Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6136–6141
Marseille, 11–16 May 2020

The rJokes Dataset: a Large Scale Humor Collection

Orion Weller and Kevin Seppi
Brigham Young University
3361 TMCB, Campus Dr, Provo, UT 84604
{orionw, kseppi}@byu.edu

Abstract

Humor is a complicated language phenomenon that depends upon many factors, including topic, date, and recipient. Because of this variation, it can be hard to determine what exactly makes a joke humorous, leading to difficulties in joke identification and related tasks. Furthermore, current humor datasets are lacking in both joke variety and size, with almost all current datasets having less than 100k jokes. In order to alleviate this issue we compile a collection of over 550,000 jokes posted over an 11 year period on the Reddit r/Jokes subreddit (an online forum), providing a large scale humor dataset that can easily be used for a myriad of tasks. This dataset also provides quantitative metrics for the level of humor in each joke, as determined by subreddit user feedback. We explore this dataset through the years, examining basic statistics, most mentioned entities, and sentiment proportions. We also introduce this dataset as a task for future work, where models learn to predict the level of humor in a joke. On that task we provide strong state-of-the-art baseline models and show room for future improvement. We hope that this dataset will not only help those researching computational humor, but also help social scientists who seek to understand popular culture through humor.

Keywords: Computational Humor, Social Media, rJokes Corpus

1. Introduction

Understanding humor has been a large area of research for fields such as psychology, linguistics, and even health research (Wolff et al., 1934; Norrick, 1993; Knapp et al., 1992). It is an intrinsic aspect of our nature, yet often fickle and difficult for even humans to understand. Computational research in the field of natural language processing has focused on understanding (Hempelmann, 2008), classifying (Zhang and Liu, 2014; Chen and Soo, 2018), and generating (He et al., 2019; Valitutti et al., 2013) humorous text in order to better understand the structures that create humor (López and Vaid, 2017; Mihalcea and Strapparava, 2005).

We see this focus on humor illustrated in recent SemEval competitions, which include pun detection in English (Miller et al., 2017) in 2017 and a humorous headline regression task in 2020 (Hossain et al., 2019). However, these tasks (i.e. identifying humor in a dataset full of humorous and non-humorous instances), as well as other recent tasks in humor identification (Yang et al., 2015; Chen and Soo, 2018; Weller and Seppi, 2019) focus on small datasets with between 10-100k instances, usually on specific areas of humor, e.g. puns (Yang et al., 2015), Ted Talks (Chen and Soo, 2018), or TV shows (Purandare and Litman, 2006).

We see a similar rise of popularity in the area of humor generation, with recent models including GANs (Luo et al., 2019), RNNs (Yu et al., 2018), and retrieval models (He et al., 2019). Despite widely acclaimed advances in natural language generation from models such as GPT-2 (Radford et al., 2019), this success has not translated into the humor generation area, which has been a more difficult area of research. Here we find state-of-the-art (SOTA) work showing that humor generation is preferred to actual jokes 7 times out of 100 (Luo et al., 2019), with retrieval based generative models generating puns 31% of the time (He et al., 2019). One main concern of those papers is the lack of large corpora available for the training of these systems.

Having large corpora has been seen as a contributing factor to the success of other sub-fields, such as machine translation or language modeling, where they take advantage of datasets such as GigaWord, BooksCorpus, and the Common Crawl (Parker et al., 2011; Zhu et al., 2015). These datasets allow modern machine learning techniques to glean insight from the massive amounts of textual data they contain. However, in the areas of humor classification and generation we find much smaller datasets, due to the complexity of humorous natural language. Having substantially sized corpora would make large scale techniques feasible, instead of being a major obstacle to overcome.

Our contribution to this area consists of scraping, processing, and filtering the largest humor dataset to date, consisting of 573,335 jokes in English.[1] This dataset includes features such as labels for the body and punchline of the jokes, the level of humor present, and the date of posting. In this paper, we will provide analysis of basic features, yearly trends, and strong baselines for a new humor prediction task. We envision this dataset being used in pursuit of the following research areas:

- Understanding what creates humor: analyzing the structure and context of these jokes
- Examining reactions to public events, as measured through humor
- Using the corpus as a resource for training on different but related tasks, such as irony or satire
- Generating humor using this corpus as a resource for large scale training

2. Dataset Construction

The rJokes dataset was compiled from scraping the subreddit of r/Jokes. Information was retrieved from Red-

[1] Our code, analysis, and datasets are available at https://github.com/orionw/rJokesData.

