COMP 646 Project Proposal
Kai-Po Lin - kl72
Hannah Lei - hyl3

Idea: AI Lie Detector on Statements

Datasets:
1. https://www.kaggle.com/thesergiu/truth-detectiondeception-detectionlie-detection?select=politifact_clean_binarized.csv (Mostly related to politics)
2. https://codeocean.com/capsule/5416660/tree

Properties of datasets.

Dataset	#Total data	#Fake news	#Real news	Avg. length of news articles (in words)	Topic(s)
LIAR	12791	5657	7134	18	Politics
Fake or real news	6335	3164	3171	765	Politics (2016 USA election)
Combined corpus	79548	38859	40689	644	Politics, economy, investigation, health, sports, entertainment

=> Paper: https://reader.elsevier.com/reader/sd/pii/S266682702100013X?token=781536D91491492CF4E0734073ADAE1D1CCEF70997566670CAFF482420CA31969E9E0D55C49A00B793336712A1D55261&originRegion=us-east-1&originCreation=20220205032452

3. https://github.com/manideep2510/siamese-BERT-fake-news-detection-LIAR

Blueprint:
Our plan is to create a model (currently tends to build a model based on BERT) that can predict how accurate a statement is. We want to implement a deep learning model that will understand the language and text in our dataset.

In most of these datasets, we have text data which means we are planning to do language processing in our model. However, we were unable to find sufficient datasets containing video that could be incorporated into our model training.

Our current idea only involves mostly text processing, but we want to find a way to incorporate image or video processing too. However, with images, there is not much data or information to be obtained other than what the image is showing at face value. Therefore, there may not be much we can gain from involving image processing. This may be a little out of our scope and abilities, but it would make our model more robust and able as far as what kinds of data can be fed to it.

We are open to ideas and feedback for ways of incorporating vision and images into our plan.