International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

Detecting Text Anomalies in Social Networks using Different Machine Learning Algorithms

N.Deepika, N.Guruprasad

Abstract— Today information is very powerful tool and this information is highly unreliable and wrong information can be very manipulative. It can cause harm to life as well. This in today's world is called fake news or yellow journalism. Media sometime tries to manipulate and take advantage of innocent viewers by making them believe something that is not real. This has come to publicity recently due to the US presidential election and how media was used to operate the results of the election. Sometime news channel want to garner attention hence they intentionally put this news and as it is our human tendency we always go for the bad apple and this spreads faster than a normal true news would have. The news source got what they needed but they did so by making a fool out of millions of people. So how do we make this problem go away? We train different machine learning algorithms to recognize between fake and real and we know that a machine cannot be subjective and its decision will always be fair and mathematical. Hence in this project we create an API which can tell whether given news is real or fake based of our training data.

Keywords: Machine Learning, Naive Bayes, Random forest, SVM, ANN;

I. INTRODUCTION

By using Social Media and Internet the information can be transferred from one place to other place very rapidly by anyone. This information may be true or false. Proper ways to verify the information is very popular in now a day's research. One of the major sources of false information is fake news. This is when a certain website or someone provides certain incorrect information mostly on purpose [3][4]. The disadvantage is that this "news" gathers much more interest than its true counterpart and spreads like wildfire. The news will have real consequences ranging from altered election outcomes to mob lynching where life faces threats [5]. Eradicating these fake news and finding credibility of the information has become one of the utmost important tasks for tech giants like Twitter and Face book[6] tops the list in amount of fake news as expected but countries like Turkey that are the major victims of this menace. Therefore in this paper we aim to try different algorithms to device an accurate system to detect fake information in social networks.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
N.Deepika*, Research Scholar, Department of CSE,VTU, Karnataka, India.
Dr.N.Guruprasad, Professor, Department of CSE,GAT, Bangalore, Karnataka.

Data gathering is an important stage of any research. Getting data from Social network such as Facebook is even more tedious when compared to other networks [28].

II. LITERATURE SURVEY

For developing any software Literature survey is most important. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, then next steps is to find the OS and which language. After this step we the programmers require lot of outside support. This support can be obtained from existing research works, senior programmers [16] or from different books or from many websites. Before building the system the above consideration is taken into account for developing the proposed system. We have to analysis the different techniques in Machine Learning.

A. Machine Learning

It is a division of algorithm that facilitates a software program to improve itself to become more and more accurate in predicting outcomes without actually being explicitly programmed. The fundamental concept of a machine learning algorithm is to receive data as input[1][2][3] and apply statistical analysis on them to foretell an outcome as well as updating itself for new data .The mechanism involved in this type of system is similar to that of data mining and another concept called predictive modeling. These two techniques use data set to realize patterns and improvising the program accordingly. One of the widely seen usages of machine learning is targeted advertisements [18].This is when we search something online and then minutes later we are shown ads for the same of similar products. Here the system is adjusting itself in real-time.

The system used here is called recommendation engine. Other than that we see them being used in fraud detection, spam detection and filtering, creation of personalized news feeds. There are mainly two types of algorithms [19][20]. The first one is supervised learning and the other one is unsupervised learning. Along with these we have semi supervised and reinforcement learning as well. Supervised learning requires preprocessing of the data before training the model. Data analyst analyses the data and determines which variables are important and which are not. The variable thus selected must help in the prediction. Also in supervised learning the outputs and inputs both are required to train the model. Unsupervised learning does not require expected output to be given explicitly. But the drawback of this method is it requires large sets of data. These large sets of data are used to find pattern and subtle relations between different variables.

B. Steps in ML

1. Figure out different data sets and prepare them for analysis.
2. Determine which machine learning algorithm is best suited for the data.
3. Build an analytical model based on the algorithm that we have chosen.
4. Train the model on test data and adjust [7] it accordingly.
5.Run the model to generate scores and get other findings.

