Prediction of Fake News with Model Comparison

Harsha Vignyan Ayaluri

11904067 KM018

CODE

Abstract

The fake news on social media and various other media is wide spreading and is a matter of serious concern due to its ability to cause a lot of social and national damage with destructive impacts. A lot of research is already focused on detecting it. This paper makes an analysis of the research related to fake news detection and explores the traditional machine learning models to choose the best, in order to create a model of a product with supervised machine learning algorithm, that can classify fake news as true or false, by using tools like python scikit-learn, NLP for textual analysis. This paper applied machine learning on a Kaggle dataset to predict whether an article of news was real or fake. We applied three different classifiers, all yielding promising results.

1. Introduction

Fake News contains misleading information that could be checked. This maintains lie about a certain statistic in a country or exaggerated cost of certain services for a country, which may arise unrest for some countries like in Arabic spring. There are organizations, like the House of Commons and the Crosscheck project, trying to deal with issues as confirming authors are accountable. However, their scope is so limited because they depend on human manual detection, in a globe with millions of articles either removed or being published every minute, this cannot be accountable or feasible manually. A solution could be, by the development of a system to provide a credible automated index scoring, or rating for credibility of different publishers, and news context.

Search engines are one of the leading forces in tackling fake news. For example, Google has implemented various methods to stifle the spread of fake stories [1]. To appear on Google News, a source must first meet Google's guidelines and applies for inclusion.

This paper proposes a methodology to create a model that will detect if an article is authentic or fake based on its words, phrases, sources and titles, by applying supervised machine learning algorithms on an annotated (labelled) dataset that are manually classified and guaranteed. Then, feature selection methods are applied to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results. We propose to create the model using different classification algorithms. The product model will test the unseen data, the results will be plotted, and accordingly, the product will be a model that detects and classifies fake articles and can be used and integrated with any system for future use.

2. Background

The machine learning approach implemented in this study was a simple binary classification architecture. By analyzing the language of thousands of articles in the dataset, the model theoretically should pick up on the text's nuances, which it can then use to classify a piece of news as real or fake. A problem solved in this fashion is spam email detection. The two issues are remarkably similar: spam emails, like fake news articles, are designed to lure and mislead people while appearing in obnoxious quantities and filling up a user's feed. Sharaff et al. compared the performance of four different types of classifiers on labeling emails in a binary fashion: spam or legitimate [3]. Their best performing model achieved an accuracy exceeding 0.93, which shows promise for a binary architecture in predicting the legitimacy of news pieces. Other similar areas of study include a different method of categorizing fake news called stance detection. Thota et al. focused on implementing a neural network architecture that determines the stance of an article in regards to a given headline [5]. The article could agree, disagree, be neutral, or be completely unrelated. By comparing the article's stance and the mainstream stance on the headline, their model would predict whether the article was fake or real. Their models achieved accuracies of over 0.94 when tested on the Fake News Challenge [2] data set. Other previous studies factor auxiliary information besides the text itself into their predictions, such as sentiments [6] and the social engagement and profiles of viewers [7]. In this specific study, we will be focusing on a binary classification architecture, considering just the text provided in the dataset.

3. Approach

The dataset used is available on Kaggle, and contains one set of real news and another set of fake news. All the models here were created through Jupyter Notebook and its available software libraries, including SciKit-Learn [4], Numpy [8], Pandas [9], and more.

4. Data Description

Train.csv: A full training dataset with the following attributes:

- id: unique id for a news article
- title: the title of a news article
- author: author of the news article
- text: the text of the article; could be incomplete

label: a label that marks the article as potentially unreliable

- 1: unreliable
- 0: reliable

test.csv: A testing training dataset with all the same attributes at train.csv without the label.
submit.csv: A sample submission that you can

5. Evaluation Metric

$$accuracy = \frac{correct\ predictions}{correct\ predictions + incorrect\ predictions}$$

6. Work Flow

6.1 Import the necessary libraries

We import the necessary librarires for the project like numpy, pandas, re and nltk. For the classification algorithms we use logistic regression hence, we import that library as well along with MultinomialNB for the multinomial naïve baiyes and also Passive Aggressive Classifier. We also import some of the libraries which are required for plotting graph and dealing with the data throught the project like Classification report, Confusion Matrix, Term Frequency – Inverse Document Frequency.

