Identifying tweets on COVID-19 for spreading fake information

Submitted by

Priyanshu
2019CH10115

under the guidance of
Prof. Anil Verma

Thesis presented to the
Indian Institute of Technology, Delhi
in partial fulfillment of the
award of the degree of
Bachelor in Technology

Department of Chemical Engineering
New Delhi, November, 2022

Thesis Certificate

This is to certify that the thesis titled **Identifying tweets on COVID-19 for spreading fake information**, submitted by **Priyanshu**, to the Indian Institute of Technology, Delhi, for the award of the degree of Bachelor of Technology, is a bonafide record of the research work done by him under my supervision.

Prof. Anil Verma
Department of Chemical Engineering
Indian Institute of Technology, Delhi

Contents

Acronyms

BERT idiectional Encoder Representations from Transformers

COVID-19 Corona Virus Disease 2019

FN Fake News

ICMR Indian Council of Medical Researc

IFCN International Fact Checking Network

LSTM Long Short Term Memory

NLP Natural Language Processin

RNN Recurrent Neural Network

UI User Interface

URL Uniform Resource Locator

Corona Virus Disease 2019 (COVID-19) International Fact Checking Network (IFCN) User Interface (UI) idiectional Encoder Representations from Transformers (BERT) Natural Language Processin (NLP) Uniform Resource Locator (URL) Recurrent Neural Network (RNN) Long Short Term Memory (LSTM) Indian Council of Medical Researc (ICMR) Fake News (FN)

List of Tables

List of Figures

Abstract

With the advent of the COVID-19 pandemic in 2020, there was also an onset of an Infodemic due to overabundance of all types of information on the social media platforms by self proclaimed experts. During the Lockdowns, Screen times of people across the globe increased manifolds, leading to a large exposure to all types of fake news spreading on the web.

There has been a discussion of implementation of various Natural Learning Processing techniques recently for solving many of the real-world problems we are facing today. During 2020, Twitter also rolled out calling out tweets for spreading fake information. However, humans can't process the vast ocean of data manually. So the same NLP techniques like transfer learning of BERT were used to classify tweets as "fake" or "real." The train dataset consisted of 6420 samples. Two variants of the BERT: small and standard, were used in the training process. The output from the BERT models were given as input to a light neural network layer consisting of a dropout layer. Finally, their F1 scores were **0.88** and **0.86**, respectively.

The *small-bert-en-uncased-L-4-H-512-A-8* model was saved and **deployed** on the Hugging Face interface with the help of Gradle library in Python with a clean user interface for testing and educational purposes.

Chapter 1

Introduction

During the recent and ongoing COVID-19 pandemic, there has been an overflow of information on all the social media platforms from people of all varieties of backgrounds. This has caused an overabundance of people interacting with them and leads to misinformation in a lot of cases. Such fake information in the cases of Covid-19 can be fatal for some and needs to be dealt with.

A study conducted by the International Fact Checking Network (IFCN) in 2020 lists symptoms, causes, and cures; morphed videos and pictures; comments from politicians; and conspiracies that blame particular groups, countries, or communities for the spread of the virus as examples of fake news. Some nations have experienced economic collapse as a direct result of the widespread dissemination of false information on social media. In certain nations, for instance, vegetarianism has gained popularity after false reports suggested that COVID-19 was spreading via non-vegetarian diets. The sales of meat and other animal products took a major hit as a result, threatening the livelihoods of many people in several nations.

Conspiracy theories like a particular country's biological weapon, water treated with lemon or coconut oil that may fight the virus, or medications that, while being authorised for different uses, could potentially be beneficial in prevention or treatment of COVID-19 all contribute to the possible impact of FN. The term "Infodemic knowledge" was used to describe the results of the epidemic of illness information sharing (Hua and Shaw 2020).

Misinformation detection on social media is crucial but difficult to implement. Because it requires time-consuming evidence collecting and meticulous fact verification, even humans have a hard time distinguishing fake from real news. With the rise of social media and the prevalence of false stories, it is crucial to develop automatic frameworks for spotting fake news. In this paper, I detail my technique for automatically determining if a tweet posted on social media is authentic or not using state of the art Deep Learning techniques like Transfer Learning and BERT algorithms. The models with the best performance were integrated with UI functionality in a web app which can take tweet text inputs and output the relevant truth probabilities.

