

NAME OF THE PROJECT

Fake News Detection Project

Submitted by:

HARSHITA SEHGAL

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

I got various research papers which gave me a better understanding of the theoretical aspects of the problem statement

Research papers (Pdf’s)- https: //pdf.sciencedirectassets.com/

Analysis of Classifiers for Fake News Detection VasuAgarwala , H.ParveenSultanaa ,SrijanMalhotraa \* , AmitrajitSarkarb

Other sources that helped me understand the practical aspect of the problem(this was a covid news fake or not detection problem)-

aptrinhx.com/covid-19-fake-news-detection-using-naive-bayes-classifier-2997726107/

INTRODUCTION

Business Problem Framing

We basically have data relating to various news articles, their writers, headlines and the target variable (label) 0 stands for not fake news and 1 stands for fake news.We have to build a model to predict whether the news is fake or not.This problem is a very practical problem in our real life scenarios ,as in today’s time there is a huge spread of rumours/fake news which can lead to economic,global and political consequences.Also the news here in our dataset is majorly related to the recent us elections,which was quite a hot topic for last few days.

Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

Herein the two major domains are

Machine learning-Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

Machine learning is a broad concept we will be focusing on the concept of classification for this problem

In Machine Learning and Statistics, Classification is the problem of identifying to which of a set of categories (sub populations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known

&NLP-Natural language processing is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

Now the next question is How these will both work ,what role will classification will play and what role will NLP play ,this will be discussed further in this report

Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

For the purpose of building this model,I did detailed research on the following sub concepts of machine learning classification problem,

Studied the concept of machine learning

Various machine learning algorithms

And as far as NLP is concerned I did research on how to clean the text regrex functions,lemmatization,building a corpus and finally tf-id vectorizer.These concepts are discussed in analytical problem framing of this report

Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

So as i discussed earlier,how in our daily lives the spread of fake news has serious repercussions,as the fake information may propgate and can mislead the public and also the system as a whole,this specific dataset is related to news regarding the elections,so hereby for such an important matter as election a fake news can have serious political economic consequences.So all this motivated me to try my hands on this project so that i can build a model to predict and conclude whether the news is fake or not,thereby controlling such unwanted circumstances of crime,threats and other issues which may arise due to this spread of fake news

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

Now as it as a nlp (text related problem),so before building our classifier we have to work on our text data.These are the steps listed below

1.Tokenize Text-tokenization is the process of turning a meaningful piece of data, such as an account number, into a random string of characters called a token that has no meaningful value if breached. Tokens serve as reference to the original data, but cannot be used to guess those values.

2.Word Tokenization-Word tokenization is the process of splitting a large sample of text into words. This is a requirement in natural language processing tasks where each word needs to be captured and subjected to further analysis like classifying and counting them for a particular sentiment etc.

3.Stop Word Removal-The process of converting data to something a computer can understand is referred to as pre-processing. One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words.

4.Stemming and Lemmatization-Stemming and Lemmatization both generate the root form of the inflected words. The difference is that stem might not be an actual word whereas, lemma is an actual language word. Stemming follows an algorithm with steps to perform on the words which makes it faster.For my dataset I have used lemmatization

5.And the most import tf-idf vectorization, TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction, tf-idf consider overall documents of weight of words.

Now once we have our text converted to numbers we can proceed with machine learning,for the purpose of machine learning,

Analysis and understanding of following concepts is neccseary-

Selecting the right algorithm which includes giving considerations to accuracy, training time, model complexity, number of parameters and number of features.

Choosing parameter settings and validation strategies.

Identifying underfitting and overfitting by understanding the Bias-Variance tradeoff.

Data Sources and their formats

What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

The data was in a excel sheet,I converted my excel file to a csv file and uploaded it on jupyter notebook.Now before uploading I filtered my data as the label column was not organized so for that purpose I filtered the data keeping only that data which has label 0 and 1.Label 0 has value counts of and label 1 has value counts of .About the dtypes,

id object

headline object

written\_by object

news object

label int64

Shape of my dataset (20721, 6)

Talking about the length of the news

Origian Length 91925618.0

Clean Length 67280435

Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

For the purpose of cleaning

lower

Regex functions

str.replace()

Lemmatization

Now as this was a huge dataset so i faced Unicode and decoding issues

For that purpose I used this command,

Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

So now concerning the input and output relation we observe two things,

Symbolic Approach: The symbolic approach to natural language processing is based on human-developed rules and lexicons. In other words, the basis behind this approach is in generally accepted rules of speech within a given language which are materialized and recorded by linguistic experts for computer systems to follow

.

