Spam Detection Project

Submitted by:

VIVEK KUMAR SINGH

ACKNOWLEDGMENT

I take great pleasure to thank and acknowledge the help provided by **Flip Robo Technologies**. I extend whole hearted thanks to Mr. Shwetank Mishra who become my Mentor and with whom I worked and learned a lot and for enlightening me with her knowledge and experience to grow with the corporate working. Her guidance at every stage of the Project enabled me to successfully complete this Project which otherwise would not have been possible without her consent encouragement and motivation. Without the support it was not possible for me to complete the report with fullest endeavour.

INTRODUCTION

- Business Problem Framing

  The authenticity of Information has become a longstanding issue affecting businesses and society, both for printed and digital media. On social networks, the reach and effects of information spread occur at such a fast pace and so amplified that distorted, inaccurate or false information acquires a tremendous potential to cause real world impacts, within minutes, for millions of users. Recently, several public concerns about this problem and some approaches to mitigate the problem were expressed. In this paper, we discuss the problem by presenting the proposals into categories: content based, source based and diffusion based. We describe two opposite approaches and propose an algorithmic solution that synthesizes the main concerns. We conclude the paper by raising awareness about concerns and opportunities for businesses that are currently on the quest to help automatically detecting fake news by providing web services, but who will most certainly, on the long term, profit from their massive

- Conceptual Background of the Domain Problem

  In the news industry, in particular, but also in society at large, fake news detection has become a central discussion topic, as the need to permanently assess the veracity of digital content has been raised by the constant spread of false news / information. Information veracity is a long-term issue affecting society both for printed and digital media. The sensationalism of not-so-accurate eye catching and intriguing headlines aimed at retaining the attention of audiences to sell information has persisted all throughout the history of all kinds of information broadcast. On social networking websites, the reach and effects of information spread are however significantly amplified and

occur at such a fast pace, that distorted, inaccurate or false information acquires a tremendous potential to cause real impacts, within minutes, for millions of users. Societal issues are being raised about the individuals' ability to tell apart what is fake and what is authentic, while surfing and actively engaging in information overloaded networks. As reported by Anderson , youngsters are known to be tech-savvy when compared to their parents, but when it comes to the ability to tell if a news piece is fake or not, they seem as confused as the rest of the society and 44% have confirmed it in a research conducted by Common Sense Media. The same research also indicates that 31% of kids aged 10 to 18 have shared online at least one news story that they later found out was inaccurate or fake. This situation raises a whole new dimension of concerns related to digital literacy that go beyond the mere ability to access and manage technology. Together with societal challenges, there is a dramatic and inconspicuous situation happening in the media landscape, the public sphere and journalism industry that requires debate and examination, pointing out two main aspects. The first one relies on the fact that news publishers have lost control over the distribution of news, which are presented to Internet users by obscure and unpredictable algorithms. Also, news market newcomers (such as BuzzFeed, Vox and Fusion) have built their presence by embracing these technologies, undermining the long-term positions occupied by more traditional news publishers. The second aspect relies on the increasing power that social media companies, such as Google, Apple, Facebook and Amazon, have gained in controlling who publishes what to whom, and how the publications are monetized. In the above context, establishing the reliability of online information is a daunting but critical current challenge 3 , demanding the attention, regulation and active monitoring of digital content spread by

the major parties involved in sustaining how the information is presented and shared among people over the Internet, including search engines and social networking platforms. The fake news subject has become so prevalent that the Commons Culture, Media and Sport Committee is currently investigating concerns about the public being swayed by propaganda and untruths 4 . The curation of high-quality journalism is also at stake, since an increasing proportion of the adults are getting their news from social media and fictional stories are presented in such way that it can be very difficult to tell them apart from what is authentic.

