FAKE NEWS DETECTION SYSTEM

a fOUNDATION OF dATA sCIENCE aSSIGNMENT

Rohit Sanjay – 2018a7ps0010u

2021

**FOUNDATION OF DATA SCIENCE**

**Objective:**

The main goals of this project is to do the following:

* To detect fake news using NLP and other methods.
* To showcase the potential of repetition and right extraction methods to get right information.

**Intro:**

News has been around for past 300+ years and it has been transforming ever since. It is the medium through which series of event are kept in account, awareness and opinions are shared through. As once quoted “newspaper is a window to the outside world”. As we take the last couple of decades, with explosion of digitalization and the internet the credibility and trustworthiness of the press or media has been questioned. The rise in Deep-Fakes and fake news spread through social media websites like Facebook, twitter, Instagram, and WhatsApp are on the rise. Combating such false information that flow around disturb the perception of false reality.

The issue with news itself is theres a lot of speculation and a lot of data present across that validating and rectifying/ removing the false ones are going to take years if not decades to clean out. Companies like Facebook and WhatsApp have implemented “point of origin” algorithms where if a false news is reported it tracks using linking partners to traceback the first message ever spent regarding the topic itself. Even thought this is not 100% effective these are small steps that we must implement to tackle this giant hurdle. Through my readings, I have understood that right now the biggest implementation of fake news detection is supervised or reinforcement learning algorithms. A simple approach suggested by most researchers are that using sentimental analysis and NLP for such tactics. But in one form of project called the “FakeBox Project”, it took a different approach that yield promising results. Its approach was NOT to find/detect fake news but rather find/detect REAL NEWS. This approach gave the author close to 95% accuracy for a large enough dataset.

In the next we shall see a dataset found by me that shall be used throughout the next phase of this project.

**Dataset:**

The data set that is used for this project is from Kaggle. The URL is given below for any further details. [Fake News | Kaggle](https://www.kaggle.com/c/fake-news/data)

The composition of the dataset is:

* ID: Unique numbering for each news article
* TITLE: The title of the news article
* AUTHOR: The Author of the news article
* TEXT: The content of the article (Could be incomplete also)
* Label: A tag to show if the article is potentially unreliable
  + 1: Reliable news
  + 0: Unreliable news

There are 3 files in the dataset:

* Submit.csv: This file contains only the id and label as columns and 5201 rows worth of data.
* Train.csv: This file contains 25117 rows and 5 columns.
* Test.csv: This file contains 5881 rows and 5 columns.

**Research Papers:**

In this assignment all the papers were taken from IEEE. The following research papers and a small gist has been laid out below.

* **Deep Diffusive Neural Network Based Fake News Detection from Heterogeneous Social Network:**

*The paper takes a novel approach towards fake news detection using a FAKEDETECTOR which is based on explicit and latent features extracted from textual information. The FAKEDETECTOR builds a deep diffusive network model to learn from articles, creators, and subjects. The FAKEDETECTOR achived best performance compared to other bi-class labels of news of an average 14.5% better results and in the multi-class it achieved 40% average better results.*

* **Exploiting Multi-Domain Visual Information for Fake News Detection:**

*The proposed methodology in this paper is to make use of a framework called Multi-Domain Visual Neural Network (MVNN) which will use visual information of frequency and pixel domains to detect fake news. Later is again processed by a CNN\_RNN model to extract features from different sematic levels in a single pixel domain. The real-world tests gave this model atleast 9.2% increase in accuracy and for multi-model fake news detection by 5.2%.*

* **Fake News Detection on Social Media using K-Nearest Neighbor Classifier:**

*The author of this paper decided to take a simple approach by implementing a k-nearest neighbor classifier in their model. The model was able to present a proper 79% accuracy which was tested with Facebook news posts dataset.*

* **Fake News Detection System using Article Abstraction:**

*This article was a Korean fake news detection system which uses a fact DB that is built and updated with human’s direct judgement after collecting facts. The paper suggests we choose a deep learning model, in this case a Bidirectional Multi-perspective Matching for NLS(Natural Language Sentence) which is powerful enough to match sentences. The conclusion was that the model was able to achieve AUC of 0.782.*

* **Fake News Detection using Disclosure Segment Structure Analysis:**

*The author of this paper takes a simple approach by using a deep-learning network that uses discourse-level structure analysis to formulate the structure that differentiates fake from real news. This model upon evaluation gave 74% accuracy.*

* **Fake News Detection System using Grammatic Transformation on Neural Network:**

*This paper uses grammatical transformation with 4 layers: Word embedding layer where words in proposition are embedded into a word vector, then a contact generation layer which enters into a LSTM layer and generate context vector, followed by a matching layer in which a attention vector is generated from the contextual embedding vector in the previous layer computing the weighted sum. The LSTM and attention vector are compared through a matching operation generating a sentence which has the same meaning but different form.*

* **Fake News Detection with Generated Comments for New Articles”**

*In this paper, the author uses a fake news generator model. This model is trained to generate comments using a dataset which consist of news articles and their social contexts.*

* **Semi-Supervised Learning and Graph Neural Networks for Fake News Detection:**

*Our work proposes a graph-based semi-supervised fake news detection method based on graph neural networks. It gave back a 67% accuracy.*

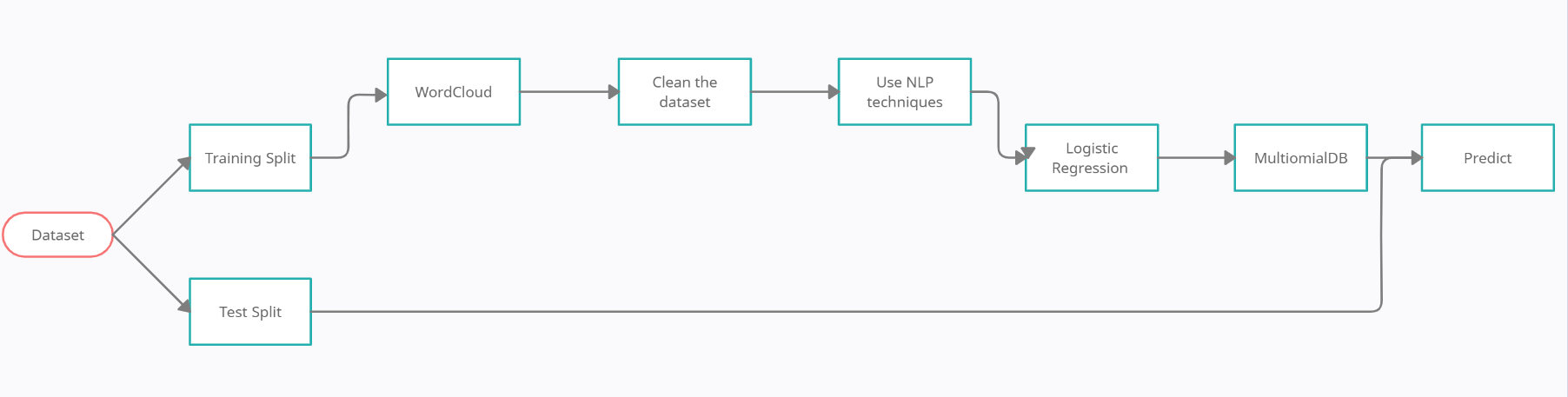
* **Fake News Detection Model based on Unified Key Sentence Information:**

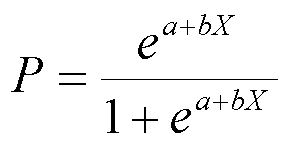
*In this paper the author uses a new fake news detection model that uses unified key sentence information which efficiently performs sentence matching. One unified word vector for the key sentence of article by extracting them and merging them somewhere else. This model was able to give back an accuracy of 64%.*

* **Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator:**

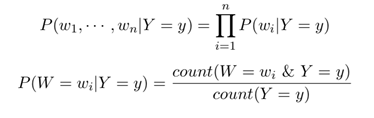
*In this paper, the author uses Natural language process and SciPy toolkits that were developed to find fake news detectors and later matched with a Bayesian classified to get the likelihood that the given article is a fake or not. The model was able to give back a accuracy of 63.33%.*

**Methodology:**

In this assignment we will take the Kaggle set and put it through different packages and linear transformation. We first try to split the data and then create a WordCloud, A Word Cloud visual is creating the topmost recurring words in each dataset with its corresponding intensity. After this step we clean the dataset using Regex, tokenization and find stop words. After the first step we use NLP techniques mainly 3 of them, TfidTransformer, CountVectorizer and TfidVectorizer. afterextracting the features we implement logistic regression and multiomialNB and finally we can implement the model on the final test dataset and get the predicted value.

 The regression equation is:

The equation used for the text-identification is:

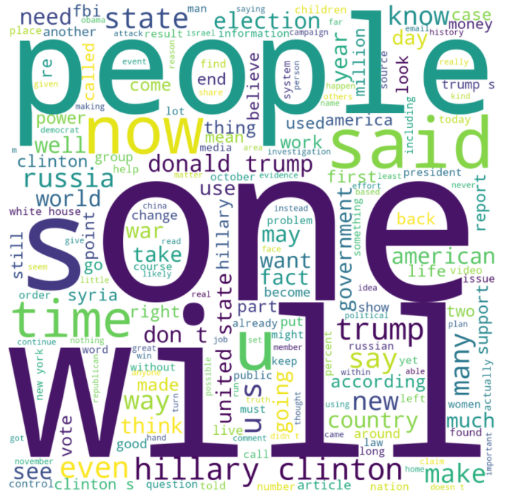


**Experimental Results:**

* Experimental setup:

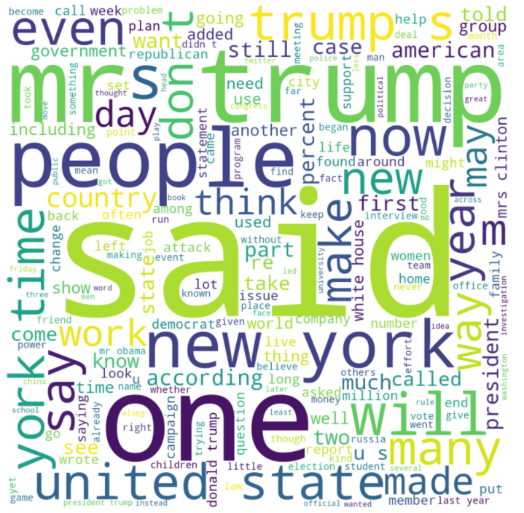
In the fake detection system, we use the Kaggle dataset that has 30,998 rows. We start of with cleaning the data and then we create a word-cloud for both the train and test dataset to interpret the type of words that can be found and to understand maybe the topic of all the news collected in the dataset. From that we start removing stop-words in the sentences while implementing lemmatization and tokenization to make the sentence only have meaningful words. From the NLP techniques we use feature extraction. And create a model using logistic regression and then implement a pipeline. We also create a confusion matrix to quantify our results.

* Statistical Results:

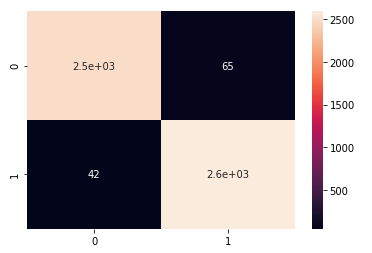


This is a word-cloud formed from the real-words operation to find out which words are the most used and occurs most throughout the dataset while making sense of the data.

In this we can see Hillary Clinton and people being use with Donald trump etc. implies this dataset is from a political campaign in 2016.



In this word-cloud it can see that Donald Trump, New York and people have been one of the most common words used. This can show use that this dataset has Donald Trump’s political campaign showed.



In this confusion metrics from the prediction, we did the true positive and false positives are detected perfectly and accurately while false positive and true negatives are lower. This shows us the detection algorithm is working better and can give almost accurate results.



The accuracy of our model using Logistic Regression.



The accuracy of our model using MultinomialNB.

**Conclusion:**

From the above experimentation it is seen that implementation of a fake news detection system is a long process of cleaning and making sense of words and sentence structure to create the perfect system. In my system a couple of flaws are well as better implementation is possible via html and pipeline for a proper system. The accuracy and prediction were better than expected. Better implementations can be done and cleaning of the dataset itself can be improved. Overall, the project showcased a method to detect fake news from a already existing dataset.

**References:**

* *Deep Diffusive Neural Network Based Fake News Detection from Heterogeneous Social Network:*
* *Exploiting Multi-Domain Visual Information for Fake News Detection:*
* *Fake News Detection on Social Media using K-Nearest Neighbor Classifier:*
* *Fake News Detection System using Article Abstraction:*
* *Fake News Detection using Disclosure Segment Structure Analysis:*
* *Fake News Detection System using Grammatic Transformation on Neural Network:*
* *Fake News Detection with Generated Comments for New Articles”*
* *Semi-Supervised Learning and Graph Neural Networks for Fake News Detection:*
* *Fake News Detection Model based on Unified Key Sentence Information:*
* *Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator:*