Statistical Approach: The statistical approach to natural language processing is based on observable and recurring examples of linguistic phenomena. Models based on statistics recognize recurring themes through mathematical analysis of large text corpora. By identifying trends in large samples of text the computer system can develop its own linguistic rules that it will use to analyze future input and/or the generation of language output.

Like the use of tf-idf ,this will generate a meaningful numercal data which will in turn have a relation to our label variable 0 and 1.

State the set of assumptions (if any) related to the problem under consideration

Here, you can describe any presumptions taken by you.

Now for building my final model,I took news column ,we had two text columns headlines and news but i preffered news column as it had more data than headlines ,also my model gave a good accuracy

Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

Libraries and packages used

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

As discussed in the assumption I took two columns label and news ,

Firstly converting the news into tf idf

Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

Logistic Regression

mnb

Ada boost classifier

K-Nearest Neighbours

Decision Tree

Random Forest

Svm

Run and Evaluate selected models

Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

Logistic Regression

It is the mostly used algorithm across industry ,we will import libraries such as Logistic Regression from sklearn.linear\_model, importing confusion matrix and classification report we get the following results

precision recall f1-score support

0 0.95 0.95 0.95 2649

1 0.94 0.95 0.95 2382

accuracy 0.95 5031

macro avg 0.95 0.95 0.95 5031

weighted avg 0.95 0.95 0.95 5031

. Decision Tree Classifier

A decision tree classifier is a tree in which internal nodes are labelled by features,We have two criterion ‘gini’ and ‘entropy’

Importing DecisionTreeClasifier from sklearn.tree

recision recall f1-score support

0 0.88 0.89 0.89 2649

1 0.88 0.86 0.87 2382

accuracy 0.88 5031

macro avg 0.88 0.88 0.88 5031

weighted avg 0.88 0.88 0.88 5031

3.Random Forest Classifier

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

precision recall f1-score support

0 0.89 0.96 0.92 2649

1 0.95 0.87 0.91 2382

accuracy 0.92 5031

macro avg 0.92 0.92 0.92 5031

weighted avg 0.92 0.92 0.92 5031

4.Ada Boost Classifier

An AdaBoost [1] classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

0.8660306102166567

[[2237 412]

[ 262 2120]]

precision recall f1-score support

0 0.90 0.84 0.87 2649

1 0.84 0.89 0.86 2382

accuracy 0.87 5031

macro avg 0.87 0.87 0.87 5031

weighted avg 0.87 0.87 0.87 5031

Key Metrics for success in solving problem under consideration

What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

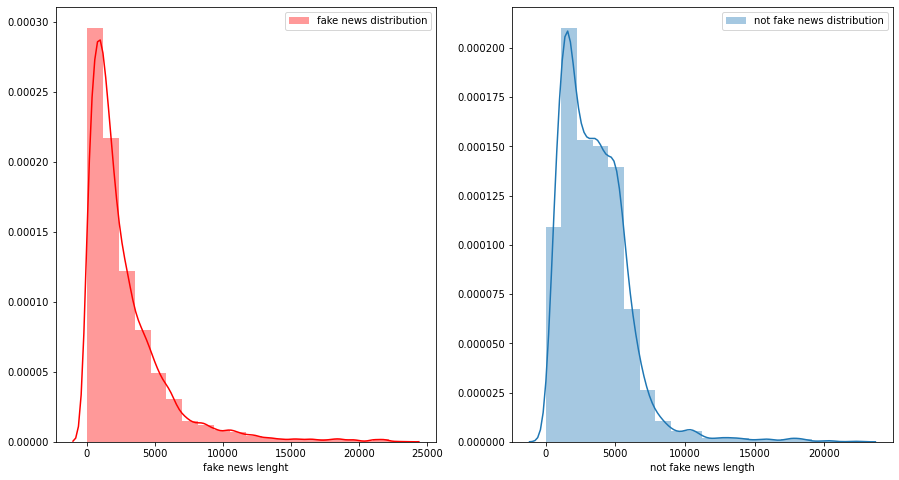
Svm and logistic regression works pretty well with 95 accuracy and even great f1 scores,recall and precision.

