# Group 10
# SemEval-2021 Task 7: Detecting and Rating Humor and Offense

**Alex Christian**
Student ID: 1133629

**Chintan Sutariya**
Student ID: 1149533

**Mounik Patel**
Student ID: 1144127

**Ruchitkumar Vora**
Student ID: 1150522

## Abstract

Humor is the most intriguing component of interpersonal communication. Due to multiple word senses and different cultural understanding, humor and offense are highly subjective in nature. Detecting humor and offense is a challenging task and can lead to controversy if same sentence evokes different feeling of humor and offense in various demographic groups. We implement a DistilBERT model, which is lighter and cheaper version of BERT, to detect and rate humor and offense in SemEval-2021 Task 7. The proposed model employing DistilBERT shows high performance and outperformed other baseline models in all four subtasks.

## 1 Introduction

Humor is the most appealing aspect of human communication. It is a highly intellectual and communicative activity that encourages people to laugh and enjoy themselves (Chen and Soo, 2018). It requires a good amount of imagination and intelligence to embark humor as well as recognizing it in a conversation. If a machine is trained well enough to detect humor, it can facilitate human-computer interaction and can also assist computers in understanding human conversation and making suitable decisions (Fan et al., 2020).

Humor is difficult to comprehend in texts since it is heavily influenced by human emotions and thought process. Detecting humor in texts is a challenging task since words can have various meanings and, depending on the context, the same words can elicit offensive sentiments for different demographic groups (Smadu et al., 2020).

SemEval-2021 Task 7 (Meaney, 2020) is the first humor detection test that considers the subjective nature of humor and offense in various demographic groups. Users of all ages and genders annotated the data with their opinions on the text's humor and assigned a score to it (Gupta et al., 2021). There are four subtasks in SemEval-2021 Task 7: Detecting and Rating Humor and Offense, shown in Fig.1:

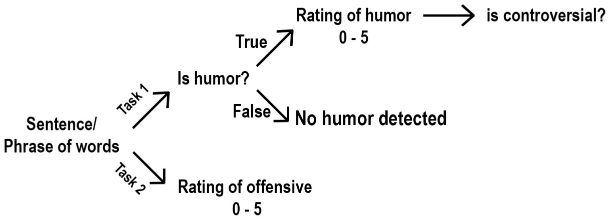

Figure 1: Workflow of SemEval-2021 Task 7

- Subtask 1(a) is a binary classification task to detect whether the text is humorous.
- Subtask 1(b) is a regression task to predict how humorous a text is on a scale of 0 to 5.
- Subtask 1(c) is also a binary classification task with the goal of predicting whether the humor-classified text will cause controversy. In this case, controversy means that a humorous text for one demographic group may be offensive to another.
- Subtask 2 is a regression task to predict how offensive a text is on a scale of 0 to 5.

The rest of the paper is organized as follows: Section 2 reviews previous research on the task of detecting humor, with a focus on transfer learning methods. The dataset that was utilized for the competition is described in Section 3. Section 4 discusses various baseline models as well as our proposed model. The experiment details are presented in Section 5, and the evaluation metrics and experiment results are described in Section 6. Finally, Section 7 brings this paper to a conclusion.

## 2 Related Work

There has been a lot of research done in the past to try to detect humor in text. The related work primarily consists of experiments conducted using various models that leverage Transformer architecture in any form based on transfer learning methods and attention mechanism (Vaswani et al., 2017).

Weller and Seppi (2019) pioneered the use of Transformer architecture in humor detection, outperforming state-of-the-art algorithms on a variety of datasets. The results were also compared to human performance. Pre-trained BERT was selected as the base model because of its success in recognizing and attending to the most significant words in given texts. Dataset was upsampled during training and downsampled during testing to avoid class imbalance. BERT model was evaluated by comparing with CNN model and human performance. Even though the CNN model used a variety of techniques to extract the best features from the dataset, self-attention layers were more successful in extracting the important features. The results also show that this architecture can better predict the level of humor for a specific audience than for a general audience.