Chapter 2

Dataset Description

2.1 Features

The dataset being used in this paper was crawled using Twitter API. For real tweets it depends majorly on big and certified news providers like WHO(World Health Organisation), ICMR(Indian Council of Medical Research) etc. Other than that all of the dataset was verified by humans using fact checking.

The entire dataset is divided into three parts: Train, Validation and Test each consisting of 6420, 2140 and 2140 examples respectively.

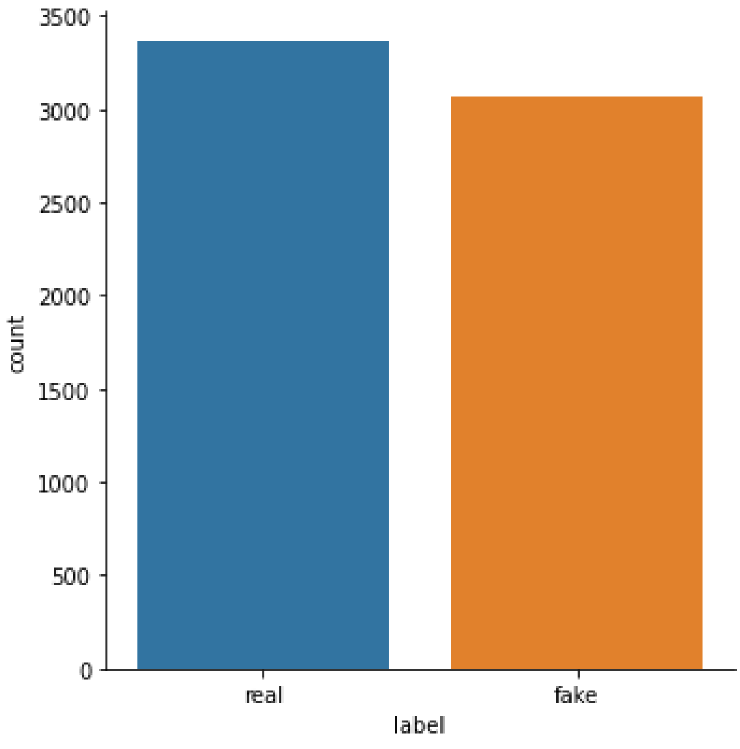

Figure1: Distribution of Train Data between Real and Fake news

As can be seen from Figure 1, our dataset is not biased towards any particular label and consists of an almost same number of examples from ”Fake” and ”Real” news.

Attribute	Fake	Real	Combined
Unique Words	19727	22915	37501
Average Words Per Tweet	21.7	32.9	27.1
Average Characters Per Post	143.3	218.4	182.6

Table 2.1: Dataset statistical analysis

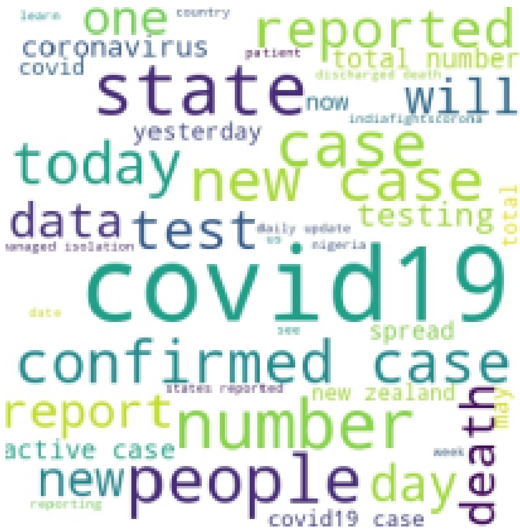

Figure 2: Word cloud of the dataset labelled "Real"

