International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

A Novel Technique for Fake News Detection using Machine Learning Algorithms and Web Scrapping

K S Veda[1], Sudarshana K[2], Moulya M[3], Amulya[4], Chandana N S[5]

Department of Information Science and Engineering, AIET, Moodbidri, 574225, India

Abstract: ***In today's world we all can see the growth of social media. Social media is one of the common platforms where large number of people interacts with each other. Social media is a communication bridge between the people to interact, share, and to provide information. The same social media is facing some sort of issues like fake news. Fake means which is not real instead they are spreading fake news as real news and unfortunately people are not able differentiate them and are believing that news. In this paper we would like to provide a software that would differentiate between fake and real news. In this model we had come up with four algorithm and web scarping technology.***

Keywords: Fake news, Web Scrapping

1. Introduction

As of Now we can see lot of improvement in technology through internet world and one among them is social media. Social media has made life easier where we can interact with our close ones and even unknown people or a different people who is from different places. But through social media some people are spreading or posting fake news which might affect people or make people to believe unreal facts. This is one of the major problem that the social media is facing. Facebook and Google have been taking measures for the same problem. Facebook is using tools to help users to find out the fake news by flagging them as fake. Google is using hoax sites. They are using fact checking labels in Google news.

People should be aware what to believe and what not to believe but in some cases it can't be differentiated which is true and which is fake news.

Since the news spread faster and wider it could be difficult to predict them. So, our model could help them out to predict these kinds of news. Some people wontedly miss uses the social media to spread unreal facts or a fake news to mislead the people to believe such kind of people and let us try not to believe such kind of fake news. To find out the news is fake or real we can use machine learning algorithms.

These machine learning algorithms train system to predict the news to be real or fake based on text, words used and stop words, etc.

Our Model is a text based fake news detection application. There are variety of Fake news that may be on text, image, sound etc. There are plenty of ways to spot fake news and they might be fact checking, go through the news deeply and search in trusted websites, or if it is forwarded from one person to other to so on then back propagation should be performed etc. America as researched and said how much percent people will react to the fake news in what way. 64% of people are greatly confused by the fakes, 24% of people are somewhat confused because of the fake news and 11% of people are not that confused towards fake news. Americans are confident in spotting fake news following percent. 39% of people can great confidently spot the fake news, 45% of people are somewhat confident to spot the fake news, and 15% of people are not at all confident to spot fake news. There are many researches and projects that is been taking place to stop this fake news or help people to spot fake news so that people won't be mis leaded believe in fake news spread by some people.

2. Literature Survey

As many projects and researches has been taken place in spotting fake news many different machine learning algorithms also have been used. Most of the researchers has built a model or done a project based on Naïve Bayes, SVM algorithms.

Mykhailo Granik, Volodymyr Mesyura said they had achieved about 74% accuracy for predicting real or fake news using Naïve Bayes classification.

Akshay Jain, Amey Kasbe had said that by using n_grams better accuracy can be obtained. The n_grams is nothing but the combination of title and text. Using this method they had obtained a better results comparatively.

This Fake news becomes a famous issue because of the attention it got in American Presidential Selection in the year 2016. During this election time, much more of different fake news are discussed and posted in social medias. Some posts are even concluded that Trump had won the president referendum due to influence of fake news [4] [5]. Due to uncontrolled excitement created by the social media, after this some interest shown on fake news and its problems and also concerns has been raised about the bad effects of wide growth of false news.

The concern on boosting up of faux news is sensible, taking into account the wide growth climb of the web therefore the quick usage of social medias like blogs, Twitter, micro blogs and Facebook and WhatsApp , which has to the served in creation and transmission of stories as well as knowledge, thus giving a very big effect on the expansion of access to stories through the social media platforms. The report on fake news released in the year 2017 [6] reveals a rapid growth

of accessed stories via social media rather than any newspapers available in market. Along with this report of the year 2018[7] Some countries still believe newspapers for news instead of social medias. In Malaysia the tendency of Facebook usage is 6 t0 64%.

