The Naughtyformer: A Transformer Understands and Moderates Adult Humor (Student Abstract)

Leonard Tang, Alexander Cai, Jason Wang

Harvard University
leonardtang@college.harvard.edu, alexcai@college.harvard.edu, jasonwang1@college.harvard.edu

Abstract

Jokes are intentionally written to be funny, but not all jokes are created the same. While recent work has shown impressive results on humor detection in text, we instead investigate the more nuanced task of detecting humor *subtypes*, especially of the more adult variety. To that end, we introduce a novel jokes dataset filtered from Reddit and solve the subtype classification task using a finetuned Transformer dubbed the Naughtyformer. Moreover, we show that our model is significantly better at detecting offensiveness in jokes compared to state-of-the-art methods.

Introduction

The field of humor detection has received much interest over the years. Early work attempted to leverage N-grams (Taylor and Mazlack 2004), and Random Forest classifiers acting on Word2Vec embeddings (Yang et al. 2015) with limited success. More recently, researchers have explored deep learning-based approaches for humor detection, in particular Transformer inspired architectures (Weller and Seppi 2019).

Despite this progress and introduction, no prior work exists on the more nuanced task of classification amongst humor subtypes. Indeed, we are the first to solve this task.

A Humor Subtype Dataset

To train the Naughtyformer, we introduce a dataset of 92,153 total jokes across three categories of 1) Wholesome Jokes, 2) Dark Jokes, and 3) Dirty Jokes.

In particular, Wholesome jokes are defined to be jokes that are inoffensive in nature. Meanwhile, Dark Jokes are described as humor that is viewed as dark, morbid, cruel, offensive to some, or graphic in nature. Finally, Dirty Jokes are obscene, indecent jokes that primarily consist of vulgar, sexist, racist, and discriminatory content. Critically, Dark and Dirty Jokes are distinctly different subtypes of humor.

We source all three joke types from Reddit, alongside news articles that act as non-jokes. We scrape Thomson Reuters as the source of our news articles due to its reputation as a neutral, inoffensive media outlet. Table 1 lists the statistics of our final dataset after collection and processing.

Statistic	Clean	Dark	Dirty	News
Examples	7450	79230	5473	10710
Avg. Length	31.47	24.64	55.24	778.84
Std. of Length	46.21	78.67	91.14	292.34
Avg. Upvotes	87.30	105.11	38.85	NA
Std. of Upvotes	175.24	589.35	50.23	NA

Table 1: Statistics of our jokes dataset scraped from Reddit. Post length is measured via the Penn Tree Bank tokenizer.

Subreddits as Natural Data Reddit is a social news website featuring socially curated feeds within well-defined communities aggregating specific content. We choose Reddit as our source of jokes precisely for these communities, also known as *subreddits*. Subreddit users control the popularity of a post by contributing upvotes to it. Notably, these subreddits feature explicit forum rules that gatekeep the type and content of posts that are allowed to appear in the forum. Moderators carefully determine if posts abide by the forum rules and fit the ethos of their given subreddit. Due to their siloed nature and community-specific content, subreddits thus act as a natural manifestation of clearly separated content categories. In particular, the subreddit r/cleanjokes contains Wholesome Jokes, while r/darkjokes contains Dark Jokes, and r/DirtyJokes contains Dirty Jokes.

Scraping Reddit We scrape the above three subreddits to obtain our joke dataset. Unfortunately, the official Reddit API is insufficient for our purposes, due to the fact that it limits user access to only the 1000 most recent listings in a given subreddit. To circumvent this, we use the Pushshift.io API as an effective surrogate. Sending paginated requests, sleeping between queries, and continuously saving queried results, multiprocessing, we successfully obtain data across our three joke categories.

Data Processing After obtaining the scraped Reddit posts, we prune our dataset by removing posts that have a deleted or empty body of text, as well as duplicate posts. We accomplish this via RegEx to ensure robust removal.

