Natural Language Processing Project Report

Mohamed Amine Rguig
Université Paris-Dauphine

Introduction

As social media platforms have become increasingly prominent, online discussions have evolved into a fundamental aspect of the internet. However, this transformation has also exposed a dark underbelly, where abusive content can inflict significant harm on both individuals and society as a whole. Our project focuses on the vital task of toxic comment classification, from the Kaggle challenge [1]. Our primary objective is to accurately differentiate toxic comments from constructive and respectful discourse. To accomplish this, we have employed three distinct models: Fine-tuning DistilBERT, implementing TF-IDF (Inverse Document Frequency) for comment vectorization and Classification, and Finally a Bidirectional LSTM model.

1 Dataset Description

The dataset, derived from Wikipedia's talk page edits, is designed for a Kaggle challenge that aims to classify comments into six distinct categories: 'toxic,' 'severely toxic,' 'threat,' 'insult,' 'identity hate,' and 'non-toxic.' It consists of approximately 150,000 comments for both training and testing, with labeling conducted via crowdsourcing. This diverse dataset presents a rich variety of linguistic styles and toxicity levels, reflecting the complexities of online discours.

2 Finetuning DistilBERT

Transfer learning from large-scale models has gained widespread traction across various NLP tasks. In our project, we opted for fine-tuning DistilBERT, a distilled version of BERT renowned for its efficiency without sacrificing accuracy. By fine-tuning DistilBERT, we achieved remarkable results, with our model attaining an impressive 98.2% accuracy on the test set. The training process was conducted on 2 Nvidia-V100 GPUs, employing a carefully selected learning rate of 0.00001. The model was trained for a total of 4 epochs, and we implemented an early stopping mechanism based on validation loss to prevent overfitting.

3 TF-IDF + Classifiation

Our second model utilizes TF-IDF (Term Frequency - Inverse Document Frequency) for comment vectorization, a technique specifically crafted to distill the semantic essence of sentences from their constituent words. The TF-IDF process consists of two pivotal stages: an initial computation of term frequency and a subsequent weight assignment based on a word's presence within a particular document. Subsequently, after transforming the comments into TF-IDF vectors, we employed Logistic Regression with chained multi-label classification, capitalizing on label correlations. This strategic approach resulted in a noteworthy 1% improvement when contrasted with the binary classification of individual labels.

4 Bidirectional LSTM

For our third model, we harnessed the capabilities of LSTM (Long Short-Term Memory) networks—a specialized architecture for sequence data. Our LSTM model, consisting of 1 LSTM layer followed by 2 dense layers, was meticulously crafted to grasp the intricate

contextual dependencies within comments, enabling it to discern subtle linguistic patterns. This section will provide a deeper insight into the architecture of our LSTM model, its training methodology, and its impressive performance in classifying toxic comments, achieving a commendable accuracy of 95% . This result underscores the significant role played by toxic language in the classification task.

5 Results

As we expected the most performing model was Finetunned DistilBERT, showing very good results, showing 98.2% Accuracy, this showed the power of transfer learning compared to training a model from scratch, however our two other models did not shy away as our second model got 92% Accuracy, and the bi-LSTM model with 95 % Accuracy, showing that the task does not require complex models, maybe because of the repetitions of injure words. Some of the best solutions did not require pretrained models, but the best models often used various blends from models, a thing we did not experience as our goal to obtain the specific best solution with the perfect blend .

References

[1] cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, Will Cukierski.

Jigsaw Toxic Comment Classification Challenge.

https://kaggle.com/competitions/jigsaw-toxic-comment-classification

[2] Victor Sanh, L.Debut J. Chaumond and T.Wolf. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*

[3] Jason P. C. Chiu , E.Nichols . *Named Entity Recognition with Bidirectional LSTM-CNNs.* arXiv preprint arXiv:1906.09436, 2016.