BERT-BiLSTM model

Swetha Vemulapalli, Pallavi Kailas, Krish Didwania

August 2023

1 Abstract

This paper attempts at humour classification based on the the SemEval-2021 Task-7 as attempted by Team KGP. The data-set runs through a 12 layer pre-trained Bert model and a BiLSTM layer. The outputs of the BERT layer and the BiLSTM layer both undergo fusion before being passed through 3 linear layers.

You can find the repository for the implementation on GitHub:
https://github.com/pallavikailas/finaltask.

2 Introduction

Automatic offense and humour detection has notable significance in chatbots and virtual assistants. But identifying them within texts represents a considerable challenge as it is very significantly subjective and depends on various factors such as age,gender,socio-economic status,race,religion,and more. In the realm of Natural Language Processing, it is therefore considered a very complex field to tackle.

This paper is concerned with humor detection (binary classification) primarily. The first sub task is to check if a text is humorous.We also have to predict whether the text is generally offensive (binary classification task).The dataset consists entirely of English texts.

The traditional methods deployed earlier for humor detection include Support Vector Machine(SVM) with RBF kernel, Random Forest Classifier, and SGD with Logical Classifier, all of which provide modest results. Recently more state-of-the-art transformers have provided better results. The system presented in this paper is finetuning one of the best and most popular state-of-the art models, Bidirectional Encoder Representations from Transformers . We have used pretrained BERT embeddings to represent the words and used the features of the last layer of the 12- layers BERT Model to detect both humor and offense. BERT can model complex interactions between different levels of hierarchical information.

3 Dataset

The datasets used in this paper are from Task 7 of SemEval 2021. In this task, the organizers collected labels and rating from a balanced set of age groups from 18 to 70. The annotators also represented a variety of genders, political stances, and income levels. The training set consists of 8000 texts while the evaluation and testing sets contain 1000 texts each. Each text is represented by a unique ID and each subtask has a separate column for each text. The annotators were asked:

1. Was the intention of this text to be humorous? (0 or 1)

2. If it is intended to be humorous according to the rater, how humorous was it considered? (1-5)

In cases of a tie, the class was given the value 1 while the humor rating label was given the average rating. The annotators were also asked:

3. Is this text generally offensive? (0 or 1)

4. If the rater found it to be offensive, how offensive was it? (1-5)

In our model, we focused primarily on classification based on the questions 1 and 3.

4 Experimental Setup and Evaluation Metrics

Preprocessing: The train, evaluation and test datasets were preprocessed to remove irrelevant features, stop-words or emojis (if any), and the output of the pipeline is lower-case stemmed word sequences. Any NaN values were replaced by 0 or 1 randomly.

Adam Optimizer: Adam combines the best properties of the AdaGrad and RMSProp algorithms to provide an optimization algorithm than can handle sparse gradients on noisy problems and works well on deep neural networks.

Experimental Tools: We used Kaggle to run the experiments as it offers GPU time to run our model. We also used transformers from the PyTorch HuggingFace API as well as Scikit-learn package.

Evaluation Metric: The metric we used to evaluate our classification model was F1 score.

$$F1 = \frac{2.Precision.Recall}{Precision + Recall}$$

5 System

BERT has proven promising for many NLP tasks. Our system implements fine tuning strategies on pre-trained BERT architecture. It is a bidirectional transformer pre-trained using a combination of masked language modeling objective and next sentence prediction on a large corpus comprising the Toronto Book Corpus and Wikipedia.

BERT base model:

We have experimented with the BERT-base model from PyTorch Hugging Face API. It is the bare BERT Model (BertModel) transformer outputting raw hidden-states without any specific head on top. The 768 hidden features are extracted from the last layer of the 12-layered BertModel.

6 Classification Task

In our model,we have used a BiLSTM approach for the classification tasks.We have experimented with the BERT-base model from PyTorch HuggingFace API. It is the bare [**?**] BERT Model (BertModel) transformer outputting raw hidden-states without any specific head on top. The 768 hidden features are extracted from the last layer of the 12-layered BertModel..This vector is passed through a BiLSTM layer whose output size is set to 256. Futhermore,we use a linear layer after the BERT output and a linear layer after the BiLSTM output and concatenate the results to form a vector size of 12.Then,we use a linear layer to reduce the dimension to 1 which can be used as classification target.The output layer is a softmax layer for the classification job (for both approaches). The reason behind experimenting with the Bi-LSTMmodel is that it fully considers the context information and can better obtain the text representation of the comments .

7 Results and Analysis

For each subtask, we trained our data on the entire training set of 8000 texts. We used the development set of 1000 texts as cross-validation data. The results tabulated in this section are reported on the gold test set of 1000 texts. Each proposed model for all the subtasks was run for 10 epochs with a batch size of 32 and use a constant learning rate of 1e-5.

	Metrics	
	Accuracy	**F1-Score**
is_humour	0.89	0.91
humor_controversy	0.52	0.54

Table 1: Accuracy and F1-Score (Original Implementation)

	Metrics	
	Accuracy	**F1-Score**
is_humour	0.68	0.55
humor_controversy	0.5	0.45

Table 2: Accuracy and F1-Score

8 Conclusion

In this paper, we describe a system developed to address the SemEval Task-7. The task has four subtasks, viz., detecting humor, detecting offense, predicting humor rating, and predicting offense rating. Our system is able to perform humor and offense detection quite well with the help of BERT-BiLSTM model.

Our system deploys linear layers and bi-LSTM layers independently to process the features produced by the BERT model. Based on observations, the train(F1 score of 0.95 and 0.90) and validation (F1 score of 0.91 and 0.51) exceeded the test F1 scores which indicates that the model may tend to overfit.In the original implementation,they got F1 scores of 0.91 and 0.54 in is humor and humor controversy respectively compared to our work of 0.55 and 0.45.