Vanroy et al. (2020) describes two methods for estimating the level of humor generated in edited headlines. The first system was a feature-based machine learning system that combined different linguistic information in a Nu Support Vector Regressor (NuSVR), whereas the second system was a deep learning-based approach that learned latent features in news headlines to predict humor using the pre-trained language model RoBERTa. During the evaluation of the models, it was observed that the low-level annotator introduced a layer of noise, resulting in weak correlation (opposite annotators depicting better performance). Deep learning system outperformed feature-based machine learning system, according to the results.

Annamoradnejad and Zoghi (2020) extended the use of BERT models to propose an automated approach for humor classification based on an accepted linguistic structure of humor. Existing humor detection datasets combine formal text and informal jokes with incompatible statistics making it more likely to detect humor without comprehending the underlying hidden connections and rendering it susceptible to overfitting. As a result, a new dataset of short texts was created specifically for the goal of detecting humor. The proposed method employs BERT to build sentence embeddings for a given text and uses them as inputs of parallel lines of hidden layers in a neural network. Finally, these lines are concatenated to predict the target value. The suggested method takes advantage of the linguistic characteristic of humor to split sentences and retrieve mid-level information using hidden layers. Experiments show that the proposed method can accurately detect humor in short texts, with an F1-score of 98.2%.

Gupta et al. (2021) employs Large Language Models (LLMs) and their ensembles to capture the perplexity associated with humor/offense detection and rating. LLMs are used for each subtask separately by adding classification or regression layer over pretrained models like BERT, RoBERTa, ERNIE-2.0, DeBERTa and XLNet, ensuring that the model learns features solely relevant to the task. Above models were also used as ensemble for Classification (Voting) and Regression (Weighted Aggregate) outputs. Adding classification/regression head allows the model to fine-tune on a small dataset with few epochs, avoiding overfitting and improving generalization. Data augmentation technique utilizing the Masked Language Model also helped to overcome the problem of overfitting. However, this resulted in a lot of noise in the dataset and a poor model performance during training. The study highlights some of the inherent difficulties that arise because of the subjective nature of humor and the offense detection task.

## 3 Dataset Description

The dataset utilized for this project is fetched from SemEval-2021 Task 7. The dataset contains sentences or phrases of words in each row. Each row has been annotated with text as unique identifier. The problem dataset comprises of a training dataset (8000 labeled texts). Every text input is labelled as 0 or 1, depending on whether it is humorous or not, and rated with the offensiveness score on a scale of 0-5. If a text is identified as humorous, it is further annotated with a humor rating (0-5) and classified whether it is controversial or not (labelled 0 or 1). We will train our model on the train dataset. We will further test our humor detection model using a test dataset (1000 labeled texts).

## 4 Methodology

### Baseline Method

1. **Logistic Regression** : It is important to note that Logistic Regression should only be utilized when the target variable is divided into discrete categories. Logistic Regression, like other classifi-

cation algorithms, does not use a linear function. It uses a logistic function known as the sigmoid function, as the name implies. When plotted on a graph, the sigmoid function (also known as the logistic function) forms a ”S” shaped curve. It takes values between 0 and 1 and ”squishes” them towards the top and bottom boundaries, labelling them as 0 or 1. The equation for sigmoid function is as follows:

$$f(x) = \frac{1}{1 + e^x}$$

2. **LinearSVC** : In the LinearSVC model, we assumed that training examples plotted in space. There should be a noticeable difference between these data points. A straight hyperplane dividing two classes is predicted. When designing the hyperplane, the main goal is to minimise the distance between the hyperplane and the nearest data point of either class. A maximum-margin hyperplane is the drawn hyperplane.

3. **Naïve Bayes Classifier** : Naive bayes is a classification technique that is based on the Bayes theorem and assumes that features are conditionally independent of each other. A Naive Bayes classifier, in simplistic words, assumes that the presence of one feature in a class is unrelated to the presence of any other feature. The Naive Bayes model is simple to develop and is especially useful for large data sets.

4. **Random Forest** : Random Forest is a supervised learning algorithm. Random forest is a learning algorithm that is supervised. It generates a ”forest” out of several decision trees, which are usually trained using the ”bagging” approach with replacement N samples from the training dataset. The bagging method’s basic principle is that combining different learning models improves the overall output. Random Forest selects attributes based on the Gini Index. Each instance is passed down to all the decision trees generated by the algorithm, which are then voted on by simple majority from all of the outputs to assign the class with the best appearance.