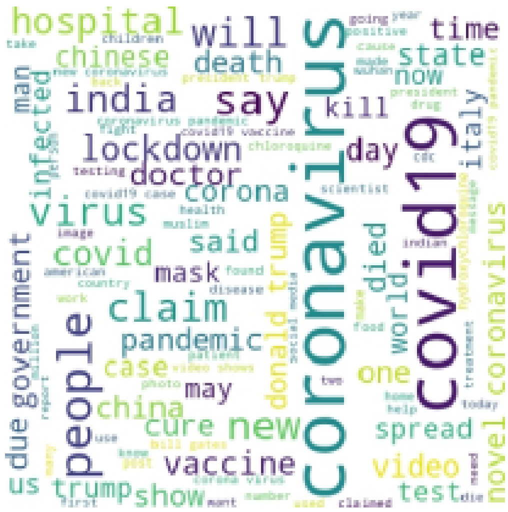

Figure 3: Word cloud of the dataset labelled "Fake"

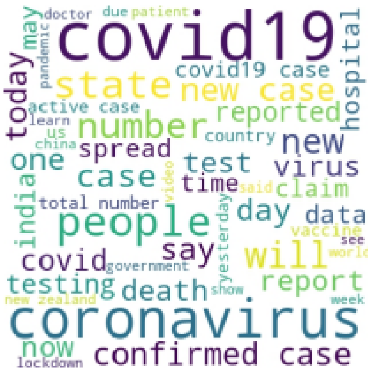

Figure 4: Word cloud for the whole dataset

2.2 Examples

The table below lists some of the example tweets from the dataset

Tweet Examples	**Labels**
"Football player Cristiano Ronaldo turned all his hotels into hospitals to help corona virus patients and is paying doctors and the staff."	Fake
"Protect yourself and others from #COVID19 when using public transportation .Practice social distancing avoid touching surfaces and practice hand hygiene"	Real
"Florida Governor Ron DeSantis Botches COVID-19 Response - By banning Corona beer in order to flatten pandemic curve."	Fake
"The new coronavirus causes sudden death syndrome."	Fake
"This is about giving the hospitals that capacity they need to get people through the doors. Health minister Edward Argar explains the additional £150 million funding to expand and upgrade hospital A&Es in England. #KayBurley"	Real

Table 2.2: Some Tweets and their respective labels.

Chapter 3

Literature Review

Natural Language Processing (or NLP) in Computer science refers to the technique by which we aim to analyse human text samples for understanding their meaning and sentiment with the help of computing power of machines and state of the art Machine Learning and Deep Learning techniques. With the onset of the Internet era, there has been an overflow of data of all types, be it text, photo or video.

It has been almost impossible to moderate the content that is being circulated over the web. As discussed in the problem statement, the need for this increased multi-fold during the COVID-19 pandemic.

3.1 Pre-Processing

First step in the process involves pre-processing for organising and increasing the machine readability of our tweet dataset. It involves removing any parts of our input that do not contribute significantly to our model training process like any URLs, emojis and other symbols like "&amp".

In this paper, however, other hard pre-processing techniques like stop-words removal or reducing elongated words were not implemented to preserve the original language as these factors help the transformer based learning algorithms to establish a meaningful relation between different words of a sentence.

Example tweets like these do not contribute to the training process and make our model to learn the wrong way:

"Where Are They Now: Covid-19 https://t.co/j6AAcjvhg9"

"https://t.co/JDv5GVMioP is the most comprehensive body of data on the effects of COVID19 on patients with cancer. https://t.co/Zk8FaSHKIM CCC19 @covid19nccc"

The "Regular Expressions" module in python can be used to form an expression which can remove URLs in our data.

3.2 Models

3.2.1 Transformers

Transformers are a type of complex artificial neural network architecture developed by a team of engineers at Google in 2017. It uses self-attention methods to analytically weigh the importance of different parts of input sentences in Natural language.

Transformers allow for parallelisation of the architecture giving faster training times compared to RNNs (Recurrent Neural Networks) like LSTM (Long Short Term Memory).

The Self-Attention part in this model consists of three components: Key(K), Value(V), Query(Q) These (K,Q,V) pair is generated for each word in the input sequence and is used to calculate the self-attention term.

