PROJECT PROPOSAL

REAL AND FAKE NEWS CLASSIFICATION

Team:
Anisha Yidala - G01223768
Sravya Sangaraju - G01328406

Introduction:
News articles, bulletins, and headlines i.e., information from across the world is so rapidly available than ever in the history and it is strongly impacting the society because of various factors like social media, news channels and other resources. Today the way any piece of information is projected has the power to make or break an issue from being heard. Fake news, deliberate disinformation, hoaxes, parodies, and satire are various ways to mislead people to gain financial or political benefits. Because of this extremely powerful nature we see a lot of misuse of this too. In today's world infinitely large amounts of data is available to us just a click away and it has become a boon and bane at the same time. Unlike the olden times when the resource of our information can be verified and the correctness of any data can be easily crosschecked, today because of the vast amounts of data it has become tremendously difficult to verify the information we have.

Objectives:
The objective of this project is to predict if any news article is reliable or unreliable as accurately as possible. We will be analyzing this information and extracting several key features which will help us train various classifiers and build a model that can efficiently predict unseen data. If we succeed in getting maximum accuracy on real time data, it can later be developed into an automatic machine learning model (beyond the scope of this project) which works as an efficient way to combat the widespread dissemination of unreliable news.

Methodology:
We intend to preprocess the huge corpus of news articles that is available in the form of text data and perform numerous classification models on it to decide if any given news can be considered reliable or not. For the preprocessing phase, we would be using certain libraries to execute removal of stop words, stemming and tokenization of cleaned data. To complete feature extraction, we would be converting the tokenized vector into term frequency – inverse document frequency (tf-idf) matrix. Then on the preprocessed data, use different classifications algorithms and assign labels to each row. Later we would perform model selection to acquire the maximum accuracy. The dataset that we would be using has been downloaded from Kaggle repository and has entries of several thousands of news articles with information about the author and text of the news articles and reliability.

Improvisation idea:
Build a second model that clusters news articles into different categories such as politics, sports, technical etc. Then combine both models to predict news category in addition to whether it's fake or real and then get insights like which category has more fake news, proportions of fake/real news by category.

Literature Search:
On essential search on this problem, we came to an understanding that fake news, fraudulent information, and unreliable data has been in the spotlight in the recent times, and it has been a major concern during the testing times of the pandemic and the recent presidential elections. We came across several solutions utilizing data mining itself that have tried to predict the reliability of news in these contexts. But all these solutions are streamlined or directed to work best on a particular single area of interest whereas we intend to draw inspiration from these and build a generic model that can work best on all data from multiple diverse unrelated topics.

Milestone/Timeline:
Nov 12th – Finish data preprocessing
Nov 20th – Run all the classification models and Model selection
Nov 25th – Clustering
Nov 30th – Cross Validation and Accuracy
Dec 3rd – Video Project Presentation
Dec 10th – Prepare project report.