University of Sunderland

The Detection of Fake News with Data Science

Student Name: Adigun Abdul Azeez Oladipupo
Student Number: 219143011

Introduction and problem area

Looking at the term Fake news, it has been defined in various ways, (Gelfert, 2018) define Fake News as the intentional declaration of deceptive statements as news, in which the allegation are designed to mislead. Recently, Fake News is seen as a very important issue nationwide even though it has been going on for several years. Both the political election and Brexit in 2016 were strong instances to see the consequences of Fake News in this new era (Rogers and Bromwich, 2016). Considering the characteristics of the internet, someone could easily circulate false or biased News, which makes the attempt to avoid the creation of Fake news nearly impossible. Hence, the most appropriate step to take is finding new method of identifying and separating false news from genuine news. A method of determining authenticity is through fact checking, although it is tedious and needs abilities which are not common with people. Therefore, a really good idea is automating the detection of Fake news through the utilization of the various Data Science techniques and processes.

Data Science is a discipline which aims at finding patterns within data, relating that to this research, it therefore could support the world in distinguishing whether the news is fake or not. Additionally, with the enormous datasets combined with algorithms, Data science assist in giving the required understand to discover patterns from the data which could have never been realized or would have taken a long period to discover. Particularly, the utilization of both Machine Learning and Artificial Intelligence helps in detecting patterns characterizing Fake News which are not so visible to human sight (Rashkin *et al.*, 2017).

Over the years, several approaches were attempted for detecting Fake news using Artificial Intelligence, in which most the studies focus more on brief reports rather than the whole news articles. Great number of the studies done by researchers paid more attention to short statements with various length, and these statements were gotten from social media such as twitter or text messages. A commonly known dataset, LIAR originated from a statement database Politifact, and it has six stages of truth values including the writer's data (Wang, 2017). They are limited datasets available for complete articles, and this is due to the easy nature of labelling a statement unlike a whole article (Lazer *et al.*,2018).

FakeNewsNet is a modest dataset that contains supplementary data, and amongst the unique datasets that holds a whole article. The collection of the data is done by gathering articles that were published on twitter and it contains detailed information about the person posting the article including any additional social media context (Shu *et al.*, 2020). BS DETECTOR is also a dataset that assist with websites and label listing. Although, it does not accept articles, just URLs are required in which a lot of these websites might not function anymore or be accessible (Horvath, Bray and Krigsman, 2019). Hence, it is complex or impractical collecting articles within the dataset's sources. An additional Fake news dataset that holds more than 40,000 articles is called ISOT, ISOT Fake News Dataset is considered to be the largest dataset existing currently (Ahmed, Traore and Saad, 2018). Although, the entire articles within the dataset with a label "true" are linked to Reuters. The "true" articles distort the data since Machine Learning Algorithms could discover the way Reuters authors therefore "learn" to categorize the news to be "not fake" if the pattern aligns with it.

The Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) are the various types of classifiers scientist use to detect fake news (Kim, 2014). Although, the Convolutional Neural Network (CNN) is created for computer vision applications, it has the tendency of being relatively useful. Nevertheless, The Long Short Term Memory units (LSTM) displayed better performance than the Convolutional Neural Network (CNN) on few occasions (Long *et al.*, 2017). LSTM have the capability to "Forget" some information aimed on or "remember" other important information. Therefore, they perform efficiently with substantial amount of text data.

(Rubin *et al.*, 2016) were able to design a classifier attaining 90% accuracy and could differentiate if it is a "legitimate" or a "satire" news. Although, their concentration was mainly on the "satire news" due to how misleading it is, even though it is not as misleading compared to that of Fake News. Satire is expected to be obviously false unlike Fake News with pure intention to mislead. Rubin et al based their research around satire instead of Fake News since detecting satire is easier.They made use of little amount of dataset containing 90 test articles and 290 training articles. They utilized the Support Vector Machine classifier because it is perfect to classify binaries, however it not suitable looking at multiclass classification.

Project plan and data/simulated data

This report aims at discovering the correlating pattern of possible fake news, without a doubt, every classification analysis will need the involvement of people overtime, even if attaining 100% precision might not be achievable, detecting the similarities common in fake news can be seen as a progress. Therefore, the research aims at gathering huge volume of data that has been confirmed to be Fake News then attempt training a model which would relate some news sample to check whether it is fake news or not.

