**FAKE NEWS DETECTION**

A project report

Submitted by

Niranjan Naidu V-1700333C204

Kavyasree Chatakonda- 1700301C204

Jahnavi Angati-1700317C204

Batch 2017-ECE

Course

MACHINE LEARNING



SCHOOL OF ENGINEERING & TECHNOLOGY

BML MUNJAL UNIVERSITY

DECEMBER 2020

**ACKNOWLEDGEMENT**

This project that is detecting fake news has given us a great opportunity to gain good practical experience. By doing this project we learned new topics not only theoretically but also practically. This project gave us a good opportunity to build models or implement different classification algorithms to gain practical exposure.

We sincerely thank our course faculty **DR.Shuvabrata Bandopadhaya**  to give this opportunity for doing this project and also for helping out whenever his assistance is needed even when he is busy with their duties.

**ABSTRACT**

Late political occasions have led to an expansion in the notoriety and spread of counterfeit news. As shown by the across the board impacts of the huge beginning of Fake news, people are conflicting if not by and large poor finders of Fake news. With this, endeavour's have been made to mechanize the procedure of Fake news location. All things considered, the objective of this undertaking was to make an apparatus for recognizing the language designs that portray Fake and genuine news through the utilization of Machine learning. The after effects of this undertaking show the capacity for Machine learning to be valuable right now. We have constructed a model that gets numerous instinctive signs of genuine and fake news just as an application that guides in the representation of the order choice.

**KEYWORDS -**

Accuracy, confusion matrix, Euclidean distance, f1 score, K-nearest neighbour, Logistic regression, Machine learning, Manhattan distance, passive aggressive classifier, Random forest, Radial basis classifier, Support vector classifier, Tf-IDF vectorizer

**INTRODUCTION**

The ascent of fake news during the 2016 U.S. Presidential Election featured not just the risks of the impacts of fake news yet additionally the difficulties introduced when endeavouring to isolate fake news from genuine news. Fake news might be a moderately new term however it isn't really another phenomenon. Fake news has in fact been around at any rate since the appearance and prevalence of uneven, factional papers in the nineteenth century. None the less, propels in innovation and the spread of news through various kinds of media have expanded the spread of fake news today. Accordingly, the impacts of fake news have expanded exponentially in the ongoing past and something must be done to keep this from proceeding later on. The people are moreover are addicted to social media and will be seen different types of news about politics, sports, films and many more but as we seen we are getting different types of news in which majority are fake news than real news. The fake news is becoming more popular than the real news and many channels and many pages are also spreading fake news because of their spreading and for their popularity. Nowadays all of us are more concentrated on fake news as it is becoming more popular.

The technique we are using to solve this problem (distinguish fake news) is by using the classification algorithms on the dataset. The classification algorithms classify the dataset into the genuine and fake. The classification algorithms we are using in this project are logistic regression, random forest algorithm and passive aggressive classifier.

**LITERATURE REVIEW**

For our project we had made a background check mainly about what is fake news and what are the algorithms we need to use in order to construct our project. And these are the papers we checked in order to background check

1. <https://www.ijeat.org/wp-content/uploads/papers/v9i1/A2633109119.pdf>

Here is one of the papers where we found the fake news detector is used based on the logistic regression in this model. They have used that logistic classifier in order to distinguish the fake news from the real news. Logistic regression has a good predictive value so it is best used in solving binary classification

Logistic regression works for short as well as long input text. And the accuracy by using the logistic regression in in the range of 79% - 89% The TF-IDF is a best feature to use on text pre-processing in the data, it is used as feature for the text- vectorizer so that text of strings can be changed into the required format which help to suit in the data.

2. <http://dspace.library.uvic.ca:8080/bitstream/handle/1828/8796/Ahmed_Hadeer_Masc_2017.pdf?sequence>(university of victoria)

In this paper they used N-gram Analysis, here the advantage is that the N-gram method will work on multiple types of the fake content. The reason for that is it relies on the text of the fake content and it ignores the behaviour of the spammers. It can detect the fake news, fake articles, fake journals with best efficiency.