Figure 1: Word Cloud of the most common phrases, after stemming and stop word removal

dit's servers about initial creation times, textual information, and upvotes (which we will interchangeably refer to as *scores*). The dataset contains information from January 2008 through December 2019, with varying rates of posts. We note that as of the current time of writing, the r/Jokes subreddit has over 17 million members and more than 1 million posts. This forum was chosen as our source of jokes because of several unique qualities: its (mostly) textual-only content, tags for body and punchlines, and crowd-sourced rating system. Although the subreddit includes some posts containing links or videos, the amount of those posts pales in comparison to the amount of purely textual postings. We found that most other joke forums or sites, especially those in other languages, contain large proportions of videos or memes. The rJokes thread provides a large community where reactions to jokes can be quantitatively measured, with humorous jokes being *upvoted* and disliked jokes being *downvoted*. Although this is not a perfect measure of humor, it does provide some level of humor contrast: a post with only one upvote is likely not as funny to the population as one with 10,000 upvotes.

We acknowledge that humor varies from person to person and from group to group. Thus, we do not claim that these jokes are representative for the human race, however, they are representative of a multi-million member group of Reddit users. We hope that this sub-sample of the population can provide insight that can help generalize humor to future groups.

In order to provide a clean resource for others to use, we employed several techniques to ensure the validity of our data. We first scraped the entire subreddit into a dataset with 1.1 million posts. The dataset was then processed, removing posts whose text had been deleted or removed due to the posting user or Reddit moderators. We further removed instances whose content contained pictures or

Statistic	Value	SD
Joke Token Count	239.78	501.57
Punchline Token Count	47.84	25.91
Body Token Count	191.93	502.55
Unique Tokens	256,619	N/A

Table 1: Mean and standard deviation for joke token counts. Unique tokens were calculated from the entire dataset

Percentile	0	10	25	50	75	90	100
Score	0	0	1	5	20	103	136353

Table 2: Score Percentiles for the rJokes dataset

videos in order to limit posts that included no real textual content. In total, this constituted removing 463,707 posts from the original dataset, most of which were marked as "deleted" or "removed." We note that we purposefully left in ungrammatical spelling, strange formatting, and other similar noisy data features, in order to preserve the structure and humor of the joke (i.e. newlines for emphasis or purposeful misspellings). We see in Table 2 the percentiles of the upvote score for the processed dataset, with around 20% of all jokes failing to earn a single upvote.

Some example joke instances[2] are the following (separation between body and punchline sections are indicated with the — symbol):

- Man, I was so tired last night; I had a dream I was a muffler... — and I woke up exhausted (276 upvotes).
- I told my teenage niece to go get me a newspaper... She laughed at me, and said, "Oh uncle you're so old. Just use my phone." — So I slammed her phone against the wall to kill a spider. (28315 upvotes).
- Just got my ticket to the Fibonacci conference! — I

[2]We do not endorse the jokes found in this dataset.

Year	1st	2nd	3rd
2008	American	Billionaire	Bush
2009	Elmo	Michael Jackson	Tickle
2010	Husband	America	Dad
2011	American	Indian	Mexican
2012	American	God	Mexican
2013	American	God	Mexican
2014	American	God	Mexican
2015	American	Mexican	God
2016	Trump	American	Clinton
2017	Trump	American	Russia
2018	Trump	American	God
2019	Trump	American	God