Some of the widening shows that fake news found on social media is transferring very fast from people to people. Media platforms have changed the style of news it posts in media [8]. In present, a little bit stories can be simplifying the hyperlinks or a tweet which is of a 280 character long in size. Not for only journalist but non-journalist also posts the articles in social media, to encourage people with non-framework category of news area to supply reporter contents on the online portal [9].

Fake news may be a powerful if it is not misleading means it is not used to mislead the people in the areas like cinema, politics, sports and many other areas. The side effects of faux news are often very toxic to society. The wide spread of misleading information may result in serious damages like affecting the reliability of the eco-system of news, damaging the reputation of any person or any other political organization, or causing fear among the people in public which may broke the stability in the society [10], [11].

There are many other types of effects thanks to the wrap and rapid growth on social media. From famously referred to as the Pizzagate debacle [2], [12] to TheStorm right conspiracy theory [13] that created quite stir in US"s political scene, fake news went viral with none of oversight. There are serious cases in Asia caused by the viral of false stories, too. for example, during the Jakarta"s gubernatorial campaign which was held in the year 2016, the incumbent governor of Jakarta; Basuki Tjahaja Purnama who happen to be a Chinese Indonesian Christian fell victim to an altered video featuring him criticizing the utilization of a Quran verse as a ground to oppose non Muslim"s role in leadership [14]. The video went viral on social media, resulted with public protests that led to Basuki being reported to the police. Then, in May 2017, Basuki was convicted on charge of blasphemy and sentenced to 2 years in prison [11],

I.

Clearly this shows how fabricated content could become a dangerous tool to tarnish one"s reputation and to induce public"s rage on sensitive issues like ethnicity and non-secular. Another occurrence of trauma caused due to the wide spread of fake news at Malaysia in the year 2018. When Tengku Mahkota Johor Tunku Ismail Sultan Ibrahim spent RM1million by footing the bills of some lucky shoppers during his appearance in one among the hypermarkets in Johor, rumors began to spark afterward claiming that the royal prince would offer cash aids for shoppers at other hypermarkets also. The hoax viral on social media causing a crowd of individuals to flock to the mentioned hypermarkets and began to refill their trollies hoping for a few luck. Police was alerted for the massive gatherings and later confirmed that the viral stories of Tengku Mahkota Johor's cash aids were actually fake [16], [17]. Unlikely chaos, but public got into the deception trap of the viral, unreliable news.

These days, fake news issues has also drawn great attention from the tutorial side, prompted more studies particularly on developing fake news detection mechanisms to counter the matter. A fake news detection mechanism may be a technique or system that assists users with the tools and functions in predicting deceptive news content. The mechanism works with algorithms and measures which will classify and verify information or news. However, developing a fake news detection mechanism is ever challenging effort that needs deep comprehension on different aspects regarding news consumption on social media.

As shown in many researches, an efficient fake news detection requires reliable veracity assessment techniques as ferreting out deceptive cues from the dubious content with diverse topics and designs on social media, is technically challenging considering false information usually contains deliberate deception and implicit ambiguity of language [18], [19]. In fact, these characteristics of faux news are linguistically and technically very demanding to be understood, interpreted, extracted and analyzed [10], [20]. Thus, some recent works had also proposed on exploiting auxiliary information just like the knowledge domain as essential measure to reinforce the deception prediction, although acquiring quality data are often critical as data verification by existing knowledge repositories could also be lacking thanks to insufficient of authenticated evidences [10].

On the opposite hand, some studies also suggested that a reliable fake news detection mechanism is additionally required to be ready to mine and exploit the extra information of the social context features derived from the social engagement within the network, particularly of the news consumption on the web communication platforms as these additional information would be useful to reinforce the prediction of faux news [21]. However, mining this additional information could also be demanding since users" social engagements with fake news cave in wide, partial and unstructured data [21].