6.2 Download stop words

In English grammar there are many stop words and to make our work easy in this project we first download the stop words using nltk and print them later for the user understanding.

6.3 Pre Processing Data

We will import the data and preprocess it by replacing the null values with empty string and reading the final data again, later we merge the author name and the news title followed by sepearting the data and label into x and y.

After preprocessing the data it looks like:

	id	title	author	text	label
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

After merging the author name and the news title:

```
0        Darrell Lucus House Dem Aide: We Didn’t Even S...
1        Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2        Consortiumnews.com Why the Truth Might Get You...
3        Jessica Purkiss 15 Civilians Killed In Single ...
4        Howard Portnoy Iranian woman jailed for fictio...
                               ...
20795    Jerome Hudson Rapper T.I.: Trump a ’Poster Chi...
20796    Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797    Michael J. de la Merced and Rachel Abrams Macy...
20798    Alex Ansary NATO, Russia To Hold Parallel Exer...
20799              David Swanson What Keeps the F-35 Alive
Name: content, Length: 20800, dtype: object
```

6.4 Stemming

A stemming algorithm is a process of linguistic normalisation, in which the variant forms of a word are reduced to a common form. In simple words, Stemming is the process of reducing a word to its Root word. Examples:
actor, actress, acting --> act
eating, eats, eaten --> eat

6.5 Converting the textual data to numerical data

TF-IDF stands for “Term Frequency – Inverse Document Frequency ”. TF-IDF is a numerical statistic which measures the importance of the word in a document. Term Frequency : Number of time a word appears in a text document. Inverse Document Frequency : Measure the word is a rare word or common word in a document.

Example:

```
(0, 15686)    0.28485063562728646
(0, 13473)    0.2565896679337957
(0, 8909)     0.3635963806326075
(0, 8630)     0.29212514087043684
(0, 7692)     0.24785219520671603
(0, 7005)     0.21874169089359144
(0, 4973)     0.233316966909351
(0, 3792)     0.2705332480845492
(0, 3600)     0.3598939188262559
(0, 2959)     0.2468450128533713
(0, 2483)     0.3676519686797209
(0, 267)      0.27010124977708766
(1, 16799)    0.30071745655510157
(1, 6816)     0.1904660198296849
(1, 5503)     0.7143299355715573
(1, 3568)     0.26373768806048464
```

6.6 Splitting

Now we will split the data into train data and test data representing them with x and y assigning a particular test size and random state.

6.7 Logistic Regression

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables[10]. For example,
To predict whether an email is spam (1) or (0)
Whether the tumor is malignant (1) or not (0)

Model

Output = 0 or 1

Hypothesis => Z = WX + B

hΘ(x) = sigmoid (Z)

Sigmoid Function

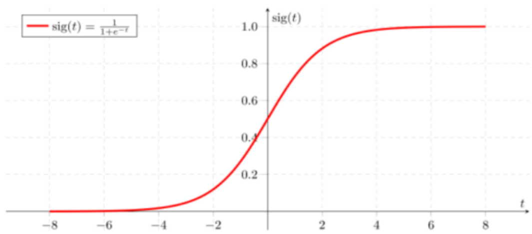


Figure 2: Sigmoid Activation Function

If ‘Z’ goes to infinity, Y(predicted) will become 1 and if ‘Z’ goes to negative infinity, Y(predicted) will become 0.

6.8 Confusion matrix

The confusion matrix is a 2 dimensional array comparing predicted category labels to the true label.

		Classifier Prediction	
		Positive	Negative
Actual Value	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Confusion Matrix for Binary Classification

