**Assigment 3 Report**

**Submitted by:-**

Vishal Deo

**Under Guidance of :-**

Professor Dr. Weimao Ke

Professor Dr. Heejun Kim

Index

|  |  |  |
| --- | --- | --- |
| Sr.no | Topic | Page no |
| 1 | Data Prepartation | 3 |
| 2 | Probabilities and Zipf | 4 |
| 3 | Text vectorization | 6 |
| 4 | Classification | 7 |
| 5 | Conclusion | 9 |

1. **Data Preparation**
2. Include a link to your data or submit your data.

* <https://www.kaggle.com/rchitic17/real-or-fake>

1. Include a brief description of your data, attributes, and instances.

* The topic of fake news became common in connection with recent US presidential elections. The topic has become so hot that several crowdsourced groups composed a dataset that has published on various web-sites and categorized them between real and fake.

Data attributes :-

Attribute 1 :- Integer (Character count of the article)

Attribute 2 :- String [Title] (Title of the News Article)

Attribute 3 :- String [Text ] ( Content of the News Article)

Attribute 4 :- String [Label] ( Classification of the article in true or false]

1. Discuss the classification task on the data and objectives.

* This article has its own data set of about six thousand English language news, marked as fake or true. The assignment objectives are to implement clustering and to predict wether the articles are fake or real. Aim of the assignment is to explore algorythms to help identify fake news from real news

1. **Probabilities and Zipf**

B.1. Class Distributions and Probabilities

**🡪**

|  |  |  |
| --- | --- | --- |
| Class c | Instances nc | Probability p(c) |
| Real | 3171 | 0.5 |
| Fake | 3164 | 0.5 |

B.2. Term Probabilities and Zipf's Law

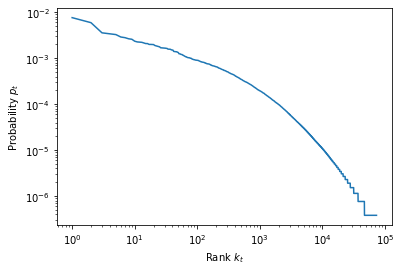
#### B.2.1. Rank, Frequency, and Probability

Rank terms by frequency and show the top five with probabilities

|  |  |  |  |
| --- | --- | --- | --- |
| **Term No.** | **Term** | **Frequency** | **Probablity** |
| **0** | **trump** | **19937** | 0.007644 |
| **1** | **clinton** | **15503** | 0.005944 |
| **2** | **new** | **9318** | 0.003572 |
| **3** | **state** | **8854** | 0.003395 |
| **4** | **president** | **8536** | 0.003273 |

#### B.2.2. Probability vs. Rank Plot

Produce a probability pt vs. kt plot on **log-log** coordinates.



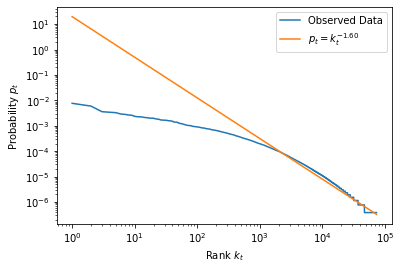
#### B.2.3. Regression line

#### 1) Identify the coefficient value from the regression

🡪 -1.59902822274253

2) Plot the regression line on the pt vs. kt

🡪



3)Do your data follow Zipf's law? Discuss the visual pattern, the fitted coefficient, and the regression line

🡪

Zipf's law is used to compress indices for search engines based on word distribution .It is used to predict the number of times the n most popular word occurs in a corpus, or the fraction of words with frequency n. If the word count data follows Zipf’s law, the data points will follow the line. Most of the points follow a straight line, but they do not follow the comparison line. So, while my data don’t follow Zipf’s law, the distribution isn’t completely dissimilar. That similarity probably explains why the first glance at the data in the last post didn’t reveal any obvious differences from the expected distribution.

Virtually all of the data points on the graph are above the line. A point above the line indicates that the observed frequency at the corresponding rank is higher than the frequency that Zipf’s law predicts. In other words, as rank decreases, frequency doesn’t decrease as rapidly as expected.

So the fitted line is roughly:

Not exactly as the Zipf's law suggests.

## Text Vectorization

### C.3. Terms and Conditional Probabilities:-

1. What do these probabilities mean?

These conditional probalities mean that the terms which are evaluated for the conditional probablities show the probability of occurrence of the terms in the two classes.

For example the Term :- p(Clinton|real) = 0.0062961

p(Clinton|fake) = 0.0047562

shows the probality of the occurrence of the term Clinton in news Classified as real 0.0062961 where as in news classified as fake it is 0.0047562 so there are higher

chances that word clinton is more likely to be associated with Real news

1. Are the above probability values reasonable (sensible)? Why or why not?

The values generated by the conditional probablities requires further analysis to see wether the values are sensible or not as there seems to be certain ambiguity terms and their associated values.