Our project aims to provide a review on the different techniques for fake news detection using machine learning and web scrapping. We are trying to build an algorithm which gives high accuracy in detecting fake news.

Datasets

Dataset is force apart into three parts that is labeled data training data, and test data sets. Test data set would contain certain expectations so that the machine would understand the prediction method and could be get trained according to the datasets that we had provided. Training data set contains set of sentences which should also satisfy Test data set conditions.

Labeled data set is same set of both the test and training data set. But with that it is also labeled for all the sentences as Real and Fake news so that machine could be able to identify and then get trained to automatically predict the news.

3. Model Implementation and Its Results

We had used four machine learning algorithms to measure it accuracy and find out which could be best algorithm in one among. We can even see the processing of these algorithms and how it is giving us the results.

a) Support vector Machine

In 1963 SVM was invented by Hava Siegelmann and Vladimir Vapnik. This applies the strategy of support vector. The current standard SVM method was published in the year 1995 by Corinna Cortes and Vapnik.

Figure 1: Model of Graphical User Interface

b) Naïve Bayes Algorithm

It is the Bayes classifier is the relationship of probabilistic, based on the probability of the particular characteristic or words it classifies that to one of the class where the probability is more. In Naïve Bayes features are not related one another. so that presence or absence of that particular feature won't affect any other appearance and non-appearance of the features. In our model Naïve Bayes algorithm provides second highest accuracy compared to another. In our model as a result Naïve Bayes classifier archives 0.8923 accuracy.

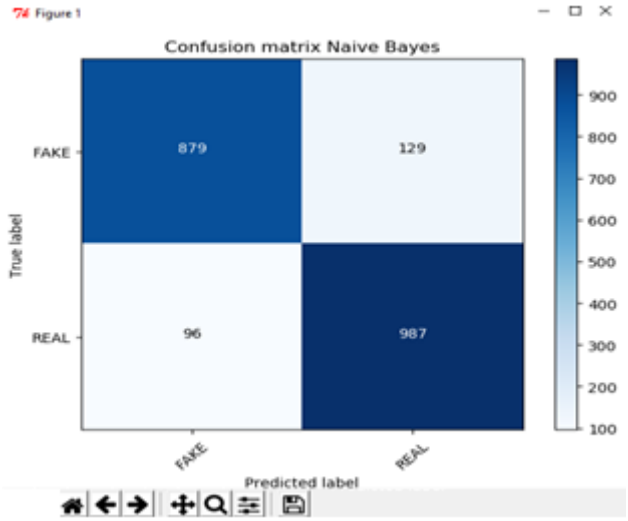

Figure 2: Confusion Matrix of Naïve Bayes

SVM is a linear model for classification and regression problems. This separates the data into different classes and then classifies the problem and predicts the outcome of the problem. SVM secures the highest accuracy in our model that is 0.9024. SVM is one of the best algorithms in machine learning for classification.

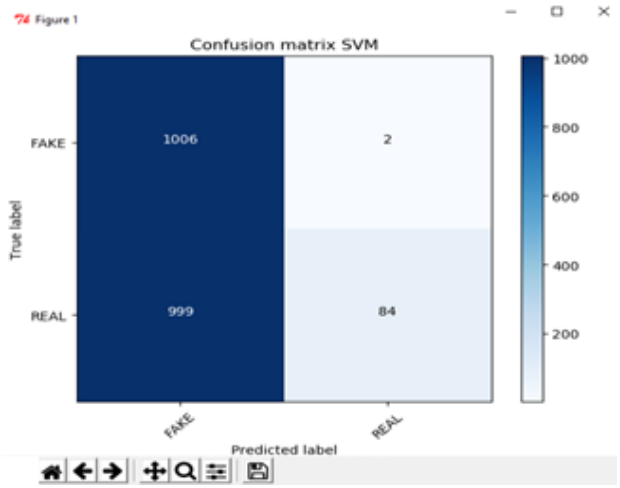

Figure 3: Confusion matrix of SVM

c) Logistic Regression

Quetelet and Verhulst invented logistic regression in 19th century. This gives the probability of two classes. This can be mostly used in the text news prediction because this would be very helpful to predict text news. We had used Logistic Regression as a part of our model and it has given us a result of 0.5212. And logistic regression might be a good algorithm for predicting text base news. Hence we had included logistic regression in our model.

