A mini project report on

Fake News Classifier

Submitted in partial fulfilment of the requirements for the machine learning laboratory during 6th semester of

Bachelor of Technology
in
Computer Science & Engineering

Submitted by :

01FB16ECS405: SURAJ ARALIHALLI
01FB16ECS402: SUMANTH V RAO
01FB16ECS419: TEJAS PRASHANTH
01FB16ECS403: SUMEDH BASARKOD

January – May 2018

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the mini project entitled

Fake News Classifier

Submitted by :

01FB16ECS405: SURAJ ARALIHALLI
01FB16ECS402: SUMANTH V RAO
01FB16ECS419: TEJAS PRASHANTH
01FB16ECS403: SUMEDH BASARKOD

In partial fulfilment for the machine learning laboratory during sixth semester in the Program of Study Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period Jan. 2019 – May. 2019.

Signature
<Guide Name>
< Designation>

Signature
Dr. Shylaja SS
Chairperson

Name of the Examiners Signature with Date

1. ____________________

2. ____________________

1.0 Literature Survey

The reliability of news articles is a pressing issue in today's technology driven world. Due to the prevalence of social media in our everyday lives, the spread of news from one corner of the world to the other has never been faster. However, this has also resulted in the spread of deceptive news articles on web-based platforms and social media, which can lead to misjudgements and formation of extremely biased opinions. For example, researchers claim that biased and false news influenced the outcome of the 2016 US presidential elections.
Thus, the factors that influence the reliability of news articles primarily are the style of writing used and the source of the news article. Firstly, a **styled based approach** can be used to predict the reliability of news articles. For example, the intentional use of exaggeration and catchy phrases in an article and in its title entices a reader's attention. As a result, such phrases may indicate an unreliable news article [1]. This is also evident due to hyper-partisan styles of writing [1], which is a style of writing that indicates extreme behaviour towards a particular political party. Such styles of writing depict a strong objective for creating and spreading fake news. Moreover, research has shown that a deep-syntax model [1] can be used to develop a set of Probabilistic Context Free Grammar(PCFG) rules, which can be used to identify deceptive phrases. Secondly, the **source of the news articles** plays a crucial role in determine its reliability. For example, during the results of 2016 US presidential elections, the most-visited website on Google was an unreliable website called "70news"[2] made a biased statement that President Donald Trump had won both popular vote and the Electoral College vote. **Metadata about an article** such as its creator and subject can be used for determining its reliability. For instance, a **deep-diffusive network model**, a variation of the conventional neural network model, [3] is constructed by using explicit and latent feature extraction techniques to construct features. The primary assumption behind the model is that there exists a strong correlation between an article's creator and the corresponding article, since articles written by a reliable source are less likely to be unreliable. Explicit feature vectors are constructed for articles, authors and their subjects using trivial techniques such as word counts of words in the vocabulary. In addition, latent feature vectors are constructed using a Recurrent Neural Network. Feature vectors from both sources are used in the diffusive unit model. Furthermore, a **Convolutional Neural Network**(CNN) has been used to determine the veracity of political news[4]. The features used are word embeddings (Word2Vec and GloVe) from statements along with metadata on author of the statement. For instance, if the author was an election candidate, then metadata would include employment details, party supported and a history of the number of correct and incorrect statements made by the creator. In this way, the reputation and reliability of a creator can be built, which can be used for determining the reliability of their corresponding articles and statements. Lastly, the presence of quantitative measures affect the reliability. News articles today consists of various facts and statistical information, which needs to be evaluated. As a result, there is a strong need for a model capable of selecting facts that need to be evaluated and a fact checking system that verifies selected information from external sources, as described in [1].

2.0 Environment Requirements

The dataset used for training was obtained from the paper Automatic Detection of Fake news [5]. The dataset consists of various categories of news such as technology, sports, business and politics, which each article present in a separate file. Each article has two version, a fake version as well as a real version. Both the versions are used for building classifiers.

