TECHSSN at SemEval-2021 Task 7: Humor and Offense detection and classification using ColBERT embeddings

Rajalakshmi Sivanaiah, Angel Deborah S, S Milton Rajendram, Mirnalinee T T,
Abrit Pal Singh, Aviansh Gupta, Ayush Nanda
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Chennai 603 110, Tamil Nadu, India
{rajalakshmis, angeldeborahs, miltonrs, mirnalineett}@ssn.edu.in
{abritpal18007, aviansh18028, ayush18031}@cse.ssn.edu.in

Abstract

This paper describes the system used for detecting humor in text. The system developed by the team TECHSSN uses binary classification techniques to classify the text. The data undergoes preprocessing and is given to ColBERT (Contextualized Late Interaction over BERT), a modification of Bidirectional Encoder Representations from Transformers (BERT). The model is re-trained and the weights are learned for the dataset. This system was developed for the task 7 of the competition, SemEval 2021.

1 Introduction

Natural language processing faces the challenges working with humor as it is a highly subjective phenomena and the age, gender and socio-economic status are known to have an impact on the perception of the joke. It usually involves multiple word senses and cultural knowledge to appreciate humor to its best. Now a days chatbots and virtual assistants require automated humor detection systems for a better interaction with the user by understanding what a human-like approach to humor is. It is crucial to understand the real motive of the user and provide appropriate answer to have a better experience of the user with the virtual assistants (Chen and Soo, 2018). Based on the general linguistic structure of humor, we propose an approach for detecting humor in short texts using ColBERT in this paper.

We have developed a system in the name of TechSSN for previous SemEval tasks (Sivanaiah et al., 2020; Logesh et al., 2019) for offensive language detection using various machine learning and deep learning networks. In SemEval 2021, we participated in subtask-1a and 1c of humor and offensiveness detection in task 7- HaHackathon (Meaney et al., 2021).

2 Related Work

Continuous research is going on in this field of humor detection and the systems are getting better every year. De Oliveira and Rodrigo (2015) developed a model for humor detection in Yelp reviews and used convolutional networks with a maximum of 81.57% accuracy.

The system developed by Ortega-Bueno et al. (2018) used UO_UPV, a Attention-based Long Short-Term Memory Network. The model consists of a Bidirectional LSTM neural network with an attention mechanism that allows to estimate the importance of each word and then, this context vector is used with another LSTM model to estimate whether the tweet is humorous or not. The F1 score for this system is approximately 0.78 with accuracy, 0.84.

Mao and Liu (2019) developed a system with BERT, a multi-layer bidirectional transformer encoder which can help to learn deep bidirectional representations, and the pretrained model is fine-tuned on training data. Their best F1 Score on the test set is 0.784.

Risch et al. (2020) explained the need and various methods used for offensive language detection. BERT model is used with transfer learning for the offensiveness detection by (Liu et al., 2019) with F1 score as 0.8286 and accuracy as 0.8628.

3 Methodology

ColBERT base model is chosen for humor text classification which has 8 layers with the last layer's activation function as the sigmoid func-

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1185–1189
Bangkok, Thailand (online), August 5–6, 2021. ©2021 Association for Computational Linguistics

tion, as it is performing binary classification. The remaining layers have ReLu activation function.

3.1 Model Architecture

This classification model uses a separate line of hidden layers especially designed to extract features from each sentence. The used model is a neural network that includes two parallel lines of hidden layers: One to view text as a whole and another to view each sentence separately. Figure 1 displays the architecture of the proposed method.

First, to assess each sentence separately and extract numerical features, the sentences are separated and are tokenized individually. To prepare these textual parts as proper numerical inputs for the neural network, they are encoded using BERT sentence embedding (Devlin et al., 2018). This step is performed individually on each sentence and also on the whole text (shown in Figure 1). As we get the BERT sentence embedding for each sentence, they are fed into the parallel hidden layers of the neural network to extract mid-level features for each sentence (could be related to context, type of sentence, etc). The vector size obtained from this layer for each sentence is 20.

