Assignment 1

Seyedali Shohadaeolhosseini
Master's Degree in Artificial Intelligence, University of Bologna
s.shohadaeolhosseini@studio.unibo.it

Abstract

Today, as some datasets and corpora are being gathered basically based on crawling data from the Internet (Wikipedia, 2020; Kili-Technology), the necessity of supervision of these data, most of which are user-generated contents and convey different perspectives as well as sexism speaking, is more important than any other time. This project conducted a detailed exploration on the problem of classifying text as either sexism or not. This project introduces a BLSTM model capable of classifying a given unseen tweet taken from the test set as sexism with an F1 score of 0.72. Additionally, we have tested the test set on the RoBERTa transformer and it achieved the score of 0.83 on the test set.

1 Introduction

Internet is known as a place where user's can talk openly. When you crawl the Internet to build corpora, you are crawling data of people, experts, researchers with different thinking, talking, and believes. In these crawled data, like other different points of views, we also have the sexism point of views which if we don't take care of them carefully before training models, will lead us to model that also gives some probability to texts and words in these context.

The sexism detection is not a new problem. (Jhakal et al., 2023) have worked on the same topic. They tested a variety of pre-trained transformers including RoBERTa, BERT, BERTweet and XLNet.

In this research, we will study the effectiveness of using a Bidirectional Long Short-Term Memory (BLSTM) model. We will also, use the pre-trained RoBERTa transformer to analysis how well it is performing in this problem, compared to the BLSTM.

We have tested the aforementioned models on the (EXIST) dataset.

2 System description

In this project, we have developed two BLSTM models. One, which is the based model, has one layer of a BLSTM with 64 units. The other model has one more BLSTM layer with 32 units. Moreover, to check the ability of transformer on this task, we have checked the pre-trained RoBERTa transformer.
Our BLSTM base model starts with an embedding layer, then it proceeds with a Bi-directional LSTM layer with 64 units, then it predicts the output by sending the output of this layer to a one-unit dense layer which uses the sigmoid as activation function. Our second BLSTM model only has one more BLSTM layer with 32 units immediately after the 64 unit BLSTM.
The embedding layer uses the 6B-100D version of GloVe embeddings which is freezed and is not being trained. Batch size is set to 64 and the maximum length of the input is set to a length of sequences with more than 90 percent occurrence. We have chosen word tokenization approach and the vocabulary is created on the union of all the token type exists in the training data and GloVe embeddings.

3 Data

In this project we will use the (EXIST) which is the same dataset used by (Jhakal et al., 2023). However, different than them, who used the entire dataset for the sexism classification task, we have filtered our dataset to contain only English texts. Moreover, the Original dataset contains other columns like whether the tweet is reported Judgmental or not, which are not our interest in this project. As the results, we only kept the "id EXIST", "lang", "tweet", "labels task1" columns.
On the other hand, the column "labels task1" has a list of True and False, which is the opinions of the annotators on each tweet whether it is sexist

or not. We have changed this column and choose the label of that tweet based on majority approach. In continue, since the dataset is not balanced and the number of non-sexism tweets is more, we have created another training data from this unbalanced data that is balanced. To balanced the data, we have down-sampled the majority label, non-sexism class is down-sampled. This balancing helps the model to not get bias on any class.
In addition to these pre-settings of the dataset, as the data cleaning step, using the regex tool, we have cleaned the tweets from having emojis, hashtags, mentions, URLs, special characters and symbols. At the end, we finished the data processing part by applying a lemmatization function to each tweet.

4 Experimental setup and results

As we have two training data, balanced and unbalanced, and two BLSTM models, we have for models to train. Moreover, for the seek of rubostness, we have choosen three seeds. Hence, in the training section of BLSTM we will perform 12 training each has 15 epochs. The table 1 shows metrics of our four models, as we have two different architecture with two training data. The numbers are an average on the three seeds that we use to train each model. Moreover, the table 2 shows the results of the transformer on the defined two datasets. The results are averaged on the seeds.

Metrics	Val F1-Score	Test F1-Score
Model B	0.7630	0.6929
Model BA	0.7629	0.6894
Model U	0.7647	0.7172
Model UA	0.7603	0.7173

Table 1: B corresponds to usage of balance data for training. U corresponds to usage of unbalanced data for training. Models B/U are has base architecture, and Models BA/UA is the model with additional BLSTM layer.

Metrics	Test F1-Score
Balanced Data	0.8312
Unbalanced Data	0.8328

Table 2: These are the scores that transformer achieved on the test sets.

5 Discussion

Shown diagrams in the notebook demonstrates that training models more than 7, 8 epochs leads to over-fitting. Moreover, the over-fitting and divergence happens faster in the second model with the additional BLSTM layer. This is because of the less amount of data that we have which provides less information to learn for the model.

The models trained on the unbalanced data, were significantly performing better prediction for the class 0, compared to the prediction of class 1. Which is due to the availability of more data of class 0. As mentioned before, the performance of model with additional BLSTM layer was poorer on predicting the class 0. This is because, for the same amount of data, the model is being trained more, in other words, a greater network is being used which helps the model to converge on the data faster and overfit. This could be tackled by adding Dropout layers or decreasing the number of epoch or providing more data for the model. Other techniques such as up sampling the class with fewer data instances can also be a good strategy.

On the other side, the models trained on the balance dataset, unexpectedly were performing better on the class 0. However, they weren't as powerful as models trained on the unbalance dataset in predicting the class 0. The distance of correct prediction of the class 0 and class 1 is less here for these models which is reasonable. These models were performing better prediction than the previous two models on predicting the class 1. Which is because the model does not have any bias on other class. Like what happened in the previous two models' total predictions, here also the model with additional LSTM layer is performing a poorer prediction on predicting the class 0.

6 Conclusion

Quality and quantitative defines a models abilities in working better. An state-of-the-art model architecture cannot solely change everything. It is the data matters the most. In such projects, our aim must be on improving the data. If we have less data, then we need to synthesize some or choose a simpler datasets.

References

Clef EXIST. Exist: sexism identification in social networks. Webpage. Accessed: January 9, 2025.

Chirayu Jhakal, Khushi Singal, Manan Suri, Divya Chaudhary, Bijendra Kumar, and Ian Gorton. 2023. Detection of sexism on social media with multiple simple transformers. In *CLEF 2023: Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece. Chirayu Jhakal and Khushi Singal are Corresponding authors.

Kili-Technology. Open-sourced training datasets for large language models. Webpage. Accessed: January 9, 2025.

Wikipedia. 2020. Gpt-3. Webpage. Accessed: January 9, 2025.