The accomplished exactness delineates that fake substance will utilize unexpected words in comparison to genuine content clients. Moreover it had the option to recognize counterfeit substances composed by various clients and fake substances composed by the same clients.

By looking down these papers we are trying to build a model that will be better than those techniques that are used in the above two papers.

N-gram

It is a method of predicting the n-th word from the n-1 words. If you are predicting the third word from the previous two words it is a trigram. If you are predicting the second word from the first word it is called a bigram.

Example –

He is going to \_\_\_\_\_\_\_ .

In the above example we are predicting the last word from the previous 4- words, this is pentagram.

Bigram example -

Let us take some random text

<S> I am a human</S>

<S>I am not a stone</S>

<S>I I live in Lahore</S>

Let us find the probability of the sentence “I I am not”

P(I I am not)=P(I/<S>)\*P(I/I)\*P(am/I)\*P(not/am)

=(3/3)\*(1/4)\*(2/4)\*(1/2)

= 0.065

As we are using conditional probability and multiplying all of them this will surely produce smaller results to get at higher values as the representation of the probability we use a mathematical method called as perplexity.

Perplexity-> pp(W)=P(W1.W2.----W3)-1/N

N – no.of words.

pp(W)=(0.0625)-(1/4)

=2

Perplexity and the probability are inversely proportional to each other.

The problem in this method is

When we take the same sentence from the data set surely it will have more probability.

P(I am a human)=P(I/<S>)\*P(am/I)\*P(a/am)\*P(human/a).

= 3/3\*2/4\*1/2\*1/2

= 0.125

Now let us find the probability of a sentence just by removing a word from the above sentence that is ‘a’

P(I am human) = P(I/<S>)\*P(am/I)\*P(human/am)

= 3/3\*2/4\*0/2=0

As we removed just one word from the sentence we got the probability as 0

To remove this problem we use laplace smoothing.

The procedure of this is simply adding 1 to all the conditional probabilities so the zero will not occur.

P(X/Y)=c(XY+1)/c(Y)+V

V=vocabulary (count of distinct words)

(3+1)/(3+11)\*(2+1)/(4+11)\*(0+1)/(2+11) = 0.00439

Due to this problem we used some other algorithms to detect the fake news.

**METHODOLOGY**

Model Evaluation

In evaluating a model, the produced output of the model is measured. The performance of the model is checked by using some performance metrics like F1 score, accuracy rate and precision which is produced by the confusion matrix.

Model Integration

Here as we are not having any idea in keeping our model in web portal, if possible we will try to build a model with all specifications and the model is passed through API services

Model Management

In the information board stage, a lot of information (dataset) is gathered which is a set of news stories. Once the dataset is gathered, the information is investigated to show signs of improvement information on its structure and significance. The dataset is then should have been purified which implies so as to make it helpful, through vectorization procedure or information pre-preparing stage. TF-IDF is utilized right now to convert every one of those sentences of articles into an organized arrangement. At the point when it is washed down, the information is going to be set up for stacking into programming condition. At long last, the information will be isolated into training and approval subsets.

Model Training

After the information is appropriately arranged, the ML model is fit to be prepared. In this step all the ways need to be considered and the learning task is decided which is a prediction task, Available training dataset is then studied. At that point, a proper algorithm is chosen to prepare the model. In this model we take many algorithms from those we will choose the algorithm which gives highest accuracy.

Model development

This is the last step,in this step we will apply our algorithm to the new data and the results are taken and observed clearly the model can still be modified if necessary.

These are the basic steps that are followed to complete any project. Now let us see how we implemented our project.

**Step-1**: We need to collect the data and train the models, so that we can detect if the news is fake or not. And we got the data from internet resources.

**Step-2:** Now we need to do data preprocessing as we can only numericals on any classification algorithms but our news is of English text so we need to apply data pre-processing and the technique used is Tf-IDF vectorizer.

Now let us know what is TF-IDF:

Tf-IDF means term frequency-inverse document frequency. And this is used as a weighting factor for the features that means it is mainly used for converting text that is in English to numbers. And this can be implemented using sklearn library and the method used is tf-idf vectorizer. Generally the weight increases as if the words are more repetitive, Now let us see the working of tf-idf.