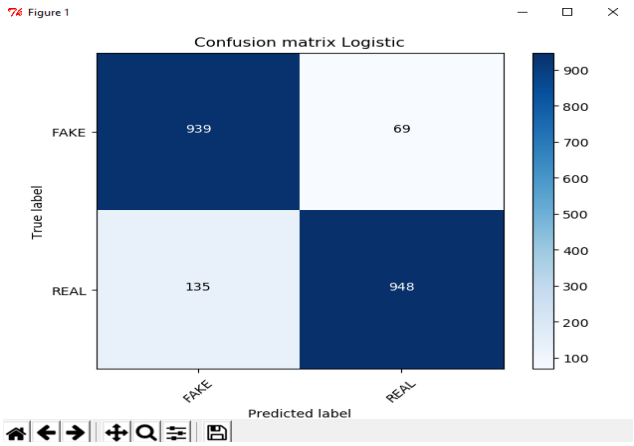

Figure 4: Confusion Matrix of Logistic Regression

d) Random Forest

Tin Kam Ho proposed the random forest in 1995. He designed by thinking forest trees splitting into oblique hyper plane so that it could give a good accuracy. Random Forest can give accuracy without suffering by overtraining as long as they are restricted to some sensitive selected features.

We had used random forest so that it could give us improved accuracy and it could give good results in prediction of fake and real news. As results we obtain the random forest accuracy of 0.8565.

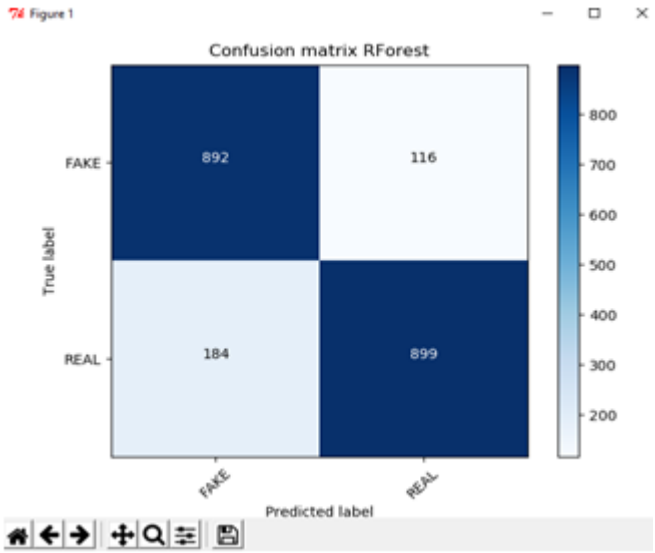

Figure 5: Confusion Matrix of Random Forest

e) Web Scrapping

Web scrapping is nothing but grabbing or harvesting information from different web sites or screening of the page and then it stores that information into our database. We include this web Scrapping to be up to date or to be always updated. And this might help us to predict even current news and the past news.

4. Results and Future Enhancement

Most of the time social media is facing problems from fake news. People are getting mis leaded most of the time because of the fake news. Many existing works has been done on detecting fake news. This shows how it is needed to find out false news in the social media and how much this news has been generating a problem in today's world. Motivated by these facts, we had developed our model. See the comparative results obtained by all four algorithms. SVM gives a best accuracy among all those algorithms. Web scrapping enables us to be up to date. Using all four algorithms we could get better prediction and our model might be trained easy to predict by any of the four algorithms and as we had given a longer data sentences for training it gives us a good result.

