

Fake News Detection system

Anudeep S. Nerwal, Pushwinderjit S. Mandahar, Manjot Gill

W0785695@myscc.ca, W0780371@myscc.ca, W0770110@myscc.ca

Data Analytics for Business, July 2022.

St. Clair College, Windsor, ON, N9G 3C3.

Abstract

With the recent boom in social media, the spread of fake news has become a big problem for everyone. It has been used to manipulate public opinion, to influence elections - especially the 2016 US presidential election, to incite hatred and unrest. According to a 2018 MIT study, fake news spreads six times faster than real news on Twitter. The news media's credibility and trust are at an all-time low. It is becoming increasingly difficult to distinguish between real and fake news. In this study, we attempted to achieve this using a passive-aggressive classifier and natural language processing.

Dissemination of misleading information daily access to media such as social media sources, news blogs and online newspapers has done it difficult to identify credible news sources the growing need for computational tools that can provide insight into the reliability of online content. In this paper, we focus on the automatic identification of fake content in news articles. First, we introduce the data set for the task of detecting fake news. We describe the preprocessing, feature extraction, classification, and prediction process in detail. We used linguistic processing of logistic regression fake news classification techniques. Preprocessing functions perform some operations such as tokenization, emerging and exploratory data analysis as a response variable distribution and data quality control. A simple bag of words, n-grams, and TF-IDF is used as feature extraction techniques. Logistic regression the model is used as a classifier to detect fake news probability of truth.

Introduction

Fake news is false information presented as news. It often aims to damage the reputation of a person or entity or to make money through advertising revenue. Once common in print, the prevalence of fake news has increased with the rise of social media, particularly the Facebook News Feed. During the 2016 US presidential election, various kinds of fake news about the candidates were widely spread on online social networks, which can have a significant impact on the election results. According to a post-election statistical report, online social networks accounted for more than 41.8% of fake news traffic in the election, far more than traditional TV/radio/print media and online search engine traffic shares. respectively. Detecting fake news is becoming increasingly difficult because people with bad intentions write fake news so convincingly that it is difficult to separate it from the real news. What we've done is a simplified approach that looks at news headlines and tries to predict whether they might be fake or not.

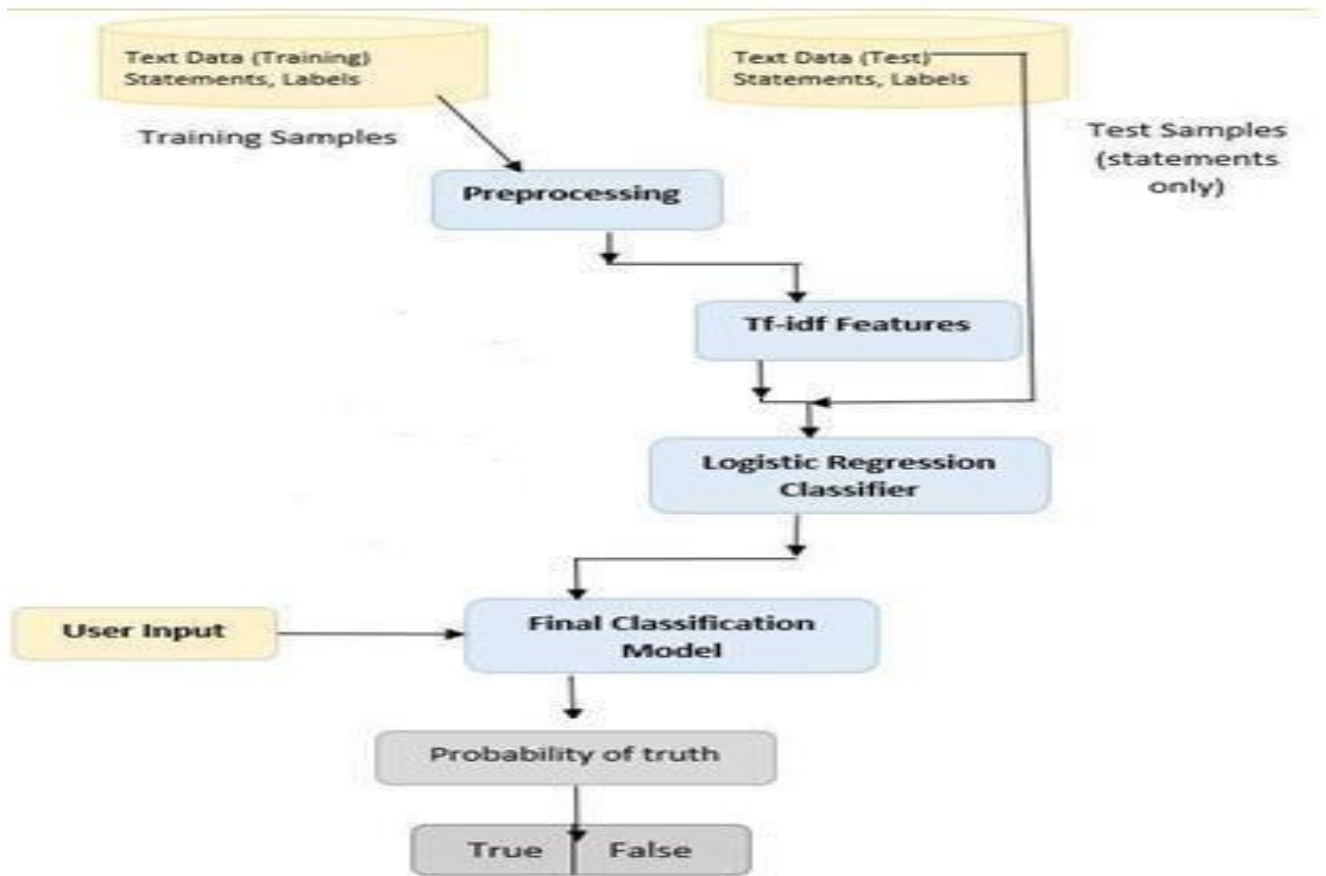
Fake news can be intimidating because it attracts more viewers than usual. People use them because it can be a very good marketing strategy. But the money earned may not make up for the fact that it can harm people.