**Tf-idf(t,D) = tf(t,d)\*idf(t,D)**

tf(t,d) = frequency of given term in the document

idf(t,D)=log(N/|{d belongs D : t e d}|)

Here TF= term frequency.

IDF = inverse document frequency

t = term or word

d = document

D =Data set

N = no of documents

For example, consider two documents.

D1- the sky is blue.

D2- the sky is not blue.

N = 2

|  |  |  |  |
| --- | --- | --- | --- |
| terms | Tf | Idf | Tf-idf |
| the | D1 D2  1 1 | 0 | D1 D2  0 0 |
| sky | 1 1 | 0 | 0 0 |
| is | 1 1 | 0 | 0 0 |
| Blue | 1 1 | 0 | 0 0 |
| not | 0 1 | 0.301 | 0 0.301 |

By the above technique we can convert the English words into numerical.

Step-3: Now we need to split the data into two parts so that we can test the model that is trained and fitted. For that purpose, we use the train test split method and for this we need to import the sklearn library. In our project

We partitioned 70 percent of data for training the model and 30 percent for testing the model so that we can check the accuracy of the model.

TOTAL DATA SET

|  |  |
| --- | --- |
| Training data - 70 percent | Testing data – 30 percent |

Step-4: As our project is to classify the fake news and real news, we need to apply different classification algorithms. In our literature review we observed that the fake news detection is classified by using only one technique or one algorithm. But in our project, we implemented five different algorithms and chosen the best algorithm which has high frequency.

The implemented algorithms are as follows.

1. Logistic Regression.

2. Random forest classifier.

3. K- nearest neighbour classifier

4. Support vector classifier and its types.

5. Passive aggressive classifier

1. **Logistic Regression:** It is a machine learning algorithm generally used for binary classification. The working of logistic regression is first we apply linear regression on the data and the line is passed through a sigmoid function which gives the probability of an event and we set a threshold like p = 0.5 and then if p<=0.5 then it belongs to class 0 and if p>0.5 it belongs to class 1. By this technique we can classify whether the news is fake or real. And the function used is sigmoid function is as below.

**P(y) = e^(β0X) / (1+ e^ (β0X))**

2. **Random forest classifier:** Random forest classifier is an ensemble technique that means combining multiple models. And we use bagging (bootstrap aggregator) for classification or regression using random forest algorithms. To implement a random forest classifier, we need to import sklearn.ensemble library. Now let us see the working of a random forest classifier. We use many models to train the data and get the result and these models are called base learners. Let us consider a dataset D with m columns and d records or rows and from the whole data we use row sampling and feature sampling technique and extract and pass as data to base learners in case of random forest the base learner used is decision tree let the data passed to models is d’.

The conditions need to be followed are d’< D. Similarly we do row sampling and feature sampling with replacement that means d’ is not equal to the data sent to the next model. This is done for the number of models we will select. This process is known as boot-strapping and the model gets trained and gives the output that means if there are 4 models namely m1,m2,m3,m4.

As our problem is a binary classification problem, we get the output for each base learner as 1,0,1,1 respectively. Now we use the majority vote to give a final result that means the highest frequency will be the result of the random forest algorithm in the above example 1 is the output of the algorithm (for one record).

**3. K nearest neighbour:** K nearest neighbour is an supervised algorithm mainly used for non-linear data. And it is used for both classification and regression purposes. Now let us know the working of knn. We consider dataset D that has 2 categories they are real or fake and then we select a k value so that they are no of neighbours for an instance if we select a data point and then we calculate the distance between the point and the k nearest neighbours where the k value(odd) is pre-defined. And the distance can be calculated using two methods they are

1. Euclidean distance

2. Manhattan distance

1. Euclidean distance - the distance between two points will calculated using the formula – sqrt((x2-x1)^2-(y2-y1)^2)
2. Manhattan distance - the distance between two points can be calculated using an intermediate point that forms right angled triangle and the distances summed for final distance

Any one of the techniques can be used to calculate the distance between all the neighbours and datapoint these are done according to k-value.