For Example:- As discussed in the above word clinton is more likely to be associated with Real news whereas the word Hillary p(Hillary|real) 0.0016952 p(Hillary|fake) = 0.0035570. So while the first name Hillary probality shows is less to appear in a real news. On other hand the last name clinton is more likely to be associated with Real news. But a name is reffered as First and Last Name together and its related with one idenitity but the results are different which show that further analysis is required to get complete insight

3 Do you think they will be helpful in the classification task?

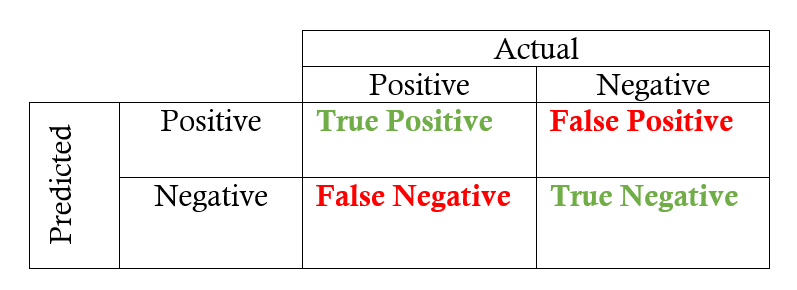
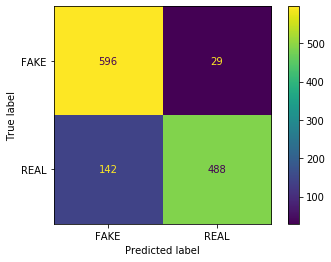
The conditional probalities of terms help in classification task was when the terms are occur in the content of a news article it will be very useful to predict that wether the article is real or fake

**Classification**

D.1. Probabilistic Naive Bayes Model

1. Discuss, for your task and objectives, which number(s) in the confusion matrix is most important, that you wish to minimize or maximize? Why?

A confusion matrix is typically computed in any machine learning classifier such us logistic regression, decision tree, support vector machine, naive bayes etc. to calculate a cross-tabulation of observed (true) and predicted classes (model)



1. Discuss, for your task and objectives, **which number(s)** in the confusion matrix is most important, that you wish to minimize or maximize? Why?

So here as we can see the value with 596 are the True Positive Label means the news articles which are Real and classified as Real

Then there are 29 False Positives means the News article which are fake but are classified as real

Also there are 142 False Negative which means these are the news article which are real but classified as fake

And Finally True Negatives with 488 which means a news article which is fake and is successfully classified as Fake by the system

In the case of news articles both False Positives and False negatives can be harmful but more focus should be given on false positivies to be negative as a Fake news labelled as True news can be much more harmful.

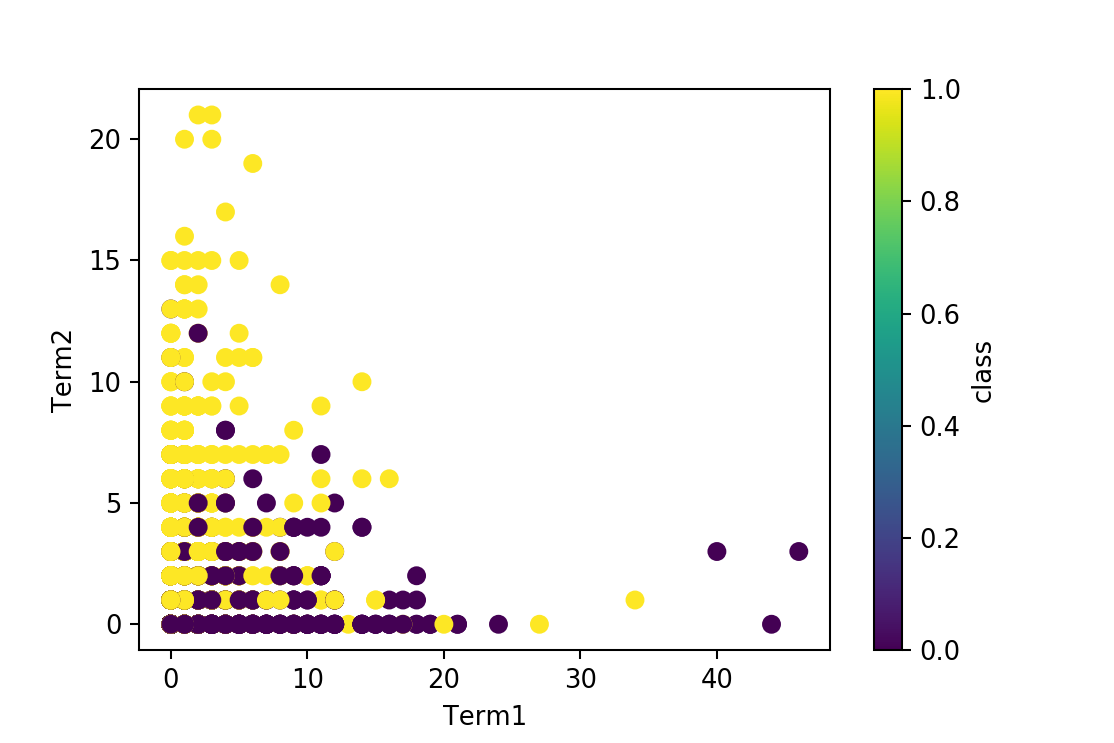
Also it is observed that the acuraccy increases if the value of aplha parameter is closer to 0