5. **Linear Regression** : Regression is a type of predictive analysis in which one or more independent variables are compared to a dependent variable. It works in the same way that functions do: a value of x is entered, that value is manipulated using coefficients such as slope and intercept, and then another value is outputted. When the structure of the dataset very closely follows a straight line, linear regression is used.

**Transformers**

1. **DistilBERT** : DistilBERT is a distilled version of BERT which is smaller, faster, cheaper and lighter. This method leverage knowledge distillation during the pre-training phase and show that it is possible to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities (Sanh et al., 2019). It was self-supervised pre-trained on the same corpus using the BERT base model as a teacher. This means it was pre-trained on raw texts only, with no human labelling (which is why it can use a lot of freely available data), and then used the BERT base model to generate inputs and labels from those texts.

## 5 Experiment Details

### 5.1 Data Preprocessing

Missing values are a typical occurrence in datasets. Missing values can arise during data processing or because of a data validation rule, but they must be considered anyway. For example, some features in our dataset, such as humor rating and humor controversy have missing data. The missing value in our dataset is not random. If the text is not humorous, then there will be no need of humor rating and humor controversy and hence this is the reason for missing values. We have imputed the missing values with 0.

We performed the text preprocessing by eliminating special symbols using regular expression. In addition, a text was cleaned and reduced to collection of word stems joined by white space. Because DistilBERT contains an intrinsic tokenizer that cleans and tokenizes the text, this text preprocessing was only done for the baseline models.

### 5.2 Baseline Models

Initial experiments to train and test the baseline models for detecting and rating humor and offense were conducted. All the baseline models were trained using train dataset and tested using gold-set dataset. Different baseline models were used for classification task and regression task.

### 5.3 DistilBERT Model

We have also used DistilBERT as another approach to increase the model performance. We have utilized the Tensor vector for input to the model as vector representation. 'DistilBertTokenizerFast' has been used for text preprocessing and end-to-end tokenization. We have used 'distilbert-base-uncased' as pre-trained model to make use of pre-trained model for weights.

## 6 Evaluation and Result

### 6.1 Evaluation Metrics

We will evaluate the model based on the confusion matrix, which can help us to further define various evaluation metric such as accuracy, precision and recall based on the count of occurrences of the four scenarios (true positive, false positive, true negative, false negative). We have used Accuracy for baseline models and F-1 score for DistilBERT model as evaluation criteria for subtask 1(a) and subtask 1(c) which are classification problems, whereas Regressor score for baseline models and Mean Square Error (MSE) score for DistilBERT model was used for subtask 1(b) and subtask 2 which are regression problems.

### 6.2 Results

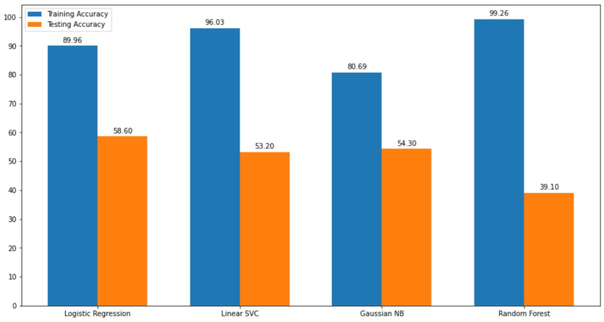

Figure 2: Accuracy of different baseline models for humor detection

All the baseline models were evaluated on the basis of accuracy score in the initial experiment for subtask 1(a). The comparison of different baseline models with respect to accuracy score is shown in Fig. 2. We can observe from the plot diagram that training accuracy as well as testing accuracy is almost same for all different approaches. We can also see that testing accuracy is not good comparably.

We have also used same algorithms for subtask 1(c). The Fig. 3 is showing comparison of different algorithm's performance. We have used Linear Regression for subtask 1(b) and subtask 2, but the model did not work well in both scenarios.