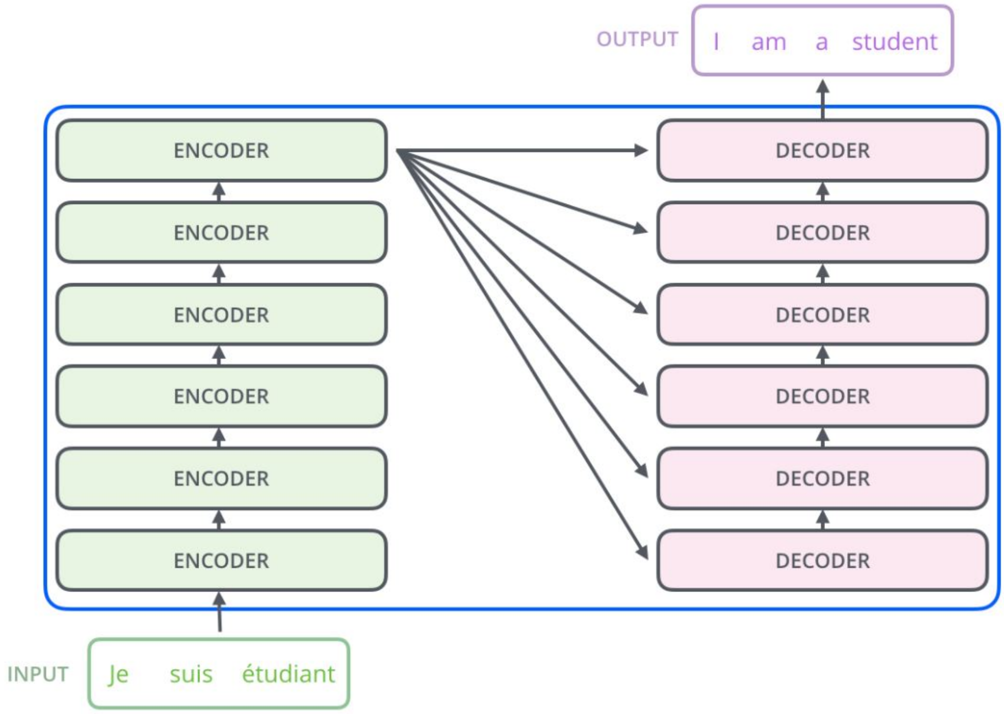

Figure 5: Low level representation of working of a Transformer

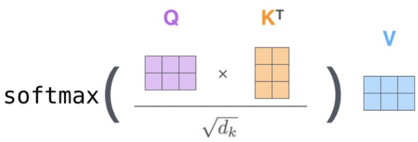

Figure 6: Self-Attention calculation using (Q, K, V)

BERT(Bidiectional Encoder Representations from Transformers) is a pre-trained Transformer based architecture model. This is almost identical to the original Transformers architecture.

It was Pre-trained on **Language modelling** and **Next Sentence Prediction** tasks. This process requires specific machine specifications and requires lot of time. We need to only fine-tune the model based on our requirements which can be done on our machines

3.3 BERT

For this paper, I have used the following models from Tensorflow Hub to implement Transfer Learning:

1. small-bert/bert-en-uncased-L-4-H-512-A-8
2. bert-en-uncased-L-12-H-768-A-12

For transfer learning, these pre-trained BERT models were attached finally to a light neural network and Dropout layer was also used for preventing over-fitting of our model to the provided dataset. The neural network is as follows:

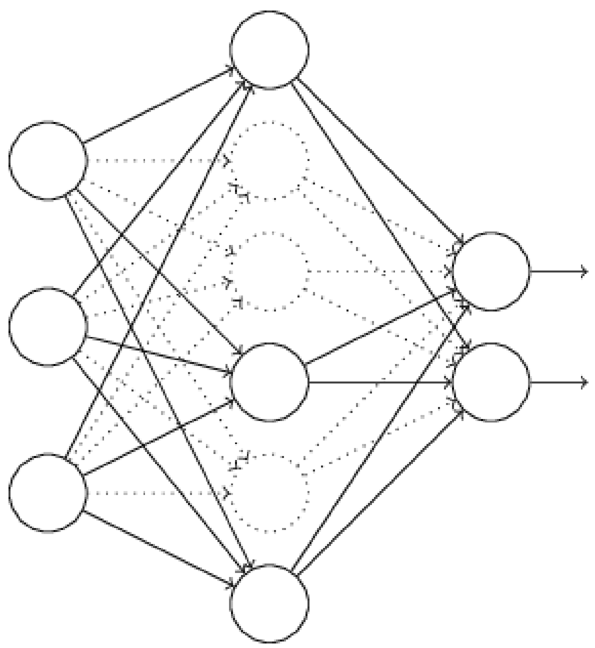

Figure 7: Neural Network Layer on top of our pre-trained BERT model

Chapter 4

Results and Discussion

Outputs of both of our models from the dense layer of neural network were used for predicting the final output probabilities. F1-score was considered as the ultimate metric for a model's usability. It is calculated using the following formula:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Figure 4: F1-score

Precision	Recall	Accuracy	F1-score
0.9428	0.8328	0.8710	0.8844

Table 4.1: Evaluation score Matrix for Model 1

Similarly for the **BERT standard**, results obtained were as follows:

Precision	Recall	Accuracy	F1-score
0.9526	0.7874	0.8406	0.8622

Table 4.2: Evaluation score Matrix for Model 2

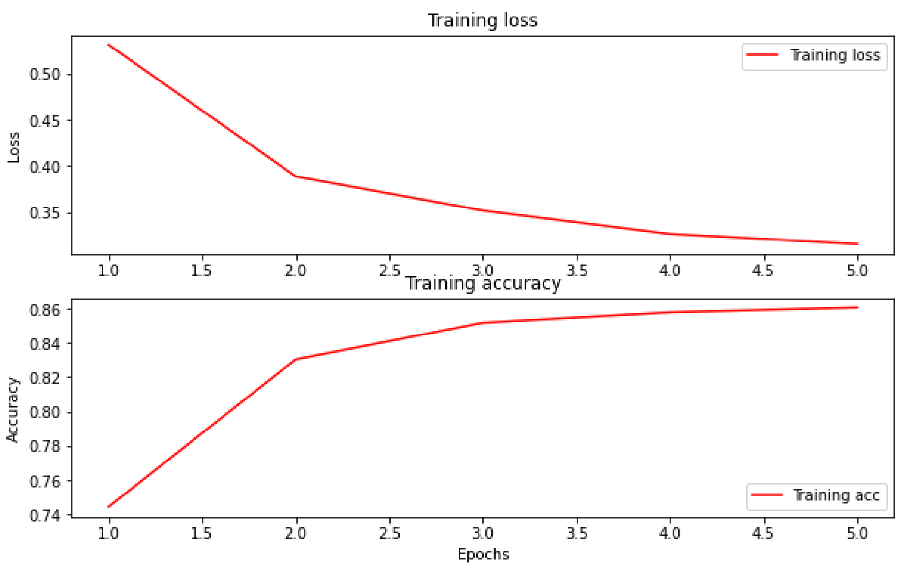

Figure 8: Training Loss and Accuracy for Model 1

4.1 Deployed Model Predictions

The *small-bert-en-uncased-L-4-H-512-A-8* model was saved and deployed on **HuggingFace** interface with the help of *Gradle* library in python for getting outputs on unseen dataset by our model. Click Here to try our deployed model.