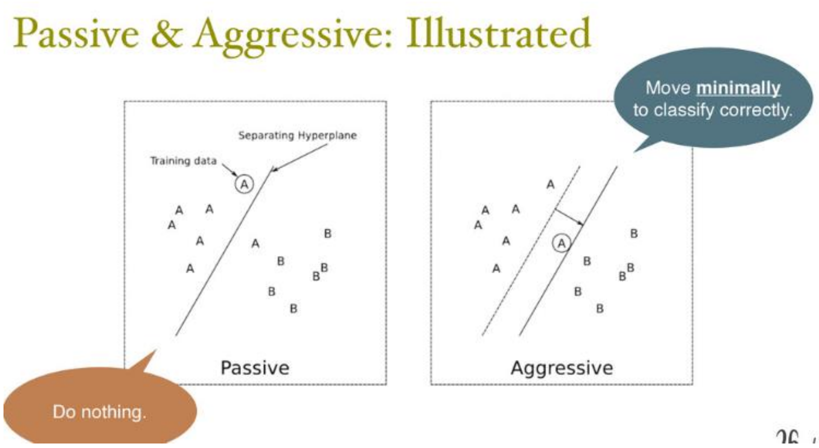
6.9 Multinomial Naive Bayes

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). Multinomial Naïve Bayes uses term frequency i.e. the number of times a given term appears in a document. Term frequency is often normalized by dividing the raw term frequency by the document length.

$$\boxed{P(C_1 \mid x_1 \cap x_2 \cap x_3 \cap x_4) = \frac{P(x_1 \mid C_1) * P(x_2 \mid C_1) * P(x_3 \mid C_1) * P(x_4 \mid C_1) * P(C_1)}{P(x_1) * P(x_2) * P(x_3) * P(x_4)}}$$

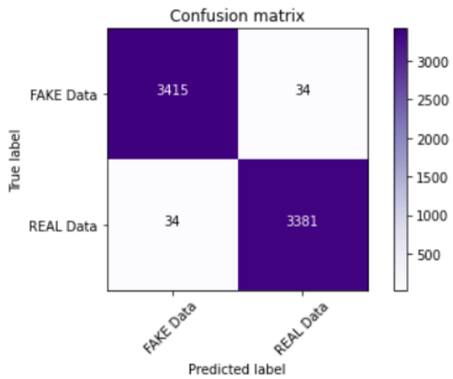
6.10 Passive Aggressive Classifier

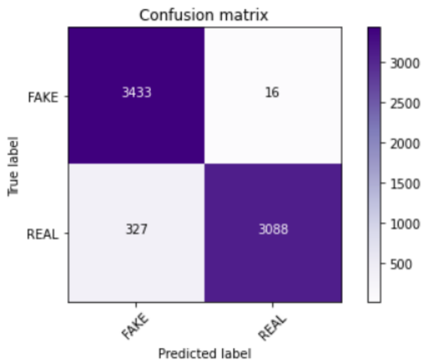
Passive Aggressive Classifier works by responding as passive for correct classifications and responding as aggressive for any miscalculation.



6.11 Building a predictive system

Building a predictive system inorder to find that the inital word in the dataset is real or fake using LogisticRegression model, Multinomial Naive Bayes, Passive Aggressive Classifier.





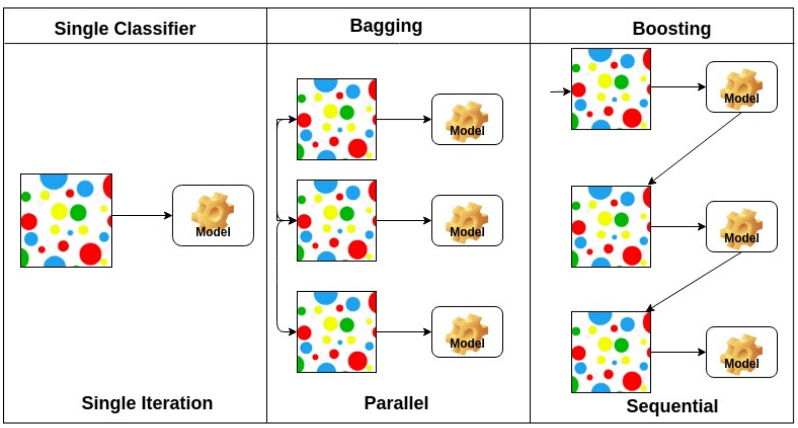
6.12 Classification report

A Classification report is used to measure the quality of predictions from a classification algorithm. It is used to show the precision, recall, F1 Score, and support of your trained classification model. Later we do model comparision for all the three algorithms.

