RedwoodNLP at SemEval-2021 Task 7: Ensembled Pretrained and Lightweight Models for Humor Detection

Nathan A. Chi
De Anza College
Cupertino, CA
chinathan@student.deanza.edu

Ryan A. Chi
Stanford University
Stanford, CA
ryanchi@stanford.edu

Abstract

An understanding of humor is an essential component of human-facing NLP systems. In this paper, we investigate several methods for detecting humor in short statements as part of Semeval-2021 Shared Task 7. For Task 1a, we apply an ensemble of fine-tuned pre-trained language models; for Tasks 1b, 1c, and 2a, we investigate various tree-based and linear machine learning models. Our final system achieves an F1-score of 0.9571 (ranked 24 / 58) on Task 1a, an RMSE of 0.5580 (ranked 18 / 50) on Task 1b, an F1-score of 0.5024 (ranked 26 / 36) on Task 1c, and an RMSE of 0.7229 (ranked 45 / 48) on Task 2a.

1 Introduction

Humor detection is the process of identifying sequences of text that are amusing—an important task, as such sequences are present in most channels of communication. Although humor detection comes naturally to humans, it is difficult for artificial systems to do the same. Part of the challenge is that it is debatable what constitutes humor; what one reader finds funny may be found utterly prosaic by the next. The problem is only complicated when demographic factors come into play; now, the element of offense is also a factor.

SemEval-2021 Shared Task 7 attempts to address some of these open problems (Meaney et al., 2021). Rather than definitively labeling text as humorous or not, Task 1a aims to determine whether the author intended for the sentence to be humorous, Task 1b predicts its humor rating by the average user (its first moment), and Task 1c attends to whether the variance of its humor ratings (its second moment) exceeds the median. Task 2a, meanwhile, considers the text's average offensiveness score, a metric that often correlates with whether the author meant the text to be humorous and—perhaps equally importantly—affects whether the joke would be considered acceptable. Overall, training models to perform well on these tasks is of central importance to developing systems that are responsive to a wide range of input, whether in complete jest or meant to be taken at face value.

2 Dataset

We train and validate our models on the SemEval-2021 Task 7 training set (Table 6). Each English sentence is annotated for the following four labels, with continuous annotations labeled using a Likert scale from 1 to 5.

#	Description	Label
1a	Is the intention of this text to be humorous?	Binary
1b	How generally humorous is the text for the average user?	Continuous
1c	If the sentence is humorous, is the the humor controversial? [1]	Binary
2a	How generally offensive is the text?	Continuous

Table 1: Annotations/subtasks with their descriptions. We submit separate models for each of these tasks.

2.1 Train-test split

The dataset has a total of 10000 examples, split 8000–1000–1000 between train, validation, and test sets. However, the official development set lacked labels until the last phase of the competition, so we created our own held-out validation set for our experiments. Consequently, our train set has 6,400 examples, and our validation set has 1,600 examples. In our paper, all "validation set" performance is reported on this internal held-out set.

[1] In gold standard labels, an example is deemed controversial if its variance exceeded the median variance of all examples.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 1209–1214
Bangkok, Thailand (online), August 5–6, 2021.

3 Methods

3.1 Task 1a: Humor Prediction

The goal of this task is to model whether a given text is intended to be humorous. Hypothesizing that pretrained language models could effectively model the presence of humor in statements, we investigate the following models:

- **BERT** (Devlin et al., 2019) is a pretrained masked language model. We use BERT-Large in our experiments (335M parameters).

- **RoBERTa** (Liu et al., 2019) is a robustly optimized BERT pre-training approach that utilizes changes including a larger pre-training dataset and a dynamic masking pattern strategy. We use RoBERTa-Large (335 parameters).

- **ELECTRA** (Clark et al., 2020) is a pretrained model that uses a discriminative replaced-token identification loss rather than a demasking objective, resulting in greater data efficiency. We use ELECTRA-Large (336M parameters).

Ensemble We also investigate an ensemble which incorporates one BERT-large, one RoBERTa-large, and nine ELECTRA-large models. Our models were averaged with equal weights. Each ELECTRA model was trained with a different random seed from 100 to 900; in the row corresponding to the ELECTRA model's performance, we have only included the result from the best seed (200).

Pretraining details We trained with binary cross-entropy loss for 3 epochs, using a learning rate of 1×10^{-5} and batch sizes of 16 (ELECTRA) and 8 (BERT, RoBERTa).