Following are some example outputs from our model on the interactive web-page:

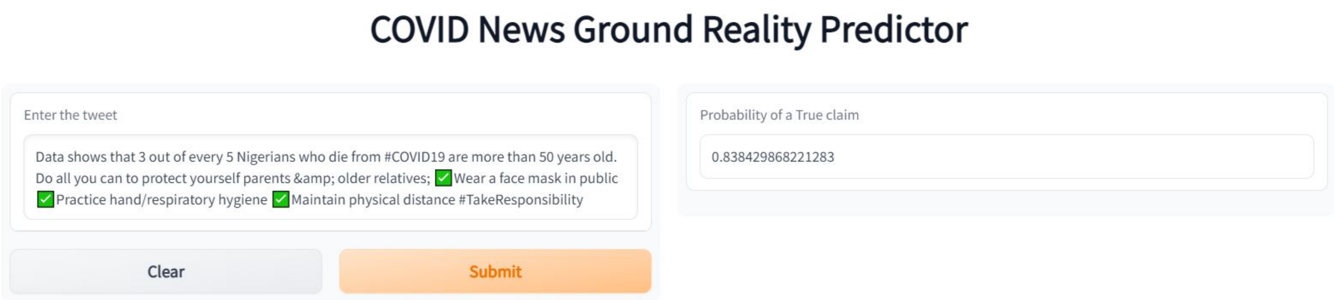

Figure 9: Result 1

Figure 10: Result 2

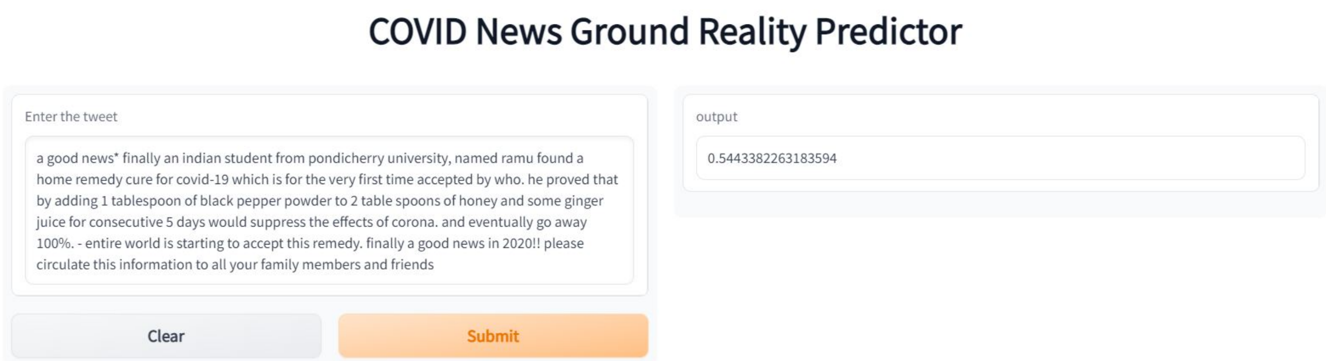

Figure 11: Result 3

Chapter 5

Conclusion

Consider the following two simple tweet inputs:

1. "COVID-19 first case was found in 2019"
2. "COVID-19 first case was found in 2018"

The output probability of their truth by our model came out to be 0.44 and 0.48 respectively. Comparing these results to the reality it does not make much sense. Probability of the first tweet should have been much greater than the second one, but we in fact obtain a lesser probability compared to second.

The results obtained on our testing dataset are great(F1 score: 0.87, 0.84), however examples like above state the obvious limitation of the final model to any random COVID-19 tweets from the Web. The given dataset apart from limited training data (6240), also has skewed distribution of non-statistical or fact based data under the "Real" label.
Hence, in cases like above, our BERT model is only able to establish relation between sentence parts but it needs more data to be able to tackle any type of tweet inputs.