- Logistic Regression : Accuracy is 0.98
- Multinomial Naive Bayes : Accuracy is 0.95
- Passive Aggressive Classifier : Accuracy is 0.99

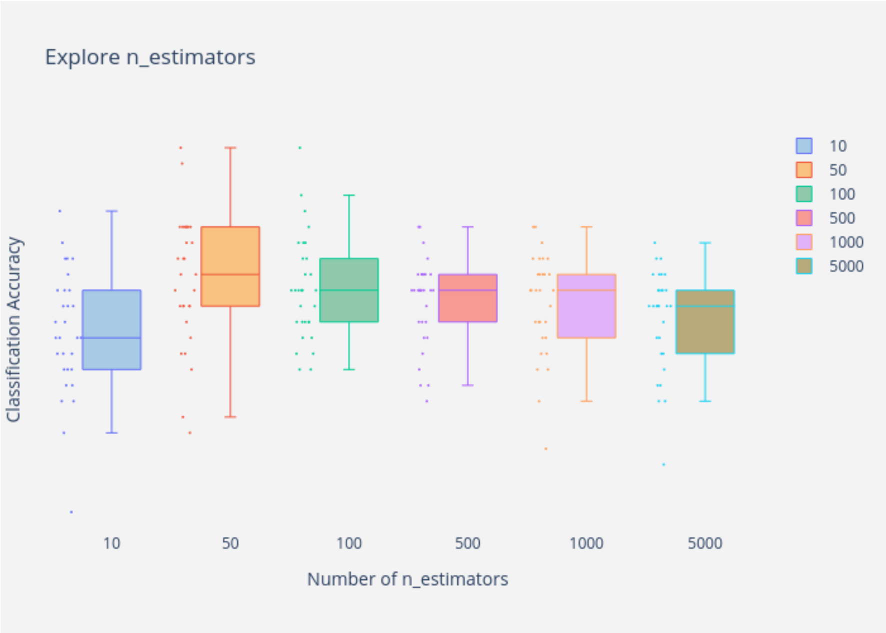
6.13 Ensemble Model

In this project we used boosting method with the help of AdaBoostClassifier. A weak learner is a model that is very simple, although has some skill on the dataset. Boosting was a theoretical concept long before a practical algorithm could be developed, and the AdaBoost (adaptive boosting) algorithm was the first successful approach for the idea.



6.14 Hyper Parameter Tuning

Often by changing the number of base models or weak learners we can adjust the accuracy of the model. The number of trees added to the model must be high for the model to work well, often hundreds, if not thousands. Afterall the more is the number of weak learners, the more the model will change from being high biased to low biased.



7. Conclusion

The research in this paper focuses on detecting the fake news by reviewing it in two stages: characterization and disclosure. In the first stage, the basic concepts and principles of fake news are highlighted in social media. During the discovery stage, the current methods are reviewed for detection of fake news using different supervised learning algorithms. We did the prediction using three classifiers and created an ensemble model finding the best classifier with high accuracy.

8. References

[1] Bowden, I. (2016, December 1). How Google is tackling fake news, and why it should not do it alone. Search Engine Land. https://searchengineland.com/google-tackling-fake-newsnot-alone-264058

[2] http://www.fakenewschallenge.org/

[3] Sharaff A., Nagwani N.K., Dhadse A. (2016) Comparative Study of Classification Algorithms for Spam Email Detection. In: Shetty N., Prasad N., Nalini N. (eds) Emerging Research in Computing, Information, Communication and Applications. Springer, New Delhi. https://doi.org/10.1007/978-81-322-2553-9_23

[4] https://scikit-learn.org/

[5] Thota, Aswini; Tilak, Priyanka; Ahluwalia, Simrat; and Lohia, Nibrat (2018) "Fake News Detection: A Deep Learning Approach," SMU Data Science Review: Vol. 1 : No. 3 , Article 10. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/10

[6] O. Ajao, D. Bhowmik and S. Zargari, "Sentiment Aware Fake News Detection on Online Social Networks," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 2507- 2511, doi: 10.1109/ICASSP.2019.8683170.

[7] K. Shu, S. Wang and H. Liu, "Understanding User Profiles on Social Media for Fake News Detection," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, 2018, pp. 430-435, doi: 10.1109/MIPR.2018.00092.
[8] https://numpy.org/
[9] https://pandas.pydata.org/
[10] https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a#:~:text=A%20confusion%20matrix%20is%20a,the%20true%20values%20are%20known.