International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

Fake News Identification using Machine Learning

Priya S. Gadekar
Department of Electronics and Telecommunication,SGBAU Amravati University

Abstract: Fake news is nothing but junk news also we can say hoaxes that consists of disinformation spread via social media. Fake news is written and published usually with the intent to mislead in order to damage an organization or a person. Fake news can be spread via traditional print, broadcast news. Nowadays Controlling of fake news is become important issue. For that various methods are used. It is difficult to minimize 100% hoax from social media, but at least there is a technology which can reduce the hoax in social media. With the help of different classifier one can identify fake news. In this work I used two different classifiers which are SVM classifier and Naive Bayes Classifier With this classifiers I achieved 60.97% accuracy with the SVM classifier and a 59.76% with the naïve bayes classifier. This Paper shows the operational comparison of both the classifiers.

Keywords: SVM, NB, SVC, SGD, Hoaxes

I. INTRODUCTION

Fake news is now viewed as one of the greatest threats to democracy, journalism, and freedom of expression. It has weakened public trust in governments. The goal of the Fake News Challenge is to explore how artificial intelligence technologies, particularly machine learning and natural language processing, might be leveraged to combat the fake news problem. Fake news is increasingly becoming a menace to our society. It is typically generated for commercial interests to attract viewers and collect advertising revenue.

II. TYPES OF FAKE NEWS

There are differing opinions when it comes to identifying types of fake news. However, when it comes to evaluating content online there are various types of fake or misleading news we need to be aware of. These include:

A. Clickbait

Clickbait stories use sensationalist headlines to grab attention and drive click-throughs to the publisher website, normally at the expense of truth or accuracy.

B. Propaganda

Stories that are created to deliberately mislead audiences, promote a biased point of view or particular political cause or agenda.

C. Satire/Parody

Lots of websites and social media accounts publish fake news stories for entertainment and parody. For example; The Onion, Waterford Whispers, The Daily Mash, etc.

D. Sloppy Journalism

Sometimes reporters or journalists may publish a story with unreliable information or without checking all of the facts which can mislead audiences. For example, during the U.S. elections, fashion retailer Urban Outfitters published an Election Day Guide, the guide contained incorrect information telling voters that they needed a ‘voter registration card’. This is not required by any state in the U.S. for voting.

E. Misleading Headings

Stories that are not completely false can be distorted using misleading or sensationalist headlines. These types of news can spread quickly on social media sites where only headlines and small snippets of the full article are displayed on audience newsfeeds.

F. Biased/Slanted News

Many people are drawn to news or stories that confirm their own beliefs or biases and fake news can prey on these biases. Social media news feeds tend to display news and articles that they think we will like based on our personalised searches.

A variety of sources for news exist, and the most widely used are listed below.

1) *Radio:* There are many types of radio broadcasters, including commercial and non commercial, that deliver news to listeners. It is common for local radio stations to exist as affiliates of larger broadcast networks.
2) *Newspapers:* The highest circulating print media is newspaper.
3) *Television:* The largest broadcast US television news networks include NBC, CBS, ABC, and Fox. Local television news channels carry national news programs as well as local news programs.
4) *Word of Mouth:* News that you hear from others by word of mouth is also a source of news!
5) *The Internet:* Since its invention and widespread use, the Internet has rapidly become a major source for news. Many news sites correspond to radio, newspaper, and television networks, but some news organizations, which mostly focus on political news, only appear online such as Politico, Real Clear Politics, and The Huffington Post.
6) *Social Media:* Social media is another source of news users visit sites like Facebook and Twitter to follow and share news. For mobile use, social media sites also develop and release apps for phones or other mobile devices that users can download to access news.

