

## CIS 522 Project Proposal

**a) Team Name:** Ground Zero

**b) Team Members:** Arvind Balaji Narayan, Gopik Anand, Vasanth Kolli

**c) Project Statement:** Classification of News from Social Media Platforms into Real and Fake using suitable NLP Techniques

**d) Dataset:** Firstly, the dataset we will be using is the **Fake and Real News Dataset** from Kaggle which consists of around 39000 samples of both real and fake news put together. We will then transfer the learning from that dataset and later proceed on to scrap data from Twitter or Reddit or other social media handles, preprocess it to the right form, combine the title and content of the post and then classify the resultant news text into real or fake.

**e) Methods:** Our project work involves the Implementation of standard GPT-2, BERT, LSTM and other transformer model architectures (AlBERT, RoBERTa, DistilBERT, etc.) scheduled for the task of text classification into real/fake. This will be followed by evaluating the performance of these models on common performance metrics (Accuracy, F1-score, etc.) and reporting results to compare and identify the best performing model for the given data under consideration.

**f) Innovation:**

1) We will be making a data contribution in terms of scraping data from social media handles like Twitter, Reddit, etc. This data is bound to be in a different format and fashion when compared with the data seen in the current dataset which has been scraped from news articles. So, we believe that this is a novel effort to get data from a different source and train NLP models to get good accuracies. This is our primary goal given that many people are not drifting from traditional newspapers and articles to the news that comes instantly on social media.

2) As a stretch goal and a second innovation, we are planning to make the model more inherent to the actual message in the post so that it understands the nuances of the message rather than it being used as a statistical model that is trained and then asked to predict whether the news is real or fake. We believe the class of the news depends on the state of the external world and the labels do not offer a lot of real information as such. This is a stretch goal that we would like to achieve after completing our primary objective.

**g) Social Impact:** In a real-life setting where people largely consume news from social media handles, there is a huge possibility that we might be consuming fake news without knowing it and spread this out to friends, relatives and others which might lead to confusion and even pave the way to serious repercussions later on. In such cases, It would be great if people can actually have a system/setup to filter out the real news from the fake ones so that they can take the right and appropriate action (whatever it might be). They can also share the right news with others.

This is where Ground Zero plays its part. By being able to distinguish between real and fake news seamlessly, we can provide a setup where people access the right content always.

**h) TA Preferences:**

- 1) Mrigank Saksena
  - 2) Kevin Liang
  - 3) Harry Wang
-