SemEval-2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense

2209678 NLP Systems Final Project

Krittin Leewanich
Kuntida Kongkad
Chidapha Phongkhahabodi
Chulalongkorn University
{6534402523, 6534407723, 6534408323}@student.chula.ac.th

Abstract

This project involves training a model to detect and score humor and abusive content. We divided the work into three subtasks: Task 1a, 1b, and 1c. The goal is to predict whether a given text is considered humorous. If it is, the system should then predict how funny it is and whether the humor score is likely to cause controversy. The system was developed using a pre-trained RoBERTa model. We also used other models for comparison. Our model completed the task 1a with F1-score at 0.9653 using ROBERTA-BASE model. Task 1b with RMSE of 0.5490 and task 1c scoring the F1-score at 0.6453.

1 Introduction

In our daily lives, we face increasing challenges in dealing with sarcasm, bullying, and various forms of abuse. Sarcasm and bullying, complex and challenging topics for major companies and institutions, are increasingly prevalent. Artificial intelligence and text processing techniques effectively detect these problems in texts and images.
Sarcasm and abuse involve attacking individuals or groups through unintended jokes or directly affecting their psychological well-being. Irony and offensiveness use humor to conceal the opposite meaning.

The SemEval-2021 "HaHackathon: Detecting and Rating Humor and Offense" task presents two main tasks: task-1 with three sub-tasks 1a, 1b, and 1c. The objective is to predict whether a text is humorous and, if so, estimate its humor rating and determine if it is perceived as controversial. Task-2 aims to predict how users perceive a text as offensive. Our solution, has achieved interesting results. The approach utilizes a dataset containing 10,000 rows of text data

2 Related works

Early humor and offensiveness detection relied on handcrafted features—lexical cues, punctuation, and user metadata Deep learning models later improved performance by learning text representations. Transformer-based fine-tuning set new benchmarks applied BERT and XLNet to sarcasm and humor tasks, while various BERT variants tackled hate speech with imbalance-aware losses.

At SemEval-2021 Task 7, the SarcasmDet team models on the HaHackathon dataset across four subtasks (binary classification and regression). Their RoBERTa-large ranked 4th in humor detection, 10th in humor rating, 3rd in controversy detection, and 10th in offensiveness regression.

3 Our Approach

We aim to explore how advanced pretrained language models and task-specific strategies can be effectively used to detect humor and offensiveness in short texts. The objective is to understand the strengths and limitations of current approaches when applied to subjective and context-sensitive NLP tasks, and to evaluate whether combining classification and regression outputs can improve robustness across subtasks.

The NLP tasks addressed are (1) humor classification (predicting whether a text is humorous or not), (2) humor rating regression (predicting the funniness score of a humorous text), and (3) offense classification (predicting whether a text is offensive). These tasks require a nuanced understanding of tone, semantics, and cultural context, and are tackled using transformer-based models, majority voting, and logic-based postprocessing methods. The project builds on the framework proposed in the SemEval-2021 Task 7 challenge on task 1a, 1b, and 1c.

4 Experiment setup

As for the dataset, this project will use the monolingual dataset provided for SemEval-2021 Task 7 for task 1 and its three subtasks, consisting of:

- Training set: 8000 samples. Which contains 3000 samples of humor content. And 5000 non-humor content.
- Testing set: 10000 samples.

There was no need to implement preprocessing methods for the dataset of task1a. However, the dataset for task1b and task1c contain null values. Therefore, we attempted to convert all null values into zeros. In task 1b, since the dataset where is_humor is 1 always contains a humor rating, it is reasonable to replace the null humor rating with 0. Since the label will contain is_humor, or 0, indicating that it is not a joke, setting the humor rating to 0 is appropriate, and there is no need to filter it out. Same as in task 1c, since labels that has humor_controversy (controversy indicator) value is null, is_humor value will always be 0. Therefore, humor_controversy can be set to 0, as the purpose of this task is to identify whether a joke (with is_humor of 1) is controversial or not, not to find any sentences that are controversial.

In this project, we will conduct experiments with multiple pre-trained models to determine the most suitable model for the tasks. We will evaluate the performance of each model and make an informed decision based on its capabilities and suitability. Then we would use the model to hard-vote.

As a baseline, the Semival2021 organizers employed a BERT-base classification /regression model which was run for one epoch, with a batch size of 16 and a learning rate of 5e-5, for all sub-tasks. As this system out-performed the linear benchmarks on all sub-tasks, we refer to this as the baseline in the rest of the paper.