For suppose if k=5 if 3 points are in one category and 2 points are in another category so that point belongs to category with 3 point as it has max points

**4. Support vector classifier:** Support vector classifier is used for classification of two classes. It is basically used to identify the point extensively that means it tells which class datapoint belongs to when the datapoint is complex. That means for example if we want to classify a dog or cat but the point or the object is cat but has the features of a dog or it is a dog but has the features of a cat. This is the

Algorithms used to overcome the problem. this algorithm works in multi-dimensional also and it used for both linear data and non-linear

We used three types of support vector classifiers in out project they are

1. Linear support vector classifier

2. Poly support vector classifier

3. Rbf support vector classifier

**Linear support vector classifier:** If the data is linearly separable we use this method that means the data is separated using a hyperplane using a linear function(dot product) and two vectors are used as a support so that the region is used to classify so complex points like the above example of cat and dog. The function that is basically used is

**K(x ,y)= x\*y x,y= vectors**

**Poly support vector classifier**: This type of svc is used for non-linear separable data and kernel function is used to train the model and fit non-linear decision boundary for higher dimensional data. Basically, this kernel function works like dot product of vectors to transform non-linear data to linear data

**Kernel function- K(x,y)=(x\*y+1)^d**

**Radial Basis Function support vector classifier:** This algo is similar to poly svc but the only change is the kernel function and the kernel function used is

**Kernel function- K(x,y)=e^(-(x-y)^2/2(s.d)^2))**

**5. Passive Aggressive classifier:** It is an online algorithm used for classification purposes. These algorithms will be passive for correct classification but for any wrong classification it will become aggressive. And it updates the wrongly classified data but updating or changing the weight vector.

Step-5: We need to construct a confusion matrix for every algorithm. This is used for knowing the no of classifications or no of mis-classifications. The structure of confusion matrix is as below

|  |  |  |
| --- | --- | --- |
|  | Predicted yes | Predicted no |
| Actual yes | True positive | False negative |
| Actual no | False positive | True negative |

Step-6: To measure the performance we have to produce classification report for every algorithm. So that we can know which one is the best algorithm. The properties we get from classification report are precision, accuracy, recall, F1 score and many more but the above are the important features.

**PERFORMANCE**

In this part of the report let us see the performance of every classification algorithm. And the performance metrics used are

1.) Accuracy

2.) precision

3.) Recall

4.) F1-score

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| algorithms | accuracy | precision | | recall | | F1-score | |
| Logistic  regression | 91 | fake | real | fake | real | fake | real |
| 0.94 | 0.89 | 0.90 | 0.93 | 0.92 | 0.92 |
| RFC | 90 | 0.91 | 0.91 | 0.92 | 0.90 | 0.91 | 0.91 |
| PAC | 93 | 0.93 | 0.93 | 0.93 | 0.92 | 0.93 | 0.92 |
| Linear SVC | 93 | 0.94 | 0.92 | 0.92 | 0.94 | 0.93 | 0.93 |
| Poly SVC | 85 | 0.97 | 0.72 | 0.78 | 0.96 | 0.87 | 0.82 |
| RBF SVC | 92 | 0.95 | 0.90 | 0.91 | 0.95 | 0.93 | 0.92 |
| KNN | 65 | 0.59 | 0.98 | 0.99 | 0.28 | 0.74 | 0.44 |

