INVESTIGATING FAKE NEWS DETECTION USING MACHINE LEARNING ALGORITHMS: SCIKIT-LEARN

[*]K. Parameswara Rao & [#]Dr. M. Seshashayee

[*]PG Student, MCA, GIS, GITAM, Visakhapatnam. parameshkarumuri1997@gmail.com

[#]Assistant professor, GIS, GITAM, Visakhapatnam. mseshashayee@gmail.com

Abstract: *The main objective is to detect the fake news, which is a classic text classification problem with a straight forward proposition. It is needed to build a model that can differentiate between "Real" news and "Fake" news. This paper shows a simple approach for fake news detection using Scikit-Learn classifier. This approach was implemented as a software system and tested against a data set of Twitter posts. We achieved classification accuracy of approximately 93% on the test set which is a decent result considering the relative simplicity of the model. This result may be improved in several ways that are described in the article as well.*

Keywords: Supervised Learning Algorithms, Scikit-Learn, Multi Nominal Naive-bayes classifier, Testing Linear Models.

1. INTRODUCTION

Now a day's fake news is creating different issues from sarcastic articles to a fabricated news and plan government propaganda in some outlets. Fake news and lack of trust in the media are growing problems with huge ramifications in our society. Obviously, a purposely misleading story is "fake news" but lately blathering social media's discourse is changing its definition. Some of them now use the term to dismiss the facts counter to their preferred viewpoints.

Twitter posts have been at the epicenter of much critique following media attention. They have already implemented a feature to flag fake news on the site when a user sees' it; they have also said publicly they are working on to distinguish these articles in an automated way. Certainly, it is not an easy task. A given algorithm must be politically unbiased – since fake news exists on both ends of the spectrum – and also give equal balance to legitimate news sources on either end of the spectrum. In addition, the question of legitimacy is a difficult one. However, in order to solve this problem, it is necessary to have an understanding on what Fake News is. Later, it is needed to look into how the techniques in the fields of machine learning, natural language processing help us to detect fake news.

2. CLASSIFICATION MACHINE LEARNING TECHNIQUES: SCIKIT-LEARNING

2.1. Scikit-Learning Algorithm

Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms. It's built upon some of the technology you might already be familiar with, like NumPy, pandas, and Matplotlib.

a) **Numpy**- NumPy is the fundamental package for scientific computing with Python. It contains among other things: NumPy is a module for Python. The name is an acronym for "Numeric Python" or "Numerical Python". It is an extension module for Python, mostly written in C.

b) **Pandas**- pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive.

c) **Matplotlib-** matplotlib.pyplot is a collection of command style functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a

figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

2.2. Building Vectorizer Classifiers

Now that training and testing data, we can build the classifiers. To get a good idea if the words and tokens in the articles had a significant impact on whether the news was fake or real, you begin by using `Count Vectorizer` and `Tfidf Vectorizer`. This removes words which appear in more than 70% of the articles. Also, the built-in `stop_words` parameter will remove English stop words from the data before making vectors.

a) **Tfidf Vectorizer-** Transforms text to feature vectors that can be used as input to estimator. Vocabulary is a **dictionary** that converts each token (word) to feature index in the matrix, each unique token gets a feature index. Each sentence is a vector; the sentences you've entered are matrix with 3 vectors.

b) **Count Vectorizer-** The Count Vectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary. It can be use as follows: Create an instance of the Count Vectorizer class.

c) **Hashing Vectorizer-** Another vectorizer used sometimes for text classification is a `Hashing Vectorizer`. `Hashing Vectorizer require less memory and are faster (because they are sparse and use hashes rather than tokens) but are more difficult to introspect. It can read a bit more about the pros and cons of using `Hashing Vectorizer`.

2.3. Classifier (multinomial naïve bayes classifier)

Combining probability distribution of P with fraction of documents belonging to each class.

3. ABOUT DATA-SET AND ATTRIBUTES

There must be two parts to the data-acquisition process, “fake news” and “real news”. Collecting the fake news was easy as Kaggle released a fake news dataset consisting of 13,000 articles published during the 2016 election cycle. Now the later part is very difficult. That is to get the real news for the fake news dataset. It requires huge work around many Sites because it was the only way to do web scraping thousands of articles from numerous websites. With the help of web scraping a total of 5279 articles, real news dataset was generated, mostly from media organizations (New York Times, WSJ, Bloomberg, NPR, and the Guardian) which were published around 2015 – 2016.

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL

Fig.1. Fake News Dataset

4. RESULTS AND ANALYSIS

Confusion Matrix: A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused

when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

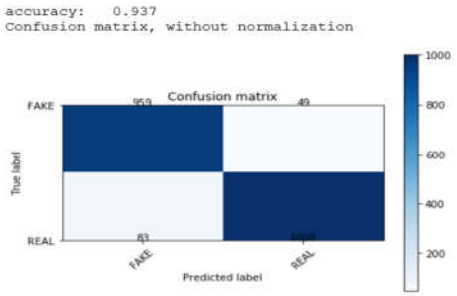

Fig.2. Confusion Matrix

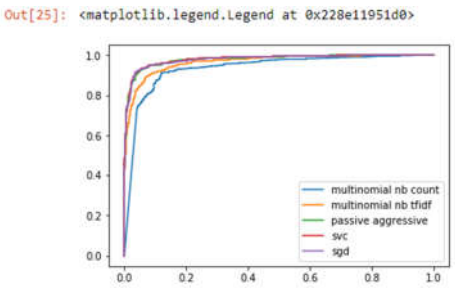

Fig.3. Block Diagram

5. CONCLUSION

The Scikit-Learn Algorithm giving 0.937 Accuracy for the fake news detection. The problems of fake news and disinformation play an important role on nowadays life. This is because the advanced level of technology and communication methods we have enabled information spreading among people without any verification. This is a reason why researchers started searching for solutions to stop fake news and disinformation from spreading easily. However, it is well known that controlling the flow of information is impossible.

6. FUTURE SCOPE

As future work, we plan to better study the combination between the feature extraction methods and the classifiers as we will be able to choose the text representation model that performs best with the classifier. Moreover, to achieve a higher accuracy. It can be implement a more sophisticated algorithm which may use data mining technologies with big data, because creating a big dataset including more types of news articles with more class variables (labels) will help raising the accuracy score.

REFERENCES

Journal Articles

[1] *N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.*

[2] *S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Association for Computational Linguistics, 2012, pp. 171–175.*

[3] *Shlok Gilda,Department of Computer Engineering, Evaluating Machine Learning Algorithms for Fake News Detection,2017 IEEE 15th Student Conference on Research and Development (SCOReD).*

[4] *K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," SIGKDD Explor. Newsl., vol. 19, pp. 22–36, Sept. 2017.*

[5] *R. J. Sethi, "Crowdsourcing the verification of fake news and alternative facts," in Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT '17, (New York, NY, USA), pp. 315–316, ACM, 2017.*

[6] *V. Rubin, N. Conroy, Y. Chen, and S. Cornwell, "Fake news or truth? using satirical cues to detect potentially misleading news," in Proceedings of the Second Workshop on Computational Approaches to Deception Detection, (San Diego, California), pp. 7–17, Association for Computational Linguistics, June 2016.*

[7] *A. Singhal, "Modern information retrieval: a brief overview," BULLETIN OF THE IEEE COMPUTER SOCIETY TECHNICAL COMMITTEE ON DATA ENGINEERING, vol. 24, p. 2001, 2001.*

[8] *A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop clickbait: Detecting and preventing clickbaits in online news media," in Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, pp. 9–16, IEEE, 2016.*