We chose to go with our proposed approach, and after implementing the proposed approach (DistilBERT model) for all of the subtasks, we obtained high performance. We have used 'distilbert-base-uncased' as pre-trained model for our implementation. We got F-1 score of 1.0 for both subtask 1(a) and subtask 1(c) which are classification problems, and we got 1.973e-15 and 5.770e-16 as MSE score for subtask 1(b) and subtask 2 respectively which are regression problems during testing of DistilBERT model.

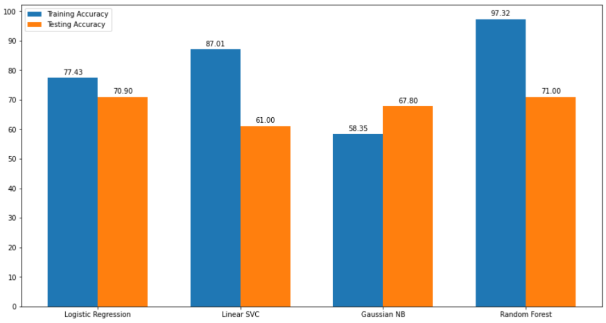

Figure 3: Accuracy of different baseline models for controversy detection

## 7 Conclusion and Future Work

In this paper, we presented DistilBERT model, which is lighter and cheaper version of BERT, for implementing all four subtasks of SemEval-2021 Task 7. We also implemented several baseline models for detecting and rating humor and offense, to compare their performance with our approached model. We observed from the experiment results that DistilBERT outperforms baseline models for the task of detecting and rating humor and offense. The experiment results show similar F-1 score of 1.0 for both detecting humor and controversy, and MSE score of 1.973e-15 and 5.770e-16 for rating humor and offense respectively. This depicts that DistilBERT model attained high performance compared to other models. For future work, we plan to implement the ensemble of several deep learning models with the DistilBERT model to see how well it performs in terms of detecting and rating humor and offense.

## References

Issa Annamoradnejad and Gohar Zoghi. 2020. ColBERT: Using BERT Sentence Embedding for Humor Detection.

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. Association for Computational Linguistics.

Xiaochao Fan, Hongfei Lin, Liang Yang, Yufeng Diao, Chen Shen, Yonghe Chu, Tongxuan Zhang, and Danilo Comminiello. 2020. Phonetics and Ambiguity Comprehension Gated Attention Network for Humor Recognition.

Aishwarya Gupta, Avik Pal, Bholeshwar Khurana, Lakshay Tyagi, and Ashutosh Modi. 2021. Humor@IITK at SemEval-2021 Task 7: Large Language Models for Quantifying Humor and Offensiveness.

J. A. Meaney. 2020. Crossing the Line: Where do Demographic Variables Fit into Humor Detection? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 176–181. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Razvan-Alexandru Smadu, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. UPB at SemEval-2021 Task 7: Adversarial Multi-Task Learning for Detecting and Rating Humor and Offense.

Bram Vanroy, Sofie Labat, Olha Kaminska, Els Lefever, and Veronique Hoste. 2020. LT3 at SemEval-2020 Task 7: Comparing Feature-Based and Transformer-Based Approaches to Detect Funny Headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1033–1040. International Committee for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need.

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 3621–3625. Association for Computational Linguistics.

## Appendix A Technical Resources

- **Google Colab** was utilized to write and execute Python code for the project, since it provides the Jupyter notebook experience on Google's cloud with additional advantage of Google hardware including GPU and TPU.

- Because the dataset for the project is in tabular format, the **Pandas** library was employed since it provides multiple functionalities for working with tabular data in machine learning and data analysis.

- **Scikit Learn** library was employed for implementation of various baseline models related to both classification and regression sub-tasks. It was also used for the purpose of evaluation metric.

- **Transformer** is a Python library which is used in our project for the implementation of DistilBERT model. It employs several functionalities related to data preprocessing and training of the model based on Transformer architecture.

## Appendix B Work Contribution

Each member examined different research papers related to humor detection. Following a thorough discussion, we selected few significant research papers and performed literature review for the same. The tasks related to implementation of baseline models was divided among the pair of members, where Alex and Mounik performed text preprocessing portion, and Chintan and Ruchit performed the training of baseline models. All the four individuals contributed equally for the implementation of project model(DistilBERT) since it was effective for all team members to understand it. Each member was assigned different components of the final project report so that it could be completed in a timely and effective manner.