Hate Speech and Fake News Detection

Department-Msc. Data Science

Dhirubhai Ambani Institute of Information and Communication Technology

Gandhinagar, Gujarat, India (382007)

Ravi Satvik
202018008@daiict.ac.in

Siddhant Shah
202018013@daiict.ac.in

Sambhav Gulla
202018018@daiict.ac.in

Saswat Nanda
202018029@daiict.ac.in

Omkar Chavan
202018037@daiict.ac.in

Dev Patel
202018055@daiict.ac.in

Keya Shah
202018056@daiict.ac.in

ABSTRACT

Hate speech is commonly defined as any communication that belittles a target group of people based on some characteristic such as colour, gender, ethnicity, sexual orientation, nationality, religion, race, or other characteristic. Due to the massive rise of user-created web content on social media, the numbers of hate speech is also steadily increasing. Over the past years, interest in online hate speech detection and, particularly, the automation of this task has continuously grown, along with the societal impact of the incident. This paper describes a hate speech classification model based on Distil-BERT.

During the COVID-19 pandemic, social media has become a home ground for false information. An example of its influence can be found in the fake-news outbreak that followed the COVID-19 pandemic and the events that Infected thus. To tackle this infodemic, scientific oversight, as well as a proper understanding by practitioners in crisis management, is needed. This paper describes a fake news classification model based on GloVe embeddings for feature extraction followed by training an LSTM model.

KEYWORDS

Twitter: Hate Speech: Distil-BERT Model: COVID-19: Fake News: LSTM model:

1 INTRODUCTION

Hate Speech:

In recent years, social networks (especially Twitter) have been used to spread hate messages. Hate speech refers to a kind of speech that defames a person or multiple persons based on their membership to a group, usually defined by colour, ethnicity, sexual orientation, gender identity, disability, religion, race, political affiliation, or views. Related to this, hate crimes are a type of violation of the law, whose primary motivation is the entity of prejudices regarding the victims.

Researchers claim that hate crimes are influenced by singular widely publicized events like terrorist attacks, uncontrolled migration, demonstrations, riots, etc. These events usually act as triggers, and their effect is dramatically increased inside Twitter and a source of valuable information for crime forecasting.

Twitter is filled with messages from individuals instigating punishment against different targeted groups. When these messages are collected after an event over a period of time, they can be used for the analysis of hate crimes in all the phases: climbing, stabilization, duration, and decline of the threat. Therefore, monitoring Twitter becomes a main priority for the forecasting, detection and analysis of hate crimes.

Following this necessity, the main goal of our project is to design a model to identify and classify Hate Speech in Twitter as well as.

Fake News:

Fake news sharing has become dominant in today’s digital world. This suggests that even some government officials and individuals engage in the spreading of false-formation to a large audience to suit their agenda. Thus, fake news has touched virtually every aspect of our life, and the most precarious in recent months is the Source text expansion of false content in this period of the coronavirus disease 2019 (COVID-19) outbreak.

According to a recent survey, many rumours and false-news stories and messages are circulating about the COVID-19. It is becoming more and more difficult to distinguish fake news from reports whose rectitude should not be questioned. Due to, false-information in social media has fuelled panic among members of the public regarding the COVID-19 pandemic, prompting governments and authorities to request citizens to confirm the actuality of news stories before circulating them.

In this view, research has found that as the importance to find a treatment for COVID-19 continues across the globe, fake news proliferation has become acute on social media, which many experts and higher authority believes is contributing to the threats of the pandemic. It has been found that misinformation regarding health issues constitutes a probable threat to public health.

Therefore, to comprehend the predictors of fake news sharing on social media, we developed a comprehensive model in which features were extracted using GloVe embeddings and a LSTM model was trained on these features.