Table 3: The 1st through 3rd most mentioned entities in each year

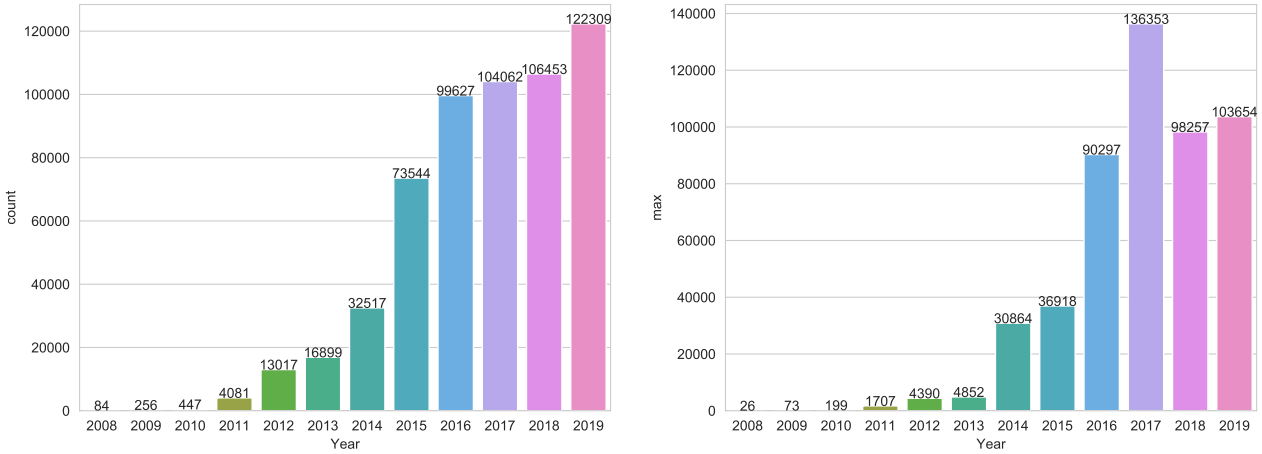

Figure 2: The left figure shows the number of posted jokes during each year, while the right plot shows the upvotes given to the highest rated joke posted in each year. For example, the highest rated joke that was posted in 2008 has only ever received 26 upvotes.

heard this year is going to be as big as the last two put together (20930 upvotes).

3. Dataset Analysis

In examining the basic dataset statistics, we find a large range in average joke length. Table 1 shows us that although punchlines tend to be short (around 48 tokens +- 26) the body of the joke has a larger standard deviation than its mean (192 tokens versus 503 tokens). This is due to the fact that some joke posts are humorous short stories, while others are simple puns. After counting all tokens in the dataset, we found more than 250,000 unique tokens.

In order to get a better sense of which words are being used, we visualize the most popular words in a Word Cloud (Figure 1), stemming and removing stop words. We see many common elements of jokes; in fact, one can almost see a joke just by reading the largest words, as it has some common phrases that one would associate with humor: "one day," "walks bar," etc.

We also see that the r/Jokes forum, which contains jokes that span from 2008 through 2019, has become increasingly popular. We see this illustrated in the left plot of Figure 2 as post submissions grow steadily until 2016. It seems that joke submissions have remained fairly stable since then, with an slight increase in posts in 2019. Because of the growing popularity trend, we find that the number of upvotes given to posts in the early years do not reach the levels of the those in the last five. For example, the most upvoted post in 2008 has only ever received 27 upvotes, whereas the top posted joke in 2019 has received more than 100,000 (right plot of Figure 2). This difference in upvotes is something that would have to be accounted for when designing tasks that attempt to predict the level of humor in a joke (see Section 3.1 for more details).

We also computed the top cited entities in joke posts during each year. We manually exclude generic joke names that consist of only one word, e.g. *John*, but allow non-generic entities like *John Deere* or *Trump*. We would imagine that different news events would affect the types of jokes being posted, perhaps for example, a U.S. election. Table 3 shows us the top three entities from each year, after removing stop words and generic names. We see that almost all years have Americans as a common joke theme, with several other ethnic groups being popular topics. However, in recent years Trump, Clinton, and Russia show up in the top three mentioned entities, perhaps alluding to the current U.S. news at the time. Interestingly, the elections of 2008 and 2012 do not seem to have the same effect, but this is likely due to the lack of community at the time (as the subreddit did not hit a stable level of posts until around 2016). We note that this paper looks only at entity trends over years, but similar analyses could be done with seasons, topics, or other parts of speech.

As Reddit is prone to all sorts of humor, we examine the sentiment scores of jokes over the years. We use Spacy's