We created simple, linear benchmarks using sklearn for the classification tasks which consists of a Naive Bayes classifier with bag-of-words features.

Next, on to the experimentation. We experimented with multiple pre-trained models: albert-base-case, roberta-base, xlnet-base-v2 and phayathaibert. We implemented the system using a simple Transformers library during the development phase. In the evaluation phase (test phase), we improve our system's capabilities by experimenting with various hyperparameters for the task 1a to see which is the best model. Then we use said model for task and 1c. And because 1b is a humor rating regression task. We use BERT-base model to compute the task, but we train with different batch size.

5 Result and analysis

5.1.1 Task 1a

In this task, from our experiment with varios models, roberta-base outperformed in all categories with accuracy, precision, recall and F1-score at 0.9575, 0.9751, 0.9575 and 0.9653, respectively.

MODEL NAME	ACCU-RACY	PRE-CISION	RE-CALL	F1-SCORE
ALBERT-BASE-CASE	0.9506	0.9719	0.9473	0.9482
ROBERTA-BASE	0.9575	0.9751	0.9575	0.9653
XLNET-BASE-V2	0.9425	0.9735	0.9321	0.9523
PHAYATHAIBERT	0.8838	0.9167	0.8925	0.8780
BASE-LINE (BERT)	*0.9110*	-	-	*0.9283*
BASE-LINE (LINEAR)	*0.8570*	-	-	*0.8840*

Table 1 : Comparison between pre-train models and its performance in task 1a.

5.1.2 Task 1b

This was a humor rating regression task. We have to predict the average rating given to texts ranging from 0 to 5. Texts that were not labeled as humorous by our annotators did not have a humor rating, and predictions for these texts were not included in the final score by our scoring system. The metric used for this task was the root mean squared error (RMSE). RMSE can be calculated by follow:

$$RMSE = \sqrt{\sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{N}\right)^2}$$

We use BERT-base and experiment with different batch sizes, at 8 and 16. Here are the results.

Batch Size	RMSE
BERT-large-cased (batch 8)	0.5490
BERT-large-cased (batch 16)	0.5472
BERT (classification model)	1.9297
baseline (BERT)	*0.8000*
baseline (SVM)	*0.8609*

Table 2: Result in RMSE from different batch sizes in task 1b.

5.1.3 Task 1c

Similar to task 1b, we use Roberta-base to compute task 1c. Here are the results.

Model	Accuracy	Precision	Recall	F1-score
Roberta-base	0.6900	0.4963	0.5447	0.6453
Baseline (BERT)	*0.4731*	-	-	*0.6232*
Baseline (SVM)	*0.4374*	-	-	*0.4624*

Table 3: Performance result from task 1c, showing accuracy, precision, recall and F1-score for different models.

5.2 Comparison with other system

This section presents a comparation between our system and other participating teams in the SemEval-2021 Task 7 (HaHackathon). We highlight differences in performance metrics across subtasks, discuss model choices and techniques employed by top-ranked teams, and examine how our approach aligns with or diverges from theirs. This comparison helps contextualize our results and identify potential areas for improvement based on successful strategies from other participants.

5.2.1 Task 1a

Team	Acc	F1
PALI	0.9820	0.9854
stce	0.9750	0.9797
DeepBlueAI	0.9600	0.9676
OUR TEAM	**0.9575**	**0.9653**
EndTimes	0.9570	0.9655
MagicPai	0.9570	0.9653
mmmm	0.9560	0.9647
baseline (BERT)	*0.9110*	*0.9283*
baseline (Linear)	*0.8570*	*0.8840*

Table 4: Performance comparison for Task 1a.

Our RoBERTa-base model performed moderately across all metrics. This is because RoBERTa-base excels at binary classification, particularly when presented with substantial datasets, which contributes to its robust performance. Nevertheless, top teams have employed larger models (such as RoBERTa-large, ensemble models, or additional data) giving them a slight edge. Limited task-specific fine-tuning or data augmentation could have hindered your system's ability to attain the top accuracy/F1 levels. RoBERTa might have encountered challenges in comprehending nuances in humor types, particularly culturally or contextually dependent humor, which demands more than syntactic understanding.

5.2.2 Task 1b

Team	RMSE
abcbpc	0.4959
mmmm	0.4977
Humor@IITK	0.5210
YoungSheldon	0.5257
CS-UM6P	0.5401
SarcasmDet	0.5446
OUR TEAM	**0.5472**
baseline (BERT)	0.9283
baseline (Linear)	0.8840

Table 5: RMSE comparison for Task 1b.