D.2. Linear Model

Discuss whether:

1)The two classes are (roughly) **separable** on the plot?

🡪

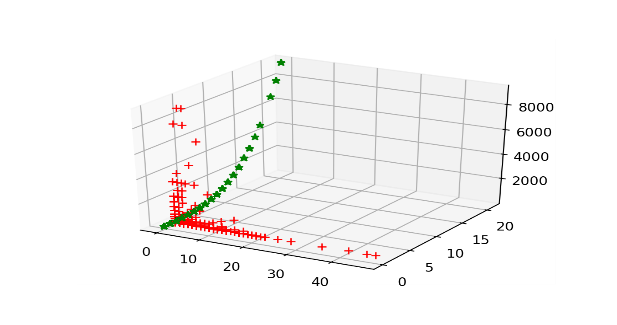


The two classes are (roughly) **linearly separable** on the plot as we can see on row which corresponds to term 1 and term 2 0 are mostly having class 0 which means Fake class but as there are few values with Real class as well we can say that the the classes are roughly linerarly seprarable

2)The two classes are **linearly** separable on the plot?

🡪 The two classes are not linearly sepearable on the plot as there are some overlapping values

1. (Optional) It is possible they are linearly separable in a higher dimensional space with **all term features**?



🡪As we can observe from the 3d plot that the True values are visibly from the False values hence it is possible these are linearly separable in a higher dimensional space with these features.

E.Conclusion

1. **Compare the results in section D**

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | Accuracy | Kappa | Hamming Loss |
| Probabilistic Naïve Bayes Model [BernoulliNB] | 0.86 | 0.73 | 0.14 |
| LinearSVC[Support Vector Classifier] | 0.93 | 0.85 | 0.07 |
| DecisionTreeClassifier | 0.83 | 0.66 | 0.17 |

Meaning of the Metrics Used :-

**Accuracy** is the percentage of correctly classifies instances out of all instances

**Cohen's Kappa** is much like classification accuracy, except for being normalized on ones dataset at random chance baseline.

**Hamming Loss** is the ratio of wrong labels to the total number of labels

Explaination of the models

1. The Bernoulli naive Bayes classifier assumes that all our features are binary such that they take only two values. The best results are achieved when the value of alpha is minimum
2. The second model uses Linear SVC, a Linear SVC (Support Vector Classifier) attempts to suit the data we have, returning a "best fit" hyperplane that separates the data or categorizes them.
3. A decision tree is constructed from a root node downwards and involves partitioning the data into subsets containing instances of related values. The decision tree provides better results when the parameter is set to entropy.
4. **Review your classification task and objectives.**

The classification models are implemented to predicit based on the training dataset which of the following content in the test data set falls under the real or fake news.

1. **Which model gives you the best result so far? Why?**

|  |  |  |  |
| --- | --- | --- | --- |
| LinearSVC[Support Vector Classifier] | 0.93 | 0.85 | 0.07 |

The model with best result is LinearSVC which accurately predicits the tems 93% and a high Kappa Value which means with random data also the accuracy is higher and the lowest observed Hamming Value (third metric) which shows that there are least false values. As in higher dimmension we can see that the data points are seperable and LinearSVC approach uses datapoints on a hyperplane to classify the variables hence the best results in approach are with the LinearSVC.

Also different setting different parameters generate different results in the model such as if the paramenter c (classificiation) should be minimum to achieve the best results.

1. Thoughts on future work on the data?

* The widespread of fake news on the social media and other platforms have been an cause of concern in elections and also creates unrests and sometimes even situations of animosity in the society.

Based on the data and predicitions carried out on the dataset we can use the model on unseen news articles to ascertain that what is the possiblity wether the articles is in the category of real or fake news. Therefore adding more dataset covering other situations such as fakenews related to natural disaster and civil unrest and compiling it one data and bulding a model to predict unseen articles wether the article falls in real or fake classification will be benfitial to the cause.