3.0 Proposed Approach

3.1 Feature Extraction

Prior to feature extraction, the input text had to be preprocessed.This included converting the input text to lowercase characters and removal of stop words. Stop words are commonly used words such as “the”, “a”, “an”, “in” etc.

After the input text is preprocessed, a number of feature extraction techniques are applied. Broadly the feature set takes into account the following

1. **TF-IDF**
2. **Readability**
3. **Punctuation usage**
4. **Polarity of the text.**

TF-IDF

To begin with, the preprocessed text was tokenized using the scikit-learn CountVectorizer module. An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of times each word appeared in the document. This count matrix was further transformed to a tf-idf (term frequency- inverse document frequency) matrix. The number of times a term occurs in a document is called its *term frequency.* Because some terms are way too common, term frequency will tend to incorrectly emphasize documents which happen to use the these terms more frequently, without giving enough weight to the more meaningful terms . Hence an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely.

Readability

The 9 Metrics that score understandability of text which are based 9 different formulae are mentioned below

- Linear Write Index

- Dale-Chall Readability Score
- Automated Readability Index
- SMOG Index
- Flesch Reading Index
- Flesch Kincaid Grade
- Linear Write Index
- Difficult Words

4.9 or lower	Average 4th grade student or lower
5.0 to 5.9	Average 5th grade student or 6th grade student
6.0 t0 6.9	Average 6th grade student or 7th grade student
7.0 to 7.9	Average 7th grade student or 8th grade student
8.0 to 8.9	Average 8th grade student or 9th grade student
9.0 to 9.9	Average 9th grade student or 10th grade student

A score of 6.5 conveys that the given text can be effortlessly comprehendable by a 6th grade or 7th grade student.

These 9 different metrics are computed for the text which will be used in the final feature set for the fake news classification.

Punctuation Usage

The usage of punctuations can be a promising feature when engineered wisely. The usage of punctuation gives a lot of insights about the professionality of the source of the news thereby the reliability of the news. The news with grammatical or spelling mistakes is highly likely to come from unprofessional source and unlikely to be genuine. The same goes with improper or over utilisation of punctuation marks like @,!,? Etc.

Some of the punctuation metrics are

- Usage of upper-case alphabets

- Normalized count of punctuations which include ! ? - , . $ etc
- Normalized Word Count
- Count of blank spaces (“ “) normalized by count of full stops (.)

Polarity

The polarity score gauges negativity, positive and neutrality of a statement. It provides insights about the given news in a way if the news is biased in any form or neutral.

Consider the following example and the scores marked for this text

Like you, I am getting very frustrated with this process. I am genuinely trying to be as reasonable as possible. I am not trying to "hold up" the deal at the last minute. I'm afraid that I am being asked to take a fairly large leap of faith after this company has screwed me and the people who work for me.

- Negative : 0.093
- Neutral : 0.836
- Positive : 0.071

4.0 Results

After the feature extraction stage, several classifiers are independently used as models the implementation provided in sklearn and their performance is measured using their accuracy. The following table describes the classifiers used

Classifier used	**Accuracy**
Multi Layer Perceptron	69%
Logistic Regression	65%
Support Vector Machines	54%
Decision Trees	58%
Random Forest Classifier	52%
K Nearest Neighbours	58%

Multi Layer Perceptron(MLP) served as the best classifier for the given dataset. Details of the classifiers are given below:

1. **Multi-Layer Perceptron**

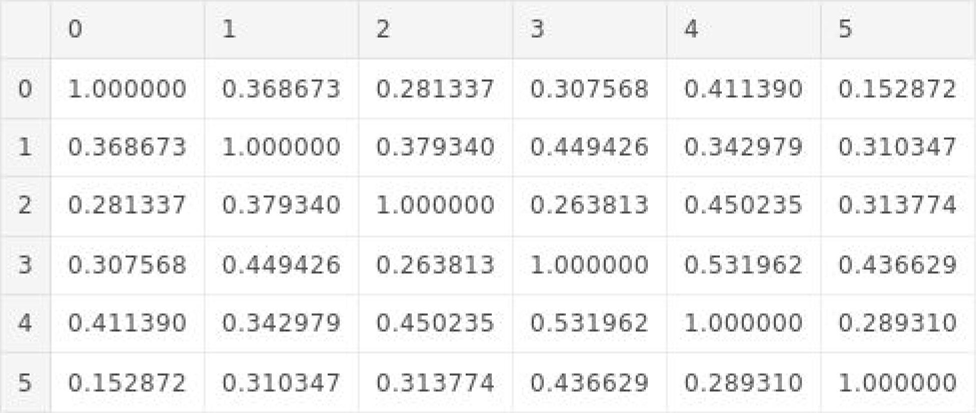
a. Regularization rate(alpha): Higher values of alpha indicate greater amount of regularization, which in turn helps prevents overfitting.
b. Optimizer(solver): The optimizer used for convergence is Limited Broyden-Fletcher-Goldfarb-Shannon(LBFGS)
c. Learning rate: The learning rate is adaptive, which results in a decrease in learning rate after training loss does not increase after two consecutive epochs

2. **Random Forest Classifier**
   a. Number of estimators: Number of trees considered in the forest is 100
3. **K Nearest Neighbours**
   a. K: After testing with various values of K, K was chosen to be 7 since it provided the highest accuracy

Note: Classifiers that are not mentioned use default parameters provided in sklearn

4.1 Ensemble Learning

Since the classifiers obtained are weak classifiers, ensemble learning was attempted in order to improve its accuracy. To elaborate, the method of stacking is used in order to aggregate the results of various classifiers obtained for each training sample and prepare a dataset consisting of hypothesis of various classifiers for each training sample. Thereafter, a meta classifier is run on the dataset of hypothesis, with labels representing the actual training labels from the original dataset. The meta classifier used is Multi Layer Perceptron(MLP), which results in a training accuracy of 90%, but a validation accuracy of 52%. This is due to the lack of diversity of classifiers, as depicted in the table below.



	0	1	2	3	4	5
0	1.000000	0.368673	0.281337	0.307568	0.411390	0.152872
1	0.368673	1.000000	0.379340	0.449426	0.342979	0.310347
2	0.281337	0.379340	1.000000	0.263813	0.450235	0.313774
3	0.307568	0.449426	0.263813	1.000000	0.531962	0.436629
4	0.411390	0.342979	0.450235	0.531962	1.000000	0.289310
5	0.152872	0.310347	0.313774	0.436629	0.289310	1.000000

5.0 Conclusion

In summary, various models are built using two stages-Feature Engineering and Classification. During the feature engineering phase, various features such as tf-idf,readability,punctuation usage, polarity and so on are extracted from the textual data. Thereafter, during the classification phase, various binary classifiers such as Multi Layer Perceptron, Decision Trees, are built using the extracted features in order to classify text as being fake or not, with the Multi Layer Perceptron providing the best accuracy

6.0 Future Work

Among the various classifiers attempted,the *Multi Layer Perceptron*(MLP classifier) provided the best accuracy of nearly 70%. However, several techniques can be used in order to improve the accuracy of the models. Firstly, the amount of training data considered can be increased. The models are trained on a dataset of nearly 790 samples, which can be replicated in order to improve the accuracy. Moreover, variants of Recurrent Neural Network(RNN) are known to provide a better accuracy with large amounts of textual data. Specifically, Long Short Term Memory(LSTM) can also be implemented for fake news detection when provided with large amounts of training data.

7.0 References

1. Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. SIGKDD Explor. Newsl. 19, 1 (September 2017), 22-36. DOI: https://doi.org/10.1145/3137597.3137600
2. http://libraryguides.vu.edu.au/c.php?g=460840&p=5330649
3. Zhang, Jiawei, et al. "Fake News Detection with Deep Diffusive Network Model." *arXiv preprint arXiv:1805.08751*(2018).
4. https://arxiv.org/abs/1705.00648
5. @article{Perez-Rosas18Automatic, author = {Ver\'{o}nica P\'{e}rez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea}, title = {Automatic Detection of Fake News}, journal = {International Conference on Computational Linguistics (COLING)}, year = {2018} },