Regression task can be more challenging than classification, particularly for subjective scores like humor ratings. Our model lacked specialized techniques for regression tasks. Additionally, RMSE is sensitive to outliers, even a few mispredicted humorous/non-humorous scores can significantly impact the metric. Furthermore, other teams might have utilized multi-task learning, ensembling, or regression-specific architectures to more effectively reduce prediction error.

5.2.3 Task 1c

Team	Acc	F1
OUR TEAM	**0.6900**	**0.6453**
PALI	0.4943	0.6302
mmmm	0.4699	0.6279
SarcasmDet	0.4699	0.6270
YoungSheldon	0.4780	0.6210
EndTimes	0.9570	0.9655
Humor@IITK	0.4520	0.6209
RoMa	0.9560	0.6197
baseline (BERT)	*0.9110*	*0.9283*
baseline (Linear)	*0.8570*	*0.8840*

Table 6: Performance comparison for Task 1c.

The RoBERTa architecture is good at spotting subtle patterns, especially those that are controversial or offensive, when train it well. These kinds of controversies often have specific words, sentences, or meanings that pre-trained language models can pick up on easily. It's also possible that the training data preprocessing and class weighting strategies worked better for this task than the other two.

5.3 Error analysis

In task 1a, despite a strong F1-score of 0.9653, our RoBERTa-base model occasionally misclassified non-humorous texts as humorous and vice versa. Upon further inspection, we saw that the model frequently mislabeled sarcastic or ironic statements lacking explicit humor markers as humorous. For instance, a text like "*All that I*

know is I'm breathing, and all I can do is keep breathing now." was predicted as humorous despite its dry tone and lack of comedic intent.

Task 1b required predicting humor ratings on a continuous scale from 0 to 5. Among the tested models, BERT-large-cased performed best as a regression model, yielding an RMSE of 0.5472 (batch size = 16). However, when using BERT as a classification model, treating humor scores as discrete classes (e.g., mapping score 0.0–0.5 to class 0, 0.5–1.5 to class 1, etc.), performance significantly dropped, with an RMSE of 1.9297. Which is more than threefold increase in error. This difference highlights a crucial limitation of applying classification frameworks to inherently continuous targets. The classification model could only predict discrete categories, this can cause quantization errors.

Task 1c showed the lowest performance (F1-score = 0.6453), and our analysis showed that model confusion often rooted from imbalanced labels and ambiguous cases. Most false positives (non-controversial jokes predicted as controversial) contained political references, gender topics, or profanity. For instance, a non-controversial joke like "I told my wife she was drawing her eyebrows too high. She looked surprised." was occasionally flagged as controversial due to the model's over-sensitivity to gendered terms like "wife" or perceived sarcasm. False negatives (controversial jokes predicted as non-controversial) included jokes involving religion, race, or mental health, often written with subtlety or euphemism. For example, a joke like "He has two speeds: slow and 'are you okay?'" may be interpreted as making light of mental health issues, but if written without explicit offensive language, the model may overlook the controversial implications.

6 Conclusion

This project aimed to detect and score humor and offensive content through three subtasks: humor classification (Task 1a), humor rating regression (Task 1b), and controversy detection (Task 1c). By experimenting with various pre-trained transformer models, we discovered that RoBERTa-base was the most effective model for Task 1a, achieving an impressive F1-score of 0.9653. In contrast, for humor rating in Task 1b, BERT-large-cased demonstrated the best performance, with an RMSE of 0.5472. This suggests that humor rating should be treated as a regression problem rather than a classification task. However, Task 1c proved to be more challenging, with the best F1-score reaching 0.6453. This could be attributed to label imbalance and the subtlety of controversial humor.

These results underscore the strengths of transformer-based models in capturing nuanced linguistic features. However, they also highlight the difficulties involved in detecting subjective content such as humor and offensiveness. Notably, the controversy detection task suggests the need for models with enhanced sensitivity to cultural, social, and contextual cues.

Reference

[1] J. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy, “SemEval-2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense,” 2021. Available: https://aclanthology.org/2021.semeval-1.9.pdf

[2] D. Faraj and M. Abdullah, “SarcasmDet at SemEval-2021 Task 7: Detect Humor and Offensive based on Demographic Factors using RoBERTa Pre-trained Model,” 2021. Accessed: Apr. 24, 2025. [Online]. Available: https://aclanthology.org/2021.semeval-1.64.pdf

[3] J. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy, “SemEval-2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense,” 2021. Available: https://aclanthology.org/2021.semeval-1.9.pdf