Oh my God: The linguistic influence of the TV series Friends

CSS Lab Holiday Paper Series 2020*

David Garcia, Hannah Metzler, Janna Lasser, Max Pellert, Anna Di Natale
Graz University of Technology

December 18th, 2020

Among English speakers, the phrase "Oh my God" might be the most frequent way to express surprise with soft blasphemy. In this first paper of the Computational Social Science Lab Holiday Paper Series, we look into how "Oh my God" became popular with the TV series Friends by analyzing the scripts of the series and other datsets of spoken and written expression across media and languages. We also study the demographics of who says it through Twitter data and explore how AI language models have learned to say it.

The expression "Oh my God" is specially prevalent in US television as part of the valleyspeak sociolect (Wikipedia 2020) (see any episode of "Keeping up with the Kardashians" to have an idea). Beyond the OMG expression, valleyspeak is characterized by the frequent use of *uptalk*, a speech pattern in which sentences end in a question-like tone, *vocal fry*, another speech pattern in which utterances end in a vibrating sound similar to what a goat makes when queezed, and the extreme use of the word *like* as a filler word. While original examples of valleyspeak originate in the San Fernando Valley in California, it can now be heard across the US and specially in mass media.

"Oh my God" is the most characteristic reaction of the characters of Friends, which is one of the most successful sitcoms in history, even more than 15 years after it ended (Economist 2019). The Honest Trailer of the series points out how frequently the characters of the series react with this phrase. For example, Rachel says it several times in just one scene in reaction to the sudden appearance of a pidgeon in her kitchen. The phrase is so typical of the series that it has even motivated a data science blog post about which Friends characters use it the most (Loscalzo 2018). Beyond this phrase, Friends is becoming a common example for popular data science analyses, including social network analysis (Albright 2015; Sahakyan 2019), finding the most popular character (Sohoye 2019), and of course, sentiment analysis (Bhattacharyya 2019).

Our aim with this article is to understand the popularity and meaning of "Oh my God", especially in relation with the TV series Friends. We will start by analyzing the scripts of Friends, comparing the use of "Oh my God" in Friends with contemporary TV shows and movies. We then continue by analyzing the expression through Google Books, inspecting what could have been the role of Friends in the use of the phrase and how it compares to similar phrases in other languages. We then analyze the current use of the phrase in social media through the analysis of a Twitter dataset, paying special attention to its use across genders and states in the US. Finally, we explore how AI language models like BERT and GPT-2 have learned the phrase and which meanings we can associate with the phrase through these models. The code, data, and detailed results of all these analyses can be found online in our github repository.

*The CSS Lab Holiday Paper Series is a tradition in which we publish a white paper about a fun research topic without thinking about peer-review, journal interests, grant agencies, nor anything else besides our passion for research. To learn more about the Computational Social Science Lab at the Graz University of Technology, visit csslab.at

1 The one with the "Oh my God"

We dowloaded the scripts of all Friends episodes from this Github repository and processed the text, converting it to lower case and matching the regular expression `"oh[:punct:]*\\s*my[:punct:]*\\s*[:punct:]*(god)"`. This regular expression counts the instances of "Oh my God" with a soft rule that allows various punctuation and spaces in between words, but we do not count other variations such as "Oh my fricking God". We denote the count of matches of the regular expression as *OMG*, defining this way the unit of our analyses. In total we found 1039 OMG the 229 episodes of the series (double episodes are merged into single files). After counting words with the tm package (Feinerer and Hornik 2020; Feinerer, Hornik, and Meyer 2008), we found that Friends has 1476.8 instances of OMG per million trigrams (i.e. sequences of three words).

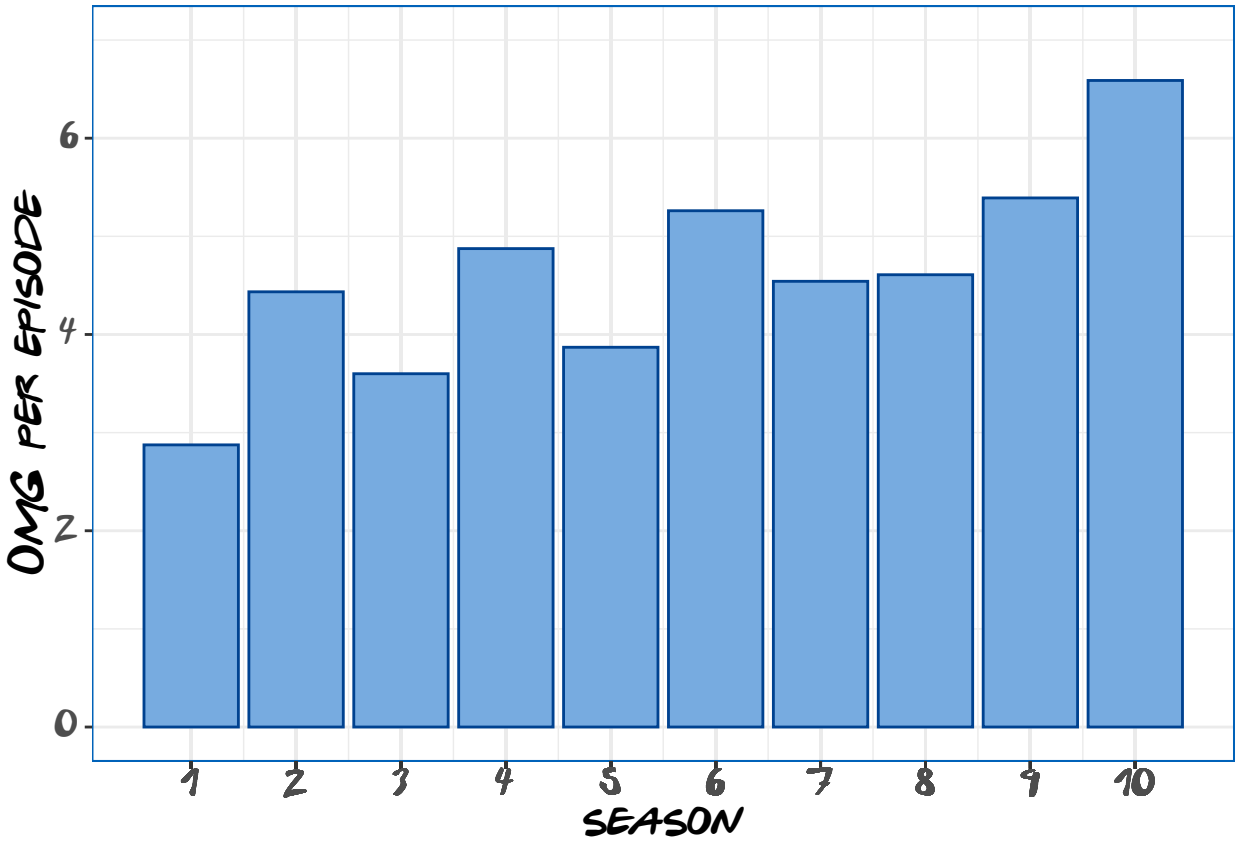


Figure 1: OMG per episode in each season of Friends

Figure 1 shows the OMG per episode in each season of the series, There is a tendency to more OMG over the lifetime of the series, from less than 3 OMG per episode in the first season to more than 6 in the last one. To compare Friends to contemporary TV shows and movies we applied the same analysis to the 2020 edition of the Corpus Of Contemporary American English (COCA) (Davies 2010). In our github repository we share the final yearly counts of our analyses, as we are not allowed to share the raw text of the corpus by the terms to access it.

