

SOCIAL MEDIA TROLLS IDENTIFICATION USING MACHINE LEARNING

2024

PRAVESH SHARMA

January 29

Tech batch-1

ASSESMENT

## **AIM AND OBJECTIVES**

## AIM:

To use “dataset-for-detection-of-cyber-troll” data to identify whether a social media comment/tweet is a troll or not by implementing fundamentals of Machine Learning Techniques(Natural Language Processing).

OBJECTIVES:

The main objectives of this research are:

1. To prepare data for our modelling experiments using Data Manipulation and Data Analysis.
2. To build a Baseline Model to set a base accuracy score and try to beat it by creating other Machine Learning/Deep Learning Models.
3. To use Pretrained Embedding Layer (Universal Sentence Encoder) to build a model and check its training (is it Overfitting or not).
4. To build a Hybrid Model (Pretrained Token Embedding + Character Embedding) to reduce the problem of Overfitting and visualize its Training and Validation Curves.
5. To evaluate our best model results and use our best model to make predictions on Test Data.

**LITERATURE REVIEW**

The relevant literature to this research is ***PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts***

**ABSTRACT**

Frank Dernoncourt and Ji Young Lee proposed a new dataset based on PubMed for Sequential Sentence Classification. This dataset consists of approximately 200,000 abstracts of trials in medical field, totaling of 2.3 million sentences and each sentence is labelled with their roles. Classes are: BACKGROUND, OBJECTIVE, METHODS, RESULTS AND CONCLUSION.

Purpose behind this was to increase the size of dataset that will help to develop more accurate algorithm in Short Text Classification.

**INTRODUCTION**

Short text classification is an important task in many areas of Natural Language Processing, such as sentiment analysis, question answering, etc. The dataset they presented in that paper i.e. PubMed 200k RCT, each short text we consider is one sentence.

They used this for classifying sentences in medical abstracts/paragraph available in the RCT (Randomized Control Trials) and they called this task as Sequential Sentence Classification Task.

Purpose behind this task of skimming paragraph into sentence and classifying them as: BACKGROUND, OBJECTIVE, METHODS, RESULTS AND CONCLUSION was to make it easy for the medical invigilators to pinpoint the information they are looking for.

So, this can significantly reduce the time to locate the desired info in the paragraph.

**DATASET CONSTRUCTION**

1. Abstract Selection: This Dataset was constructed upon the MED-LINE/PubMed

Baseline Database published in 2016. It consists of about 24,358,442 records.

1. Dataset Split: There are two datasets:

There are two datasets:

1. PubMed 200K consisting of a validation set containing 2500 abstracts, test set consisting of 2500 abstracts and a training set consisting of 190,654 abstracts.
2. PubMed 20K consisting of a validation set of 2500 abstracts, test set consisting of 2500 abstracts and a training set consisting of 15000 abstracts.

**PERFORMANCE BENCHMARKS**

They created four models namely:

1. First Baseline is a classifier based on Logistic Regression (LR) which obtained the F1 Score of 83.1 on PubMed 20K and 85.9 on PubMed 200K.
2. The second model uses the Forward Artificial Neural Networks (Forward ANN) which obtained the F1 Score of 86.1 on PubMed 20K and 88.4 on PubMed 200K.
3. The third model was the Conditional Random Field that uses n-grams as features which obtained the F1 score of 89.5 on PubMed 20K and 91.5 on PubMed 200k.
4. The fourth model was bi- ANN model which consists of a token embedding layer(bi-LSTM), a sentence label prediction layer(bi-LSTM) and a label sequence optimization layer. It obtained the F1 Score of 90.0 on PubMed 20K and 91.6 on PubMed 200K.

**CONCLUSION**

They concluded that the 4th model was performing the best on both the Datasets which is the Tri-Brid model consisting of pretrained token embedding layer, character embeddings and positional embeddings (We will try to replicate this model and modifying it according to or problem).

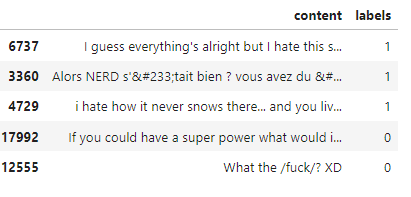
**METHODOLOGY**

DATA PREPARATION

1. The data which we have used for this research was named as “dataset-for-detection-of-cyber-trolls” and it is publicly available on Kaggle. Firstly, we have loaded the data in the raw format (it was available in JSON format) and converted it to normal Data Frame using “.apply(lambda x : method)”.