When we read a sentence or a paragraph, we can interpret the words concerning the entire document and understand the context. In this project, we teach a system how to read and understand the differences between real and fake news by employing natural language processing, NLP, and machine learning concepts, as well as prediction classifiers such as the Logistic regression, which predicts the truthfulness or falsity of an article.

Problem Statement

Nowadays it is extremely difficult to decide whether the news we come across is real or not. There are very few ways to verify the authenticity and they are all sophisticated and not available to the average person. There is an urgent need for a web-based fact-checking platform that harnesses the power of machine learning to provide us with this opportunity.

Methods



Dataset Description:

We used two datasets: one from the Kaggle competition and another from the ISOT fake news dataset. Both datasets primarily contain World News, World Politics, Government, North American News, and other types of news.

We have 5 attributes and 20800 rows in the Kaggle dataset.

- 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

• News	Number of articles
• Real-News	10387
• Fake-News	10413
• Total	20800

The ISOT fake news dataset is made up of two CSV files. "True.csv," contains over 12,600 reuter.com articles. The second file, "Fake.csv," contains over 12,600 articles from various fake news outlet resources. Each article includes the following information: article title, text, and author as well as the date the article was published corresponds with the fake news data collected

for kaggle.com, We concentrated on gathering articles from 2016 to 2017. The collected data was cleaned however, the punctuation and errors that existed in the fake news were preserved in the text. The table below breaks down the categories and the number of articles in each.

News	Size (Number of articles)
Real-News	21417
Fake-News	23481
Total	44898

We have combined both datasets and created one CSV file in which we have the following attributes:

1. Id: unique id for a news article
2. Title: the title of a news article
3. Text: the text of the actual news article
4. Label: a label that marks the article as potentially unreliable and reliable
5. Datatype: Identification of dataset over Kaggle and ISOT Dataset

Count: 65659 rows and 5 Columns

Data pre-processing

This module contains all the preprocessing functions needed to process all the input documents and texts. First, we read the train, test and validation data files then perform some preprocessing like tokenizing, stemming etc. Some exploratory data analysis is carried out, such as response variable distribution and data quality checks, such as null or missing values.

Stemming: Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base, or root form—generally a written word form—in linguistic morphology and information retrieval. The stem does not have to be the same as the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not a valid root in and of itself.

Tokenizing: Tokenization is the process of replacing sensitive data with unique identification symbols that retain all the essential information about the data without compromising its security. Tokenization, which seeks to reduce the amount of data a company must keep on hand, has become a popular way for small and medium-sized businesses to improve the security of credit card and e-commerce transactions while reducing the cost and complexity of complying with industry standards and government regulations.

