BUDT758T Project Proposal

Group S

Group members: Haoming Wang, Xuetian Xie, Shuo Yu, Shuoxin Yu

03/26/2023

Speaking of what kind of impact that fake news has brought to our society, misinformation was one of the most popular topics among all the studies . According to WHO, the negative effects of misinformation change people’s behavior, create anxiety, and become one of the main factors delaying the vaccine process which results in a legacy in the healthcare process. Other than its impact on the healthcare field, fake news on different media platforms also creates disasters for society, because the more fake news appears, the less trustworthy the media platforms are. For this project, our data sources are the title, text, and content of fake news that were collected from Kaggle, Mclntire, Reuters, Buzzfeed Political, and Twitter. The reason why our data source is very collaborative is because we would like to minimize the bias between different platforms and different sources of data in order for our conclusion to be compatible with all the fake news on the internet as much as we can.

Our dependent variable is the label of whether each news sis fake or real, and our independent variables are the words after text mining. Our key findings for this project is to successfully spot the fake news, and help the society and internet environment be realistic.

We would lemmatize and tokenize the fake news content during the initial state of the text-mining process, after that we will filter out the stop words and remove unnecessary information such as URLs in the text content. Next based on the frequency of term appearance (td-idf) we would continually set up our model constructions. For the ML models, we had chosen to apply KNN, Naive Bayes, Random Forest, Support Vector Machine and Xgboost, and analyze their performance measures, to find out the best model that could be helpful in the process of predicting fake news. Besides the text mining and machine learning models that we mentioned earlier, one of our challenges is to retrieve tweets ids with Postman Agent and get the tweets by using the twitter API.

The reason we find the project interesting is because in this project, we are trying to build several machine learning models that we learned in class to dig out the truth beneath confusing titles, texts, and content, and finding the best model for the use of predicting fake news. Furthermore, we will explore the reasons why fake news spreads further, faster, deeper, and more broadly than real news. In the end we could present a case study with a couple examples of the fake news that we spotted using our model, and to see how ridiculous they are, at the same time, how realistic they are.

Work Cited

<https://www.who.int/europe/news/item/01-09-2022-infodemics-and-misinformation-negatively-affect-people-s-health-behaviours--new-who-review-finds>