/r/jokes

An analysis of the jokes festering in the internet’s most addictive corner

Kevin Trickey
November 6, 2019

Purpose

This report attacks the general question, “What makes a Reddit joke funny?” I analyze jokes posted to https://www.reddit.com/r/Jokes/, a subreddit with 17.4 million members at the time of writing. Computational methods are used to study word frequencies, joke lengths, subjects, sentiment analyses, and other aspects in relation to each joke’s Reddit “score,” the calculated difference between upvotes and downvotes.

Data Source

The primary data source from this project is a complete set of 194,553 jokes obtained from /r/Jokes on February 13, 2017 by Taivo Pungas (1). References to Reddit jokes throughout this report should be qualified by this date. The data was cloned from Pungas’ public Github repository at https://github.com/taivop/joke-dataset on November 1, 2019. I claim no ownership of this data and use it purely academically.

One questionable assumption I make throughout this report is that higher scores are correlated with funnier jokes. In reality, scores are also determined heavily by exposure, with some jokes receiving broader exposure than others; this effect is also multiplicative, since more upvoted submissions will be seen more by new users. “Funny” in this report should therefore be understood as “Widely upvoted” instead.

Content Warning

Reddit is not known for its content discretion. Some of the jokes analyzed and/or presented in this report may be offensive or unsavory, and this report will include occurrences of sexual content, profanity, and more. All jokes are included in the analysis set to minimize bias in analyses. Inclusion or presentation of a joke here does not imply condonance of its content.

Introduction

These are the five all-time funniest jokes on Reddit, as of 2017:

Joke	Score (Net Upvotes)
Breaking News: Bill Gates has agreed to pay for Trump's wall On the condition he gets to install windows.	48526
I found a place where the recycling rate is 98% /r/Jokes	45500
Steve jobs would have been a better president than Donald Trump. But its a silly comparison really, its like comparing apples to oranges.	39570
My girlfriend told me to take the spider out instead of killing it. We went and had some drinks. Cool guy. Wants to be a web developer.	36421
For every upvote this gets, my girlfriend and I will try one thrust of anal sex. Please don't upvote. Her strap-on is huge.	35772

So, does Reddit just find Donald Trump hilarious? Or is it the wordplay on computers and technology that propels these jokes to the top? And what about the last one—is Reddit just full of schadenfreude users who delight in large strap-ons?

These are questions inspired by a qualitative glance over just 5 Reddit jokes, but they are all at least partially answerable with computational methods (post-hoc statistics aside) with the added advantage of unlimited throughput.

The following report is arranged in a series of sections, each structured around answering one or more bulleted questions about Reddit jokes. Sections and topics are rather arbitrarily determined by me, but hopefully possess some interest value to you.

Trump Jokes

The introductory jokes above suggested almost overwhelmingly a fascination by Reddit jokers over Donald Trump. The first questions, then, are these:

- *Are Trump jokes funnier than average?*
- *Is the orange/Trump combination funnier than Trump alone?*

To answer these I performed a couple of two-sample t-tests, using Welch t-statistics instead of the classical test due to vastly different sample sizes between groups. Although score distributions on jokes are heavily skewed, the large sample sizes still validate use of the t-test.

It turns out that, on average, jokes using the word “Trump” are funnier than non-Trump-related jokes (mean score 258 vs. 117, p=0.0016). However, average Trump-joke scores are not significantly higher when they include the word “orange” compared to non-orange Trump jokes (mean score 1074 vs. 237, p=0.27). At the time of data collection, only 57 jokes had been submitted with both “Trump” and “orange” in them, limiting the statistical power for this second test. It is also true that because these tests were directly inspired after observation of a data subset (the introductory 5 jokes), those jokes should be excluded from proper analyses; however, excluding them does not change either conclusion.

- *Okay, if Trump is funny, is that just because he’s the president? Or is he a particularly funny president?*

Well, I repeated the test using “Obama” jokes from the Reddit dataset, and Obama jokes are in fact *not* statistically funnier than non-Obama jokes (mean score 149 vs. 118, p=0.45). As expected, the Trump vs. Obama joke contest goes in Trump’s favor, though also not super significantly (p=0.073). Perhaps the sample size is telling, though: only 588 Obama jokes were submitted since the page’s conception in 2008, while over 2,000 Trump jokes appeared in just the 2-year span 2015–2017.

