Fake News Detection Using Machine Learning Final Project Report

Ujjwal Yadav
2018CSB1127(IndividualProject)
IIT ROPAR

1 Introduction

Fake News are false news stories ,often of a sensational nature , created to be widely shared or distributed for the purpose of generating revenue or promoting a political movement ,public figure or a company.It is very important to recognise fake news as fake news spread rumours which may have some serious consequences. In Fake News detection we attempt to develop a model which would be trained on some dataset which would help it to predict a test news dataset as fake or real based on its training data.

My main aim is to build a model to accurately classify a piece of news as REAL or FAKE using Machine Learning Techniques.We will train our model with the training set that is randomly partitioned from the overall dataset.Based on our training model,we would predict whether our test dataset to predict if a given news is 'REAL' or 'FAKE'.
INPUT and OUTPUT :A Dataset obtained from experimentation on various real world true and fake news dataset containing news text and titles will be used as an input to train the model.A part of this dataset will be used to test the model which will predict if a news element is real or fake.
This is the link to the Dataset(news.csv) that will be used: Dataset

2 Literature Survey

Here we use the news.csv dataset which contains 7796 rows and 4 columns.The first column identifies the news , the second column is the title , third column is the text (descripton) of the title and the fourth column contains the label as fake and real for the news text.Generally it is extremely difficult to deal with the unprocessed natural language dataset , which makes it necessary to use Natural Language Processing Techniques , which help in converting the english or any other language data in mathematical form.One such Technique is TF-IDFVectorizer, which consists of two techniques TF and IDF techniques. TF stands for Term Frequency ,which means that we find all the words of the language occuring in the document ,except STOP words which are less important in determining the meaning and understandinfg of the text , and find the frequency of occurance of each term in every document divided by the total number of terms in the document .IDF stands for Inverse Data Frequency which refers to the log of the number of documents divided by the number of documents that contains a particular term x.The multiplication of the TF and IDF terms is stored as the processed data for every training data text and every term present in the document. We then employ PASSIVE AGGRESSIVE CLASSIFIER to train the model with the processed dataset . PassiveAggressiveClassifier is an online algorithm in which the training data points come sequentially one by one and the model makes prediction of the label based on the current hypothesis , the model then recieves the actual label and it tries to make the cumulative loss smaller by changing and modifying the current hypothesis to give a new hypothesis for the next data point.

The algorithm works in many rounds.On each round the algorithm observes an instance and predicts its label to be either +1 or -1.After the prediction is made, the true label is revealed and the algorithm suffers an instantaneous loss which reflects the degree to which the prediction was wrong .At the end of each round the algorithm uses the newly obtained instance-label pair to improve its prediction for future rounds.

Passive means that the algorithm turns passive when the predicted result matches with the actual result but turns aggressive and makes significant and suitable changes when the model encounters wrong prediction.There are other models like Logistic Regression Model, which can be used instead of Passive Aggressive Classifier, but the best results obtained for this dataset are from the Passive Aggressive Classifier Method.The working of this algorithm is as follows:

Now we describe the symbols used to represent the algorithm as follows. At round t we select an example from an independent and identical distributed random variable x(t), with label y(t)belonging to {-1,+1}.Based on the loss suffered we compute the update rule Q and update the weight w at every round until we reach the final index. PA algorithm aggressively makes an update whenever the loss is non-zero(even if the classification is correct).

INPUT : X, y , C>=0, where C is the positive parameter to balance the trade off between passiveness and aggressiveness

OUTPUT: w, where w is the weight update

```
begin
    w <- w0 ;
    for t=1,2,3,....N do
        recieve instance : x belonging to R(N);
        predict <- sign(w(t).x)
        correct label: y(t) = {-1,1 };
        loss =l(t) <- max {0,1-y(t)(w(t).x)};
        compute Q(t);
        update: w(t) <- w + Q.y(t).x(t);
    end
end
```

where Q(t) takes following different forms:

1. Q(t) = l(t)/(||x(t)||^2)

2.$Q(t) = min(C, (l(t)/(||x(t)||^2)))$

3.$Q(t) = l(t)/(((||x(t)||^2) + (1/(2.C)))$

-> On analyzing the dataset, we find that most of the political blame games during election times are fake news , like the given dataset contains many news about the 2016 US Presidential Election which are mostly fake news in which the candidates are accusing each other for different subjects to make their own profit.

-> While most of the scientific news about discoveries and inventions are found to be real and they mostly deal with the advancements in the science and technology.

-> It is also found that most of the news related to serious crimes are real news.

-> Many social media news are fake news which are just attention seekers and in real they do not relate to anything meaningfull.