III.SVM CLASSIFIER

For a dataset which contains features set and labels set, an SVM classifier create a model to predict classes for new examples. It classifies that examples means it assigns new data points to one of the classes. If there are only 2 classes then it can be called as a Binary SVM Classifier. There are 2 kinds of SVM classifiers:

A. Linear SVM Classifier
B. Non-Linear SVM Classifier

Svm Linear Classifier is the linear classifier model, In which we can consider that training examples plotted in space. These data points are expected to be separated by an apparent gap. It predicts a straight hyperplane dividing 2 classes. The primary focus while drawing the hyperplane is on maximizing the distance from hyperplane to the nearest data point of either class. The drawn hyperplane called as a maximum margin hyperplane.
In the real world, our dataset is generally dispersed up to some extent. To solve this problem separation of data into different classes on the basis of a straight linear hyperplane can't be considered a good choice. For this Vapnik suggested creating Non-Linear Classifiers by applying the kernel trick to maximum-margin hyperplanes. In Non-Linear SVM Classification, data points plotted in a higher dimensional space. SVMs are effective when the number of features is quite large.
It works effectively even if the number of features are greater than the number of samples.

IV.NAIVE BAYES CLASSIFER

Bayes Theorem works on conditional probability, which is the probability that an event will happen, given that a certain event has already occurred. Using this concept we can calculate the probability of any event based on the likelihood of another event
Below is the formula for calculating the conditional probability.
Where,

$$P(H|E) = \frac{P(E \mid H) * P(H)}{P(E)}$$

P(H) is the probability known as the prior probability.
P(E) is the probability of the evidence.
P(E|H) is the probability of the evidence is true.
P(H|E) is the probability of that the evidence is there.
The concept I use to classify fake news is that fake news articles often use the same set of words while true news will have a particular set of words. It is quite observable that few sets of words have a higher frequency of appearing in fake news than in true news and a certain set of words can be found in high frequency in true news. Of course, it is impossible to claim that the article is fake just because of the fact that some words appear in it, hence it is quite unlikely to make a system perfectly accurate but the appearance of words in larger quantity affect the probability of this fact.[16] Naive bayes is Very simple, easy to implement and fast and Needs less training data.

V. DATASET

The important is the collection of input data that is News in this case. Online news can be collected from different sources, such as news agency homepages, search engines, and social media websites. The dataset used to test the efficiency of the model is Buzzfeed news dataset having, 10240 news article tagged as real or fake.

VI. IMPLEMENTATION DETAILS

As mentioned above I used Buzz feed news Dataset. Every news statement is labelled as "REAL" and "FAKE" were tagged as 1 and 0 respectively so as to classifier. These dataset is then shuffled randomly and divided into two parts that is training and testing dataset. Training dataset was used to train the classifier. Test dataset was used to get result or estimation of how well the classifier performs on new data. After pre-processing Feature Extraction is performed using TF-IDF transform. Term Frequency refers to how many times a given term appears in document and Inverse -document frequency measures the weight of the word in the document i.e. if the word is common or rare in document. In short it shows the importance of words in the sentence, all the words which are less important like a, an, the, as etc are get removed after these process. Count Vectorizer allows us to use the bag-of-words approach by converting a collection of text documents into a matrix of token counts.

VII. RECEIVED RESULTS

After doing all the process I got the final result in the form of confusion matrix. A confusion matrix is a technique for summarizing the performance of a classification algorithm. It shows result Including with some parameters like precision, recall, f score etc. From the obtained confusion matrix I get the classification accuracy can be obtain as the ratio of correct predictions to total predictions made.

Classification accuracy = correct predictions / total predictions

In other way we can write Accuracy as,

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{P + N} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Where,

True positive (TP), True negative (TN), False negative (FN), False Positive (FP).

Classification Using SVM and Naïve Bayes classifier is shown in the following table:

Total News statements :10240

Tabel 1.1

Parameters in confusion Matrix	SVM	Naive Bayes
TP	2016	758
TN	4228	5362
FP	1524	390
FN	2472	3730

So, by using Support vector machine and Naïve Bayes classifier fake news can be identified with the accuracy of 60.97% with the SVM classifier and accuracy of 59.76% with the naïve bayes classifier.

VIII. CONCLUSION

In this way very first I have gathered all the necessary dataset. After that the filtering of dataset was done on the basis of various news statements and its respective Labels. Then pre-processing is done in various ways to get the data in machine readable form. Finally classification is done by two classifiers which was SVM and Naive Bayes classifier. By using Support vector machine and Naïve Bayes classifier fake news can be identified with the accuracy of 60.97% with the SVM classifier and accuracy of 59.76% with the naïve bayes classifier. Therefore by observing the results in this Experiment I conclude that SVM classifier is better choice for detection of fake news than Naive Bayes classifier. To improve the results further we can use Deep learning approach for that we required lot of learning data, so these becomes the future scope.