These results should be taken with some grains of statistical salt, because they are primarily driven by outliers in a heavily-skewed distribution of joke scores likely determined by Reddit algorithms and other strange features of social networks.

Table 1: Comparison of Trump jokes and Trump/orange jokes.

A: Trump Jokes	**# in sample**	**Mean Score**	**Std. Err.**	**Median Score**
Trump Jokes	2,239	258	44.6	1
Not Trump Jokes	192,314	117	2.08	3
B: Trump/Orange Jokes	**# in sample**	**Mean Score**	**Std. Err.**	**Median Score**
Trump/Orange Jokes	57	1,074	752	3
Trump/Not Orange Jokes	2,182	237	41.3	1

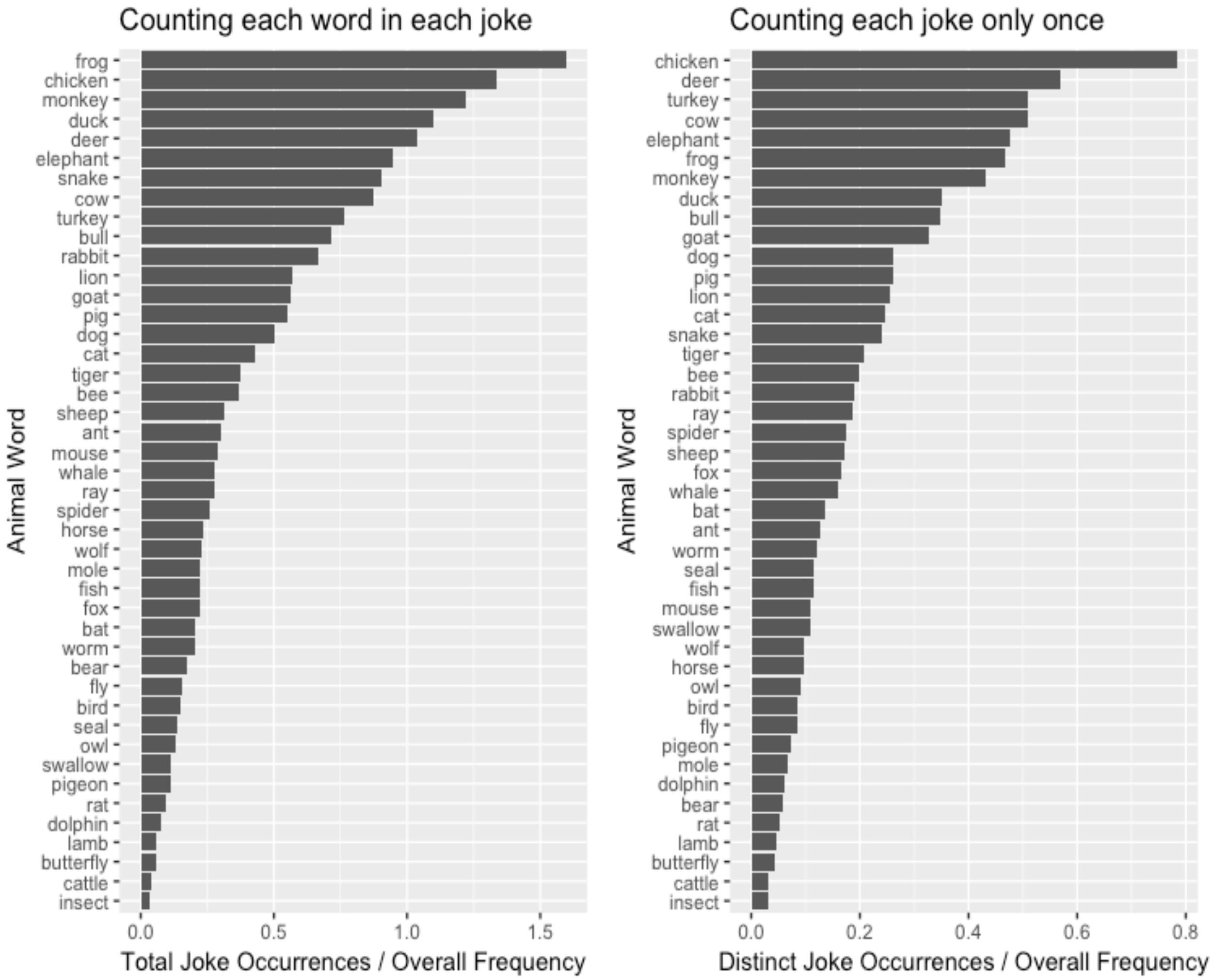
Animals

This one was inspired by the classroom revelation that ducks are supposedly the funniest animals around. But does Reddit agree?

