**Project Proposal: Fake News detection**

**(DSCI-6051-01)**

Yueyang(James) Qin

Koganti Sri Harsha

1. **Introduction:**

Based on the recent fake news in addition to increase in popularity. It currently exists hard for people to detect the truth of the news. Some of the news might have some really bad effects for people in addition to might lead people to do something wrong or harmful to the world. The project all of us are creating currently exists trying to construct an automated machine learning model, which currently can automatically extract the fake news from our world wide web in addition to making some alarm for people. The project will focus on getting the news from the world wide web in addition to extracting the dataset from the news, which includes the fake news in addition to normal news.

1. **Problem and background:**

People get misled by the fake news in addition to it might have some bad effect related to the people in the fake news. For example, people still have some worries regarding the covid19 vaccine which will let the virus continue to spread for a long time. Sometimes people are misled by fake news, they don’t trust the people, whatever person currently exists LGBTQ, in addition to won’t treat them just as normal people.

1. **Project detail:**

All of us currently are planning to collect the data from www, (including Facebook, twitter, CNN, etc). Then do the algorithm for data cleaning, data model in addition to machine learning algorithm, in addition to checking the accuracy, then make the prediction. (The project detail needs to exist decide in the future in addition to might need to change since the project’s going.)

1. **Project problem statement:**

In this project, we propose to study the fake news detection problem in online social networks. Based on various types of information sources. We aim at identifying fake news from online social networks simultaneously. We formulate the fake news detection problem as an inference problem, where the real ones will have a higher percentage while the fake ones will have a lower one instead.

1. **Project scope:**

Thinking regarding fairness, the scope should include a wide category. Tremendously enjoy the political, biology information tremendously enjoy the covid19 all of us just go through, however some people still do not trust the vaccination. All of us also need to include information from humanity. Information tremendously enjoys how to support the disability in addition to how to provide equality to men, women in addition to lgbtq people. All of us also will think regarding entertainment news, sport news, national news, international news. In addition to different types of news, which includes newspaper, mobile app news, television news in addition to radio news.

1. **Pricing, project budget, overall project cost:**

All of us will keep it free first, because every single one of the datasets currently is able to exist found on the web, in addition to most of the model building should exist done with our team. In addition to the algorithm will exist the most popular machine learning algorithm, which currently is able to treat most of the machine learning problems.

1. **Project timeline:**

The timeline currently exists going to exist the whole fall semester, based on the effort all of us might go through. There might be some exception that all of us are currently unable to not finish the project in the end, however all of us will show whatever effort all of us have made, whatever error all of us have gone through, in addition to whatever all of us have learned from the project.

1. **Project roadmap:**

Based on the course requirement, all of us currently are going to construct up the CI/CD pipeline automation for the MLOps, so the roadmap should stay in the level 2 of MLOps. Start with data extraction, follow with data validation, data preparation, model training, model evaluation, model validation, model analysis, source code, construct source repository, CI: construct, test, & package pipeline component, CD:pipeline deployment, CD: model serving, training model, prediction.

1. **Project Methodology:**

The methods that we have think about and like to try are including: (might need to change in the future work, as project development going on.)

1. Logistic regression
2. SVM
3. Multilayer perceptron
4. KNN
5. Random forest
6. Boosting ensemble classifiers
7. Voting ensemble classifiers in addition to some benchmark algorithms.

**10. Goals:**

The goal currently exists to make sure our model currently exists best trained with the CI/CD pipeline automation for MLOps in addition to make the accuracy much closer to 100%, in addition to make some prediction for the dataset, which also currently am able to exists extracted in addition to validated inside our pipeline.