Feature Selection

We used feature extraction and selection methods from the sci-kit learn python libraries in this module. For feature selection, we used methods such as simple bag-of-words and n-grams, followed by term frequency weighting such as tf-IDF

Classifier

We build all of the classifiers for predicting fake news detection in this module. Different classifiers are fed the extracted features. We used sklearn's Logistic Regression classifier. The classifier made use of all of the extracted features. We compared the f1 score and checked the confusion matrix after fitting the model. The two best-performing models were chosen as candidate models for fake news classification after fitting all of the classifiers. Finally, the model that was chosen was used for fake news detection with the probability of truth. In addition, we extracted the top 50 features from our term-frequency tf-IDF Vectorizer to determine which words are most important in each class. We have also used Precision-Recall and learning curves to see how training and test set performs when we increase the amount of data in our classifiers

Logistic regression Classifier:

It is a classification algorithm that uses Machine Learning to predict the likelihood of a categorical dependent variable. The dependent variable in logistic regression is a binary variable that contains data coded as 1 (yes, success, etc.) or 0. (no, failure, etc.). To put it another way, the logistic regression model forecasts $P(Y=1)$ as a function of X .

In the second part, we used LSTM. Here are the steps:

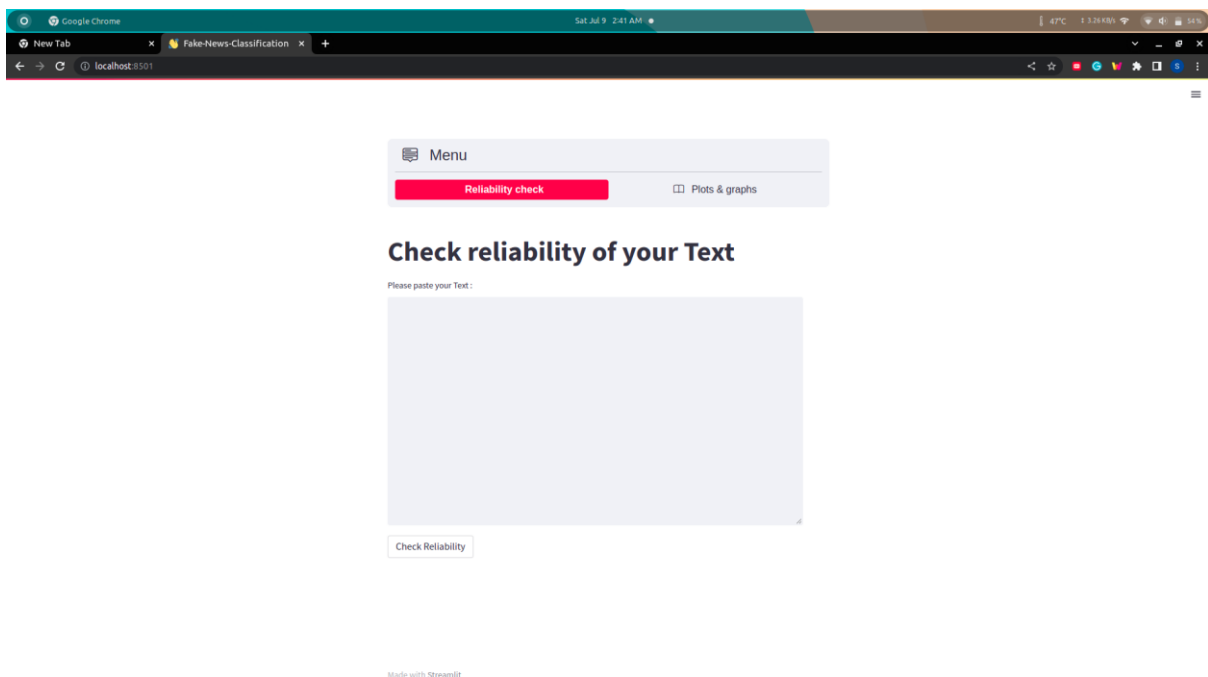
1. **Loading the data:** For this step, it is the same as the passive-aggressive one.
2. **Scanning and parsing:** Data is loaded from a CSV file. This consists of the body of selected news articles. It then contains a label field that indicates whether the news is real or fake. In this code block, we scan the CSV and clean the titles to filter out stop words and punctuation.
3. **Tokenization:** The tokenizer is used to assign indices to words and filter out infrequent words. We can now generate sequences for our training and testing data.
4. **Embedding matrix:** Apply the embedding matrix. An embedding matrix is used to extract the semantic information from the words in each title.
5. **Model Building:** Building the model and finding out the accuracy via confusion matrix. The model is created using an Embedding layer, LSTM, Dropout, and Dense layers. We are going to run the data on 20 epochs.
6. We observed that the LSTM model is vastly inaccurate in predicting the authenticity of the news. So we decided to show the output by running it through the Passive-aggressive classifier model.