2 DATASET

Hate Speech:

In this work we use the first public dataset of hate speech annotated on Internet forum posts in English at sentence-level. The dataset is publicly available on GitHub [https://github.com/aitor-garcia-p/hate-speech-dataset] .The source forum is Stormfront , the largest online community of white nationalists, characterised by pseudo-rational discussions of race, which include different degrees of offensiveness. Storm-front is known as the first hate website.

Fake News:

COVID Fake Social Media Posts:

This dataset focuses on COVID19-related fake news in English. The sources of data are various social-media platforms such as Twitter, Facebook, Instagram, etc. Given a social media post, the objective of the task is to classify it into either fake or real news. In this dataset we have a total of 6451 social media posts. 3091 are fake and 3360 are real. The dataset was taken from a competition on caodalabs. [https://competitions.codalab.org/competitions/26655#learn_the_details]

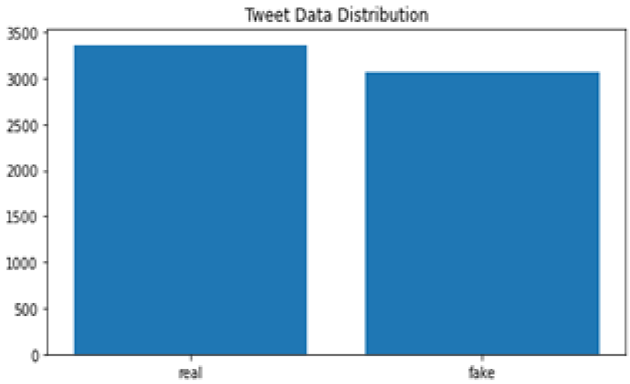

Fig 1 : Data Representation of Fake News dataset

3 METHODOLOGY

3.1 Data Preprocessing:

For Fake News:

- Remove URLs from the text.
- Remove the following characters: | : , ; & ! ? /.
- Normalize hashtags into words, so #refugeesnotwelcome becomes refugees not welcome.
- Lowercase and stemming to reduce word inflections.
- Remove user mentions from the text.
- Remove stop words from the text.

GloVe Embeddings:

Neural networks used in NLP tasks do not operate directly on texts, sentences, or words, but on their representation in the numerical form. This process of converting them into numbers is called word embeddings and it is one of the key elements enabling sentiment analysis and fake news detection. The main methods of word embeddings are ‘word2vec’, ‘glove’, and ‘FastText’. In this work, the ‘glove’ method was used.

It is a very powerful word embedding technique that has been used for text classification, where each word is presented by a high dimension vector and trained based on the surrounding words over a huge corpus. The pre-trained word embedding used in many works is based on 400,000 vocabularies trained over Wikipedia. GloVe provides pre-trained word vectorizations with 100, 200, 300 dimensions. In this work we have chosen to work with 300 dimension vectors.

For Hate Speech:

- Truncate the vector length to model max input size
- Collect list of items to one single dictionary
- Zip all input ids, masks and labels together

- For the second BERT model the same steps were repeated along with 4 different embedding strategies (explained in Model section)
- The data was fed to the model in both the models using dataloader

3.2 Models:

BERT

We first use a fine tuned Distil Bert Model. The setup at first is a simple classification setting to distinguish between hate and non-hate speech. For training the model [6] the authors recommend batch sizes of 16 and 32, learning rate 5e-5, 3e-5and 2e-5, Adam optimizer, Number of epochs 2, 3 and 4. As mentioned in [6], we train the model. We fine-tuned the model for 2-3 epochs which yielded satisfactory results.

The second step was to extract sentence embeddings from the pre-trained Bert model. We use multiple strategies to check if it will perform better than the previous Bert model. We extract the sentence embedding once from the last layer and once from the second to last layer. We perform a max pool and mean of all word embeddings on both the layers. This was trained on a slightly modified DistilBertForSequenceClassification from the hugging face library. We have removed all the computations in the model and only train the head of the model.