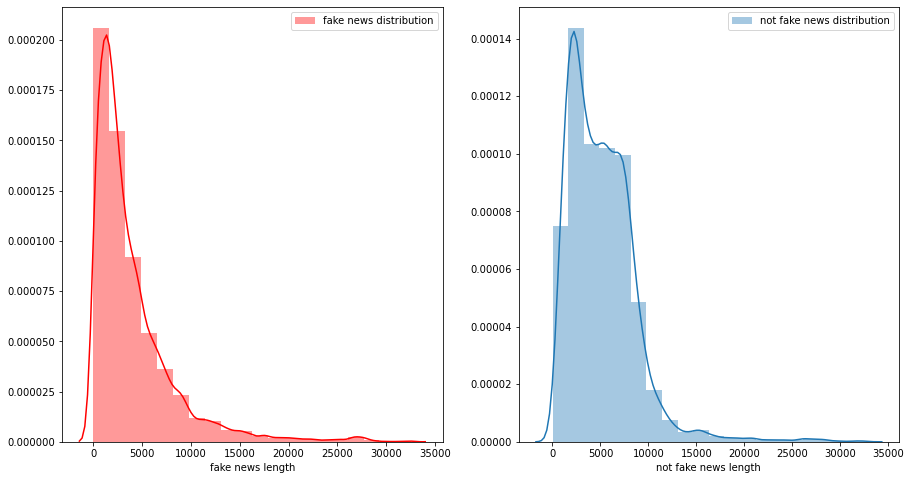
Visualizations

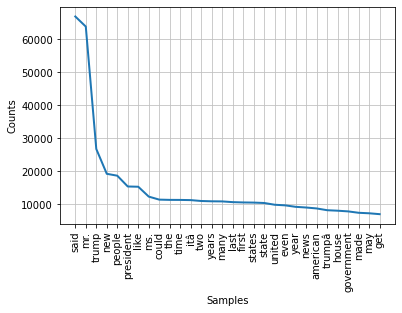
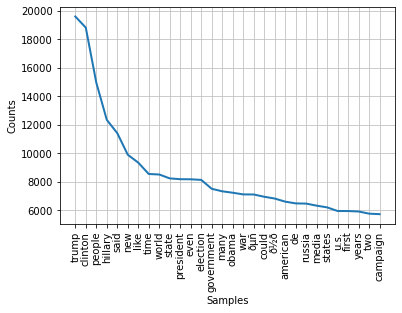
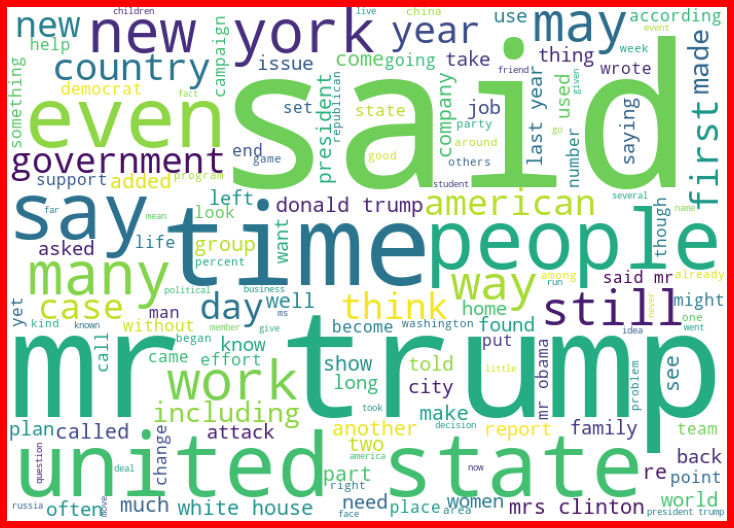
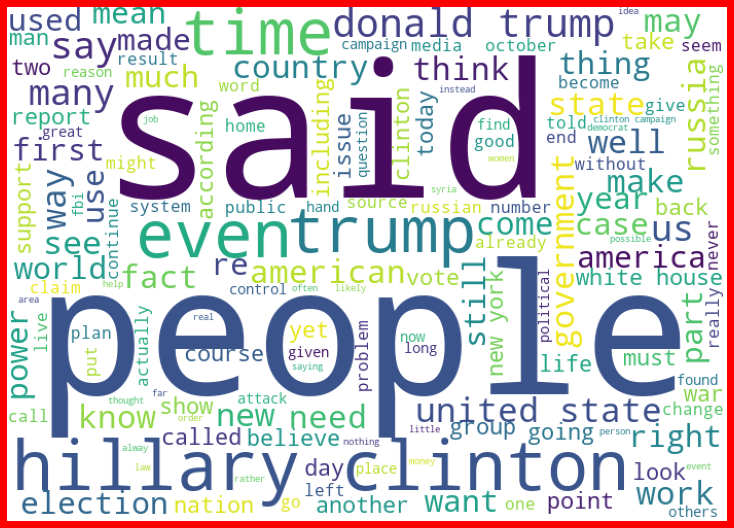
Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