storage space and communication bandwidth. End user is at risk of deleting legitimate mail by mistake. Moreover, spam also impacted the economical which led some countries to adopt legislation. 2 Text classification is used to determine the path of incoming mail/message either into inbox or straight to spam folder. It is the process of assigning categories to text according to its content. It is used to organized, structures and categorize text. It can be done either manually or automatically. Machine learning automatically classifies the text in a much faster way than manual technique. Machine learning uses pre-labelled text to learn the different associations between pieces of text and it output. It used feature extraction to transform each text to numerical representation in form of vector which represents the frequency of word in predefined dictionary. Text classification is important in the context of structuring the unstructured and messy nature of text such as documents and spam messages in a cost-effective way. A Machine learning platform has capabilities to improve the accuracy of predictions. With regard to Big Data, a Machine Learning platform has abilities to speed up analysing of gigantic data. It is important especially to a company to analyse text data, help inform business decisions and even automate business processes. For

example, text classification is used in classifying short texts such as tweets or headlines. It can be used in larger documents such as media articles. It also can be applied to social media monitoring, brand monitoring and etc. In this project, a machine learning technique is used to detect the spam message of a mail. Machine learning is where computers can learn to do something without the need to explicitly program them for the task. It uses data and produce a program to perform a task such as classification. Compared to knowledge engineering, machine learning techniques require messages that have been successfully pre-classified. The pre-classified messages make the training dataset which will be used to fit the learning algorithm to the model in machine learning studio. 3 A specific algorithm is used to learn the classification rules from these messages. Those algorithms are used for classification of objects of different classes. The algorithms are provided with input and output data and have a self-learning program to solve the given task. Searching for the best algorithm and model can be time consuming. The two-class classifier is best used to classify the type of message either spam or ham. This algorithm is used to predict the probability and classification of data outcome.

- Review of Literature

  Distorted news and “alternate facts” were not a problem in society two years ago, despite the long-term deep changes in the news market 1 . The social concern about these kinds of news has been rather deeply accelerated by the term “fake news”, coined by the US elected President, Donald Trump, conveying its origins in the political arena. For example, among other fake news that emerged during the Trump campaign one of the most popular ones consisted on the Pope Francis reported endorsement of Donald Trump for president of the US.

The news piece was advanced by the website “Ending The Fed”, managed by a Romanian youngster. BBC 4, 6 also refers to the advancement of particular (often extreme) political causes as one of the main sources of fake news, defining them as false information deliberately circulated by those who have scant regard for the truth and act under the motivation of fostering political causes or obtaining revenue out of the online traffic. In this domain, Facebook has faced an increasing criticism over its role in the 2016 US presidential election because it allowed the propagation of fake news disguised as news stories coming from unchecked websites. This spreading of false information during the election cycle was so severe that Facebook was labelled as “dust cloud of nonsense.” 7 The fact is that the presidential election year has shown how the lines have blurred between facts and speculation, with people profiting off the spread of fake news. There were more than 100 news sites that made up pro-Trump content traced to Macedonia, according to a BuzzFeed News investigation 8 . A subversive industry of fake news has been arising as an independent business opportunity in the news market, as is the case of Media Vibes SNC, a Belgium company who owns more than 180 URLs devoted to creating and spreading fake news on the web and on social networks (such as 24aktuelles.com. or react365.com). The company is also responsible for the creation of the user-generated fake news concept, by providing Internet users with an application to develop their own fake news and to spread them on their social networks. The main idea behind the business is supported by the “do-it-yourself media”, and fake news can consist of jokes, provocations, sarcasm, etc., that are written by ordinary people (c.f. react365.com). Another fake news model also worth of mention is the one based on the publication of news pieces in websites with URLs very similar to some of the most popular and well reputed news stations, such

as ABC. For instance, the official URL of the station's website is abcnews.go.com and this type of fake news are available at the URL abcnews.com.co. Among his top fake stories that had huge success in 2016 are some of the most known political fake news pieces: "Obama Signs Executive Order Banning The Pledge Of Allegiance In Schools Nationwide", "Donald Trump Protester Speaks Out: «I Was Paid $3,500 To Protest Trump's Rally»" and "Obama Signs Executive Order Declaring Investigation Into Election Results; Revote Planned For Dec. 19th". Among the main purposes of such fake news business models it's possible to highlight the interest on generating interactions on the social networks, generate web traffic to the fake news pages and earn profit through advertising or to damage someone's image and reputation.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