System Architecture

1. Static Search- The architecture of Static a part of a fake news detection system is kind of simple and is done by remembering the crucial AI measure stream. The framework configuration is displayed beneath and is self-explanatory.
2. Dynamic Search- The second search field of the location asks for specific keywords to be searched on the web upon which it provides an appropriate output for the proportion probability of that term being present in a piece or the same article with those keyword references in it.
3. URL Search- The third search field of the positioning acknowledges a specific site name whereupon the execution searches for the situating in our actual locales data set or the boycotted destinations information base. Verity sites database holds the domain names which regularly provide proper and authentic news and the other way around. If the location is 't found in either of the databases then the implementation doesn't classify the domain it simply states that the news aggregator doesn't exist.

Implementation

This is what you see when we go to the web interface. You are supposed to copy the news and paste it into the input box.



When we paste the news on the input box and click 'Predict' the model will give you the result. If the news seems authentic, the output will be 'Looking Real News. Otherwise, it will show 'Looking Fake News. That's how you can detect fake or real news via the interface.



Menu

Reliability check

Plots & graphs

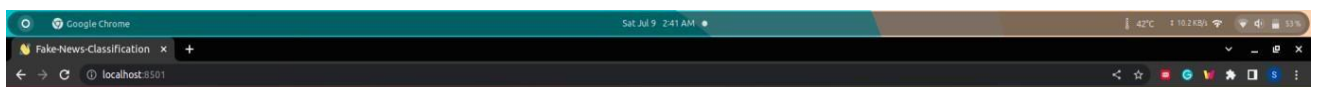
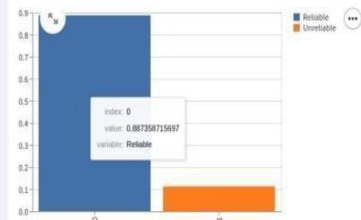
Check reliability of your Text

Please paste your Text:

held at Gitmo, 18 have been cleared for release by Obama's PRB and more than half (27) are considered "forever prisoners," or too dangerous to release. Nevertheless, Obama has already released prisoners who had been designated too dangerous to release. The remaining 10 prisoners are believed to still be undergoing war crimes proceedings at military commissions, including at least six who were facing death penalty tribunals as of late last year. On Friday, the Pentagon announced that four detainees — Salem Ahmad Hadi Bin Kanad, Muhammed Rajab Sadiq Abu Ghanim, Abdallah Yahya Yusif and Muhammad Ali Abdallah Muhammad Bwazir — had been transferred to the Kingdom of Saudi Arabia. All four detainees are from Yemen, which shares a border with Saudi Arabia. At least two individuals — Kanad and Ghanim — had been linked to the Taliban when they were apprehended and taken to the Guantánamo prison, also known as Gitmo. Meanwhile, the Bwazir has been directly linked to Osama bin Laden. also affiliated with was born in Saudi Arabia, but is a citizen of Yemen, according to the U. S. military. In December 2007, the U. S. military deemed him "a HIGH risk, as he is likely to pose a threat to the US, its interests and allies."

Check Reliability

Reliable

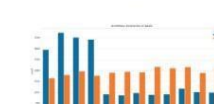
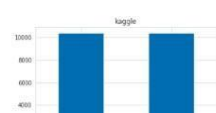
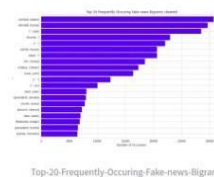
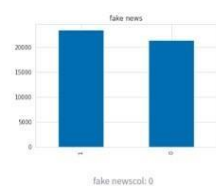
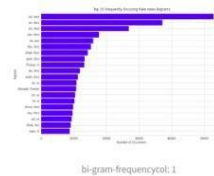
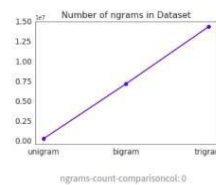


Menu

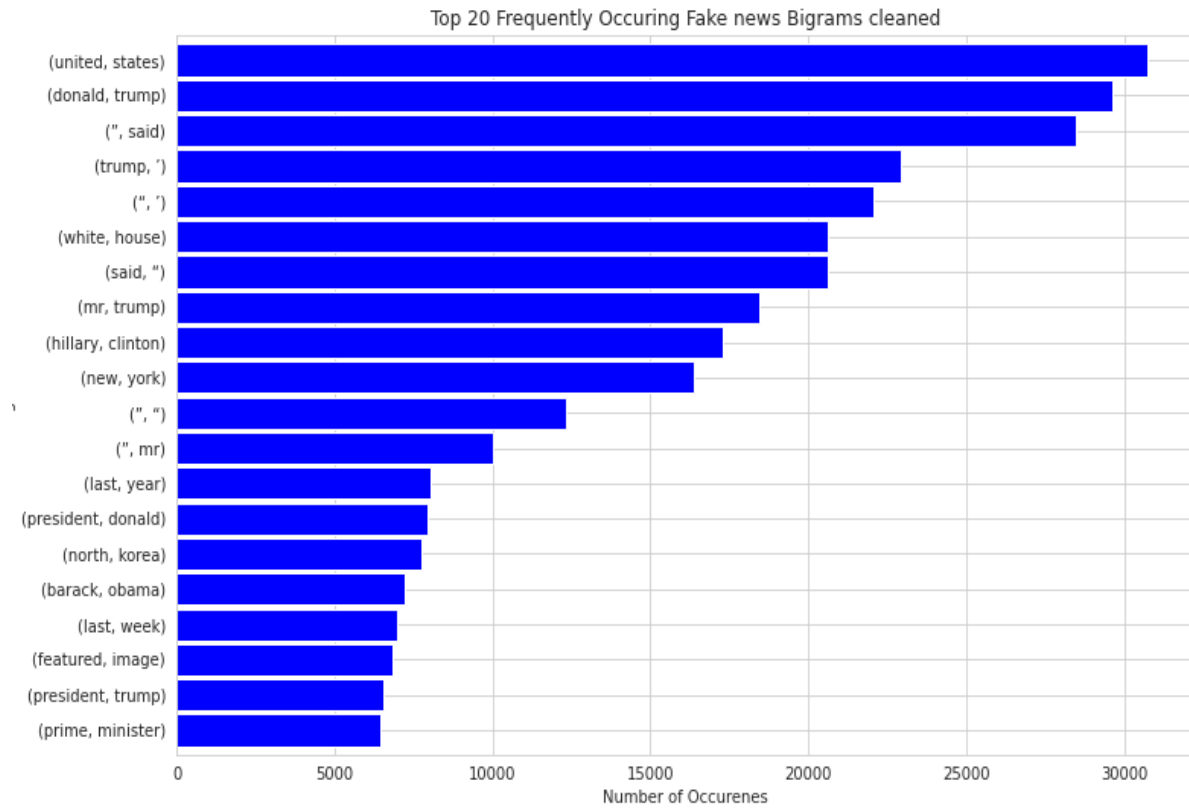
Reliability check

Plots & graphs

plots and graphs



Results

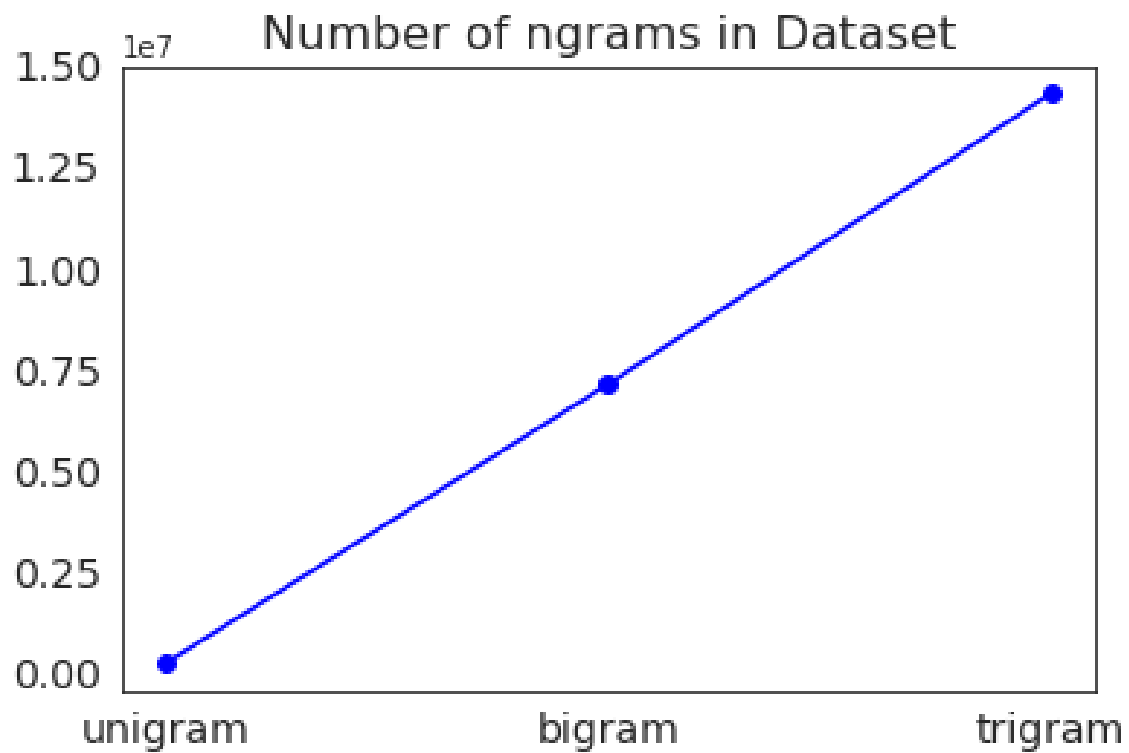


We tested our data on different count vectorizers within the Bag of Words Technique. & examined the frequency of words after using the cleaning technique.

Unigram: 223065

Bi-Gram: 7129330

Tri-gram: 14318344



According to the line graph's result, we trained the data taking Tri-gram, which has the higher values among all ngrams.

Total=10240	Logistic regression using tf-idf features	
	Fake(prediction)	True(prediction)
Fake (Actual)	1617	2871
True (Actual)	1079	4655

Discussion

Logistic Regression: It's used to estimate discrete values (like 0/1, yes/no and true/false) based on a given set of independent variables. In easy words, it predicts the probability of occurrence of a happening by fitting data to a logit function. Hence, it's also referred to as logistic regression. Since it predicts the probability, its output values lie between 0 and 1 (as expected).

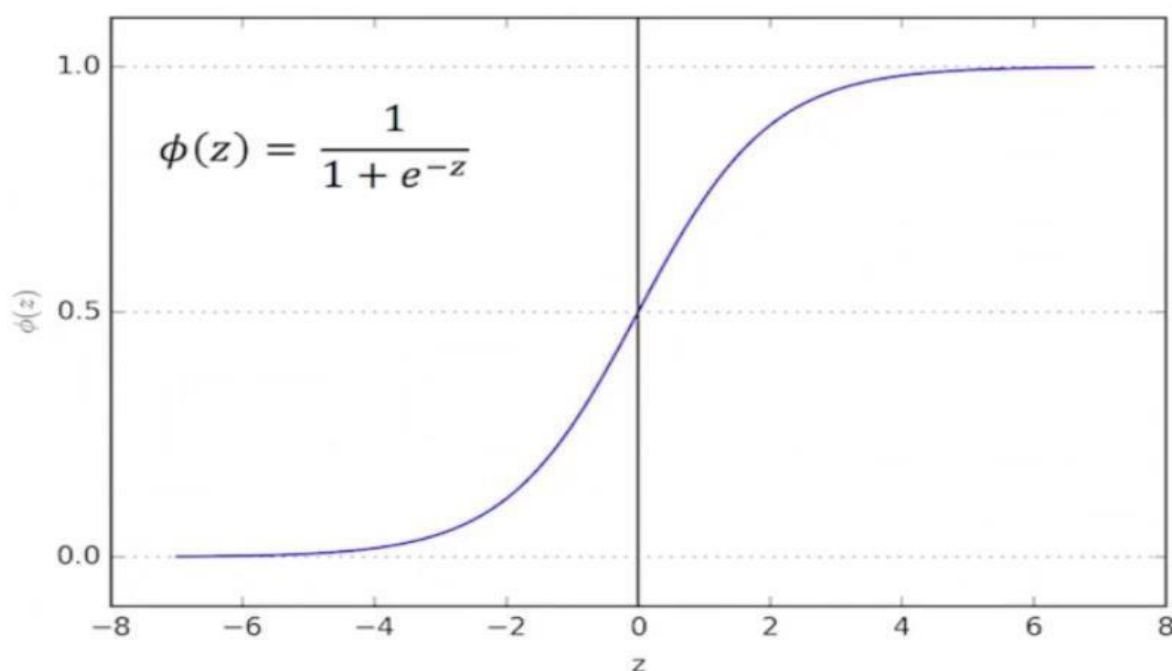
It is a classification algorithm that uses Machine Learning to predict the likelihood of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes and success) or 0 (no and failure).

They used a liar dataset for detecting fake news by Logistic Regression. They used Bag-Of-Words, N-Grams, and TFIDF. The logistic regression Train with a bag of words shows results with an accuracy of 99.005% and the Logistic regression test with a bag of words shows results with an accuracy of 95.6%.

The logistic regression train with TFIDF gave an accuracy of 99.067% and the logistic regression test with TFIDF gave an accuracy of 96.8% while testing the model. A Supervised Method for Detecting Fake News Using Machine and Deep Learning Classifiers. They used news channel data from Kaggle.com. In this paper, they used methods like LSTM. LSTM train shows the result with an accuracy of 94.53% and in LSTM test with an accuracy of 90.6%.

Mathematically, the log odds of the result are modelled as a linear combination of the predictor variables. Odds = $p/(1-p)$ = probability of event occurrence /probability of not event occurrence

$$\ln(\text{odds}) = \ln(p/(1-p)) \quad \text{Logit}(p) = \ln(p/(1-p)) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_kX_k$$



Conclusion

It is fostered that the Fake news Classifier system takes contributions from the user and characterizes them to be true or fake. To execute this, different NLP and Machine Learning Techniques must be utilized. The model is prepared to utilize a suitable dataset and performance assessment is additionally done utilizing various performance measures. The best model, i.e. the model with the most elevated accuracy is utilized to group the news headlines or articles. From the static search, the best model came out to be Logistic Regression with an accuracy of 96 %. As a result, search parameter optimization was used to improve the performance of logistic regression, which yielded a 96 % accuracy. Hence it is shown that if a client feeds a particular news article or its headline in our model, there are 96 % chance that it will be characterized in its real sense. The user can check the news article or keywords online; also check the authenticity of the website. The accuracy for the dynamic system is 93% and it increases with every iteration.

we've used a Logistic Regression classifier which will serve the model and work with the user input. We have presented a detection model for fake news using TF-IDF analysis through the lenses of various feature extraction techniques in this paper. We looked into various feature extraction and machine learning techniques. Using TF-IDF features and a logistic regression classifier, the proposed model achieves an accuracy of approximately 96 percent.

References

- [1] C. Castillo, M. Mendoza, and B. Poblete. Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5):560–588, 2013.
- [2] T. Chen, L. Wu, X. Li, J. Zhang, H. Yin, and Y. Wang. Call attention to rumours: Deep attention-based recurrent neural networks for early rumour detection. *arXiv preprint arXiv:1704.05973*, 2017.
- [3] Y.-C. Chen, Z.-Y. Liu, and H.-Y. Kao. Ikm at several-2017 task 8: Convolutional neural networks for stance detection and rumour verification. *Proceedings of SemEval. ACL*, 2017.
- [4] I. Augenstein, A. Vlachos, and K. Bontcheva. Usfd at semeval-2016 task 6: Any-target stance detection on Twitter with autoencoders. In *SemEval@NAACL-HLT*, pages 389–393, 2016.
- [5] S. B. Yuxi Pan, Doug Sibley. Talos. <http://blog.talosintelligence.com/2017/06/>, 2017.
- [6] Bahad, P., Saxena, P. and Kamal, R., 2019. Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. *Procedia Computer Science*, 165, pp.74-82.
- [7] EANN: Event Adversarial Neural Networks for Multi-Modal