Abstract

Social media plays a vital role in connecting people around the world and developing relationships. Social Media has a huge potential audience and the circulation of any information does impact a huge population.People are regularly confronted with potentially deceptive statements in the form of fake news ,misleading product reviews, or lies about activities. There have been few works on automated text based deception detection that have exploited the potential of nlp approaches.the main hurdle is the lack of interpretability and potential to understand underlying logic and classify deceptive statements.However the recent advancements in nlp have made it possible with introduction of transformer based models.This thesis proposes to make use of Bi-directional encoder representations from transformers(BERT) model for classifying deceptive statements. The BERT model for supervised learning and GAN-based BERT architecture for semi-supervised learning and finally chat-based QnA architecture for tackling lies and misinformation problems.

Motivation

The common approach of using RNN, especially LSTM archtecture for classifying text is still quite popular. LSTM makes use of feedback connections unlike traditional RNN that are prone to vanishing and exploding gradient problems.However recent studies indicates that transformer-based models have better efficiency compared to LSTM when training on larger dataset[1], But the major advantage of BERT over LSTM is understanding the context of words.
For instance the 2 sentences make use of the word “fine”. Where the context of application in both sentences are different.

I have fine hair.
I am fine.

Even bi-directional LSTMs have less efficiency as they are technically learning from left to right and right to left context separately and then concatenating them so the true context is lost.However the context of words is learned better as BERT can learn context from both directions simultaneously.
The word embeddings like word2vec, GloVe are static embeddings where each word is mapped to a single vector representation while BERT uses contextual or dynamic word embeddings where input is a sentence rather than a single word.BERT needs to know the context of the surrounding words before generating a word vector.[2]

Related work

The related work in the area of deceptive text analysis has been experimented using a bag of words and tf-idf approaches with linear support vector machines(SVM) for modelling[3]. In a bag of words, the text is represented as the bag or multiset of its words, disregarding grammar and even word order but keeping multiplicity. While in tf-idf the frequency of words to determine how relevant those words are to a given document. It's a relatively simple but intuitive approach to weighting words, allowing it to act as an inception for a variety of tasks.However, they fail to capture the grammar and actual meaning of the words.

Hu [4] used a variety of models to identify concealed information in text and verbal speech, best among them a deep learning model based on bidirectional LSTMs. Concealed information, in this context, refers to when a person has knowledge about a subject and is withholding it, as compared to Hu's definition of deception where someone fakes knowledge they do not have. Hu created a corpus of wine tasters evaluating wines and encoding in various ways such as n-grams, LIWC, and GloVe embeddings [5] based on the recordings. The LSTM model using these features achieved an f-score in identifying the presence of concealed information of 71.51, defeating the human performance of 56.28.The major problem is that bidirectional LSTMs are less efficient compared to BERT transformers as learning takes place sequentially not simultaneously.The word embeddings used are GloVe which are static word embeddings, while BERT uses newer class of algorithms and are based on dynamic word embeddings.

Approach

The proposed aim of my thesis is to tackle lies and deceptive text using deep-learning transformer based models.

Dataset-1 : consists of 40k rows of data with 20k computer generated and 20k human generated for supervised and semi-supervised learning.

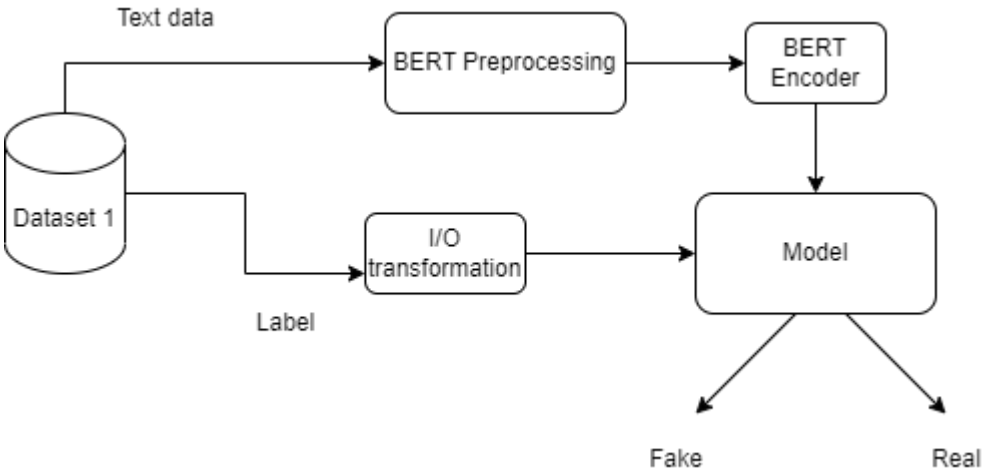
Dataset-2: consists of stock data in natural language for QnA chatbot approach which is generated using the below library. The structure of the data blocks consists of
Instruction: Question
Input: the entire block of answers
Output : the actual answer
https://github.com/Bavest/python-sdk

Validation set: live twitter textual tweets represented with labels signifying whether the user is a bot or human.
Official twitter api would be used for fetching tweets and the usernames.

The username would be validated using python botometer to check if the user is a bot or human.