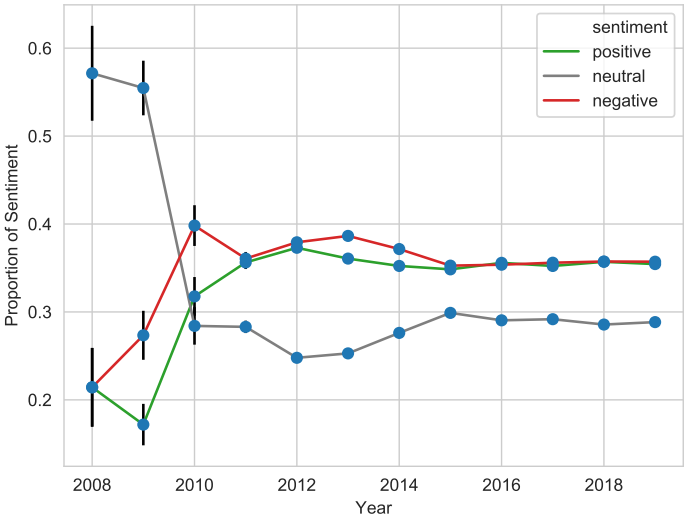

Figure 3: Proportions of sentiment for all jokes posted in a year. Vertical lines indicate standard error.

Vader (Hutto and Gilbert, 2014) as an off-the-shelf sentiment analyzer to examine these jokes because of its abilities to handle noisy data well, including emojis, acronyms, and slang. We evaluate each joke in its respective year, finding proportions and standard error for each sentiment level. These statistics are plotted in Figure 3.

Although the subreddit was more negative during its initial years, with a high amount of neutral jokes, it quickly converged to an even proportion of positive and negative sentiment as the number of postings increased (~36% each). The following are a few examples of jokes with their corresponding sentiment:

- Positive sentiment: *They laughed when I said I wanted to be a comedian... ...well they're not laughing now!*
- Neutral sentiment: *What gym did Socrates go to? The Y.*
- Negative sentiment: *Is there anything more annoying than an incomplete*

We note that most of the negative sentiment jokes were too explicit, in numerous ways, to include them in the paper. To see more examples of sentiment, please see our Github repository.

3.1. Humor Level Regression Task

A task similar to that of (Hossain et al., 2019; Weller and Seppi, 2019) can be done with this dataset, where a model predicts the level of humor found in the joke in order to examine what characterizes humor. However, due to Reddit's large scale and uneven distribution of upvotes, predicting the number of upvotes would be a sparse and difficult task. As such, we re-frame the task to predict the log amount of upvotes, reducing the scale from 0-136,353 down to 0-10. We note that we modified the log-transform slightly, leaving zero scores unchanged. Since the r/Jokes community started off slowly and did not receive a substantial amount of posts/upvotes per year until 2016, we remove all jokes posted previously in order to keep the score metric similar across years. This is a not a perfect solution, but is approximate enough, leaving 432,457 jokes for prediction.

We see that the transformed distribution in Figure 4 has a much nicer shape than the untransformed distribution, which could not plotted well, with higher scores tapering off in frequency similar to a Gamma or Zipf distribution. We thus propose this as a new task, predicting the log amount of upvotes from a Reddit post as a proxy for the level of humor in the joke. Predicting the log score will reduce noise in the dataset, allowing for jokes with nearly the same humor value to be ranked similarly (i.e. the humor difference between 100,000 upvotes and 110,000 upvotes is negligible even though the absolute scale is large).

We provide modern SOTA baselines on this task by fine-tuning three recent models: BERT, roBERTa, and XLNet (Devlin et al., 2018; Yang et al., 2019; Liu et al., 2019). We include roBERTa, despite it being a variant of BERT, because of its strong results from different training methods. These massive language models have shown strong results for transfer learning, as well as showing SOTA performance on tasks such as SQuAD (Rajpurkar et al., 2016) and GLUE (Wang et al., 2018). Due to these successes, we include them as baselines (results are with the large version of the models) in this task, using the Huggingface (Wolf et al., 2019) and PyTorch (Paszke et al., 2017) libraries. We fine-tune these models for 5 epochs, picking the checkpoint that performed the best on the dev set for final evaluation. We see the results on the test set below in Table 4: BERT/roBERTa perform similarly (scoring around 1.62 in RMSE), with XLNet under-performing in all metrics. We note that although these models provide solid results, there is still much room for improvement.