3 RESULTS

The dataset was randombly classified into training and test data in 4:1 . The model was trained using the training dataset and it was tested using the testing dataset.The foolowing results were obtained:

-> The accuracy of the model came out to be around 92% .

-> The confusion matrix obtained is as follows :

591	47
44	585

->The above Confusion Matrix shows that the number of ’True Positives’ are 591 which means that 591 of 1267 test data points are correctly classified as positives(Real)
->The number of ’True Negatives’ are 585 which means that 585 of the 1267 test data are correctly classified as negatives(Fake News).

->The number of ’False Positives’ are 44 which means that 44 of the 1267 test data points which were actually fake news are predicted as real news.

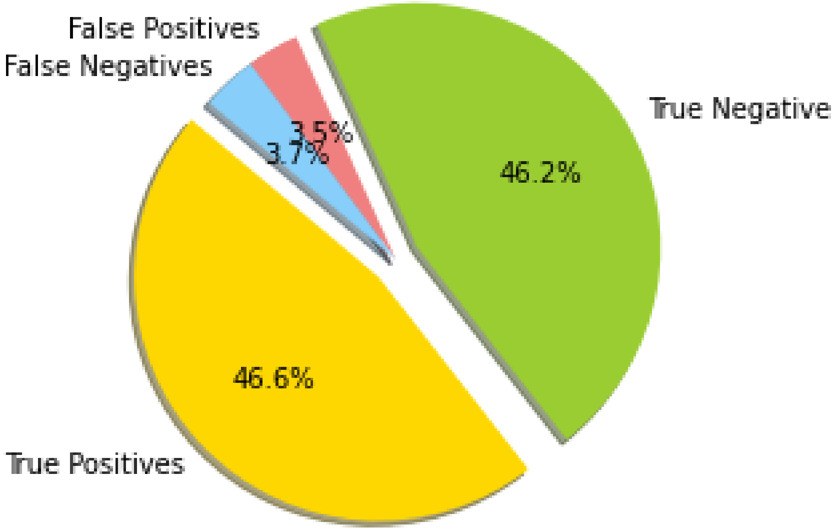
->The number of ’False Negatives’ are 47 which means that 47 of the 1267 test data pointa which were actually real news are predicted as fake news by our model.

-> The Classificatio Report is given as follows:

	precision	recall	f1-score	support
FAKE	0.93	0.93	0.93	638
REAL	0.93	0.93	0.93	629
accuracy			0.93	1267
macro avg	0.93	0.93	0.93	1267
weighted avg	0.93	0.93	0.93	1267

-> The higher values of Precision , Recall and Accuracy (around 93% each) signify that the above model works very well for our given dataset and predicts the news label as 'FALSE' and 'TRUE' with great accuracy.

-> The Pie Chart showing the distribution The Confusion Matrix is plotted below:



4 Work After MID-SEM

In the previous sections , we have used the PASSIVE AGGRESSIVE CLASSIFIER to train the model with the processed dataset with help of TF-IDFVectorizer.Such a a dataset has a huge number of columns , i.e , the number of features of each data sample are huge , which make the model very slow and there is a case of Over-fitting , with a number of similar kind of terms , which we use as an indicator to determine if a piece of news is real or fake. Thus we use the TruncatedSVD Algorithm to perform Dimenionality reduction , which is sort of similar to PCA algorithm which we have studied in our Class, except that it can be operated directly onto sample vectors,instead of Covariance Matrix , which is required for PCA. SVD stands for Singular Value Decomposition , which indicates that TruncatedSVD uses performs Linear Dimensionality reduction with the help of truncated Singular Value Decomposition.We perform Dimensionality Reduction on our TF-IDF Vectors for every sample text example , with the help of Truncated SVD algorithm.The original dataset returned by tfidf Vectorizer contains 67,351 terms of the text of dataset of news , which helps us in determining weather a news is fake or real buy using the term count or frequency information.We use a RANDOMIZED SVD Solver algorithm , which demands the number of iterations (n_iter) as parameter for functioning.We also give the n_components as input to the Truncated SVD for obtaining the data in desired number of dimensions.First we give the number of dimensions in the Truncated SVD to be equal to 100.We obtain the total data variance explained with the use of 100 components.