There are two datasets one for fake news and one for true news. In true news, there is 21417 news, and in fake news, there is 23481 news. You have to insert one label column zero for fake news and one for true news. We are combined both datasets using pandas built-in function.

```
In [2]: # All fake data in dataset:-
        df1=pd.read_csv('Fake.csv')
        df1.head()
        df1.shape

Out[2]: (23481, 4)

In [3]: # All True data in dataset:-
        df2=pd.read_csv('True.csv')

In [4]: df1['label']=0
        df2['label']=1
```

- Data Sources and their formats

Data is provided by the company Fliprobo and format of the data is in csv format.

```
[7]: # Shape of dataset.
     df.shape

[7]: (44898, 5)
```

```
[8]: # Checking null values.
     df.isnull().sum()

[8]: title      0
     text       0
     subject    0
     date       0
     label      0
     dtype: int64
```

- Data Pre-processing Done

  Data is cleaning and pre-processing:-

```
n [4]: df1['label']=0
       df2['label']=1
```

Here we create the label 0=fake and 1= True data.

```
In [6]: # What we did here we just bring both real and fake dataset into the one dataset.
        df=pd.concat([df1,df2],axis=0)
```

Concat both the dataset False and True and make single dataset.

Here , I took some steps to clean and pre- processed the dataset.

Dropping irrelevant dataset.

```
14]: # Dropping some irrelevent columns.
     df=df.drop(columns=['index','date'],axis=1)
```

```
16]: # Here i am deleting subject columns,beacuse we have to analyzed more than just a subject.
     df=df.drop(columns=['subject'],axis=1)
```

In [27]:
```
df['text_length']=df['text'].map(lambda text :len(text))
```

In [28]:
```
df.head()
```

Out[28]:

	title	text	label	text_length
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	0	2893
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	0	1898
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	0	3597
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	0	2774
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	0	2346

Length of data test is shown here .

31]:
```
df['new _text']=df['text'].apply(nfx.remove_stopwords)
```

32]:
```
df.head()
```

32]:

	title	text	label	text_length	new _text
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	0	2893	Donald Trump wish Americans Happy New Year lea...
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	0	1898	House Intelligence Committee Chairman Devin Nu...
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	0	3597	Friday, revealed Milwaukee Sheriff David Clark...
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	0	2774	Christmas day, Donald Trump announced work fol...
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	0	2346	Pope Francis annual Christmas Day message rebu...

Apply here stop words, here I am using neat text library.

In [34]:
```
df['new _text']=df['new _text'].replace('http\S+','website')
df['new _text']=df['new _text'].replace('^.+@[^\.].*\.[a-z]{2,}$','emailaddress')
df['new _text']=df['new _text'].replace('<.*?>','')
df['new _text']=df['new _text'].apply(nfx.remove_numbers)
df['new _text']=df['new _text'].apply(nfx.remove_currency_symbols)
df['new _text']=df['new _text'].apply(nfx.remove_emojis)
```

In [35]:
```
df['new _text']=df['new _text'].replace("(.*?,", "")
```

In [36]:
```
df['new _text'].iloc[-1]
```

Out[36]:
```
'JAKARTA (Reuters) - Indonesia buy  Sukhoi fighter jets worth . billion Russia exchange cash Indonesian commodities, cabinet mi
nisters said Tuesday. Southeast Asian country pledged ship  million worth commodities addition cash pay Suhkoi SU- fighter jet
s, expected delivered stages starting years. Indonesian Trade Minister Enggartiasto Lukita said joint statement Defence Ministe
r Ryamizard Ryacudu details type volume commodities negotiated . Previously said exports include palm oil, tea, coffee. deal ex
pected finalised soon Indonesian state trading company PT Perusahaan Perdangangan Indonesia Russian state conglomerate Rostec.
Russia currently facing new round U.S.-imposed trade sanctions. Meanwhile, Southeast Asia largest economy trying promote palm o
il products amid threats cut consumption European Union countries. Indonesia trying modernize ageing air force string military
aviation accidents. Indonesia,  million trade surplus Russia , wants expand bilateral cooperation tourism, education, energy, t
echnology aviation others.'
```

```
[37]: import re
df['new _text']=df['new _text'].apply(lambda X:(re.sub(r"[,.'/\:?*#]",'',X)))
df['new _text']=df['new _text'].apply(lambda X:(re.sub(r"r'http\S+', 'webaddress'"," ",X))
df['new _text']=df['new _text'].apply(lambda X:(re.sub(r"[-,)(]"," ",X)))
df['new _text']=df['new _text'].apply(lambda X:X.lower())
```

```
[38]: df['new _text']=df['new _text'].apply(lambda X:(re.sub(r"@#"," ",X)))
df['new _text']=df['new _text'].replace("@"," ")
df['new _text']=df['new _text'].apply(lambda X:lemmit.lemmatize(X))
```

```
[39]: df['new _text']
```

```
t[39]: 0        donald trump wish americans happy new year lea...
1        house intelligence committee chairman devin nu...
2        friday revealed milwaukee sheriff david clarke...
3        christmas day donald trump announced work foll...
4        pope francis annual christmas day message rebu...
                              ...
44684    brussels  reuters    nato allies tuesday welco...
44685    london  reuters    lexisnexis provider legal r...
44686    minsk  reuters    shadow disused soviet era fa...
44687    moscow  reuters    vatican secretary state car...
44688    jakarta  reuters    indonesia buy  sukhoi figh...
```

