Final Report: SemEval-2020 Task 7 - Assessing Humor in Edited News Headlines

Maisha Binte Moin
bintemoi@ualberta.ca

Uduak Ituen
ituen@ualberta.ca

Abstract

In this paper we describe our system for SemEval 2020 Task 7: "Assessing Humor in Edited News Headlines". We attempted both the subtasks using traditional machine learning approaches. For Task 1 our best RMSE score was 0.552173 and our accuracy using this model for Task 2 resulted in a 47 percent. We used pre-trained word embeddings to train the models thus obtaining better results when compared to previous work that used machine learning approaches for the task.

1 Introduction

Humor is a complex form of human language expression. It is an essential component needed for a computer to communicate effectively with humans (Binsted et al., 2006). By modelling humor generation in computer systems, we are able to better understand how humans process language and cognition (Binsted et al., 2006). Although steady progress has been made by researchers in training intelligent systems to express humor, it has not gained as much traction as it should, partly due to the complexities involved in emulating the intrinsic nature of humor (Hossain et al., 2019).

There are underlying theories and hypotheses that provide useful insight into the nature of humor. These theories have been adopted over the years as baselines for understanding the intricacies of humor. A dominant theory of humor is the Incongruity Theory (Morreall, 2016) which says that source of humor is in perceiving something unexpected (incongruous) that violates expectations that were set up by the joke. Hossain et al. (2020) have shown how the dataset supports this theory due to its humor grading scheme where the values are continuous (related to the degree of funniness) as opposed to previous binary classifications thus providing the scope to observe some of the theories of humor within the dataset.

The subtasks require the creation of systems that are not only capable of estimating the degree of humor on a scale of 0-3 (Task 1) but that can also detect variation of the degree by predicting the funnier version between two edits of original versus edited news headline (Task 2).

2 Related Work

Computational humor can be divided into two groups: recognition and generation (Docekal et al., 2020). Both humor recognition and generation prove to be challenging problems as they involve in-depth world-knowledge, common sense and various levels of understanding (Hossain et al., 2019).

For this project, we have employed traditional machine learning approaches to solve the tasks. Earlier humor recognition systems have used statistical machine learning algorithms such as support vector machines (SVMs), decision tree, k-nearest neighbors and Naive Bayes (Castro et al., 2016) and their best results have been obtained by using SVM. Ahuja et al. (2018) have also demonstrated that SVM classifier has a better accuracy compared to Naïve Bayes and logistic regression for humor classification tasks. Our results are consistent with these findings as our best model based on RMSE value from Task 1 was an SVM based regressor which we have later used for the classification task (Task 2). Docekal et al. (2020) had also tested machine learning models for the SemEval-2020 task-7, their models utilized TF-IDF word features but they have demonstrated poor performance. In our work we have used pre-trained non-contextual word embeddings

(Word2vec, GloVe) which yielded far better RMSE values for the models compared to Docekal et al. (2020).

3 Methods

We have used traditional machine learning models to approach both the tasks, as machine learning models are widely used to provide solutions in both classification and regression tasks. High-level overview of the pipeline is data preprocessing, using word embeddings for training, fitting the models, then finally selecting the best performing model based on test dataset performance for Task 2.

For the data preprocessing we have changed the original word between '</>' with the edited word to form a new column with the new headline after applying the change. Then we turn the new headline into vectors. This step has been the most difficult in the project, we had first started out by applying CountVectorizer then TF-IDF Vectorizer for this project but our model RMSE results were poor. Later, we switched to using two pretrained word embeddings that is Word2vec and GloVe and compared the results from using both embeddings separately. Using pre-trained word embeddings worked better in this context as the previous vectorization steps were threating each word as a feature, thus the number of features were large. The other steps involved lowercasing, tokenizing, removing stop words, punctuations.

The models that we used were: NuSVR, Linear Regression, Epsilon SVR, Linear SVR, Random Forest, K-Nearest Neighbors, Decision Tree, Logistic Regression.

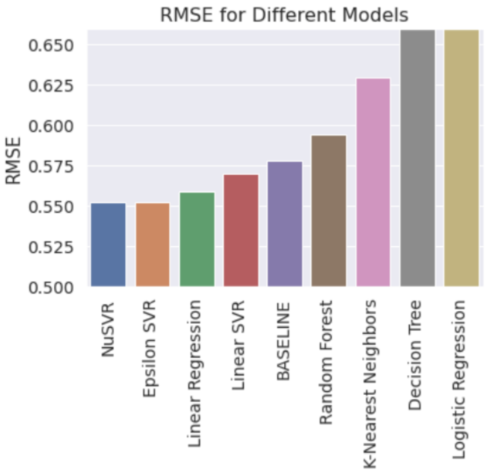


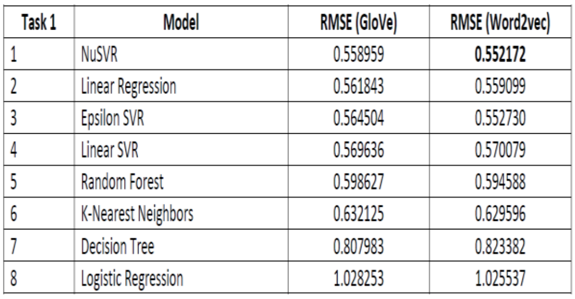
Figure 1: RMSE comparison (Word2vec)

Figure. 1 graphically show the RMSE values for the different methods. The best model from our results was NuSVR using Word2vec embedding, we have utilized this model for Task2, where we predicted the individual predictions for both the edits and compared the predictions to assign the associated labels (0 - if the values are same, 1 - if edit1 is funnier, 2 -if edit2 is funnier).