Machine Learning needs thousands of data points whereas Deep Learning on the other hand requires millions of data points. Usually the return type in machine learning is numeric like a classification or a score. But Deep Learning can return any free form elements may it be text or sound. Many processing layers are used in deep learning to find the relations. In Machine Learning different algorithms are used to arrive at the expected outcome. Based on our requirement where we know the expected outcomes that is needed we are going to stick with machine learning. The number of machine learning algorithms available has no limit. Some are simple and some are very difficult to train and understand. machine learning algorithms[1][2][3] include SVM, Random Forest-Nearest neighbor, Neural networks, Decision tree, K-means, Naive Bayes, Multiple Regression, Logistic regression and soon.

C. Natural Language Processing

With the help of Artificial Intelligence, neural networks and deep learning models, NLP can examine and extract patterns in stored data to make sense of user input avoiding certain language mistakes. It is because of NLP that we can analyze large amount of information in text documents [11][12][13] to get useful information. Here we use NLP to interpret large texts to make it comparable with information from a different source. This process is done by giving weightage to certain words and phrases and with the help of machine learning .There is few solutions available and one of them is called FakeBox by Machine Learning Enthusiast Aaron Edell [16]. After different attempts he found out that the answer didn't lie in detecting the fake news but in detecting the real news. He says that "Real news is much easier to categorize. It's factual and to the point, and has little to no interpretation. And there are plenty of reputable sources to get it from"[27][16]. He categorized every data into two separate labels one real and another not real. Satire, opinion pieces, fake news, and things that were not written in factual manner were put under not real. The existing system according to him detected with an accuracy of 95%.They too have created an API can be integrated on large scale. In FakeBox you enter the title and content and it tells you whether the news is true or fake [14].

III. PROPOSED SYSTEM

In the proposed system we have used NLP and machine learning kit from scikit [6] to arrive at the best possible solution resulting in the highest accuracy. The app will be developed as an API which can be used by any other application developed, the application returns if the given context is true or false. The frontend for the app will be android. We will use different methods of vectorizers for text like count vectorizer, tdidf vectorizer etc. Different Machine learning algorithms such as Random Forest, SVC, KNN, MLP [12][13] were trained and tested with a a data set of 7456 messages which are collected from 10741 different users. The accuracies, Recall, Precision and F1-Score of all the algorithms have been compared .Among all the algorithms SVM has outperformed with 2% difference.

A. WEB APIs

API (Application Programming Interface) is a set of subroutine definitions, tools and protocols for creating applications. It is an interface that has a set of functions that allow programmers to access certain features of applications. Web API is an API [21][22][23] over web which we can access using HTTP protocols. Here we are using the RESTful API architecture which is based on representational state transfer (REST), which is an architectural style and approach most commonly used in web services.

The Bandwidth required for REST technology [24] is minimum which makes it more suitable for the internet. It is an ideal choice for building APIs that allow connection between users and cloud services. This architecture is widely used in Amazon, Google, LinkedIn and Twitter.