- *What animals appear more often than you'd expect among Reddit jokes?*

I generated a list of 374 common animals, *Aardvark* to *Zorilla*,[1] by combining lists found online (2, 3). I counted how many times each animal occurred within the dataset of Reddit jokes, then divided by the total frequency of the word in the British National Corpus (4), limited to the words that were included in the free dataset (i.e. the most common ones—I don't think "aardvark" *or* "zorilla" made it). Chickens seem to be pretty common joke animals, as do other fowl; deer and frogs make surprise appearances high up the list; and elephants and monkeys round out the top six or so places.

[1] Otherwise known as a *striped polecat*, resembling a skunk.

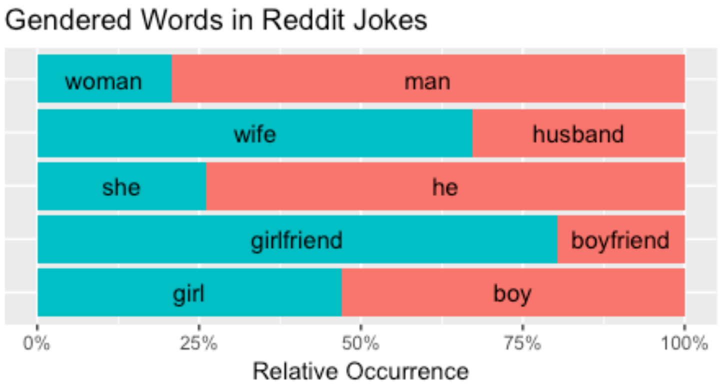


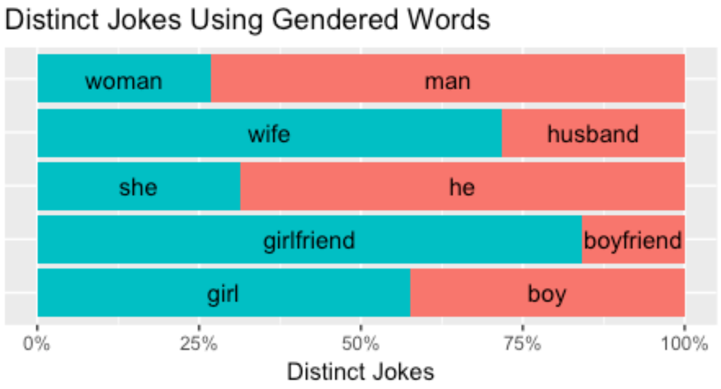
I should note that some of these words, like “swallow,” “fly,” and “ray,” are probably mostly used for their other, non-animal meanings, and I’ve done nothing to differentiate between usages.

Gender in Reddit Jokes

Gender and perceptions of humor is a widely researched subject. Studies on the topic have found people to associate the ideal humorist with a male persona (5), while women’s humor has been more often marginalized both in everyday life and theoretical models of humor (6). “The Humor Gap,” as it is sometimes termed in popular science (7), would be an interesting phenomenon to detect in this corpus of jokes.

- *Are the subjects of jokes more often female- or male-gendered?*





I selected five pairs of gendered words that might be expected to occur equally in conversation: woman/man, wife/husband, she/he, girlfriend/boyfriend, and girl/boy. I counted the number of occurrences each word made among all the Reddit jokes, and I also tracked how many distinct jokes the words appeared in. “Man” and “he” were more popular than their female counterpart pronouns, while “boy” and “girl” appeared with similar frequencies to each other. Notably, the “wife/husband” and “girlfriend/boyfriend” pairs both skewed substantially toward the female-gendered terms.