3.1.1 Results

We find that all models achieve high F1 and accuracy, with ELECTRA performing the best of any individual model. However, we achieve the highest performance using our ELECTRA + BERT + RoBERTa ensemble. Notably, the ensemble achieves a slightly superior performance to each of its individual component models. Overall, we are ranked 24th out of 58 on this task, achieving an F1-score of 0.9571.

Model	# params	F1	Accuracy
BERT	335M	0.941	0.928
RoBERTa	355M	0.952	0.940
ELECTRA	336M	0.956	0.944
Ensemble	—	**0.957**	**0.946**

Table 2: Performance of our candidate models on the official evaluation set for Task 1a (humor prediction). Out of the individual models, ELECTRA achieves the strongest results, and ensembling the predictions of multiple pretrained models slightly helps both F1 and accuracy.

3.2 Tasks 1b, 1c, and 2a: General Humor, Controversy, and Offensiveness

3.2.1 Models

Despite their success on Task 1a, we were unable to achieve strong results with pretrained language models on the other tasks. Consequently, we experimented with several other machine learning methods, using lightweight features as inputs. We examine a number of different supervised learning algorithms, implemented using the **Scikit-learn** (Pedregosa et al., 2011) framework:

- **Support Vector Machine** is a lightweight classification algorithm that employs a hyperplane that divides a dataset into two subsets.

- **Random Forest** is an supervised learning technique that utilizes independently trained decision trees that sample from a random selection of data.

- **Gradient Boosting** is a technique that ensembles a number of weak learners (typically decision trees) and optimizes based on a differentiable loss function.

- **LightGBM** (Ke et al., 2017) is a highly efficient gradient boosting decision tree that takes advantage of GOSS (gradient-based one side sampling) and EFB (exclusive feature bundling).

- **AdaBoost** (Schapire, 1999) (**Ada**ptive **Boost**ing) is an instance of gradient boosting that optimizes by re-weighting weak learners based on high-weight data points (rather than using a differentiable loss function).

- **Multilayer Perceptron** is a feed-forward deep neural network.

Model	Features	F1 (1c)	Accuracy (1c)	RMSE (1b)	RMSE (2)
AdaBoost	GloVe	0.48	0.48	0.564	1.355
CatBoost	GloVe	0.51	0.51	0.563	0.877
GradientBoosting	GloVe	0.52	0.52	0.572	0.848
LGBM	**GloVe**	0.50	0.50	0.552	**0.808**
Logistic Regression	GloVe	0.52	0.52	—	—
Logistic Regression	Manual	0.50	0.53	—	—
MLP	GloVe	0.49	0.49	0.562	0.798
RandomForest	**GloVe**	0.52	0.53	**0.548**	0.928
SVM	**GloVe**	**0.55**	**0.55**	0.551	0.874
XGBoost	GloVe	0.52	0.52	0.556	0.858

Table 3: Validation set performance of candidate models on Task 1b, 1c, and 2a (controversy classification). For tasks 1b (humor rating), 1c (humor controversy), and 2a (offense rating), the highest-performing models are the random forest model with $n_{trees} = 1000$, the support vector machine, and the LGBM, respectively. We did not run experiments for entries marked —.

- **CatBoost** (Dorogush et al., 2018) is a variant of gradient boosting that prioritizes low latency via symmetric trees.

- **XGBoost** (e**X**treme **G**radient **Boost**ing) (Chen and Guestrin, 2016) is an implementation of gradient boosting that efficiently makes use of parallel computation.

3.2.2 Features

Given that the subjectivity of humor is often correlated with the subject matter of the joke, we also examine its impact on humor controversy in an alternative approach to Task 1c. Often, a joke regarding a sensitive topic may be comical to one reviewer but downright unamusing to a second, whose sense of humor is entirely disparate from the first's.

In this approach, we use a suite of engineered one-hot features with logistic regression (Table 4). Our manual features consist of groups that are typically stereotyped: more specifically, each manual feature consists of a set of tokens, and its value is the number of times a token from its set appears in the input.

In an effort to interpret the significance of these features, we calculate logistic regression (LR) coefficients with respect to the controversy label. The results show that several features were unrelated or inversely correlated to humor controversy (most notably the "Black" feature); they also indicate that a few were strongly positively correlated (such as the "White" feature).

For our final models, we use 300-dimensional GloVe word vectors (Pennington et al., 2014) mean-pooled over each sentence.

3.2.3 Results

The official evaluation set performances for our Transformer-based models in Task 1a are listed in Table 2, while the unofficial validation set performances for our regressors and classifiers are listed in Table 3.

