*”, 2017 IEEE First Ukraine Conference on Engineering(UKRCON).

AUTHORS PROFILE

**Manish M,** student in the department of Electronics and Communication Engineering at Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal.

**Gurukiran,** student in the department of Electronics and Communication Engineering at Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal.

**Manoj,** student in the department of Electronics and Communication Engineering at Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal

**Nikil S Kumar,** student in the department of Electronics and Communication Engineering at Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal

**Sandesh Kumar,** Assistant Professor(Senior) in the department of Electronics and Communication at Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal