Fake News Detection

Milestone -1

Submitted By:

Project Group - 29

Pavitra Mohanty (110013596)
Arpit Patel (110023955)
Dipal Modi (110016791)

Related Existing Models:

- One of the most common approaches used in many of the existing models was to preprocess the data using TfidfVectorizer[1]. In which, IDF (Inverse Document Frequency) is calculated to indicate how significant a word is in the entire corpus and TF (Term Frequency) to indicate the number of times a word appears in the document. These two measures are fed to the Passive Aggressive Classifier to predict the fake news[1].
- Stance detection[2] for detecting fake news involves comparing the headline with a body of text from a news article to determine what relationship exists between the two. It employed the 4-class classification approach to predict whether the news article agrees /disagrees/discusses/unrelated to the headline[3]. This model predicts related /unrelated using the naive baseline (Jaccard similarity) and it feed-forwards the results to the neural network and applies LSTM and recurrent seq2seq recurrent model to classify correctly.
- Another model was to use machine learning and NLP techniques to determine whether the news is fake or not. Their approach was doing Data preprocessing, Generating News Feature Vector which includes methods such as Bag of Words, TF-IDF (Term Frequency - Inverse Document Frequency), Semantic Analysis, and applying different Classification algorithms such as Naive Bayes, Random Forests, Gradient Boosting for finding accuracy[4].

Step by Step Process of the task

The workflow for this project will be as below:

- One of the important things for our project is to have a proper dataset on which we will be training our model and that data we are going to collect from multiple sources like Kaggle, GitHub, and many more.
- All the dataset we will be merged to make a final dataset for analysis.
- Merged data will be stored into a NoSQL database as a document and complete the future analysis by fetching the data from the NoSQL database.
- So once completing the above steps our team will perform data related operation (data preprocessing).
- We will be checking for the null values in the dataset and will decide the next step like if there are more null values then we will perform the operation to fill the missing values and if there is a negligible amount of missing values then we will remove it as we will have enough data for building a machine learning (ML) model.
- Noisy words will be removed from the news articles. These noisy words won't help our model to increase the efficiency to classify the news.
- Stemming is a process where words are reduced to root by removing inflection through dropping unnecessary characters, usually a suffix. Stemming increases the density of your training data. It reduces the dictionary size (in the body of words) two or three times (in many languages like French, where a single tree will generate dozens of words for verbs, for example). This reduces the dictionary's size having the same corpus, but fewer input dimensions news classifier feature vector, ML algorithm will work better.
- Then after we will split the dataset into training and test dataset.
- As our dataset will be containing text data and we are going to feed this data to ML algorithm but those text data, the machine learning algorithm won't be able to understand so we need to convert those text data into a word vector and then after that need to be passed into the ML algorithms. Some of the known methods are Bag of Words, Splitting the dataset into train and test dataset.
- Once done with the above steps we are ready to develop ML models so we will be creating several classifiers like Random Forest Classifier, Multinomial NaiveBayes, XGboost, LSTM(Long Short Term Memory) and will comparing the accuracy of these models.

References:

- Filippos Dounis. (2020, April 2). Detecting Fake News With Python And Machine Learning. Retrieved July 17, 2020, from Medium website: https://medium.com/swlh/detecting-fake-news-with-python-and-machine-learning-f78421d29a06

- Fake News Challenge. (2016). Retrieved July 17, 2020, from Fakenewschallenge.org website: http://www.fakenewschallenge.org/
- Chaudhry, A. K., Baker, D., & Thun-Hohenstein, P. (n.d.). Stance Detection for the Fake News Challenge: Identifying Textual Relationships with Deep Neural Nets.
- Nirwan, P. S., Bansal, V., & Aloor, S. A. (2018). Fake News Detection. Indian Institute Of Information Technology.