When classifying news articles, the unprocessed information needs to be converted to a more valuable thing, which is also known as feature extraction. They are several forms of features extraction such as n-gram, sentiment analysis, word count and a lot more. After the extraction of the features has been completed it can therefore be use in classifying articles where the features are gotten, various features might provide unique outcome which varies on how the structure of the data is underlain. For pattern detection within the data, Various classifiers will be tested with various features. Through the determination of the most appropriate features for Fake News classification, they are increasing possibility for a programmed Fake News detector .

Computational methods to be used to analyse data and detect patterns.

To detect patterns in Fake News, the initial process is gathering and labelling the Fake news coupled with the genuine news which are both required to almost be of similar amount in other to prevent the rate of recurrence of Fake news dataset to be utilized as a deciding element for classification. Gathering adequate data is important for justifiable outcome. An adequate data with regards to this situation can be said to be a data which represent reality and can be use for generalization.

ISOT Fake News Dataset will be utilize for training the classifiers, which is the biggest dataset with a whole Fake News article (Ahmed, Traore and Saad, 2018). They are 44,898 articles in the ISOT dataset in which 21,417 of them are labelled Real while 23,481 of them are labelled

Fake, an additional dataset that contains full length article is FakeNewsNet, although, it contains just a total of 422 labelled articles (Shu *et al.*, 2020). Finally, 180 articles were gathered by author, in which 90 are Fake and 90 are Real and they will be described as Original Data. Hence, the two added dataset would be utilized for testing how accurate the classifiers that were trained are.

Every single model will undergo training containing 80% of the ISOT data, while the other 20% of the ISOT data would be utilise for testing how accurate the trained classifiers are. As previously stated, both the Original Data and FakeNewsNet would also be used for analysis, reason being that the extra analysis would confirm the detection of Fake News within the ISOT dataset and not unrelated things like the style of a certain news company.

All the articles within the ISOT dataset that has a labelling called "Real" were gathered from Reuters. The entire article that are in it were termed "Reuters", it is a noticeable sequence that is clear to both humans and machines. The method use for avoiding this problem is removing the term "Reuters" at the start of every article.

Pattern detection, discoveries that can be made

This work will be reporting on the detection found in similar research by (Horvath, Bray and Krigsman, 2019). Random Forest and Naïve Bayes are the two models being used for feature extraction, the comprehensive analysis for every feature set will be discussed below. Both the features and classifiers will be evaluated by how accurate they are, in other word, The accurate classification percentage done with the classifier.

The result of the Count Word and Count n-gram indicated that the analysis done on the ISOT data has greater accurate result compared to both the FakeNewsNet and Original Dataset. the next dataset with a high precision rate is the original dataset. Which implies that both Original and ISOT dataset have almost similar makeup compared to that of FaknewsNet. In addition, the data also indicated that Naïve bayes classifier is better with generalization compared to random Forest classifier. With count-ngram, Naïve bayes classifier has a greater accuracy level, unlike the Random Forest classifier that could not have a clear win with either count-word or count n-gram.

The result of the TFIDF-word and TFIDF-ngram indicated that the examined ISOT data display the greater precision level again, but in this case, the RF classifiers has a greater finding within ISOT dataset compared to the Naïve Bayes. Although, Naïve Bayes has stronger generalization with both FakeNewsNet and Original Dataset. The TFIDF-word showed a stronger precision rate than TFIDF-ngram. Looking at Random Forest's classification for Original Dateset.6.47% increase precision level is seen by the TFIDF-word. Original dataset has a greater classification compared to that of FakeNewsNet. When comparing TFIDF and Count, one with the strongest precision results is Count-ngram.

The ER result indicated that ISOT is performing great, naïve Bayes shows better generalization in comparison with the prior features, Although, The ER feature is inferior when it is alone. Nevertheless, A conclusion would not be drawn that ER feature is weak. Additional analysis can be made using the combination of ER along more Features Rather than ignoring ER to be a feature to detect Fake News.