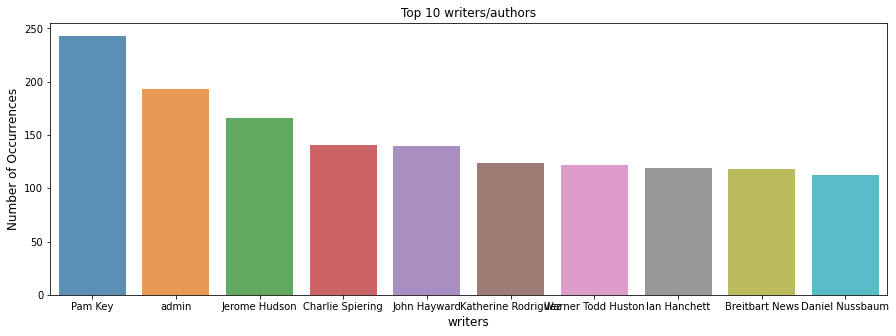
From the label countplot I see the data is equally spread,almost 50% for each

this is after cleaning the data,this is before cleaning the data



For the purpose of segregating the most frequent words by their label,I first filtered the data in excel created two dfferent data frames and then I obsereved that although the words are almost similar one change is that in the fake news mr and mrs is not used while in the not fake news mr is used,also the no of occurrences of word trump and Hillary is higher in the fake news with regard o not fake news.similar things can be observed when we plot the ]word cloud ,1st word cloud is for not fake news and the second is for fake news,also use of may ,often claim is quite evident in fake news showing the uncertainity of that news 

also we can observe the top writers who have written most of the news(top 10 writers)



Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

So following are my observations

-equal data for both the labels 0 and 1

-data for fake and not fake is very similar

-only a few things, like presence of mr and mrs in the not fake news, occurrences of word may ,often in fake news.

-data is all about elections, mostly revolving around trump, Hillary Clinton, campaigns

-the id is unique

-we can also observe thier are 4165

Unique authors and top 10 authors can be seen while plotting barplot by value counts

-in cleaning of the texts,there were many symbols,links,punctuations

CONCLUSION

Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

The performance of a classifier may vary based on the size and quality of the text data (or corpus) and also the features of the text vectors. Common noisy wordscalled ‘stopwords’ are less important words when it comes to text feature extraction, they don’t contribute towards the actual meaning of a sentence and they only contribute towards feature dimensionality and may be discarded for better performance. This helps in reducing the size/dimensionality of the text corpus and add text context for feature extraction. Also, lemmatization is used to convert words to their core meaning and this result in multiple word conversion into a single discrete representation.

Tf vec works very well for my classification problem

This is how my data is split into x and y

tf\_vec = TfidfVectorizer()

tf\_vec = TfidfVectorizer()

naive = MultinomialNB()

features = tf\_vec.fit\_transform(df['news'])

X = features

y = df['label']

print(X\_train.shape)

print(Y\_train.shape)

print(x\_test.shape)

print(y\_test.shape)

(15093, 162095)

(15093,)

(5031, 162095)

(5031,)

Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

SVM and logistic regression classifier (with accuracy of 96 and 95) have the best performance on this dataset in the model, with SVM having a slightly better performance than logistic regression classifier. The same can be perceived from the f1 scores.Also, the training data is largely based on US politics and economics news so it has been observed in our test cases, that the news statements related to US politics have been correctly classified and fake news was detected. But the test cases which have news related to technology have been wrongly predicted,which can be reflected in the confusion matrix

Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.

The limitations that come packaged with this problem is that, the data is very similar for both the labels with only a few minor changes and this means that any type of prediction model can have anomalies and can make mistakes. For future improvements, we can work more deeper and concepts like POS tagging, word2vec and topic modelling can be utilized. These will give the model a lot more depth in terms of feature extraction and fine-tuned classification.