At the risk of reading too much into preliminary data, we could propose that the relative frequencies of “man” and “he” suggest a male “base-case” for non-gender-dependent jokes (“a man walked into a bar…”), while gendered terms like “wife” and “girlfriend” that suggest intimate relationships—perhaps hinging on the gender of the joke’s subject—tend to be female-gendered instead. It is possible that “girl” and “boy,” representing a more childish or innocent character, are more balanced between the two sides. Intriguingly, the relative frequencies of *every* word-pair shift toward the female-gendered side when looking at distinct jokes instead of total words, implying that the jokes that do include gendered words tend to repeat the male form more than the female form. Interesting follow-up analyses could include comparing sentiment analyses on gendered jokes, the frequencies of possessive or other modifiers like “my,” “cute,” or “baby,” and the overlap of genders and their relationships within jokes.

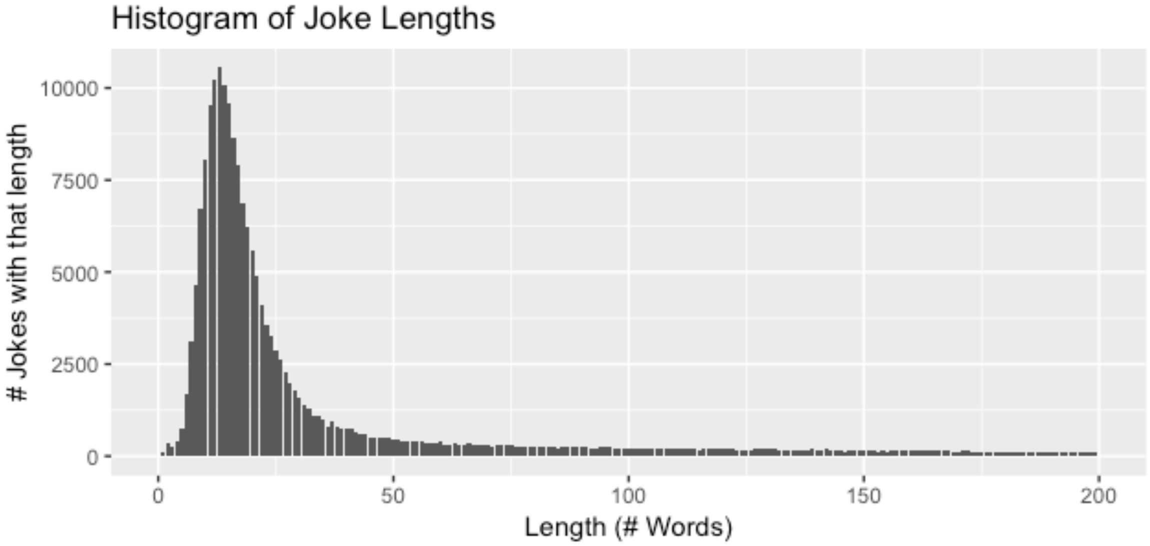
Does Size Matter? Lengths of Reddit Jokes

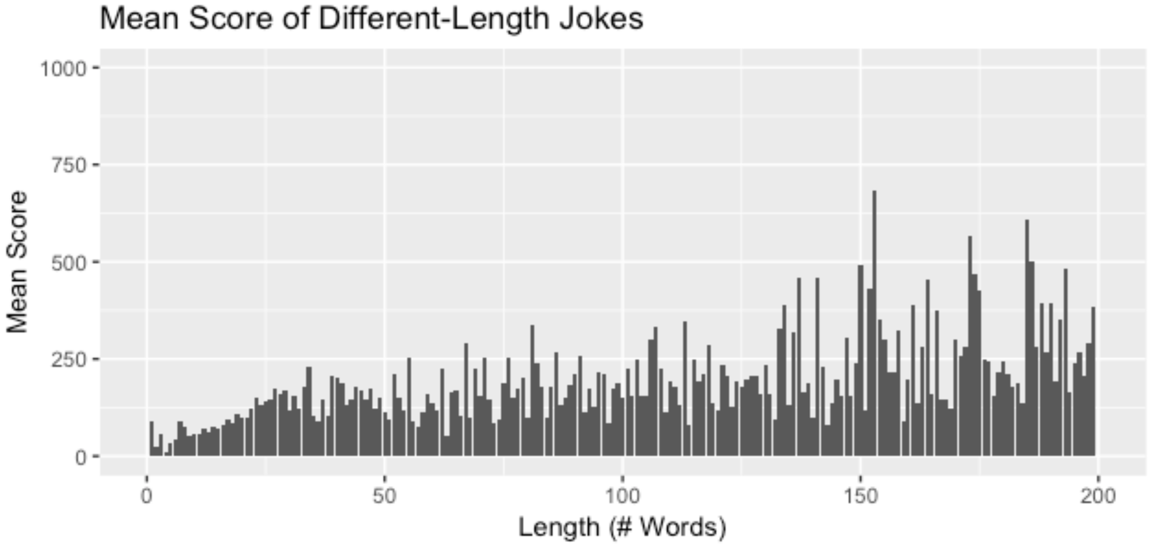
There isn’t one formula for the length of a good joke, and I was able to find surprisingly little research on joke lengths. Perhaps the closest proxy might be humor complexity, on which there has been a fair amount of research finding optimal intermediate levels of complexity (8). Perhaps the simple length of a joke is too simple for other people to report, but it’s an easy thing to calculate, so…