Table of performance metrics

**Bar graphs for performance metrics**

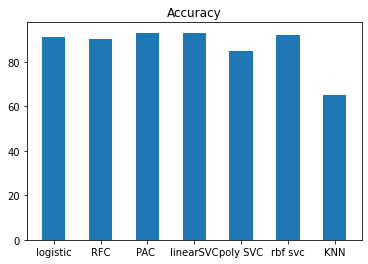


Fig: bar graph for accuracy for all algorithms

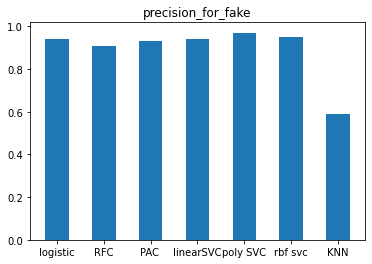


Fig: bar graph for precision for fake news of every algorithm

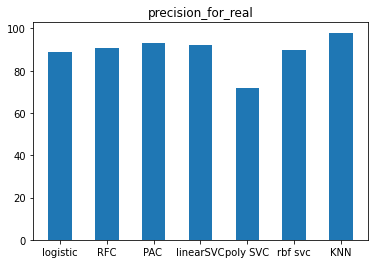


Fig: bar graph for precision of real news for every algorithm

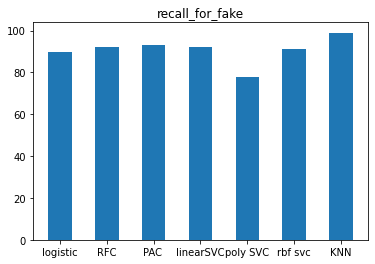


Fig: bar graph for recall of fake news for every algorithm

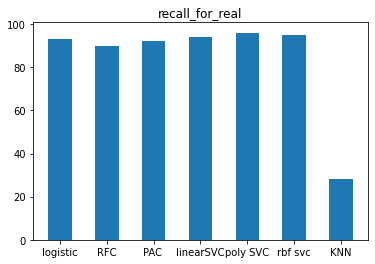


Fig: bar graph for recall of real news for every algorithm

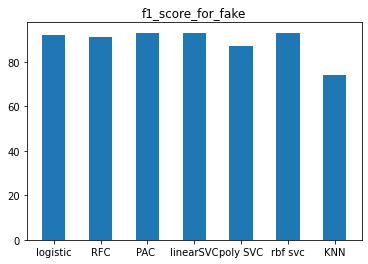


Fig: bar graph for f1-score of fake news for every algorithm

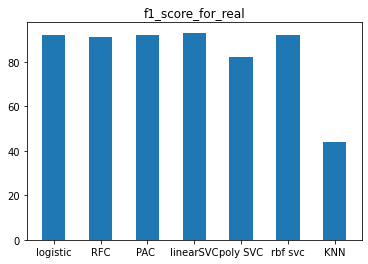


Fig: bar graph for f1-score of real news for every algorithm

**RESULTS**

From the performance metrics we can say that the passive aggressive classifier and the linear support vector classifier are best classification for detection of fake news as the accuracy of both the algorithms are high. And remaining algorithms like logistic regression, random forest, poly-svc, Rbf svc are shows good accuracy. The knn algorithm gave very less accuracy when compared to other algorithms.

**CONCLUSION AND FUTURE SCOPE**

Huge spreading of fake news through the net will convey terrible effects on the general public. Fake news will misdirect perusers also, bamboozle them to a definitive disarray in accepting some falsehood to be valid. This is the peril of fake news as individuals these days despite everything can't separate between fake news and genuine news in their every day an existence with their unaided eyes. Be that as it may, this issue can be absolutely tackled by bridling the intensity of machine Learning out how to anticipate news to be fake or not. Inside this ability, in any event will help individuals to be increasingly mindful on the news they acquired by checking it

In our future work, we might want to consolidate in our model, factual highlights, and highlights that mirror the author's styles, for example, the quantity of slang words or filler word in the content. We might want to examine the impact of including the extra highlights our model, will the precision and accuracy of the model change and in what capacity will it change.

There is absence of precisely marked genuine world datasets. In this way we might want to investigate unaided and semi-managed strategies to recognize misleading substance, even despite the fact that the accessible model that applies solo and semi-managed strategies are incapable to coordinate the presentation of managed learning strategies. I accept that looking into the unsupervised learning approach merits investigating. Audits and articles are an essential wellspring of data. There is a huge measure of information in the genuine universes. We might want to apply our models on large information accessible, to examine how it will perform. Besides, given that high dimensionality issue will emerge we might want to investigate the impact of highlight choice procedures, for example, data additions and chi-square, in our future work.

References

Dspace.library.uvic.ca. 2020. [online] Available at: <http://dspace.library.uvic.ca:8080/bitstream/handle/1828/8796/Ahmed\_Hadeer\_Masc\_2017.pdf?sequence> [Accessed 25 November 2020].

Ijeat.org. 2020. [online] Available at: <https://www.ijeat.org/wp-content/uploads/papers/v9i1/A2633109119.pdf> [Accessed 25 NOVEMBER 2020].