Classification of News into Real and Fake News

Group 6: Dhairya Amin, Parth Shah, Smit Dar

1. Description:

Fake news refers to misinformation, disinformation, or mal-information which is spread through word of mouth and traditional media and more recently through digital forms of communication such as edited videos, memes, unverified advertisements, and social media propagated rumors. Fake news spreads faster than actual news on social media. This has become a serious problem where you can post news articles within seconds, with the potential of it resulting in mob violence, suicides, etc. People then share this fake news, causing anxiety and unrest. Our team will be comparing models to predict whether a tweet containing news is real or fake to find the one with the best performance. This project will provide a basis for what model would work best for fake/actual news classification and then could be built upon in the future by news agencies and social media platforms to prevent the posting of such tweets.

2. Dataset:

We will utilize the Fake and Real News Dataset from Kaggle for this project. The dataset comprises 2 CSV files, namely, Fake.csv and True.csv. Both the files have 4 columns (features), the title of the article, text of the article, subject of the article, and date when the article was posted. In addition to that, the Fake.csv consists of 17903 unique titles out of which 39% are news-related, 29% are political related and the rest 32% are labeled as others. These articles have been posted from 30 Mar 2015 to 18 Feb 2018. In True.csv there are 20826 unique titles out of which 53% are politics related and the rest 47% are world news. These articles have been posted between 12 Jan 2016 and 30 Dec 2017. To get a large dataset that is unclassified, first, we will merge both the files, and then we will use it to train a model so that it can classify the given news as real or fake.

3. Methodology and Expected Results:

We will be using Python language for the project, on IDEs like Google Colab, JetBrains DataSpell, and Anaconda Jupyter Notebook. We will make use of libraries such as TensorFlow, Pandas, NumPy, Matplotlib, Seaborne, Natural Language Toolkit (NLTK), etc. Then, we will try different combinations of text-cleaning steps, to get the optimal results, such as:

- removing newlines and tabs/whitespaces/accented characters/links/special characters/ stopwords
- strip HTML tags
- reducing repeated characters and punctuations

- expand contractions
- correcting misspelled words
- lemmatization/stemming

Following this, we will perform exploratory data analysis using techniques like data imbalance, word cloud, statistical analysis using various plots, different n-grams analysis, feature extraction using count vectorizer, TF-IDF vectorizer, and Hashing Vectorizer, etc., to get a better understanding of the dataset. Finally, we will carry out various models and techniques for the classification task, namely, GloVe Embeddings and LSTM, Linear Models like Linear Regression, Non-linear models like MLP-Classifier (Multi-Layer Perceptron Classifier), RNN (Recurrent Neural Networks), K-Means CLustering, and even pre-trained models like BERT, etc. We will perform metric comparisons on these models to decide upon the best technique to get the highest accuracy on the available dataset. We will be using metrics such as accuracy, precision, recall, the area under the curve, etc. We aim to achieve results as high as possible, preferably above 90%.

4. Related Work:

- M. R. Hasan, M. Maliha and M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data," 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)
- Barbara Probierz, Piotr Stefański, Jan Kozak, “Rapid detection of fake news based on machine learning methods,” Procedia Computer Science, Volume 192, 2021

5. Timeline:

- Week 9 - pre-process the data and perform the cleaning of the dataset
- Week 10 - perform exploratory analysis on the dataset
- Week 11 - training/testing different classification techniques on the dataset
- Week 12 - measuring the performance of the models and tune the hyperparameters
- Week 13 - preparing the project report and final presentation

6. Responsibilities:

Overall, all three members of the group will contribute roughly proportionally to this project. All the tasks will be discussed and looked over by each member. The following is the initial task distribution:

- Pre-processing and cleaning data - Smit Dar
- Exploratory Data Analysis - Dhairya Amin
- Modeling and performance metrics - Parth Shah
- Documentation and presentation - Everyone