For Task 1b (humor rating), we achieve the highest performance using our Random Forest model. Overall, we are ranked 18th out of 50 on this task, achieving an RMSE of 0.5580.

For Task 1c (humor controversy), we achieve the highest performance using our SVM model. Overall, we are ranked 26th of 36 on this task, achieving an F1-score of 0.5024.

For Task 2a (offense rating), we achieve the highest performance using our LGBM model. Overall, we are ranked 45th out of 48 on this task, achieving an RMSE of 0.7229.

4 Conclusion

We have presented models trained to predict various aspects of humor in text: the level of intended humor, the level of humor for average users, and the level of controversy and offense of a given humorous statement.

We find that large pretrained models such as ELECTRA, RoBERTa, and BERT are effective at predicting the level of intended humor. Furthermore, we note that ensembling these models slightly improves performance. However, our ex-

Feature	Description	LR Coefficient
BLACK	Words referring to those of African descent.	-0.508
AMERICAN	The word "American."	-0.493
GENDER	Words associated with women.	-0.234
INTELLIGENCE	Words associated with stupidity.	-0.169
ISLAM	Words referring to the religion or associated institutions.	-0.102
RELIGION	All major religions not including Islam.	-0.096
RACIAL	Words referring to those of Asian, Latin American, and African descent.	-0.062
SEXUALITY	Words relating to sexuality.	-0.051
HOUSING	The word "homeless."	-0.008
BRUTALITY	Words heavily connoting violence.	0.064
COUNTRIES	Words relating to nationalities not included in "Racial" or "American" features.	0.077
BLONDE	The word "blonde."	0.112
PARTNER	Significant others or family members; controversial jokes often include words regarding female partners.	0.164
SEXUAL	Words relating to sexual activity.	0.171
VULGAR	Profanity.	0.242
WHITE	Words referring to those of Caucasian descent.	0.547

Table 4: Manual features for Task 1c (controversy classification)

periments highlight that pretrained models yield weaker results when faced with regression tasks, as well as when faced with the goal of trying to predict whether a given statement's humor rating has high controversy. This may be due to difficulty in predicting inter-rater disagreement (i.e. if the humor metric's variance exceeds the median variance).

Next, we note also that our engineered one-hot feature approach toward humor subjectivity does not perform significantly better than the baseline models. While our results do reveal a positive correlation between certain manual features and humor controversy—illustrating that humor subjectivity is to some degree affected by subject matter—our results suggest that on the whole, the effects of this relationship are limited.

Overall, our results suggest that reasonably lightweight models can achieve strong results in modelling humor in human language.

Acknowledgments

The authors would like to thank Ethan A. Chi for his support and guidance throughout this project. We would also like to acknowledge Google Colaboratory for their free compute services.

References

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154.

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*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Robert E. Schapire. 1999. A brief introduction to boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'99, page 1401–1406, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

A Reproducibility

We release our code at `https://github.com/nathanchi/hahackathon`. Additionally, we include our hyperparameters in Table 5 for reproducibility.

Model	Hyperparameter	Task 1b	Task 1c	Task 2a
AdaBoost	learning_rate	1.0	1.0	1.0
	loss	linear	linear	linear
	$n_{\text{estimators}}$	1500	1500	1500
GradientBoosting	learning_rate	0.1	0.1	0.1
	max_depth	3	3	3
	$n_{\text{estimators}}$	1000	100	500
LGBM	learning_rate	0.1	0.1	0.1
	max_depth	10	10	10
	num_leaves	22	22	22
	$n_{\text{estimators}}$	60	600	600
MLP	learning_rate	constant	constant	constant
	α	0.01	0.1	0.01
	β_1	0.9	0.9	0.9
	β_2	0.999	0.999	0.999
	hidden_layer_sizes	(100, 100)	(500, 500)	(200, 200)
	max_iter	12	200	12
RandomForest	$n_{\text{estimators}}$	1000	100	100
	criterion	mse	gini	mse
	max_depth	2	2	2
SVM	C	1.0	1.0	1.0
	degree	3	3	3
XGBoost	$n_{\text{estimators}}$	100	100	100

Table 5: Hyperparameters for lightweight supervised learning models.

Sentence	is_humor	humor_rating	humor_controversy	offense_rating
When I was in college I used to live on a houseboat and started dating the girl next door. Eventually we drifted apart.	1	2.95	0	0.25
Want to know why he disappeared? These are the most common reasons men disappear from your life.	0	0	0	0

Table 6: Examples that are intended and not intended to be humorous, respectively.