CSECU-DSG at SemEval-2021 Task 7: Detecting and Rating Humor and Offense Employing Transformers

Afrin Sultana, Nabila Ayman, and Abu Nowshed Chy
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{afrin.sultana.cu, nabila.ayman.cu}@gmail.com,
and nowshed@cu.ac.bd

Abstract

With the emerging trends of using online platforms, peoples are increasingly interested in express their opinion through humorous texts. Identifying and rating humorous texts poses unique challenges to NLP due to subjective phenomena i.e. humor may vary to gender, profession, age, and classes of people. Besides, words with multiple senses, cultural domain, and pragmatic competence also need to be considered. A humorous text may be offensive to others. To address these challenges SemEval-2021 introduced a HaHackathon task focusing on detecting and rating humorous and offensive texts. This paper describes our participation in this task. We employed a stacked embedding and fine-tuned transformer models based classification and regression approach from the features from GPT-2 medium, BERT, and RoBERTa transformer models. Besides, we utilized the fine-tuned BERT and RoBERTa models to examine the performances. Our method achieved competitive performances in this task.

Keywords: humor and offense rating, transformers, BERT, RoBERTa, stacked embedding.

1 Introduction

The exponential evolution of technologies and social platforms increases the growth of user-generated content. Therefore, detecting specific information from a pile of web data is ubiquitous. Humor, like most figurative language, poses interesting linguistic challenges to NLP, due to its emphasis on multiple word senses, cultural knowledge, and pragmatic competence. The automated humor detection and rating is one of the challenging and promising tasks for their significance in the field of opinion mining, sentiment analysis, and emotion intelligence domain.

Numerous works have been done on humorous text identification task. (Weller and Seppi, 2019) proposed pre-trained BERT architecture to identify humorous jokes from Reddit dataset, puns, and short jokes dataset. (Khatri and Pranav, 2020) used BERT and GloVe embeddings with linear support vector classifier (SVC), naive Bayes, and random forest for predicting the final label to detect sarcasm in tweets. (Badlani et al., 2019) extracted features pertaining to sarcasm, humor, hate speech, as well as sentiment and ensemble them for sentiment classification. In (Liang et al., 2020), the pre-trained BERT is adapted to the distantly supervised NER (named entity recognition) task with early stopping. Some evaluation campaigns related to automatic humor detection also has been performed. (Swamy et al., 2020) includes a logistic regression baseline, a BiLSTM + attention-based learner, and a transfer learning approach with BERT to tackle the Internet humor at SemEval-2020 Task 8.

However, most of the existing work addressed the humorous text identification problem as a binary classification task. But the humorous rating of a text can be varied at different scales based on its contents. Moreover, a humorous text may portray offensive context to other users due to a variety of users' perceptions. To bridge this gap (Meaney et al., 2021) introduced a shared task at SemEval-2021 focusing on detecting and rating humorous and offensive texts. The task comprises of four subtasks: Task 1a is to predict whether a text is humorous or not, task 1b is to rate a humorous text with a score between 0 and 5, task 1c represents the subjectivity of humor appreciation by foretelling a humorous text is controversial or not, and task 2 intend to rate a text with a value between 0 and 5 based on its offensiveness.

To tackle the challenges of this task, we employed a stacked embedding of various transformer models including GPT-2 medium, BERT, and

The first two authors have equal contributions.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1204–1208
Bangkok, Thailand (online), August 5–6, 2021.