- *What is the optimal length, in number of words, for a joke?*

In concept, this idea is pretty straightforward to test: plot each joke on 2 axes, its length and its score, and see if there’s a correlation or any other pattern. Sadly, both distributions—joke length and joke score—are extremely heavily skewed right, with long tails in the higher numbers but a vast majority at very low values; in consequence, the relationship is hard to visualize productively without doing something funky like cutting the data in weird ways.

The histogram of joke lengths seems pretty ideal: somewhat bell-shaped with a very long tail extending to one joke with 7,546 words, almost 6,000 more than this report (in fact, that particular joke is *Nate the Snake* (9), a 10,747-word story in its complete form). The bottom graph shows that jokes tend to do increasingly well until they get to about 25 words, where the data becomes pretty noisy due to small sample sizes. The initial increase is interesting, though, suggesting that the optimal joke length may be around 25 words—just enough to require some investment, but not too much to bore the reader.





Sentiment Analysis

Sentiment analysis is a neat machine learning technique to computationally quantify the positive or negative emotion associated with particular bodies of text. It is frequently used in research on tweets (10, 11), though I've also seen people do sentiment analysis projects on song lyrics (12), news articles (13), and the *Harry Potter* series (14). Well, why not jokes? This can be my contribution to the sentiment analysis pile-on.

- *What types of jokes are the most emotionally charged?*
- *Does sentiment magnitude correlate with humor?*

I linked Reddit's joke data with a dataset of 2,477 words and their average sentiments published by Finn Årup Nielson (15). Neutral words and words that did not appear in the sentiments data were assigned a score of 0, while affective words were assigned an integer from –5 to +5 based on negative or positive sentiments they typically convey. For reference, "breathtaking," "hurrah," "outstanding," "superb," and "thrilled" are the only +5-rated words in the set; "cocksucker" and "prick" are among the 16 words rated –5. Importantly, this method of single-word-based sentiment analysis does not catch sarcasm or irony, which could be particularly relevant to jokes. More complicated analysis tools are available but, well, more complicated.

Here are the 5 most positive and negative sentimented jokes, according to this method. Please note that they may not be the most tactful:

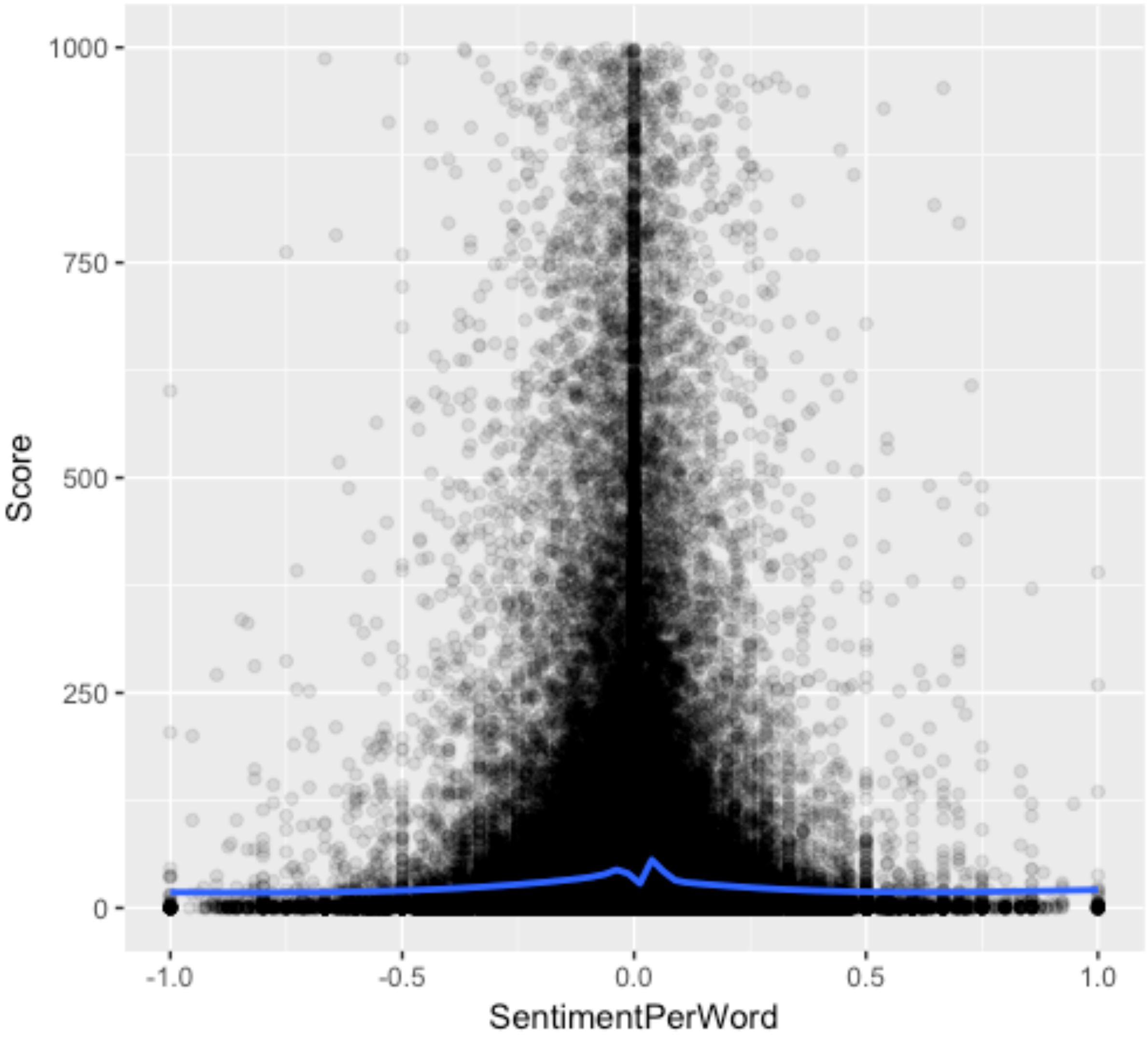
Total Sentiment	Sentiment per Word	Reddit Score	Joke
456	1.44	4	**What did the boy centipede say to the girl centipede passing by?** Nice legs. Nice legs. Nice legs. Nice legs. … *[continues]*

312	1.87	0	**Conversation between a girl and a boy** Girl: I Love You. Boy: LoL Girl: I Miss You I Miss You.Boy: LoL LoL LoL... Girl: I can give my life for you Boy: LoL LoL LoL LoL LoL ... Girl: I cant live without you. Boy: LoL LoL LoL LoL LoL LoL LoL .... *[continues]*
210	0.33	2	**A computer told me that all squares are rectangles...** I responded, "While true, it's important to remember that the converse is not true: i.e., not all rectangles are squares. *[repeats many times]*
152	0.38	2	**A few pickup lines to use on the ladies…** I like my women like I like my mattresses. Immobile and under a sheet I like my women like I like my contact lenses. Transparent and shallow I like my women like I like my four horsemen. Four of them, willing to ride *[continues; gets even worse]*
132	0.16	0	**A man walks in a bar…** He says to another man "I'm a paradox! Do you want to hear a joke?" The man is weirded out but says yes. "Ok it goes like this. A man walks in a bar... He says "I'm a paradox! Do you want to hear a joke?" The man is weirded out but says yes. "Ok it goes like this *[continues]*