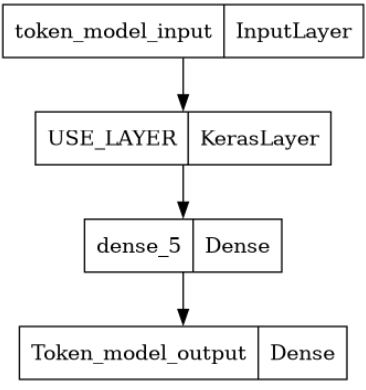
Raw Format



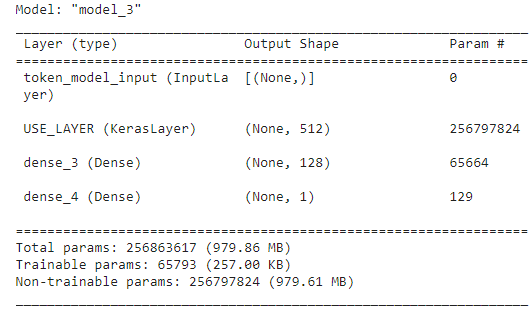
Data Frame containing only sentences and labels

1. We converted the content column to the list of sentences and labels to an array of the labels (Here, 1 = Troll and 0 = Non-Trolls). Then we split these sentences and labels into train, test sentences and train, test labels with the split size of 0.3.
2. Experiments which we are going to run includes following models:
3. **Baseline Model**: Consists of a pipeline which has a TfidfVectorizer Layer and Multinomial Naïve Bayes. This model is trained to set a baseline accuracy which we have to improve in other models.
4. **Pretrained Token Embedding Model**: Used Tensorflow Hub’s *Universal Sentence Encoder* layer followed by a Dense Layer. Model’s info is given below:

* **Token Model Structure:**

****

* **Token Model Summary:**

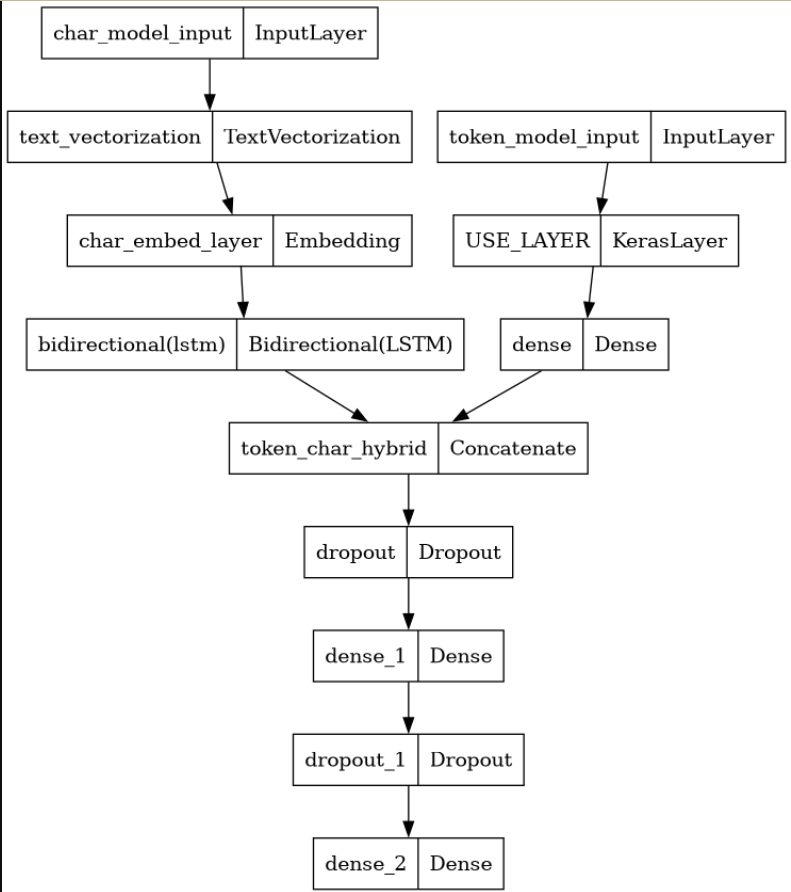
****

1. **Hybrid Model**: This model is a hybrid model i.e. made by concatenation of two different models which are pretrained token embedding model and character level embedding model.

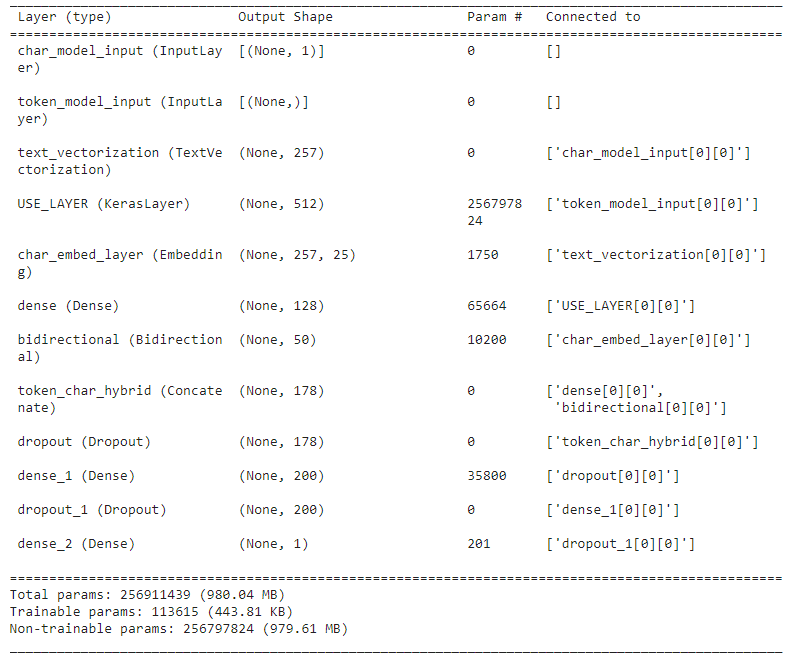
For pretrained model we use Universal Sentence Encoder layer, and for character model we have used character level data.