so, these are the steps taken for cleaning the data.

PRE-PROCESSING -

```
In [40]: from sklearn.feature_extraction.text import TfidfVectorizer
tfidf=TfidfVectorizer(max_features = 14000, stop_words='english')
```

```
In [41]: X=tfidf.fit_transform(df['new _text'])
```

```
In [42]: Y=df['label']
```

Here, I used the TFidf vectorizer to convert it into algorithm for model training and prediction.

- Data Inputs- Logic- Output Relationships

```
43]: from sklearn.model_selection import train_test_split

44]: x_train,x_test,y_train,y_test= train_test_split(X,Y,random_state=101,test_size=0.35)

45]: # Shape of train and text dataset:-
     print(x_train.shape,'\t\t',x_test.shape)
     print(y_train.shape,'\t\t',y_test.shape)

     (29047, 14000)          (15642, 14000)
     (29047,)                (15642,)
```

We split x-train,x-test and y-train ,y-test for trainig and testing the model.

First it train and we feed x_test as input data and predict and compare the y_test which is actual output.

- Hardware and Software Requirements and Tools Used

The project is done into laptop with i5 processor with quad core with 8gb of RAM with GTX 1650 GPU with Anaconda and juypter notebook.

```
[1]: # Importing important libaries:-
     import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sb
     from PIL import Image
     from wordcloud import ImageColorGenerator,STOPWORDS,WordCloud
     import re,os
     import string
     from nltk.tokenize import word_tokenize,sent_tokenize,regexp_tokenize
     from nltk.stem import PorterStemmer,SnowballStemmer,LancasterStemmer
     from nltk.stem.wordnet import WordNetLemmatizer
```

Some of important packages is imported in python for this project.

.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  As per the problem it is the supervised learning problem so, I used here supervised learning apporches.

  Problem solving approaches are :-

  - Removing special character
  - Removing emojis and emails
  - Removing punctuation
  - Removing stop words
  - Removing numbers.
  - Using lemmatize and stemming
  - Using vectorizer (tfidf vectorizer) for numerical values
  - Splitting train and test data
  - And at the end create models
  - Check the accuracy
  - Use cross validation for overfitting and underfitting
  - Saving the best performance model

- Testing of Identified Approaches (Algorithms)

  The following algorithm is used for testing .

```
]: from sklearn.svm import LinearSVC
   from sklearn.naive_bayes import MultinomialNB
   from sklearn.linear_model import LogisticRegression
   from lightgbm import LGBMClassifier
   from sklearn.linear_model import SGDClassifier
   from sklearn.ensemble import RandomForestClassifier
```

```
svc = LinearSVC()
lr = LogisticRegression(solver='lbfgs')
mnb = MultinomialNB()
lgb = LGBMClassifier()
sgd = SGDClassifier()
rf = RandomForestClassifier()
```