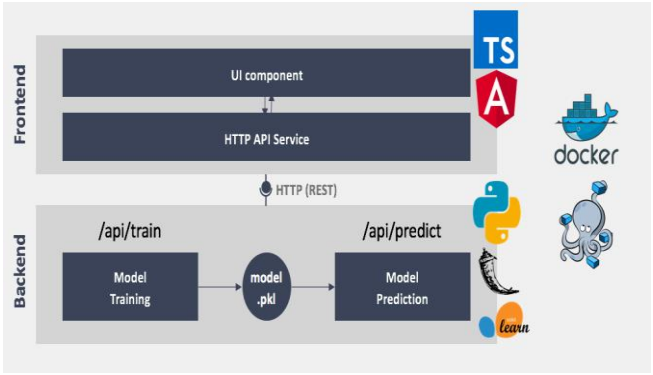

Fig. 1:Web API framework

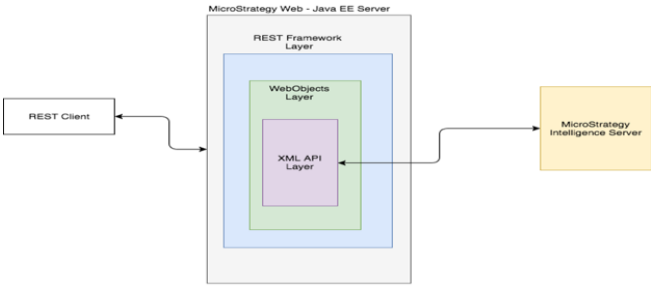

Fig. 2 :RESTful API architecture

B. WEB Scraping for Data

We all know there is a lot of information available on the internet bur you cannot go and manually add all these information. If this was for our assignments it would have been easy but because machine learning [10] requires tons of information, we have to automate this process. The process works something like, you choose the selectors like header, title, class name or id then we choose the text from it. We can provide the urls from where the information should be scraped from. It can go on a recursive loop where it keeps scrapping links after links. Web scraping is used to search the internet and gather information for our processing.

There are two web scraping options available with python. One is the beautiful soup and another is Scrapy[7][8]. They both work similarly but have their advantages. Scrapy does a better job at scrolling through web pages and beautiful soup is intelligent in recognizing incomplete html and css code. We can then save the information into a CSV file to be used later for machine learning. In scrapy you create a crawler process which you run when you want to scrape for the information.

	A	B	C	D
	title	text	label
8476	You Can Smell Hillaryâ€™s Fear	Daniel Greenfield, a Shillman Journalism Fellow at the Freedom Center, is a New York writer	FAKE
10294	Watch The Exact Moment Paul Ryan Committed Political Suic	Google Pinterest Digg Linkedin Reddit Stumbleupon Print Delicious Pocket Tumblr	FAKE
3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Monday that he will stop in Paris later this week,	REAL
10142	Bernie supporters on Twitter erupt in anger against the DNC:	â€" Kaydee King (@KaydeeKing) November 9, 2016 The lesson from tonight's Dem losses: Time	FAKE
875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners Hillary Clinton and Donald Trump are leading in	REAL
6903	Tehran, USA		FAKE
7341	Girl Horrified At What She Watches Boyfriend Do After He Lef	Share This Baylee Luciani (left), Screenshot of what Baylee caught on FaceTime (right)	FAKE
95	â€"Britainâ€™s Schindlerâ€™ Dies at 106	A Czech stockbroker who saved more than 650 Jewish children from Nazi Germany has died at t	REAL
4869	Fact check: Trump and Clinton at the 'commander-in-chief' fo	Hillary Clinton and Donald Trump made some inaccurate claims during an NBC	REAL
2909	Iran reportedly makes new push for uranium concessions in n	Iranian negotiators reportedly have made a last-ditch push for more concessions from the	REAL
1357	With all three Clintons in Iowa, a glimpse at the fire that has e	CEDAR RAPIDS, Iowa â€" â€œI had one of the most wonderful rallies of my entire career right	REAL
988	Donald Trumpâ€™s Shockingly Weak Delegate Game Someho	Donald Trumpâ€™s organizational problems have gone from bad to worse to flat-out	REAL
7041	Strong Solar Storm, Tech Risks Today | S0 News Oct.26.2016	Click Here To Learn More About Alexandra's Personalized Essences Psychic Protection Click	FAKE

Fig. 3: Sample dataset used

IV. MODEL CREATION

The procedure starts by examining data which we have gathered. This can be done using sklearn Pandas DataFrame. After through checking of data using Pandas then we have to do necessary preprocessing techniques. Once we got the DataFrame[13][15] which we intended to get, then we separate the target variables as y and predictor variables as X ,then set these as training and test datasets. For this notebook, we have used the longer texts. We have used bag-of-words and Term Frequency–Inverse Document Frequency (TFIDF) to extract features. We intentionally considered longer text hoping to allow for distinct words and features for our real and fake news data.

```
# Set `y`
y = df.label

# Drop the `label` column
df.drop("label", axis=1)

# Make training and test sets
X_train, X_test, y_train, y_test = train_test_split(df['text'], y, test_size=0.33, random_state=53)
```