4 Results

The two metrics employed for evaluation of the tasks, is the Root Mean Squared Error (RMSE) for Subtask 1 and Accuracy for Subtask 2. RMSE measures the difference between the rating generated by the system model and the benchmark rating given by crowdsourcing. While Accuracy evaluates the correctness of the model's prediction in determining the funnier headline.

Table. 1. Task-1 results

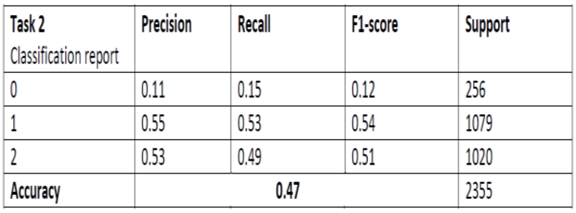


Task 1	Model	RMSE (GloVe)	RMSE (Word2vec)
1	NuSVR	0.558959	**0.552172**
2	Linear Regression	0.561843	0.559099
3	Epsilon SVR	0.564504	0.552730
4	Linear SVR	0.569636	0.570079
5	Random Forest	0.598627	0.594588
6	K-Nearest Neighbors	0.632125	0.629596
7	Decision Tree	0.807983	0.823382
8	Logistic Regression	1.028253	1.025537

From Table.1 we see the best model based on RMSE value is NuSVR with Word2vec embedding. Between using Word2vec and GloVe embedding, Word2vec had slightly better results as it considers the local property of the dataset and the model that we choose utilized Google news dataset. In general, among the models the support vector machines based regressors had the best performance compared to the baseline which was the mean of the 'meanGrade' values. Linear Regression performed better compared to the baseline as well. Random forest regressor perform better than decision tree regressor which is understandable as there were several decision trees or estimators to choose the best one from.

Table. 2 shows the classification report of using the best model to predict the funnier of the edited headlines. Our accuracy is 47%. Headline pairs with larger funniness gaps were easier to classify, thus the model had some sense of the degree of funniness of the edits.

Table. 2. Task-2 Results



Task 2 Classification report	Precision	Recall	F1-score	Support
0	0.11	0.15	0.12	256
1	0.55	0.53	0.54	1079
2	0.53	0.49	0.51	1020
Accuracy	0.47			2355

5 Discussion

This section analyses the output produced by our best model for Task 1. In general, the model does not perform well on headlines containing extreme humor and is mainly affected by general world knowledge, sarcasm, and negative sentiments as have previously been observed by Hossain et al. (2020).

We observed two error variations, classed as overestimation or underestimation. For example, headline with ID 11404 had the highest error score of difference 2.5 and the ground truth had the highest mean rating of 3 indicating the highest degree of funny. However, the model predicted a not funny score, mainly because the headline contained an extreme form of humor. Headline IDs 11927, 1805 had an underestimation in the degree of humor which shows that the model lacks understanding of world knowledge. An additional table for this error analysis is available in the runs folder located in the repository.

6 Conclusion

In general, we have trained the machine learning models using pre-trained word embeddings in order to solve the tasks and gained better results compared to some previous ML based approaches. However, the models we have used have been tested using mainly the default parameters decided by sklearn. Taking some further steps like Hyperparameter tuning and evaluating using GridSearchCV would be the next steps to find the best combination of Hyperparameters to find the best model.

Our general observation is that given the advancement of deep learning technologies we see both the utilization and success of such sophisticated models in achieving state of the art results in the SemEval tasks. For the SemEval-2020 task-7, the leading teams made use of pre-trained language models (PLMs) such as BERT, RoBERTa, ELMo, GPT-2, context independent word embeddings like Word2vec, GloVe and ensemble models (Hossain et al., 2020). We see from the submissions that as the model complexity goes down the performance also degrades. From this observation using deep learning models like BERT would have given us a better understanding as to how the results compare between machine learning models and deep learning models.

7 Repository

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

References

Vikram Ahuja, Taradheesh Bali, and Navjyoti Singh. 2018. What makes us laugh? investigations into automatic humor classification. In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 1–9.

Kim Binsted, Anton Nijholt, Oliviero Stock, Carlo Strapparava, G Ritchie, R Manurung, H Pain, Annalu Waller, and D O'Mara. Computational humor. IEEE Intelligent Systems, 21(2):59–69, 2006.

Santiago Castro, Matıas Cubero, Diego Garat, and Guillermo Moncecchi. 2016. Is this a joke? Detecting humor in Spanish tweets. In Ibero-American Conference on Artificial Intelligence (IBERAMIA 2016), pages 139–150. Springer.

Martin Docekal, Martin Fajcik, Josef Jon, and Pavel Smrz. Jokemeter at semeval-2020 task 7: Convolutional humor. arXiv preprint arXiv:2008.11053, 2020.

Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines. In Proceedings of the 2019 Conference of the NorthAmerican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 133–142, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020a. Semeval-2020 Task 7: Assessing humor in edited news headlines. In Proceedings of International Workshop on Semantic Evaluation (SemEval-2020), Barcelona, Spain.

John Morreall. 2016. Philosophy of humor. In Edward N. Zalta, editor, The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, winter 2016 edition.