ALGORITHM:

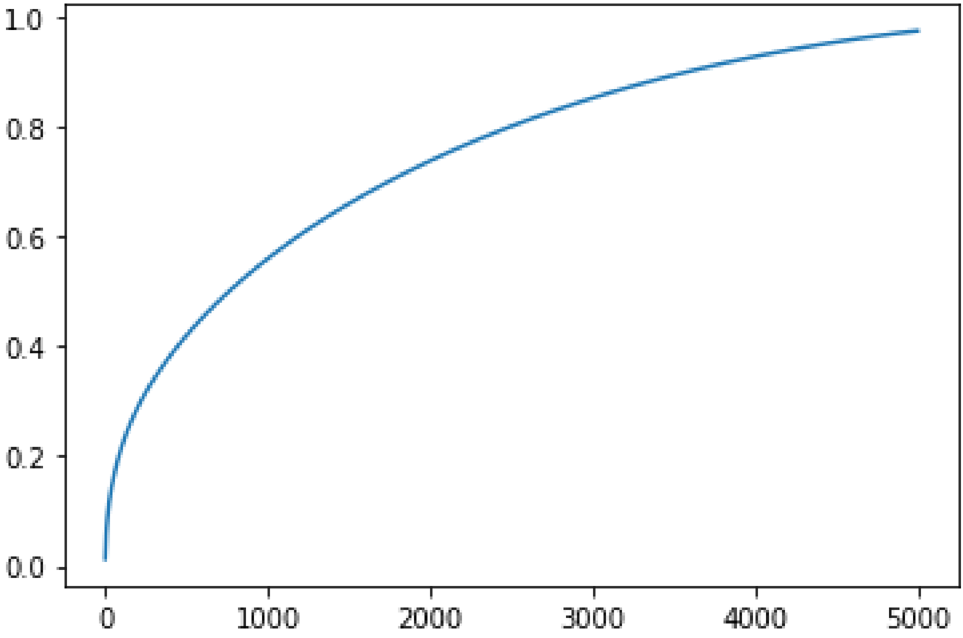
->Singular Matrix Decomposition is a Matrix Decomposition algorithm to reduce a matrix into its smaller parts for making our calculations simple and fast.
Let us take a matrix D (m*n)which is represented as:

$$D = L.S.R^T \tag{1}$$

->Here S is a diagonal matrix (n*n) , which contains singular values of D.The singular values decrease as we go down the matrix, i.e, the importance of the term,corresponding to a singular value in matrix S, decreases in determining the news as fake or real , as we move down the S matrix.
->L is a m*n matrix , whose each column represents left Singular Vectors of D.
->R is a n*n matrix whose columns represents right Singular Vectors of D.It's transpose is taken , which makes its rows as the Right Singular Vectors.
->SVD algorithm , when applied on a data matrix of dimensions m*n, decomposes or breaks it down into the 3 matrices L ,S AND R as mentioned above.
->Truncating refers to shortening or reducing something.Similarly,Truncated SVM reduces the dimensions of the data matrix by making all the singular values of matrix S zero, except the first k values , where k indicates the number of dimensions into which we want our dataset to reduce to.We also choose first k Left Vectors(k columns) of L Matrix and the first k vectors of R(first k rows of R^T),*thus giving us reduced matrix J.*

$$J = L(k).S(k).R^T(k) \tag{2}$$

->The graph showing the variation in explained variance by choosing different number of components in TruncatedSVD is as follows:



->The total explained variance by the use of 2500 components in Truncated SVD is around 79.8%.

->We apply Passive Aggressive Classifier (as discussed above in the report)on our new dataset , with reduced dimensions and compare the results with the previous results.

->The accuracy in prediction for the new dataset is around 91% , which is close to 93% ,the accuracy of the model with the original dataset without feature reduction.

->Thus features extraction help us in avoiding overfitting and also reduces the computation time and effort significantly.

->For this dataset we had around 67,000 features before feature extraction , but after feature extraction we can explain almost 99% of the data variance in around 5000 features.

->The Classification report of the model with the features reduced to 2500 by using Truncated SVD and applying Passive Aggressive Classifier is given below:

	precision	recall	f1-score	support
FAKE	0.92	0.91	0.91	638
REAL	0.91	0.92	0.91	629
accuracy			0.91	1267
macro avg	0.91	0.91	0.91	1267
weighted avg	0.91	0.91	0.91	1267

->We also obtain the Confusion Matrix which is as follows:

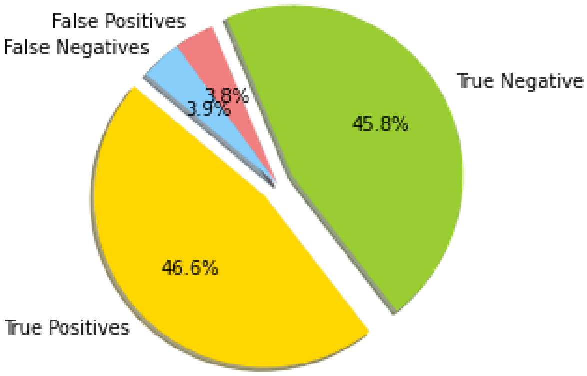
578	60
53	576

Observations:

->It shows that 578 of the 1267 news that were actually real are predicted as 'REAL',576 fake news are actually predicted as fake news.