A. Building Vector Classifiers

After splitting our data to training and testing data, we can start building our classifiers. To get a good idea if the words and tokens in the articles had [10][11]a significant impact on whether the news was fake or real, we begin by using CountVectorizer and TfidfVectorizer. For tfidf_vectorizer we can set Max threshold as 7 using the max_df argument. This removes words which appear in more than 70% of the articles. Also, the built-in stop words parameter will remove English stop words from the data before making vectors [18][19].Tfidf Vectorizer and count vectorizer usage in python language is shown below.

```
# Initialize the `count_vectorizer`
count_vectorizer = CountVectorizer(stop_words='english')

# Fit and transform the training data
count_train = count_vectorizer.fit_transform(X_train)

# Transform the test set
count_test = count_vectorizer.transform(X_test)
```

```
# Initialize the `tfidf_vectorizer`
tfidf_vectorizer = _____(stop_words='english', max_df=0.7)

# Fit and transform the training data
tfidf_train = tfidf_vectorizer._____(X_train)

# Transform the test set
tfidf_test = tfidf_vectorizer._____(X_test)
```

B. Multinomial Naïve Bayes

Out of the hundreds of algorithms out there we opt for the Multinomial Naïve Bayes which is more adapted for numerical values like the ones we have in our program [25][26][27]. Naïve Bayes calculates the probability of a certain outcome given the values is so and so. In other words the end result of Naïve Bayes is based on conditional probability. Naïve Bayes ignores the relationship between independent variables but still has proven high accuracy in many researches. Hence we have considered it as one of the algorithms for our research. Scikit makes it easy to implement the above algorithm.

C. Confusion Matrix

Confusion matrix helps in analyzing your machine learning algorithm. It's been shown in the Fig 4.1 and it talks about the below patterns:

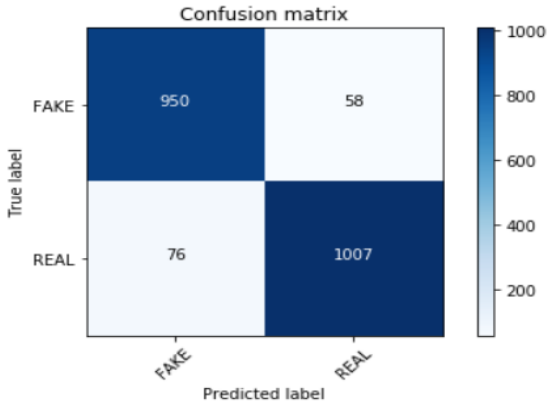

Fig. 4: Confusion Matrix

1. Your algorithm classified as true and is actually true.
2. Your algorithm classified as false and is actually true.
3. Your algorithm classified as true and is actually false.
4. Your algorithm classified as false and is actually false.

D. Support Vector Machine

SVMs can be based on classification or Regression. We have taken SVC in our proposed work [2]. It works very well with less number of hyper parameters, high dimensional space and with large amounts of datasets. In our experiments we have taken SVM as one of the algorithm [10]. We'll use y ∈ {Fake, Real} to denote the class labels and parameters w, b. For linear classifier the function is considered as $f(x)=W^Tx+b$ where W:intercept and B is bias. Scaling is very essential factor in SVC algorithm.