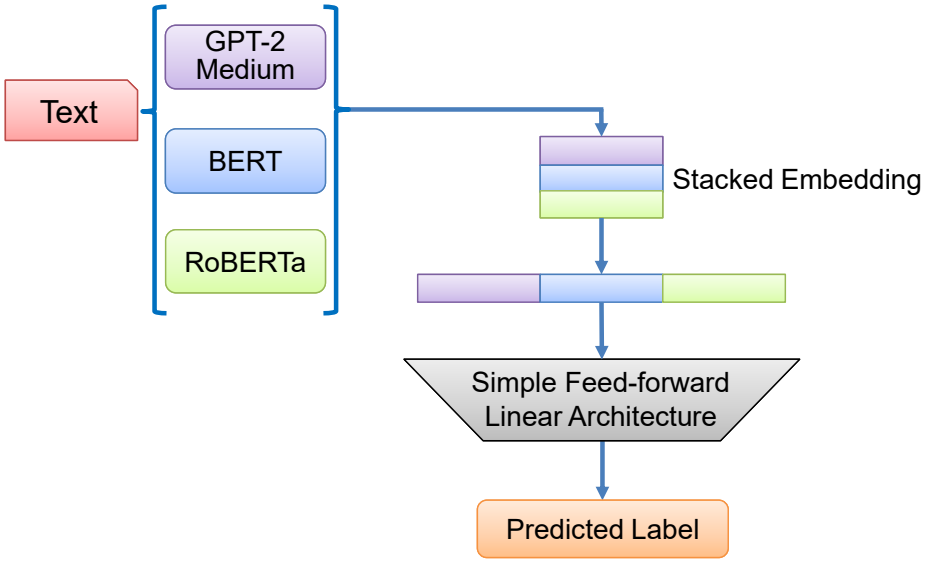

Figure 1: Proposed framework.

RoBERTa with a simple linear architecture. Besides, we conduct experiments and examine the individual performance of these three models. We used transformer-based models as they can learn the context of the sentence effectively and pushed the state-of-the-art for a wide range of downstream NLP tasks.

The rest of the paper is structured as follows: Section 2 describes our proposed framework. Section 3 illustrates our experiment and evaluation. Finally, we come to end with some conclusion and future research directions in Section 4.

2 Proposed Framework

In this section, we present the details of our proposed approach for humorous and offensive text identification and rating. The overview of our proposed framework is depicted in Figure 1.

Given a text, we have extracted embedding features from three state-of-the-art transformer-based models including GPT-2 medium, BERT, and RoBERTa. The extracted embeddings are then unified through the stacked embedding scheme and the unified feature vector is then passed to the simple feed-forward linear architecture to obtain the prediction score.

2.1 Text Encoding

GPT-2: GPT-2 (Radford et al., 2019) stands for generative pretrained transformer 2, which is trained on eight million text documents scraped from the web. It has an outstanding ability to generate coherent text from minimal prompts. We utilize GPT-2 medium version to get a 1024 dimensional feature vector from each text for sub-task 1a.

BERT: BERT (Devlin et al., 2019) stands for bidirectional encoder representations from transformers, which is a new method of pre-training sentence representations. It is trained on a large corpus of unlabelled text which includes the entire Wikipedia (that's about 2500 million words) and a book corpus (800 million words). We deploy BERT-base uncased version with fine-tuning to get a 768-dimensional feature vector for a given text. We also conduct experiments with the fine-tuned BERT-large uncased architecture to encode each text into a 1024-dimensional feature vector.

RoBERTa: RoBERTa (Liu et al., 2019) stands for robustly optimized BERT pre-training approach. An improvement on BERT which introduced as a robustly optimized method for pre-training NLP systems. BERT's language masking strategy is used in RoBERTa, wherein the system learns to predict intentionally hidden sections of text within otherwise unannotated language examples. RoBERTa modifies key hyperparameters in BERT including removal of BERT's next-sentence pre-training objective, training with much larger mini-batches, and learning rates. This allows RoBERTa to improve on the masked language modeling objective compared with BERT and leads to better performance on downstream tasks. We use the RoBERTa base model with fine-tuning to get a 768-dimensional feature vector.