- Run and Evaluate selected models

```
def print_score(y_pred,clf):
    print('classifier:',clf)
    print("Jaccard score: {}".format(jaccard_score(y_test,y_pred,average='micro')))
    print("Accuracy score: {}".format(accuracy_score(y_test,y_pred)))
    print("f1_score: {}".format(f1_score(y_test,y_pred,average='micro')))
    print("Precision : ", precision_score(y_test,y_pred,average='micro'))
    print("Recall: {}".format(recall_score(y_test,y_pred,average='micro')))
    print("Hamming loss: ", hamming_loss(y_test,y_pred))
    print("Confusion matrix:\n ", multilabel_confusion_matrix(y_test,y_pred))
    print('=======================================\n')
```

```
for classifier in [svc,lr,mnb,sgd,lgb,rf]:
    clf = OneVsRestClassifier(classifier)
    clf.fit(x_train,y_train)
    y_pred = clf.predict(x_test)
    print_score(y_pred, classifier)
```

```
classifier: LinearSVC()
Jaccard score: 0.9880528723945095
Accuracy score: 0.9939905382943358
f1_score: 0.9939905382943358
Precision :  0.9939905382943358
Recall: 0.9939905382943358
Hamming loss:  0.006009461705664237
Confusion matrix:
  [[[7305   44]
  [  50 8243]]

 [[8243   50]
  [  44 7305]]]
========================================

classifier: LogisticRegression()
Jaccard score: 0.9738784781374219
Accuracy score: 0.9867663981588032
f1_score: 0.9867663981588032
Precision :  0.9867663981588032
Recall: 0.9867663981588032
Hamming loss:  0.013233601841196778
Confusion matrix:
  [[[7255   94]
  [ 113 8180]]
```

```
classifier: MultinomialNB()
Jaccard score: 0.8717243029795381
Accuracy score: 0.9314665643779568
f1_score: 0.9314665643779568
Precision :  0.9314665643779568
Recall: 0.9314665643779568
Hamming loss:  0.06853343562204321
Confusion matrix:
  [[[6775  574]
  [ 498 7795]]

 [[7795  498]
  [ 574 6775]]]
=======================================

classifier: SGDClassifier()
Jaccard score: 0.9805013927576601
Accuracy score: 0.9901547116736991
f1_score: 0.9901547116736991
Precision :  0.9901547116736991
Recall: 0.9901547116736991
Hamming loss:  0.009845288326300985
Confusion matrix:
  [[[7284   65]
  [  89 8204]]
```

```
classifier: SGDClassifier()
Jaccard score: 0.9805013927576601
Accuracy score: 0.9901547116736991
f1_score: 0.9901547116736991
Precision :  0.9901547116736991
Recall: 0.9901547116736991
Hamming loss:  0.009845288326300985
Confusion matrix:
  [[[7284   65]
  [  89 8204]]

 [[8204   89]
  [  65 7284]]]
========================================

classifier: LGBMClassifier()
Jaccard score: 0.9941356450790413
Accuracy score: 0.9970591995908452
f1_score: 0.9970591995908452
Precision :  0.9970591995908452
Recall: 0.9970591995908452
Hamming loss:  0.0029408004091548397
Confusion matrix:
  [[[7338   11]
  [  35 8258]]

 [[8258   35]
```

```
classifier: RandomForestClassifier()
Jaccard score: 0.9928653331634603
Accuracy score: 0.9964198951540724
f1_score: 0.9964198951540724
Precision :  0.9964198951540724
Recall: 0.9964198951540724
Hamming loss:  0.0035801048459276306
Confusion matrix:
  [[[7330   19]
  [  37 8256]]

 [[8256   37]
  [  19 7330]]]
=======================================
```

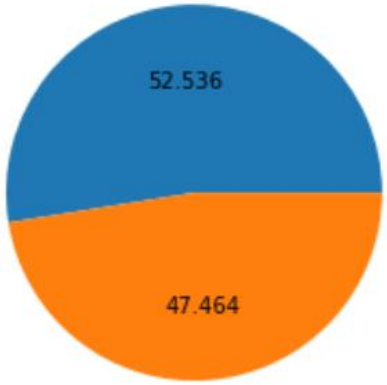
- Key Metrics for success in solving problem under consideration
  - In basis of accuracy score, matrix and cross validation. Logistic regression selected as final models.