Experiments

Metric To measure model performance, we calculate accuracy, precision, recall, and the micro-averaged F1 score.

Model	Accuracy	Precision	Recall	F1 (Micro)
bert-base-uncased	86.33%	83.91%	83.26%	83.38%
roberta-base	86.70%	84.24%	84.13%	84.17%
deberta-base	87.69%	85.32%	85.20%	85.22%
longformer-base-4096	82.58%	80.94%	78.55%	78.64%

Table 2: Results on the BERT-base, RoBERTA-base, DeBERTa-base, and Longformer models on multiclass classification of jokes.

Model	Accuracy	Precision	Recall	F1 (Micro)
twitter-roberta-base-offensive	75.84%	79.95%	71.05%	71.68%
deberta-base	92.88%	93.35%	91.86%	92.47%

Table 3: Results on the binary classification task of identifying whether or not a joke is offensive (labelled as dark or dirty) or inoffensive (labelled as not-a-joke or wholesome). We compare our finetuned version of DeBERTa with the state-of-the-art offensiveness detection model from the TweetEval becnhmark (Barbieri et al. 2020).

That is, we sum up the individual true positives, false positives, and false negatives of our system for different sets and compute the F1 score. We choose this metric in order to best reflect model performance on our uneven jokes distribution exhibiting severe class imbalance.

Models We finetuned pretrained large language models on our joke dataset. Specifically, we finetune BERT (110M params), RoBERTa (125M params), and DeBERTa (184M params). BERT (Devlin et al. 2019) is a bidirectional transformer that has set the state-of-the-art for many NLP tasks. RoBERTa (Liu et al. 2019) improves on BERT by pretraining on an order of magnitude more data. DeBERTa (He et al. 2021) improves on RoBERTa via disentangled attention.

Note the above models support context lengths of up to 512 tokens. However, 8.96% of our dataset contains examples with greater than 512 tokens, so we also evaluate the performance of the Longformer (102M params) (Beltagy, Peters, and Cohan 2020), which supports a context size of up to 4096 tokens.

Results Despite the Longformer being engineered to accommodate larger text contexts, it performs the worst out of the four architectures. Perhaps surprisingly, the BERT-based models do better despite truncating the input text to 512 tokens. We reason that this is largely because most texts in our dataset lie within 512-token limit, so the elongated context seen by the Longformer is mostly unhelpful and even potentially distracting. Our full results are shown in Table 2.

We also formulate an offensiveness detection task; since Dark and Dirty Jokes contain vulgar and inesnsitive topics, we consider them to be offensive content. Similarly, we consider Wholesome Jokes and News as inoffensive content. Effectively, we reduce the original 4-way humor subtype classification problem to a binary offensiveness detection problem. On this offensiveness detection task, our most accurate finetuned model (DeBERTa) notably outperforms the current state-of-the-art TweetEval offensiveness detection model (Barbieri et al. 2020) by a 17.04% increase in accuracy, as noted in Table 3.

Conclusion

We introduce a novel dataset and model for classifying humor subtypes of the adult variety. Furthermore, we demonstrate that our models can be successfully detect offensiveness in the context of jokes. Ultimately, we hope that our data and models open up further research at the intersection of Natural Language Processing and Computational Social Science, and that our models can be used to mitigate overly offensive humor in the appropriate settings.

References

Barbieri, F.; Camacho-Collados, J.; Espinosa-Anke, L.; and Neves, L. 2020. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (arXiv:1907.11692).

Taylor, J. M.; and Mazlack, L. J. 2004. Computationally recognizing wordplay in jokes. In *In Proceedings of CogSci 2004*.

Weller, O.; and Seppi, K. D. 2019. Humor Detection: A Transformer Gets the Last Laugh. In *EMNLP*.

Yang, D.; Lavie, A.; Dyer, C.; and Hovy, E. H. 2015. Humor Recognition and Humor Anchor Extraction. In *EMNLP*.