LSTM

For the purpose of fake news detection we use a bidirectional LSTM model. The model is preceded by a 1D Convolutional layer and a spatial dropout layer with 0.2 dropout rate. After the LSTM layer we have added one dense layer with 512 units and relu activation followed by a dropout layer with 0.5 dropout rate and another dense layer with same units and same activation function. The output layer is a 1 unit dense layer with sigmoid activation function. The use of two dropout layers curbs the problem of overfitting.

The model was trained using Adam optimizer and binary cross entropy as loss function and trained with 15 epochs.

4 EXPERIMENTAL RESULTS

This section describes the various performance metrics that we used to evaluate our models. We have also compared our models to research taken on the same datasets and provided a comparative study.

Fake News:

For the fake news classification on covid tweets, the table below summarizes the performance of our LSTM model.

	Precision	Recall	F1-Score
Real	0.88	0.91	0.90
Fake	.92	0.89	0.90
Weighted Avg	0.90	0.90	0.90

Tab 1: Performance of LSTM Model

When compared to other models in the competition for the same dataset we found that our accuracy was not far behind. The highest accuracy achieved was .98 by user parthpatwa.

The table below compares our model with the model used by the others in the competition

User	Weighted Precision	Weighted Recall	Weighted F1
parthpatwa	0.98	0.98	0.98
srishti.sahni	0.97	0.97	0.97
Our Model	0.90	0.90	0.90

Tab 2: Comparison Our Work With Other

Below are the train test accuracy/ loss given over epochs and the confusion matrix for our model..

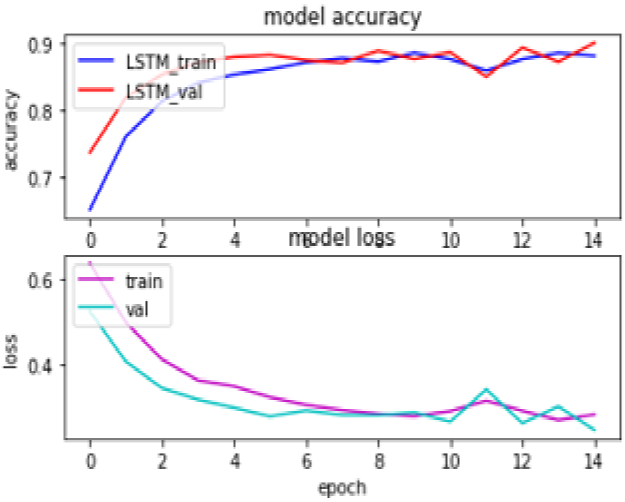

Fig 2: Model Accuracy and Loss

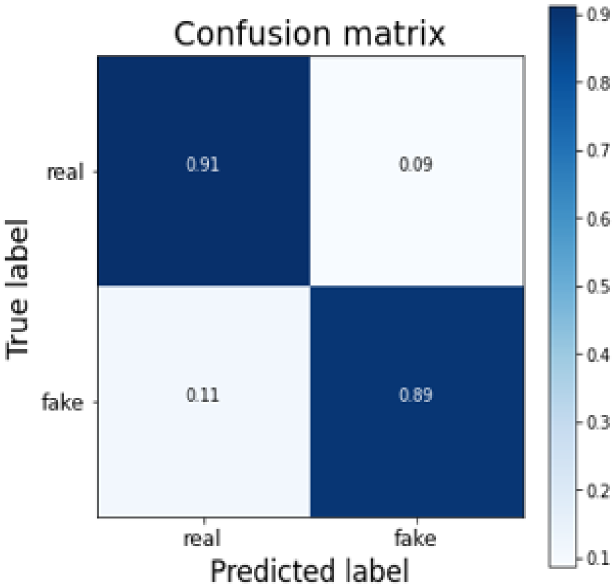

Fig 3: Confusion Matrix of a Model

Hate Speech:

For the hate speech classification on the white supremacist data, the table below summmarizes the performance of our fine tuned distil BERT model. We can see the evaluation loss for 2 epochs is the lowest. We can also see that the overall accuracy for 4 epochs is the highest but this model is clearly overfitting given the drastic decrease in eval_train_loss and increase in eval_loss.