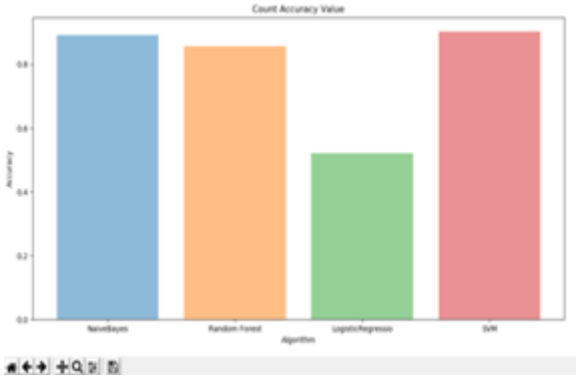

Figure 6: Comparative Result

References

[1] Fake News Detection Using Naive Bayes Classifier by Mykhailo Granik, Volodymyr Mesyura. Available: http://ieeexplore.ieee.org/document/8100379/

[2] Web Scrapping explanation. Available: https://www.webharvy.com/articles/what-is-web-scraping.html

[3] Fake News Detection by Akshay Jain, Amey Kasbe, 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Sciences.

[4] C. Dewey, "Facebook fake-news writer: „I think Donald Trump is in the White House because of me," The Washington Post, 2016.

[5] N. Newman, R. Fletcher, A. Kalogeropoulos, D. a. L. Levy, and R. Nielsen, "Reuters Institute Digital News Report 2017," 2017.

[6] N. Nic, D. A. L. Levy, and R. K. Nielsen, "Reuters Institute Digital News Report," Univ. Oxford, vol. 1, 2018.

[7] E. C. Tandoc, Z. W. Lim, and R. Ling, "Defining „Fake News": A typology of scholarly definitions," Digital Journalism, vol. 52, no. 1, pp. 1–17, 2017.

[8] S. Robinson and C. Deshano, "„Anyone can know": Citizen journalism and the interpretive community of the mainstream press," Journalism, 2011.

[9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," Proc. 25th ACM Conf. Hypertext Soc. media - HT '14, pp. 316–317, 2017.

[10] Yee and A. Yee, "Post-Truth Politics Fake News in Asia," 2017.

[11] H. Ritchie, "Read all about it: the biggest fake news stories of 2016," CNBC.com, pp. 1–9, 2016.

[12] B. Fake and N. Story, "How „the Storm" Became the Biggest Fake News Story of 2018," 2018.

[13] L. C. Hutabarat, "Ahok Jailed for Two Years," metrotvnews.com, 2017.

[14] C. A. Wijaya, "Ahok apologizes to Muslims for alleged defamation,"

[15] www.thejakartapost.com, 2017.

[16] Sebenarnya.MY, "TMJ Akan Datang Ke Econsave Pontian Untuk Belanja Pengunjung ?" 2018.

[17] Nor Azura Md Amin, "Tular TMJ ke Econsave Pontian palsu," Sinar Harian, 2018.

[18] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," Proc. 50th Annu. Meet. Assoc. Comput. Linguist. Short Pap. 2. Assoc. Comput. Linguist., no. July, pp. 171– 175, 2012.

[19] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A Stylometric Inquiry into Hyperpartisan and Fake News," no. February, 2017.

[20] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A Hybrid Deep Model for Fake News Detection," in Software, Telecommunications and Computer Networks (SoftCOM), 2012 20th International Conference on, 2017, no. SoftCOM, pp. 1–6.

[21] J. Tang, Y. Chang, and H. Liu, "Mining Social Media with Social Theories: A Survey," SIGKDD Explor. Newsl, 2014.

[22] L. Wu and H. Liu, "Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate," Proc. Elev. ACM Int. Conf. Web Search Data Min. - WSDM '18, pp. 637–645, 2018.