2.2 Stacked Embeddings

Stacked embeddings are one of the most important concepts that combine different embeddings to capture the benefits of different embedding models. Stacked embeddings concatenate embedding vectors generated from various models to form the final word vectors. The representation of stacked embedding-based framework is depicted in Figure 1. We combine GPT-2 medium, BERT-base, and RoBERTa-base embeddings using Flair's (Akbik et al., 2018) stacked embeddings approach. For each text, we extract a 1024-dimensional transfer learning feature vector from the GPT-2 medium, a 768-dimensional feature vector from the BERT-base uncased, a 768-dimensional feature vector from the RoBERTa-base model. We combine these embeddings and get a 2560-dimensional feature vector for each text.

2.3 Classification Module

The transfer learning feature vector obtained from the stacked embedding approach goes through a simple feed-forward linear architecture which applies a linear transformation to the incoming feature vectors and classifies each text either as humorous or non-humorous. Besides, we employ the fine-tuned BERT and RoBERTa based classifiers.

For the regression tasks, fine-tuned BERT and RoBERTa based regression models are utilized. We implement these models from the Huggingface transformers library (Wolf et al., 2020).

3 Experiments and Evaluations

3.1 Dataset Description

SemEval-2021 task 7 organizers (Meaney et al., 2021) provided a benchmark dataset to evaluate the performance of the participant's proposed systems. The given training set consists of 8000 texts whereas the development set contains 1000 texts and the test set contains 1000 texts. The data set contains one row for each text. Each row has a unique identifier, the text, and the values for is_humor, humor_rating, humor_controversy, and offense_rating. If the sentence is humorous the value of is_humor is 1 otherwise 0. The value of humor_rating and offense_rating defines the level of humor or offensiveness in the contents. The value of humor_controversy determines whether the humorous text is controversial or not. The value of humor_rating and humor_controversy is subject to the value of is_humor. If a given text is not humorous, humor_rating and humor_controversy contain no value for that particular text. Therefore, we converted all empty cells of humor_rating and humor_controversy to label 0. We used the train data set to train our model. The development data was used for hyperparameter tuning. Finally, we used the given test dataset to evaluate our models.

Parameters List	Sub-task 1a	Other sub-tasks
Epochs	4	20
Batch size	16	16
Learning rate	3e-5	2e-6
Maximum length	default	160
Patience	3	5
Optimizer	Adam	AdamW
Anneal factor	0.5	–

Table 1: Optimal value of parameters used in this work.

3.2 Evaluation Measures

The organizers employed different strategies to evaluate the performance of participants' systems. Standard evaluation measures including F1-score and accuracy were applied to estimate the performance for sub-tasks 1a and 1c. The regression subtasks i.e. 1b and 2a were evaluated by the metric root mean squared error (RMSE).

3.3 Experimental and Parameter Settings

We used Google Colab's GPU for training and parameter tuning of our system. We evaluated our system's performances through Codalab platform.

Now, we describe the value of optimal parameters that we've used to design our model. While evaluating for the sub-tasks, we fine-tuned our systems based on the number of epochs, batch size, learning rate, anneal factor, patience, and optimizer to obtain the improved performances. We utilize the early-stopping method to overcome the system's overfitting on the training dataset. We interchange epochs in range [4,8,15,20] with different value of patience, batch size in range [8,16,32], learning rate in range [2e-5, 3e-5, 4e-5, 2e-6, 3e-6, 4e-6] with optimum max_length. While evaluating sub-task 1a, we use Flair's (Akbik et al., 2018) framework and for other sub-tasks, we utilize transformers from HuggingFace transformers library (Wolf et al., 2020). We note that compared to BERT, RoBERTa has slightly different hyper-

parameters. In particular, RoBERTa uses weight decay with $\lambda = 0.1$ and no gradient clipping (Mosbach et al., 2020). Table 1 describes the summarized parameters settings used in this work. The default settings are used for the rest of the parameters. In this paper, we reported the results based on these settings. However, if the prediction for the rating sub-tasks is less than 0.35, the value is converted into 0.0 according to our result analysis on the development set.