Epoch	eval_train_loss	eval_loss	accHate	accNohate	Acc
2	0.150	0.438	0.807	0.836	0.82
3	0.048	0.55	0.799	0.853	0.82
4	0.013	0.70	0.861	0.820	0.84

Tab 3: DistilBert Model

The table below summarizes the performance of our Distil Bert embeddings + Distil Bert head model. Various strategies have been used here:

Emb Strategy	eval_train_loss	eval_loss	eval_accHate	eval_AccNoHate	eval_AccAll
CLS	0.54	0.57	0.78	0.67	0.73
LastMean	0.54	0.55	0.83	0.69	0.76
LastMax	0.60	0.61	0.82	0.68	0.75
Last2Mean	0.51	0.54	0.81	0.68	0.74
Last2Max	0.58	0.59	0.82	0.68	0.75

Tab 4: DistilBert For Sequence Classification from huggingface library

We can see the best strategy by eval_loss is LAST2_MEAN (0.542). None of these embeddings are better than the fine-tuned model above (with an eval_loss of 0.438) so we can conclude that fine tuning is better than simply getting embeddings. This makes sense given that when we fine tune, we are also fine tuning the embeddings themselves (the CLS token). It is worth noting that the CLS token is used as the input for the classification head, so when we fine tune we can actually use the CLS token as the best representation (embeddings) for the entire sentence.

5 CONCLUSION

Observing the results for hate speech classification using our distil BERT model, loss for 2 epochs is noticed to be the lowest. Although the accuracy is higher at 4 epochs, this model is observed to be overfitting. This can be interpreted from the table that calculates various parameters for different epoch values. Using various embedding techniques, Last2Mean method was the best strategy with the lowest eval_loss. But this is not low enough as we achieved it using fine tuning with 2 epochs. This explains that the fine tuning method fine tuned the embeddings(CLS token) itself and thus produce such results.

Now coming on to the fake news, we can conclude that the model used gave us some impressive results using LSTM as compared to other methods used in the competition. There is a minute difference between accuracies between the best method which had 0.98 and that compared to our model with accuracy of around 0.90. From the graphs, we are able to summarise that as the accuracy increases, the loss decreases for both training as well as validation data. Also, the loss decreases significantly as we increase the number of epochs. Performance of the model can be visualized using the Confusion matrix given in the results.

6 REFERENCES

[1] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam,Chris Biemann,Pawan Goyal,Animesh Mukherjee. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection,.Indian Institute of Technology, Kharagpur, India, Universitat Hamburg, Germany. (2-4) https://arxiv.org/pdf/2012.10289v1.pdf

[2] Sai Saketh Aluru.Binny Mathew,Punyajoy Saha and Animesh Mukherjee. 2020. Deep Learning Models for Multilingual Hate Speech Detection. Indian Institute of Technology Kharagpur, India.(7-10). https://arxiv.org/pdf/2004.06465v3.pdf

[3] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian,Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. Information Retrieval Laboratory, Georgetown University, Washington, DC, United States of America. https://www.researchgate.net/publication/335311146_Hate_speech_detection_Challenges_and_solutions

[4] Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, Raviraj Joshi. 2021. Evaluating Deep Learning Approaches for Covid19 Fake News Detection. Pune Institute of Computer Technology, Pune. https://arxiv.org/pdf/2101.04012v2.pdf

[5] Kai Shu,Amy Sliva,Suhang Wang,Jiliang Tang and Huan Liu. 2017. Fake News Detection on Social Media A Data Mining Perspective. Computer Science & Engineering, Arizona State University, Tempe, AZ, USA https://arxiv.org/pdf/1708.01967v3.pdf

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://arxiv.org/pdf/1810.04805.pdf