REFERENCES

[1] IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON) 90 0 "Fake News Detection Using Naive Bayes Classifier" Mykhailo Granik, Volodymyr Mesyura Computer Science Department
[2] Inggrid Yanuar Risca Pratiwi Malang State Polytechnic EE Department Malang ,Indonesia "Study Of Hoax Detection Using Naive Bayes Classifier" in Indonesian 2017 ICTS
[3] "Evaluating Machine Learning Algorithms for Fake News Detection" Shlok Gilda Department of Computer Engineering Pune Institute of Computer Technology, Pune, India 2017 IEEE 15th Student Conference on Research and Development (SCOReD)
[4] Sakeena M. Sirajudeen, Nur Fatihah A. Azmi, Adamu I. Abubakar Department Of Computer Science International Islamic University Malaysia "Online Fake News Detection Algorithm" Journal Of Theoretical And Applied Information Technology 15th September 2017. © 2005 - Ongoing J Avtoilt.9 5&. Lnlos
[5] Arushi Gupta "Improving Spam Detection in Online Social Networks" Department of Information Technology Indira Gandhi Delhi Technical University for Women Kashmere Gate, Delhi ©2015
[6] Shivam B. Parikh and Pradeep K. Atrey Albany 2018 IEEE Conference on Multimedia Information Processing and Retrieval "Media-Rich Fake News Detection: A Survey" Lab for Privacy and Security, College of Engineering and Applied Sciences University at Albany, State University of New York, Albany, NY,USA
[7] Allcott, H., and Gentzkow, M., Social Media and Fake News in the 2016 Election, https://web.stanford.edu/oegentzkow/research/fakenews.pdf, January, 2017.
[8] Kai Shu, Suhang Wang, Huan Liu "Understanding User Profiles on Social Media for Fake News Detection" Arizona State University, 2018 IEEE Conference on Multimedia Information Processing and Retrieval
[9] Sundus Hassan , Muhammad Rafi , Muhammad Shahid Shaikh NUCES-FAST, Karachi Campus "Comparing SVM and Naive Bayes Classifiers for Text Categorization with Wikitology as knowledge enrichment" 978-1-4577-0657-8/11/$26.00 © 2011 IEEE
[10] Shashank Gupta, Raghuveer Thirukovalluru, Manjira Sinha, Sandya Mannarswamy IIIT Hyderabad, India "A Community Infused Matrix-Tensor Coupled Factorization Based Method for Fake News Detection" 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) CIMTDetect.
[11] Alina campan, Alfredo cuzzocrea, Triana marious truta department of computer science "Fighting fake newsspread in online soacial network, actual trends and future reserch directions "2017 IEEE International Conference on Big Data (BIGDATA) 978-1-5386-2715-0/17/$31.00 ©2017 IEEE 4453
[12] Parag Kulkarni "Reinforcement And Systemic Machine Learning For Decision Making" IEEE Press Editorial Board John B. Anderson, Editor in Chief
[13] Based on Text Classification using SVM and SGD Agung B. Prasetijo, R. Rizal Isnanto and Dania Eridani "Hoax Detection System on Indonesian News" Sites Proc. of 2017 4th Int. Conf. on Information Tech., Computer, and Electrical Engineering (ICITACEE), Oct 18-19, 2017, Semarang, Indonesia
[14] Osvaldo Simeone (2018), "A Brief Introduction to Machine Learning for Engineers", Foundations and TrendsR in Signal Processing: Vol. 12, No. 3-4, pp 200–431. DOI: 10.1561/2000000102. Osvaldo Simeone Department of Informatics
[15] Myeongsu Kang University of Maryland "Machine Learning: Anomaly Detection" Center for Advanced Life Cycle Engineering, College Park,MD, USA
[16] Akshay Jain, Amey Kasbe Department of Electronics and Communication Engineering "Fake News Detection" 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Sciences 978-1-5386-2663-4/18/$31.00 ©2018 IEEE