3.4 Results and Analysis

We used the full training dataset for training our proposed model and the validation set for hyperparameter tuning. The comparative performance of our method based on test data against the top-ranked participants' systems in individual sub-task are presented in Table 2 and 3, respectively.

Team (Rank)	Accuracy	F1-Score
Comparative Performance on Subtask-1a		
PALI (1st)	0.9820	0.9854
stce (2nd)	0.9750	0.9797
CSECU-DSG (34th)	**0.9380**	**0.9496**
mayukh (35th)	0.9330	0.9468
Avilshmam (56th)	0.816	0.8489
Comparative Performance on Subtask-1c		
PALI (1st)	0.4943	0.6302
reynier (10th)	0.4732	0.6197
CSECU-DSG (34th)	**0.5366**	**0.4423**
GuanZhengyi (35th)	0.5593	0.4271

Table 2: Comparative results with other participants' systems in binary classification tasks.

We now report the best performing model for each sub-task. Stacked embeddings based method used for sub-task 1a and BERT's large model is used for sub-task 1c. However, for the regression subtasks 1b and 2a, BERT's large and RoBERTa's base models are employed, respectively.

Results showed that our proposed system obtained competitive results in sub-task 1a and 1c whereas in the other sub-tasks our system lags behind. Ensembling transformer's embedding achieved good performance for binary classification. In our system, we have tuned a few hyperparameters and used features from all intermediate layers of transformers. We didn't apply any approach to find out effective intermediate layers combination to obtain the best performance from transformers. Though the training set contains 8000 sentences, only 4932 sentences are humorous. That's why we get only 4932 perfect humor-rated sentences to train the system which is scanty. All these requirements limit our system performance for regression tasks 1b and 2a.

Team (Rank)	RMSE Score ↓
Comparative Performance on Subtask-1b	
abcbpc (1st)	0.4959
mayukh (4th)	0.5257
CSECU-DSG (41th)	**0.6803**
Maoqin (44th)	0.7405
JAGD (47th)	0.8847
Comparative Performance on Subtask-2a	
DeepBlueAI (1st)	0.4120
MagicPai (8th)	0.4460
CSECU-DSG (34th)	**0.5395**
Anik (38th)	0.5800
MLXG (47th)	0.9587

Table 3: Comparative results with other participants' systems in regression tasks.

4 Conclusion and Future Plan

In this paper, we have tackled the problem of identifying and grading humorous texts as defined in SemEval-2021 task 7. Achieving high performance in humor and offensive text identification or ranking is hard due to its diverse contextual form. We have presented transformer-based language models including GPT-2 medium, BERT, and RoBERTa in a unified architecture using a stacked embedding scheme. Experimental results demonstrated its effectiveness for the classification tasks. However, we have seen that finetuned transformer models performed better in the regression task.

In the future, we have a plan to focus on direct inducing topic information into the transformer-based models. We also intend to explore how to tune the parameters for regression tasks in more efficient ways, which could yield better performances.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Rohan Badlani, Nishit Asnani, and Manan Rai. 2019. Disambiguating Sentiment: An Ensemble of Humour, Sarcasm, and Hate Speech Features for Sentiment Classification. *W-NUT 2019*, page 337.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, pages 4171–4186.

Akshay Khatri and P Pranav. 2020. Sarcasm Detection in Tweets with BERT and GloVe Embeddings. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 56–60.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1054–1064.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

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*.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. *arXiv preprint arXiv:2006.04884*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Steve Durairaj Swamy, Shubham Laddha, Basil Abdussalam, Debayan Datta, and Anupam Jamatia. 2020. NIT-Agartala-NLP-Team at SemEval-2020 Task 8: Building Multimodal Classifiers to Tackle Internet Humor. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1179–1189, Barcelona (online). International Committee for Computational Linguistics.

Orion Weller and Kevin Seppi. 2019. Humor Detection: A Transformer Gets the Last Laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3612–3616.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.