The aim of SVM is to get larger margin with a separating hyper plane f(x) that divides the data space into two completely different regions thus resulting classification of the input data space into two categories Fake and Real. The hyper plane can also be represented mathematically by [14]:

$$f(x) = sgn(w^T x + b)$$

where sgn() is known as a sign function, which is mathematically represented by the following equation [65]:

$$sgn(x) = \begin{cases} 1 \; if \; x > 0 \\ 0 \; if \; x = 0 \\ -1 \; if \; x < 0 \end{cases}$$

The distance of a data point x from the hyper plane is represented by the equation:

$$D = \frac{|w^T x + b|}{|w|}$$

Then the margin is given by:

$$\frac{w}{||w||} \cdot (x_+ - x_-) = \frac{w^T (x_+ - x_-)}{||w||}$$

$$= \frac{w^T \left(\left(\frac{+1 - b}{w^T} \right) - \left(\frac{-1 - b}{w^T} \right) \right)}{||w||} = \frac{2}{||w||}$$

E.Random Forest Classifier

RF is an ensembling classification algorithm which creates sub trees with different features, and different parameters in each tree being included. RF is a great choice when we have missing values and outliers in our dataset [8][9]. The data which we gathered from Facebook will definitely have missing values during scraping process, we thought of choosing RF will be an ideal thought. To find out the parameters Random Forest won't require cross-validation because it has built-in accuracy estimation. The parameters considered for RF in training and testing phases are number of estimators as 500, max depth as 300 and maximum number of features as 200. We have understood that fine tuning of parameters in any algorithm is very important to increase the accuracy score of the model.

V. RESULTS AND DISCUSSION

In our experiments we have taken a data set of 7796 spam and ham messages from Facebook from 10,500 users and applied various supervised algorithms such as Naive Bayes classifier, Random Forest Classifier, Support vector Classification and MLP classifier and obtained results as given in the above table. Among all the classifiers SVC outperformed. The metrics used to analyze the performance are Recall, precision and F-measure. The below table compares the performance to all the algorithms.

Table- I: Accuracy, Recall, Precision and F –measure of all the four classification algorithms

Method	Metrics				
	Accuracy	***wrongly classified instances***	***Recall***	***Precision***	***F-measure***
Naive Bayes	90%	250	86%	87%	87%
Random Forest	95%	164	94%	96%	95%
SVC	98%	108	97%	99%	98%
MLP Classifier	95%	157	94%	98%	96%

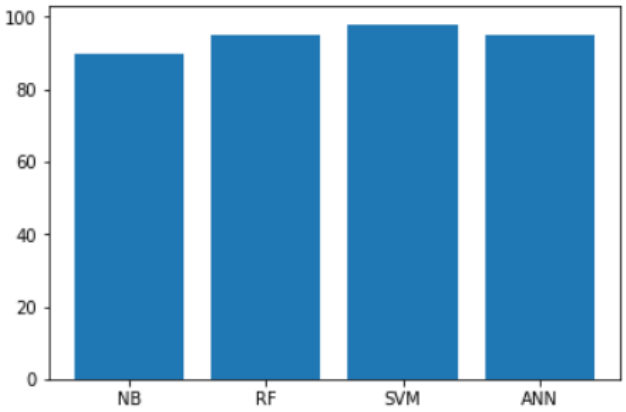

Fig. 5: Analyzing Accuracies of all the algorithms

VI. CONCLUSION

In this research work various supervised machine learning algorithms have been implemented on 7000 plus text messages from social networks and identified malicious and real text messages. The performance of all the chosen algorithms have been compared with various metrics such as Recall, Precision and F1-score, SVM has outperformed all the remaining algorithms with an accuracy difference of 3%.

REFERENCES

1. Tri Doan, Jugal Kalita,"Selecting Machine Learning Algorithms using Regression Models", 2015 IEEE 15th International Conference on Data Mining Workshops.
2. Al-Zoubi Ala'M, Hossam Paris, et al. "Spam profile detection in social networks based on public features", Information and Communication Systems (ICICS),2017 8th International Conference on, pages 130–135. IEEE, 2017.
3. Nattanan Watcharenwong and Kanda Saikaew," Spam detection for closed facebook groups", Computer Science and Software Engineering (JCSSE), 2017 14th International Joint Conference on, pages 1–6. IEEE, 2017.
4. Hailu Xu, Weiqing Sun, and Ahmad Javaid. "Efficient spam detection across online social networks.", Big Data Analysis (ICBDA), 2016 IEEE International Conference on, pages 1–6. IEEE, 2016.
5. Ghada Zakaria," Detecting Social Spamming on Facebook Platform"2017.
6. Z. Weifa, "A SVM Text Classification Approch Based on Binary Tree," 2009, pp. 455–458.