The PoS result indicated that finally, Naïve bayes is not having a stronger generalization compared to Random Forest. Likewise, the precision level is less than 50%, this therefore indicate that the utilization of this features for classification is like guessing randomly. With a precision level of 44.70%, conclusion can be made that PoS alone is not a useful feature for classifying Fake New. Although, they are possibilities which shows that if merge with different feature, PoS is capable of being a strong feature.

The result of VADER indicates VADER feature to be unfavourable toward the precision rates. Still not a strong reason for making conclusion that VADER is not capable of being useful if merge with different features. But assumptions can be made that VADER is by no means useful for Fake News Classification.

The result of Stop Word indicates ISOT to be leading the level of precision yet again, followed by the original dataset, according to the findings, Naïve Bayes classifier shows a stronger generalization compared to Random Forest classifier, but they were really close. Stop Words NLTK list is more advance in detecting Fake News than the spaCy list. According to the findings, An observation can be made that NLTKStop and spaCyStop gave more assistant to Original dataset than that of Count-Word. The assistant from NLTKStop and spaCyStop could also be detected for FakeNewNet dataset but it was not as much like that of Original Dataset.

Lastly the result of lemma also indicated that ISOT has the best precision level, followed by the Original dataset. The Naïve bayes classifier indicate a stronger generalization compared to Random Forest classifier. Lemma showed good findings compare to some other features.

According to the findings some other deduction can be drawn, the noteworthy one can be said to be that the precision level of the ISOT test data is way stronger compared to the remaining dataset. Therefore, an assumption can be drawn that the two classifiers are detecting some sequence with the ISOT dataset, and these patterns being identified by the classifiers might be sequences that were observed within Reuters articles and maybe they might be other sequences existing primarily with the ISOT dataset like the topic of the article or political leaning.

This could imply that the ISOT dataset are not excellent to be used for training. In addition, the original dataset classification is way more accurate compared to that of FakeNewsNet dataset, The articles in Original dataset are not within the Reuters, this therefore could not clarify the rapid increase in precision, thus there is a probability that the Fake News in both ISOT and Original Dataset have a close underlaying structure. Looking at the precision level,a conclusion could be drawn that both the Counts and TFIDF have a greater generalization compared to VADER, ER and PoS.

REFERENCES

Ahmed, H., Traore, I. and Saad, S., 2017, October. Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments* (pp. 127-138). Springer, Cham.

Ahmed, H., Traore, I. and Saad, S. (2018) 'Detecting opinion spams and fake news using text classification', *Security and Privacy,* 1(1), p. e9. doi: 10.1002/spy2.9.

Gelfert, A. (2018) 'Fake news: A definition', *Informal Logic,* 38(1), pp. 84–117. doi: 10.22329/il.v38i1.5068.

Horvath, B., Bray, D. A. and Krigsman, M. (2019) 'Using Data Science to Detect Disinformation', *CXOtalk*. Available at: https://www.cxotalk.com/episode/using-data-science-detect-disinformation.

Kim, Y. (2014) 'Convolutional neural networks for sentence classification', *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference,* pp. 1746–1751. doi: 10.3115/v1/d14-1181.

Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D. and Schudson, M., 2018. The science of fake news. *Science*, *359*(6380), pp.1094-1096.

Long, Y. *et al.* (2017) 'Fake News Detection Through Multi-Perspective Speaker Profiles', *Proceedings of the Eighth International Joint Conference on Natural Language Processing,* Volume 2:(8), pp. 252–256. Available at: http://www.aclweb.org/anthology/I17-2043.

Rashkin, H. *et al.* (2017) 'Truth of varying shades: Analyzing language in fake news and political fact-checking', *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings,* (August), pp. 2931–2937. doi: 10.18653/v1/d17-1317.

Rogers, K. and Bromwich, J.E., 2016. The hoaxes, fake news and misinformation we saw on election day. *The New York Times, 8*.

Rubin, V. *et al.* (2016) 'Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News', pp. 7–17. doi: 10.18653/v1/w16-0802.

Shu, K. *et al.* (2020) 'FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media', *Big Data*, 8(3), pp. 171–188. doi: 10.1089/big.2020.0062.

Wang, W. Y. (2017) '"Liar, liar pants on fire": A new benchmark dataset for fake news detection', *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers),* 2, pp. 422–426. doi: 10.18653/v1/P17-2067.