While the main idea is to detect relationships between sentences (especially with punchline), it is also essential to find out the word-level connections in the whole text (such as synonyms and antonyms). Identifying this relation will have meaningful impacts in determining congruity of the text. Similar to the previous step, we feed BERT sentence embedding for the whole text into hidden layers of the neural network. The vector size is 60. Finally, there are three sequential layers in the neural network model. These final layers combine the output of all previous lines of hidden layers in order to return the final output. These final layers are used to determine the congruity of sentences and detect the transformation of reader's viewpoint after reading the punchline.

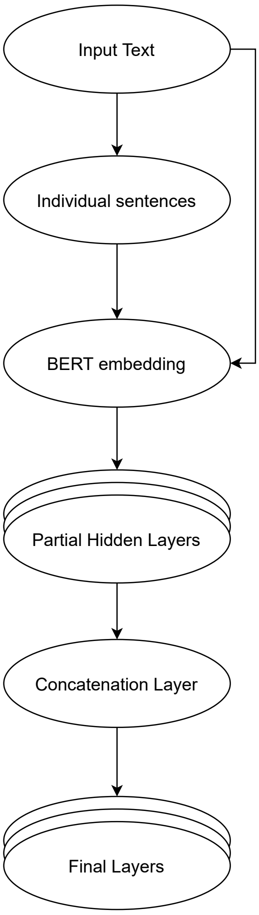

Figure 1: Model Architecture

3.2 Dataset Collection

For building the model we have used the training dataset provided by the organizers of the Hahackathon - (Meaney et al., 2021). This dataset has 8000 instances of which 4932 belong to humor class and 3068 belong to non-humor class. Out of the 4932 humor instances, 2465 are controversial and 2467 are not controversial. Each instance in the dataset has four features: is_humor, humor_rating, humor_controversy and offense_rating. We have trained our model for the classification features only; is_humor and humor_controversy. Both features have binary values.

3.3 Data Preprocessing and Tokenization

For data preprocessing the function *convert_to_transformer_inputs* is used to convert the tokenized input into ids, masks and segments for the transformer. The tokenization of the dataframe columns is done by the function *compute_input_arrays*. BERT Tokenizer pre trained on the 'BERT-base-uncased' model is used for tokenization. The maximum sequence length for reading the data is set as 200 that will be used as the input to BERT.

3.4 Model Creation

For the model we have used BERT-base. We use a function *create_model* that has the architecture which is used to fine tune BERT to our chosen dataset, and we compute the competition metric for the validation set with the help of the function spearman rank correlation coefficient.

3.5 Training, Cross Validation and Testing

The model is trained with cross validation for 3 epochs with a learning rate of 3e-5 and the size of each batch is 6. As we have performed binary classification for the humor detection task, we have set the loss function as a simple binary crossentropy.

4 Results and Discussion

The test dataset given by SemEval organisers (gold-test-27446.csv) was tested on different models and the F1 score for all these models are discussed in the following sections.

4.1 ColBERT

We have used Contextualized Late Interaction over BERT (ColBERT) (Khattab and Zaharia, 2020). ColBERT differs by providing a late interaction architecture that independently encodes the query and the document using BERT. It then uses a powerful interaction step that analyze their similarity in a fine grained mode. Eventhough the interaction learning is delayed it also maintains this fine-granular interaction in a better manner. ColBERT can leverage the expressiveness of deep language models and simultaneously gaining the ability to precompute document representations in an offline structure. This will speed up query processing since the representations are learnt in offline. Beyond reducing the cost of re-ranking the documents retrieved by a traditional model, ColBERT's pruning-friendly interaction mechanism enables leveraging vector-similarity indexes for end-to-end retrieval directly from a large document collection.

4.2 Decision Tree

Decision tree (DT) is a popular supervised technique used for classification problems. It identifies the relation between the features and form the set of rules that can be used to classify the given data into any one of the class labels. The method uses the train dataset to generate branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes.