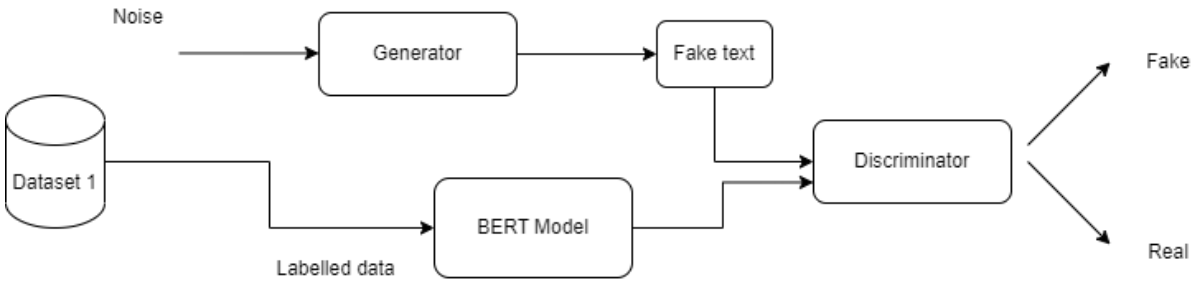
Approach 1 : Supervised learning approach using BERT model for classifying deceptive text.

- Linear approach for classifying deceptive texts, the architecture would consist of preprocessing, encoder and a model.
- BERT preprocess and BERT encoder would be used for preprocess and encoder.
- Tensorflow model for training with BERT encodings and the labels.
- Dataset 1 would be split into training and testing datasets



Approach 2: Semi-supervised approach using GAN-BERT for deceptive text. Considering a scenario with few labelled data samples, the linear BERT approach would not provide sufficient performance during fine-tuning. Hence GAN-BERT is used to extend the fine-tuning stage by introducing a Discriminator-Generator combination in the architecture.

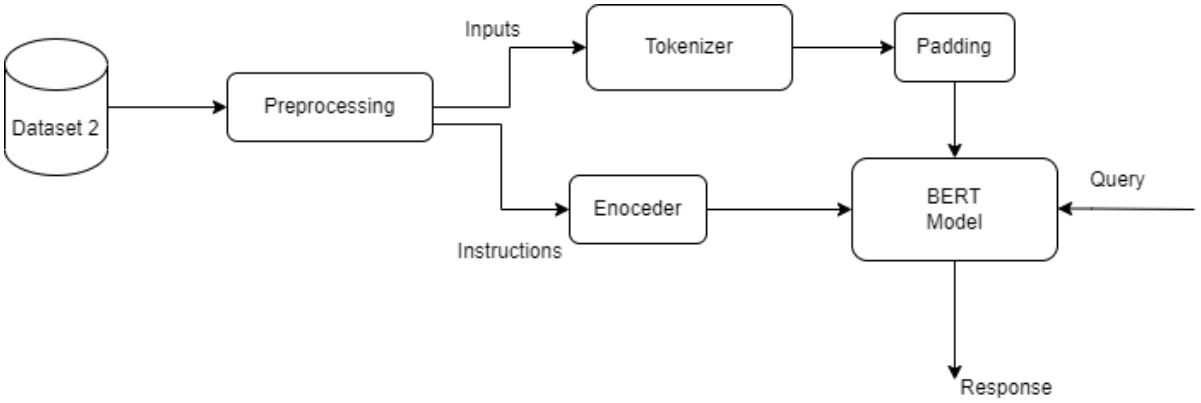
- the Generator G is devoted to produce "fake" vector representations of sentences
- The Discriminator D is a BERT-based classifier over n+1 categories.
- Small percent of dataset 1 would be used as a training data.



Approach 3: QnA approach using transformers for lie detection.

The new advancement in large language models and introduction of chatGPT, this approach takes the input statement and produces the output in natural language and corrects the user if the statement is untrue.

- The architecture starts with preprocessing of the dataset, where we categorise instructions,inputs and outputs.
- Inputs are tokenized and padded while instructions are encoded with transformer based dynamic encodings.
- Tensorflow model is used for training the instructions and inputs and provides answers for the statement or question in natural language.



References

[1] Albert Zeyer, Parnia Bahar , Kazuki Irie , Ralf Schluter , Hermann Ney, A comparison of transformer and LSTM encoder decoder models for ASR , IEEE, Feb 2020, Aachen, Germany, RWTH Aachen University, 10.1109/ASRU46091.2019.9004025.

[2] Valentin Hofmann , Janet B. Pierrehumbert , Hinrich Schütze, Dynamic Contextualized Word Embeddings, Association for Computational Linguistics, August 2021, Faculty of Linguistics, University of Oxford, 10.18653/v1/2021.acl-long.542.

[3] Amisha Sinha, Mohnish Raval, S Sindhu, “Machine Learning Based Detection of Deceptive Tweets on Covid-19”, Blue Eyes Intelligence Engineering and Sciences, International Journal of Engineering and Advanced Technology (IJEAT),June 2021, 10.35940/ijeat.E2831.0610521.

[4] S. Hu, “Detecting concealed information in text and speech,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 402–412.

[5] J. Pennington, R. Socher, and C. Manning, “Glove:Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.