Total Sentiment	**Sentiment per Word**	**Reddit Score**	**Joke**
–4,991	–2.98	0	**What to post on the R/Donnald to get banned?** Donald Trump is a Loser Loser Loser Loser Loser *[continues]*
–274	–0.26	0	**THE SHIT LIST** The Shit List: The Ghost Shit The kind where you feel shit come out, see shit on the toilet paper, but there's no shit in the bowl. The Clean Shit The kind where you feel shit come out, see shit in the bowl, but there's no shit on the toilet paper. *[continues]*
–202	–0.41	24	**This is a shit post.** *[same content as previous row]*
–180	–0.61	9	**The most functional word in English language is...** The most functional word in the English language is... Shit. That's right, shit! Consider this: You can be shit faced, shit out of luck, or have shit for brains. With a little effort, *[continues]*
–171	–0.04	3	**Drunk Jokes**

			(1)Two men are drinking in a bar at the top of the Empire State Building. *[continues for a long time]*

Okay, so sentiment analyses seem to be good at picking terrible jokes—or, in most cases, malformed jokes, errors, or other long strings of repetition. Though disappointing, it’s not surprising: this method is additive with every word, so long strings of repeated words will be more biased toward that word’s positive or negative sentiment. Dividing the total sentiment by the joke’s number of words solves this



problem, but it yields instead lots of one-word jokes and short profane slurs, which isn't much better.

Is there a sweet spot, though? Some ideal range of sentiments and lengths such that the resultant jokes might actually be good? Maybe. The plot doesn't seem to show any strong correlation between sentiment and score, although more neutral jokes may have ever-so-slightly higher scores, with a small dip in between. Slightly positive jokes seem to be a little funnier than slightly negative jokes, but not by a whole lot.

Sentiment analysis on Reddit jokes seemed like a cool idea, but without more advanced procedures I'm not sure a whole lot can be gleaned from these results. Perhaps it would be more interesting, given more time, to combine sentiment analysis with some of the earlier studies, like Trump jokes or gendered jokes, to see if the /r/jokes community tends to feel positively or negatively toward a particular subject.

Availability of Code

Code that produces the information displayed in this report is available at https://gitlab.com/kstrickey/jokes.

References

1. T. Pungas, *taivop/joke-dataset* (2019) (November 6, 2019).

2. , Alphabetical list of Animals | Alpha Lists (November 5, 2019).

3. , Animal Names: Types of Animals with List & Pictures. *7 E L* (2017) (November 5, 2019).

4. , Read-me for Kilgarriff's BNC word frequency lists (November 5, 2019).

5. S. Tosun, N. Faghihi, J. Vaid, Is an Ideal Sense of Humor Gendered? A Cross-National Study. *Front. Psychol.* **9** (2018).

6. H. Kotthoff, Gender and humor: The state of the art. *J. Pragmat.* **38**, 4–25 (2006).

7. C. Nicholson, The Humor Gap. *Sci. Am.* https:/doi.org/10.1038/scientificamericanbrain0512-66 (November 6, 2019).

8. R. I. M. Dunbar, J. Launay, O. Curry, The Complexity of Jokes Is Limited by Cognitive Constraints on Mentalizing. *Hum. Nat.* **27**, 130–140 (2016).

9. , Nate the Snake (November 6, 2019).

10. A. Kumar, T. M. Sebastian, Sentiment Analysis on Twitter. **9**, 7 (2012).

11. , Twitter Sentiment Analysis using Python. *GeeksforGeeks* (2017) (November 6, 2019).

12. , Sentiment Analysis on lyrics of popular music artists // UnInhibited Zombie // (November 6, 2019).

13. J. Yip, Algorithmic Trading using Sentiment Analysis on News Articles. *Medium* (2018) (November 6, 2019).

14. G. Rafferty, Basic NLP on the Texts of Harry Potter: Sentiment Analysis. *Medium* (2019) (November 6, 2019).

15. F. Å. Nielsen, A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *ArXiv11032903 Cs* (2011) (November 6, 2019).