4.3 SVM

Support Vector Machine (SVM) is a supervised machine learning technique used for both classification or regression problems. It uses the concepts of separating the classes using maximal margin hyperplane. We fed preprocessed data into Support Vector Classifier (SVC) with Gaussian Radial Basis (RBF) kernel for training and testing.

Table 1 shows the F1 score and accuracy for the best, baseline and our approach.

No.	System	F1	Accuracy
1	Best Approach	0.982	0.9854
2	Baseline	0.857	0.848
3	Our Approach	0.884	0.9081

Table 1: Official results of the humor detection task

Figure 2 shows the F1 score and the accuracy for the various models we tested for humor detection. ColBERT model provides better accuracy when compared to decision tree and support vector machine models.

No.	Model	F1	Accuracy
1	ColBERT	0.884	0.9081
2	SVM	0.747	0.617
3	Decision Tree	0.736	0.62

Table 2: Results for various models used for humor detection

The humor testset contains 1000 instances of which 615 are humorous and 385 are not. The confusion matrix for ColBERT model is given in Table 3.

		Actual	
		Humor	Not
Predicted	Humor	553	30
	Not	62	355
	Total	615	385

Table 3: Confusion matrix of ColBERT for humor detection

There are 279 offensive texts and 336 non offensive texts in 615 humor instances. Table 4 shows the results for ColBERT and SVM model used for humor controversy or offensiveness detection.

No.	**Model**	**F1**	**Accuracy**
1	ColBERT	0.53	0.56
2	SVM	0.487	0.530

Table 4: Results for various models used for humor controversy detection

5 Conclusion

There is a lot of demand for automatic humor detecting systems in the market as it can be used in chatbots and AI assistants to achieve human-like experience while talking to a machine. SemEval-2021 task 7 involves a subtask-1a as identifying humor in text. A modification of the BERT called ColBERT is used to classify such text sentences into humorous or not. ColBERT is a 8-layer model with 110M parameters outperforms the machine learning models we tested with a large margin, showing the importance of utilizing linguistic structure. The preprocessing techniques can be enhanced to getter better accuracy.

References

Peng-Yu Chen and Von-Wun Soo. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 113–117.

Luke De Oliveira and Alfredo Láinez Rodrigo. 2015. Humor detection in yelp reviews. *Retrieved on December*, 15:2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xiaochao Fan, Hongfei Lin, Liang Yang, Yufeng Diao, Chen Shen, Yonghe Chu, and Yanbo Zou. 2020. Humor detection via an internal and external neural network. *Neurocomputing*, 394:105–111.

Ashraf Kamal and Muhammad Abulaish. 2019. Self-deprecating humor detection: A machine learning approach. In *International Conference of the Pacific Association for Computational Linguistics*, pages 483–494. Springer.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 87–91.

B Logesh, S Harshini, B Geetika, S Dyaneswaran, S Rajalakshmi, Angel Suseelan, S Milton Rajendram, and TT Mirnalinee. 2019. TECHSSN at SemEval-2019 Task 6: Identifying and categorizing offensive language in tweets using deep neural networks. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 753–758.

Jihang Mao and Wanli Liu. 2019. A BERT-based approach for automatic humor detection and scoring. In *IberLEF@ SEPLN*, pages 197–202.

J.A. Meaney, Steven R. Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 2021. SemEval 2021 Task 7, HaHackathon, Detecting and Rating Humor and Offense. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Reynier Ortega-Bueno, Carlos E Muniz-Cuza, José E Medina Pagola, and Paolo Rosso. 2018. UO UPV: Deep linguistic humor detection in spanish social media. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, pages 204–213.

Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 137–143.

Sushmitha Reddy Sane, Suraj Tripathi, Koushik Reddy Sane, and Radhika Mamidi. 2019. Deep learning techniques for humor detection in Hindi-English code-mixed tweets. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 57–61.

Rajalakshmi Sivanaiah, Angel Suseelan, S Milton Rajendram, and Mirnalinee TT. 2020. TECHSSN at SemEval-2020 Task 12: Offensive language detection using BERT embeddings. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2190–2196.