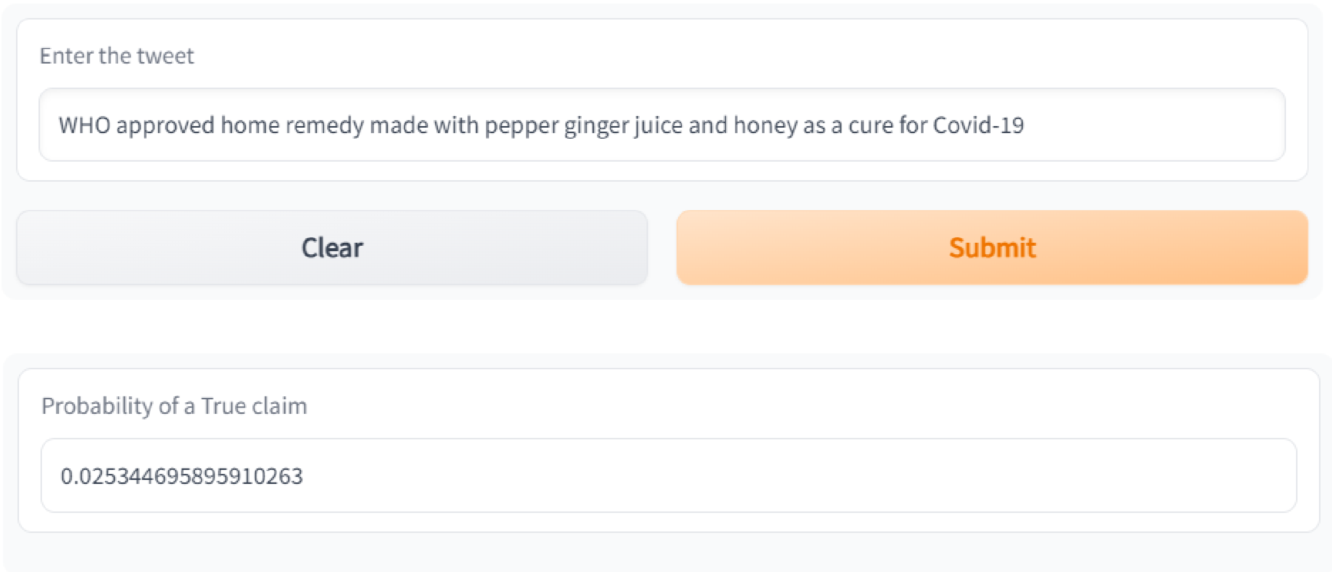

Figure 12: Result 4

Chapter 6

Future Scope

The prominent problems we saw from the results, as expected were over-fitting due to small size of our dataset. In our dataset, most of the tweets labelled "Real" are about number of cases updates around the world, Vaccination updates and other similar types of statistical data.

There are not generally any fake news regarding these types of statistical data. Most of the "Fake" tweets subset involves factual data. Because of that reason, for most of the Non-statistical inputs, we get a lower truth probability from our model.

Factual dataset can be crawled from twitter and fact checked for getting more annotated dataset. There exist other datasets, which use the Tweet-ID format, instead of plain text. Problem with these using Tweet-IDs is that, if original poster or the tweet does not exist anymore, that data row does not make any sense for our training process.

Current work and research in Natural Language Processing is mainly focused on the English language. This can be extended to other major languages of the world like Chinese, Hindi, Bengali etc. by relevant algorithms and Transfer learning approaches .

Bibliography

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[3] S. D. Das, A. Basak, and S. Dutta, "A heuristic-driven ensemble framework for covid-19 fake news detection," 2021.

[4] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Fighting an infodemic: COVID-19 fake news dataset," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pp. 21–29, Springer International Publishing, 2021.

[5] N. Nambiar, "Predicting covid-19 fake news," 2022.

[6] D. Kar, M. Bhardwaj, S. Samanta, and A. P. Azad, "No rumours please! a multi-indic-lingual approach for covid fake-tweet detection," 2020.

[7] Y. M. Rocha, G. A. de Moura, G. A. Desidério, C. H. de Oliveira, F. D. Lourenço, and L. D. de Figueiredo Nicolete, "The impact of fake news on social media and its influence on health during the covid-19 pandemic: A systematic review," *Journal of Public Health*, pp. 1–10, 2021.

[8] S. Khan, S. Hakak, N. Deepa, B. Prabadevi, K. Dev, and S. Trelova, "Detecting covid-19-related fake news using feature extraction," *Frontiers in Public Health*, p. 1967, 2022.

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] W. De Mulder, S. Bethard, and M.-F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling," *Computer Speech & Language*, vol. 30, no. 1, pp. 61–98, 2015.

[12] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," *arXiv preprint arXiv:1602.02410*, 2016.

[13] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools and Applications*, 07 2022.