- Visualizations

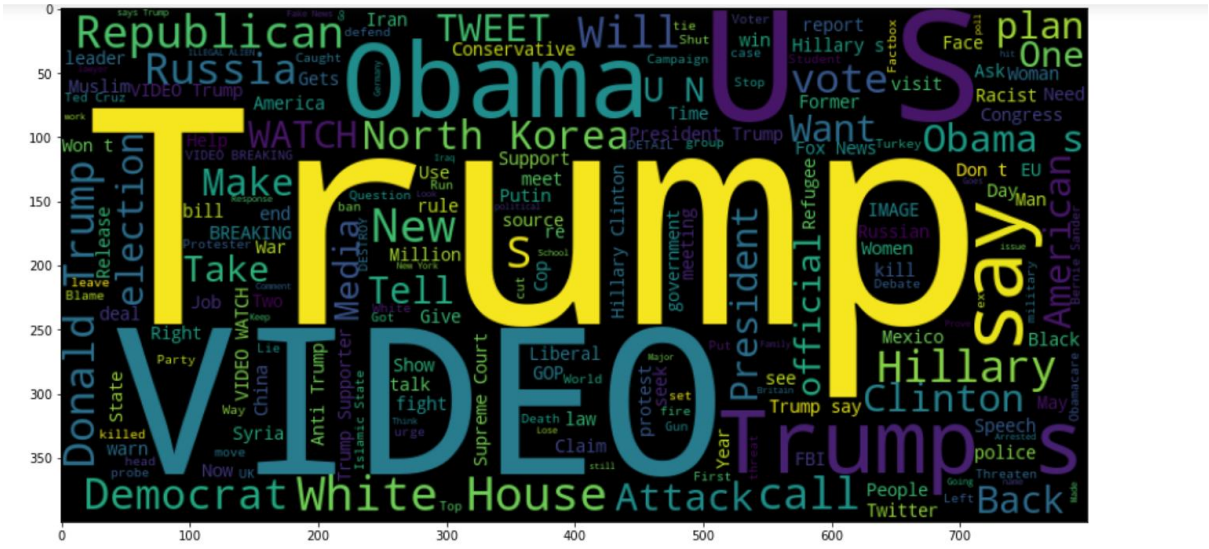
```
(df['label'].value_counts()/44689)*100
# There are no imbalance data is here.
```

```
0    52.536418
1    47.463582
Name: label, dtype: float64
```

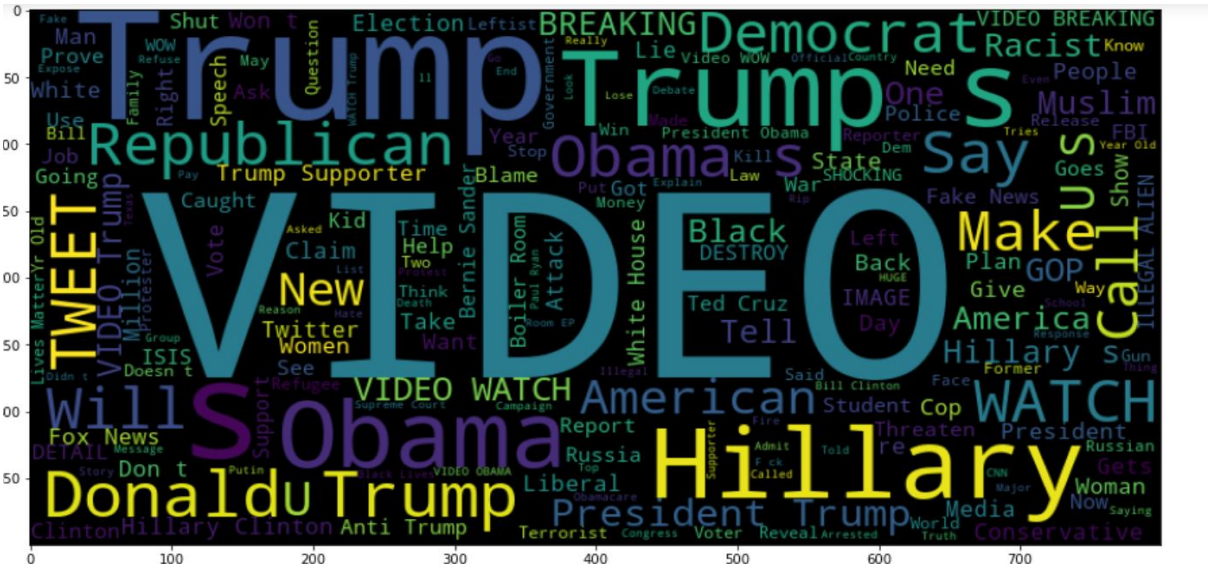
```
plt.pie(df['label'].value_counts(),autopct="%0.3f")
plt.show()
```