->Also 53 news which were actually fake are predicted as real and 60 of the real news are predicted as fake.

->The Pie Chart showing the above distribution is :



5 Comparisons of Passive Aggressive Classifier Algorithm with other Algorithms

Results for Support Vector Machines Algorithm

It helps us in classification of two or more class dataset by drawing decision boundaries in a n-dimensional space , where n is the number of features of the dataset. It finds an optimal Hyperplane which separates the two classes in an effective manner.

->We obtain the Classification Report for SVM model by using the reduced features dataset obtained by TruncatedSVD algorithm:

	precision	recall	f1-score	support
FAKE	0.92	0.92	0.92	638
REAL	0.92	0.92	0.92	629
accuracy			0.92	1267
macro avg	0.92	0.92	0.92	1267
weighted avg	0.92	0.92	0.92	1267

->The Confusion Matrix for SVM model is shown below:

590	48
49	580

Results for Gaussian Naive Bayes Classifier Algorithm

GNB assumes the Gaussian Distribution of the data and treats every feature independent of other to predict the class or label of a data point.

->We obtain the Classification Report for GNB model by using the reduced features dataset obtained by TruncatedSVD algorithm:

	precision	recall	f1-score	support
FAKE	0.68	0.63	0.65	638
REAL	0.65	0.70	0.67	629
accuracy			0.66	1267
macro avg	0.67	0.66	0.66	1267
weighted avg	0.67	0.66	0.66	1267

->The Confusion Matrix for GNB model is shown below:

403	235
190	439

Results for Logistic Regression Classifier Algorithm

The Logistic Regression Classifier is a classification algorithm which tells the probability of a data point belonging to a particular class or label.

->We obtain the Classification Report for Logistic Regression model by using the reduced features dataset obtained by TruncatedSVD algorithm:

	precision	recall	f1-score	support
FAKE	0.89	0.94	0.92	638
REAL	0.94	0.89	0.91	629
accuracy			0.91	1267
macro avg	0.92	0.91	0.91	1267
weighted avg	0.92	0.91	0.91	1267

->The Confusion Matrix for Logistic Regression model is shown below:

601	37
71	558

OBSERVATIONS:

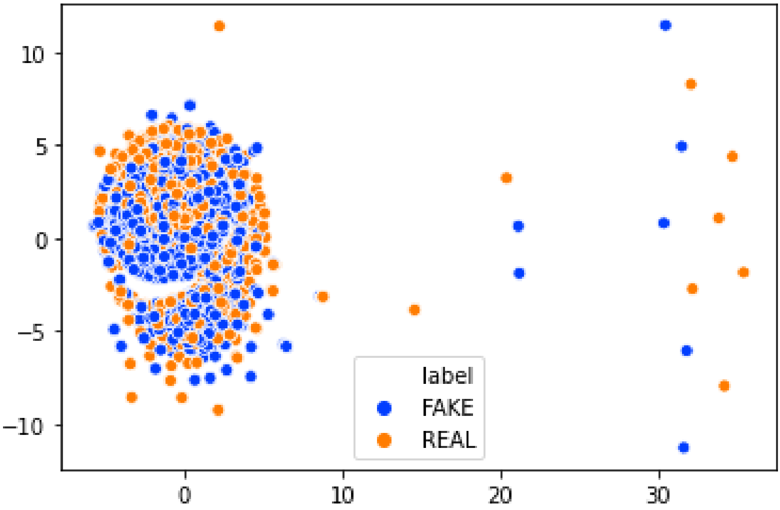
->The accuracy of prediction is around 92% for SVM model,which is almost equal to accuracy for Passive Aggressive Classifier and for Logistic Regressin Model.

->But the accuracy for Gaussian Naive Bayes Classifier is around 66% , which is low as compared to other models like SVM,Passive Aggressive Classifier or Logistic Regression.

->It shows that for the used dataset,Passive Aggressive Classifier, SVM and Logistic Classifier works best as compared to Gaussian Naive Bayes algorithm.

tSNE plot for the two class dataset:

tSNE helps us to cluster points in a low dimensional space by measuring similarity between them in the high dimensional space. tSNE helps us in better visualisation of points by projecting the similar points as cluster in a low dimensional space as visualisation in high dimensional space is very difficult.tSNE assumes a gaussian distribution around every point in the high dimensional space and assigns probability densities to every point under that gaussian and renormalize for all the points.



Observations:

->The two clusters share many overlapping points which indicates that there are similarities between real and fake news clusters.

->Also there are some points which have very different features than the rest of points from both clusters.