Towards an automated system for rating humor content in textual documents

Chirag Daryani
cdaryani@ualberta.ca

Varshini Prakash
vprakash@ualberta.ca

1 Task Description

Understanding humor is not straightforward. What can be considered humorous depends heavily on the tone of the speaker, the facial expressions, the context of the conversation, the environment in which the conversation is happening, and also the factual knowledge of the listener (Hossain et al., 2020b). Recognizing humor content and intensity becomes even more difficult in the case of textual documents since it lacks these nonverbal aspects of communication. Therefore, understanding humor from textual information is a difficult but an integral task in natural language processing as we aim to achieve true machine intelligence.

Towards this goal of developing intelligent systems that can better recognize humor content from text, the focus of our current work is on developing the solution for the SemEval-2020 Task 7 (Assessing Humor in Edited News Headlines) (Hossain et al., 2020a) Subtask 1 .

In this task, we try to predict the mean humor rating for news articles that were edited by replacing a word in their original content. Our aim would be to develop a ***sentence-pair regression model*** that takes both the original news headline and the edited news headline as input and then outputs a prediction of a mean funniness rating in the range of 0 to 3 for each edited news article. The predicted real-valued scores would represent the humor content as following: 0 representing "Not Funny", 1 representing "Slightly Funny", 2 representing "Moderately Funny" and 3 representing "Very Funny".

For example, suppose the original news headline is "Thousands of students , teachers march on White House to call for better ***gun*** control" and the edited headline formed by replacing the word **'gun'** is "Thousands of students , teachers march on White House to call for better ***homework*** control". When we pass these two inputs to our model, the prediction should be as close to 1.6 as possible because a mean funniness rating of 1.6 is the groundtruth for this observation in our training dataset.We can visualize the basic system in the figure below.

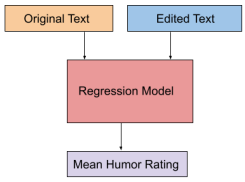

Figure 1: Basic System Architecture with Input and Output

By measuring the amount of humor content, we can then understand what exactly makes a particular text piece funny, give ranking to multiple pieces of edited texts based on their funniness rating, and then can also design recommendation systems for the same. We can even use the system in humor generation applications (Hossain et al., 2019).

Therefore, solving this task of humor content evaluation from text articles can help us gain much more understanding of the language and this is why we selected this task for our project.

2 Available data

We will be using the dataset called "Humicroedit" (Hossain et al., 2019) which was provided by the competition organizers. It was created by taking news headlines and making micro-edits i.e. replacing one or two words in the headline text in order to introduce humor in the news content. There are about 5,000 distinct news headlines, each having 3

edited versions, thus totaling about 15,000 records.

The data is partitioned into 3 subsets, the Train Dataset with about 64% records, the Dev Dataset with about 16% records, and the Test Dataset with 20% of the records. All edited versions of the same headline are in the same subset (Hossain et al., 2020a).

If we look at any one record of the training dataset as an example, we will be able to see 5 fields in the data.

id	original	edit	grades	meanGrade
12101	Iran successfully launches satellite-carrying rocket into <space/>	tree	32210	1.6

Table 1: Sample Training Data

Here, "id" is the unique identifier of an edited headline, "original" is the unedited news headline with the words to be replaced identified by the </>tag, "grades" is the concatenation of the rating given to the news headline by each judge and "meanGrade" is the mean of the ratings given to the news headline by all the judges.

3 State-of-the-art

Automatic humor assessment and humor rating have been previously solved using various approaches. The architecture proposed by Jensen et al. (2020) to understand humor combined handcrafted features, knowledge bases and a language model in a regression model. The performance of their neural model which consisted of three decoders exceeded two baselines in detecting the intensity of humor in edited headlines. Tomasulo et al. (2020) evaluated the humor of edited news titles by training the titles encoded as embeddings in a bidirectional GRU (BiGRU) ensemble with XGBoost. Mahurkar and Patil (2020) tested the language modeling and generalization abilities of BERT and its derivative models for humor grading and classification tasks through zero-shot and cross-dataset inference based approaches.

We plan to achieve high performance by performing multiple text preprocessing techniques such as cleaning, vectorization and experimenting with different types of embeddings before training an ensemble of machine learning models to understand and evaluate humor. We will also try neural network models like LSTMs as done in the paper by Miraj and Aono (2021), and some Transformer-based models that work particularly well for tasks that involve remembering sequential information which is a requirement for our humor content analyzer task. We will not limit our work to the above-mentioned and will try to experiment with many more aspects like activation functions, optimizers, feature engineering techniques, etc. during the implementation of the project.

For assessing the performance of our work, we are using the same evaluation criteria that were set by the competition organizers, which is the Root Mean Squared Error (RMSE). From the script provided by the competition organizers, the RMSE score of the baseline model (predicting every rating as the overall mean funniness grade in the training set) was found out to be 0.5783. Our aim would be to minimize this RMSE value which we calculate using the predictions of the models we'll use and the ground truth mean funniness value. In this way, we'll select the best-performing model for the task.

4 Available code

As mentioned previously, we are provided a code for the baseline model that always predicts the overall mean funniness grade in the training dataset. After execution of this code, we found that it gives an RMSE score of 0.5783 on the test set.

https://github.com/n-hossain/semeval-2020-task-7-humicroedit

Along with this, we found a code by Zhang and Yamana (2020) which is fine-tuning the transformer-based model BERT for deriving vector representations of our textual data. The trained model pickle file is provided by the authors and we were able to successfully use it to get predictions for our data. The code can be found at

https://github.com/HeroadZ/SemEval2020-task7

5 Repository URL

The Github repository of our project can be found at

https://github.com/UOFA-INTRO-NLP-F21/f2021-proj-chiragdaryani

References

Nabil Hossain, John Krumm, and Michael Gamon. 2019. " president vows to cut¡ taxes¿ hair": Dataset and analysis of creative text editing for humorous headlines. *arXiv preprint arXiv:1906.00274*.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020a. Semeval-2020 task 7: Assessing humor in edited news headlines. *arXiv preprint arXiv:2008.00304*.

Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. 2020b. Stimulating creativity with funlines: A case study of humor generation in headlines. *arXiv preprint arXiv:2002.02031*.

Kristian Nørgaard Jensen, Nicolaj Filrup Rasmussen, Thai Wang, Marco Placenti, and Barbara Plank. 2020. Buhscitu at semeval-2020 task 7: Assessing humour in edited news headlines using hand-crafted features and online knowledge bases. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 824–832.

Siddhant Mahurkar and Rajaswa Patil. 2020. Lrg at semeval-2020 task 7: Assessing the ability of bert and derivative models to perform short-edits based humor grading. *arXiv preprint arXiv:2006.00607*.

Rida Miraj and Masaki Aono. 2021. Kdehumor at semeval-2020 task 7: A neural network model for detecting funniness in dataset humicroedit. *arXiv preprint arXiv:2105.05135*.

Joseph Tomasulo, Jin Wang, and Xuejie Zhang. 2020. Ynu-hpcc at semeval-2020 task 7: Using an ensemble bigru model to evaluate the humor of edited news titles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 871–875.

Cheng Zhang and Hayato Yamana. 2020. Wuy at semeval-2020 task 7: Combining bert and naïve bayes-svm for humor assessment in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1071–1076.