7. https://www.python-course.eu/neural_networks.php
8. https://www.python-course.eu/Random_Forests.php
9. https://www.python-course.eu/Decision_Trees.php
10. https://www.pythonourse.eu/k_nearest_neighbor_classifier.php
11. Margaret Rouse ,(May 2018)"Machine Learning", https://searchenterpriseai.techtarget.co/definition/machine-learning-ML
12. Shlok Gilda, "Evaluating machine learning algorithms for fake news detection", in IEEE 15th Student Conference on Research and Development (SCOReD),2017
13. Sunil Ray,(September 2017) , "Essentials of Machine Learning Algorithms (with Python and R Codes)" ,Retrieved from https://analyticsvidya com
14. Jonathon Webster,(November 2017) , "What is Natural Language Processing (NLP), How Does it Work, and How Can Businesses Use NLP?", Retrieved from
15. https://techgenies.com
16. Aaron Edell,(January 2018) , "I trained fake news detection AI with >95% accuracy, and almost went crazy" ,Retrieved from https://towardsdatascience.com
17. Tristan Greene, (September 2018), "This fake news detection algorithm outperforms humans", Retrieved from https://thenextweb.com
18. Jason Brownie, (April 2014), "A Gentle Introduction to Scikit-Learn: A Python Machine Learning Library", Retrieved from https://thenextweb.com
19. "Comparing Python to Other Languages", Retrieved from https://python.org on December 2 2018.
20. "What is Python? Executive Summary", Retrieved from https://python.org on December 2 2018.
21. Public, (December 2018), "Flask (web framework)", Retrieved from https://en.wikipedia.org
22. Public, (December 2018), "Create, read, update and delete", Retrieved from https://en.wikipedia.org
23. "What is Web API? ", Retrieved from https://tutorialsteacher.com on December 3,2018
24. "REST API Architecture", Retrieved from https://lw.microstrategy.com on December 3,2018
25. Mykhailo Granik,Volodymyr Mesyura, "Fake news detection using naive Bayes classifier", in 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON) ,2017
26. Saranya Krishnan, Min Chen, "Identifying Tweets with Fake News", IEEE International Conference,2018
27. Shivam B .Parikh,Pradeep K Afrey "Media-Rich Fake News Detection: A Survey",IEEE Conference on Multimedia Information Processing and Retrieval (MIPR),2018 .
28. N.Deepika, Dr.N.Guruprasad "Anlysing different Machine learning approaches for malicious content detection in social media:A survey", IJPAM, 2018 .

AUTHORS PROFILE

N.Deepika is currently pursuing her PhD in the domain of Machine Learning at Visveswaraya Technological University, Karnataka. She obtained her Masters from Jawaharlal Nehru Technological University Hyderabad in the department of Computer Science and Engineering and Bachelors from Sri Venkateswara University , Computer Science of Engineering, Thirupathi. She has more than 18 years of experience in various Engineering colleges in different states of India. She is a very enthusiastic researcher who has passion to explore different machine learning algorithms in different domains. Her areas of research interests include Data Mining and Data Warehousing, Machine Learning, Computer Networks , Algorithms, Natural Language Processing etc.,

Dr.N.Guruprsad is currently working as a Professor the department of CSE at Global Academy of Technology, Bangalore. The author has more than 24 years of teaching experience in different Engineering colleges in India. He has more than 50 publications in National and international journals to his credit. Author also have more than six books published in different computer science related subjects. He is also a senior life member of various professional societies. He is a profound speaker at different Engineering colleges and industries His areas of interests include Data mining, Operations Research, Machine learning, Computer networks, Algorithms, Data structures etc.,