This will reduce the overfitting issue which occurred when we created previous models and will predict a lot better on validation datasets.

* **Hybrid Model Structure:**

****

* **Hybrid Model Summary:**

****

1. Preparation of datasets: We have prepared prefetched datasets with the batch size of 32, and fitted the model\_2 and model\_3 on *token datasets* and

*token char dataset* respectively.

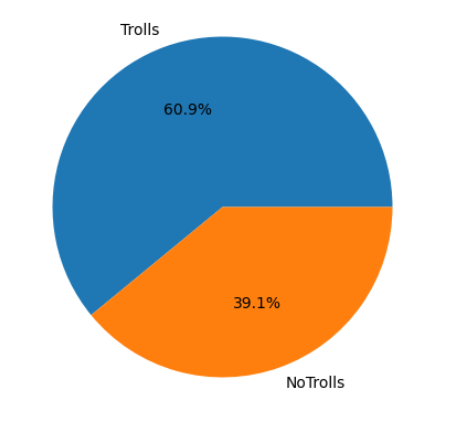
1. The evaluation metrics used for the performance of our models are :

* Accuracy score,
* Precision,
* F1 Score,
* Recall

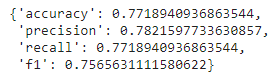
**CONCLUSION**

1. Among all the data we have, around 61% of the tweets are Troll Tweets and rest of 39% are Non-Troll tweets.

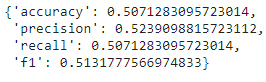
Distribution is plotted below:



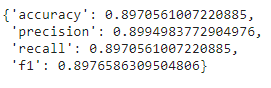
1. Our Baseline model’s performance:



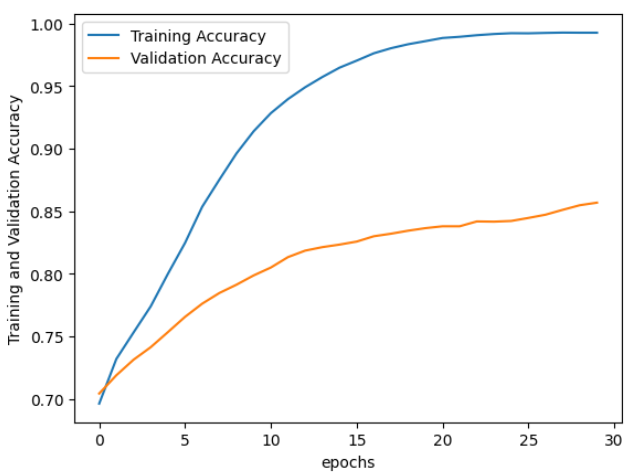
1. Our **Pretrained Token Embedding Model’s** performance: This model is getting a training accuracy of over 95% but it’s not performing so good on Validation Dataset i.e. it is overfitting.



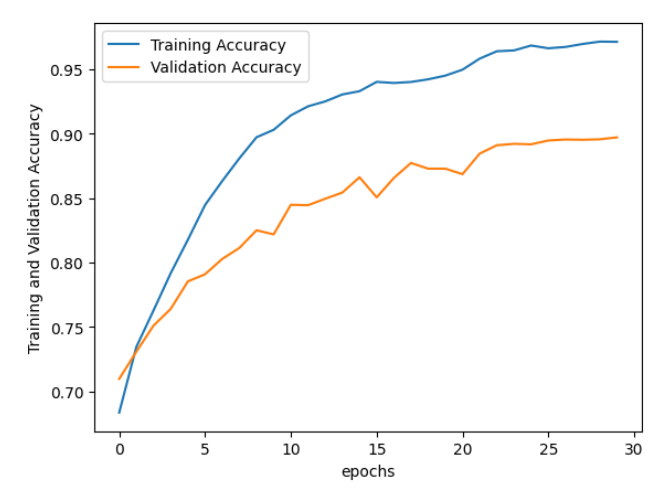
1. Our **Hybrid Model’s** performance: Clearly it is the best performance model achieving following scores-



1. Accuracy Curves of our models:
2. **Token Model:**



1. **Hybrid Model:**

****

**REFERENCES**

1. **PubMed 200K RCT**

**(Literature)**

1. **Franck Dernoncourt∗**

**Adobe Research**

1. **Ji Young Lee**

**MIT**