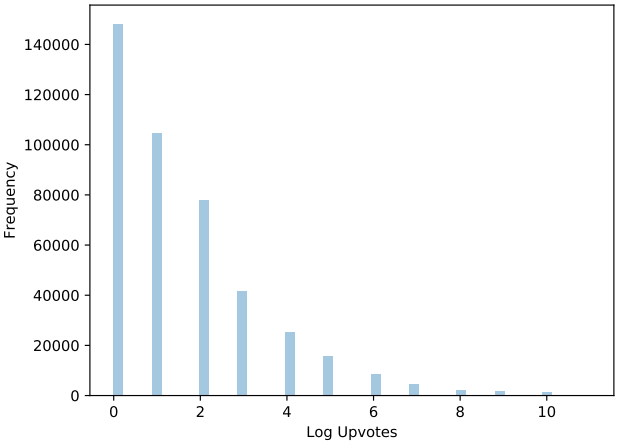

Figure 4: Histogram of the log upvote score for all jokes, used for the humor level prediction task

Model	RMSE	Pearson	Spearman
BERT	1.619	**0.471**	0.430
roBERTa	**1.614**	0.474	**0.435**
XLNet	1.739	0.457	0.411

Table 4: Baseline results on the humor level regression task for various language models. Language models were combined with a linear layer for the regression task. Best results in each column are in bold.

4. Conclusion

In this work we introduced and analyzed a novel dataset, the rJokes humor dataset, created from posted jokes over the past 10+ years and designed to enable large scale machine learning techniques for humor. We see that this language resource contains almost even percentages of negative and positive sentiment, humorous reactions to current events, and strong contrasts between non-humorous and humorous jokes. We hope that this dataset will alleviate some of the problems currently facing humor research by providing a dataset that is large enough to use with data hungry methods. We can envisage this dataset being a helpful resource to those seeking to understand the linguistic structure of humor, those examining how cultural events affect humor, and those attempting to generate humorous text.

5. Acknowledgements

We would like to thank Jordan Boyd-Graber for his insights on analyzing the distribution of upvotes.

6. Bibliographical References

Chen, P.-Y. and Soo, V.-W. (2018). Humor recognition using deep learning. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Jun.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*, Oct.

He, H., Peng, N., and Liang, P. (2019). Pun generation with surprise. 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 1734–1744, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hempelmann, C. F. (2008). Computational humor: Beyond the pun? *The Primer of Humor Research. Humor Research*, 8:333–360.

Hossain, N., Krumm, J., and Gamon, M. (2019). " president vows to cut¡ taxes¿ hair": Dataset and analysis of creative text editing for humorous headlines. *arXiv preprint arXiv:1906.00274*.

Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Knapp, P. H., Levy, E. M., Giorgi, R. G., Black, P. H., Fox, B. H., and Heeren, T. C. (1992). Short-term immunological effects of induced emotion. *Psychosomatic medicine*, 54(2):133–148.

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 preprint arXiv:1907.11692*.

López, B. G. and Vaid, J. (2017). Psycholinguistic approaches to humor. *The Routledge handbook of language and humor*, pages 267–281.

Luo, F., Li, S., Yang, P., Chang, B., Sui, Z., Sun, X., et al. (2019). Pun-gan: Generative adversarial network for pun generation. *arXiv preprint arXiv:1910.10950*.

Mihalcea, R. and Strapparava, C. (2005). Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 531–538, Stroudsburg, PA, USA. Association for Computational Linguistics.

Miller, T., Hempelmann, C., and Gurevych, I. (2017). Semeval-2017 task 7: Detection and interpretation of english puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68.

Norrick, N. R. (1993). *Conversational joking: Humor in everyday talk*. Indiana University Press.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English gigaword fifth edition ldc2011t07 (tech. rep.). Technical report, Technical Report. Linguistic Data Consortium, Philadelphia.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

Purandare, A. and Litman, D. (2006). Humor: Prosody analysis and automatic recognition for f*r*i*e*n*d*s*. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 208–215.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Valitutti, A., Toivonen, H., Doucet, A., and Toivanen, J. M. (2013). âlet everything turn well in your wifeâ: Generation of adult humor using lexical constraints. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 243–248.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Weller, O. and Seppi, K. (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 3612–3616.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Wolff, H. A., Smith, C. E., and Murray, H. A. (1934). The psychology of humor. *The Journal of Abnormal and Social Psychology*, 28(4):341.

Yang, D., Lavie, A., Dyer, C., and Hovy, E. (2015). Humor recognition and humor anchor extraction. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, September.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yu, Z., Tan, J., and Wan, X. (2018). A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1660.

Zhang, R. and Liu, N. (2014). Recognizing humor on twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge*

Management, CIKM '14, pages 889–898, New York, NY, USA. ACM.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.