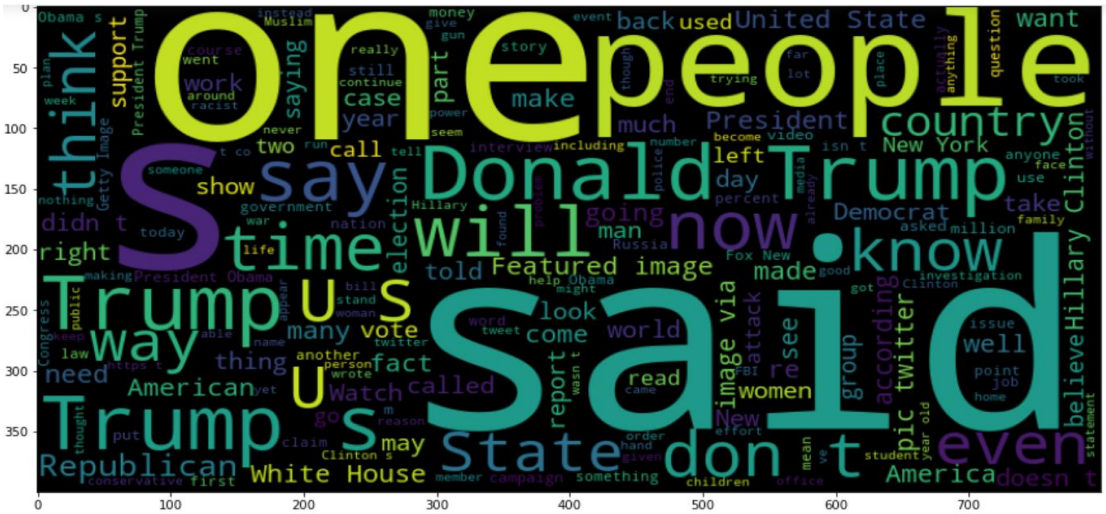
So here we see that all most fake and true data is balanced.



This is the figure before cleaning the data.



.



CONCLUSION

The quest for a system to prevent the creation of fake news collides with many democratic values like freedom of speech. However, it is possible to identify elements in news pieces (or social media posts) which can be objective, namely the “facts”, which can help to address the situation. We believe that, currently, the necessary settings and resources to attack this problem are available: we have the technology in form of algorithms (text mining, machine learning, etc), the hardware to cope with big data, access to big data for training the algorithms. We also have the context and momentum to do it because the problem is well installed in the public conscience, and we have the willpower from the major players. Still, there are some battles along this war. For example, the battle for the most used AI framework, or which company has more/best data, or which has the best infrastructure to deal with these problems. Two decades back, IBM Deep Blue – a chess-playing computer – defeated chess master Garry Kasparov in 6 matches. At the time, IBM used an immense amount of secrecy during interviews about their machine. Recently, the panorama has quite changed: last year, Microsoft won a competition whose goal was to develop an image recognition system. Microsoft team explained that they used a trained neural net comprised of more than 200 layers. It is interesting to notice that suddenly all the major companies like Amazon, IBM, Google, Facebook, Twitter, Baidu, Yahoo, and Microsoft have made their code open source and available to anyone. For instance, Google has provided the use of part of its proprietary Deep Learning TensorFlow AI for free to its commercial customers. We can roughly say that during 2016 all the major deep learning libraries became open source and freely available. This trend to develop and to highly capacitate systems based on opensource resources and cloud services, which may be freely available, or available at a small price, hands over the service-providers with an escalating huge power: the

power to choose their machine learning algorithms, their pre-trained data and, ultimately, a control over the intelligence that is built on the service provided by their systems. Therefore, the key to one problem is usually the lead to another one, however, this immense availability from major key players might be the necessary basis for fighting the proliferation of fake news.