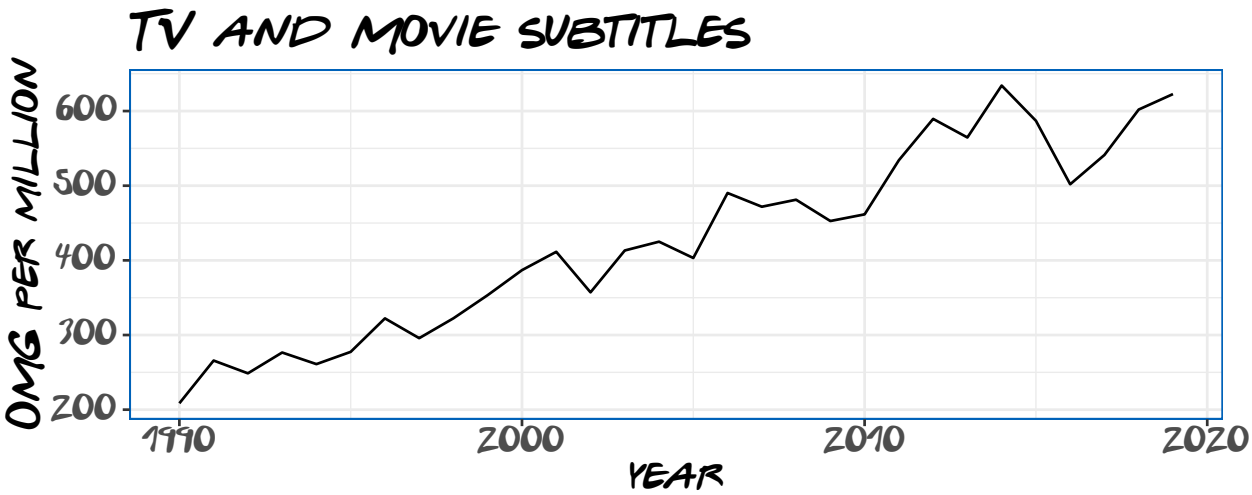


Figure 2: OMG per million trigrams in TV and Movie subtitles of the COCA corpus.

Figure 2 shows the yearly frequency of OMG per million trigrams in the TV and movie subtitles part of the COCA corpus. While it is clear that the phrase is very popular in US entertainment and its popularity is increasing, Friends had many more OMG than contemporary TV and movies. Friends had approximately 1477 OMG per million trigrams, which is 4.26 times what you would find on the typical TV shows and movies between 1994 and 2004 (300-400 OMG per million).

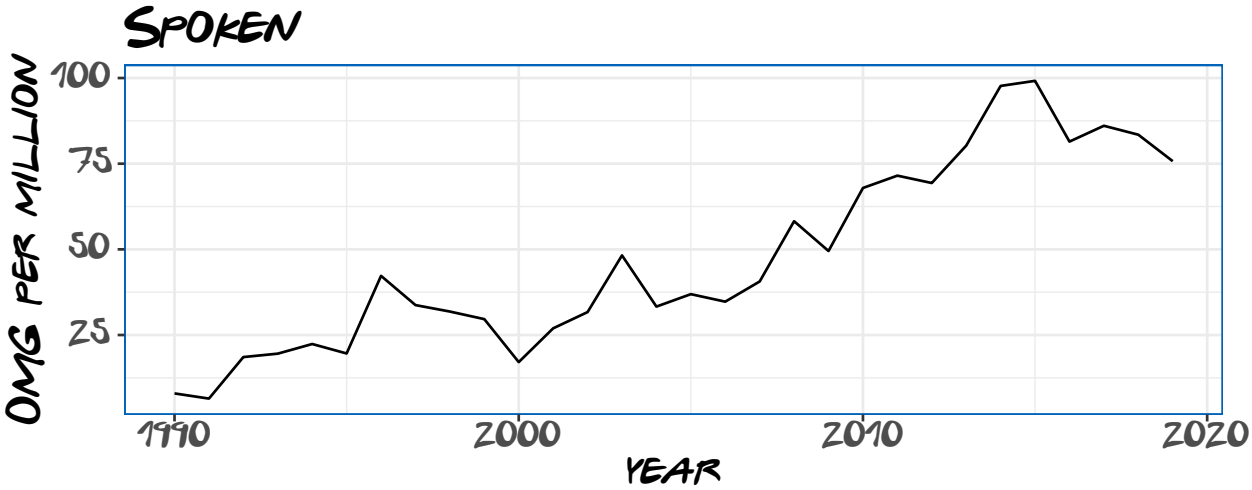


Figure 3: OMG per million trigrams in spoken transcripts of the COCA corpus.

Figure 3 shows the yearly frequency of OMG per million trigrams in the transcripts of unscripted spoken TV shows of the COCA corpus. While the frequency is about 8.45 times higher in TV and movie subtitles than in these spoken transcripts, the increasing tendency is present too. Although the source of both corpora is TV, the spoken transcripts come from talk shows and other kinds of unscripted shows. Mass media scripts seem to use "Oh my God" as a way to emphasize and elicit surprise reactions in the audience, which does not happen so naturally in live unscripted television.

2 The one with all those books

The frequency of OMG in both scripted and unscripted spoken language in TV shows and movies has been steadily increasing since the 1990s, but to test if Friends might have affected the tendency to use the phrase, we need to look at a longer time period. Inspiried by the trend previously observed in (Loscalzo 2018), we study the frequency of OMG in Google Books, one of the most comprehensive records of human written communication over several centuries and languages (Michel et al. 2011)[1]. We use the ngramr R package (Carmody 2020) to query the 2019 dataset of English fiction books to avoid known problems with non-fiction texts (Pechenick, Danforth, and Dodds 2015). We also only analyze frequencies since 1900 to avoid Optical Character Recognition errors like mistaking a long *s* for an *f* in 1600s and 1700s texts.

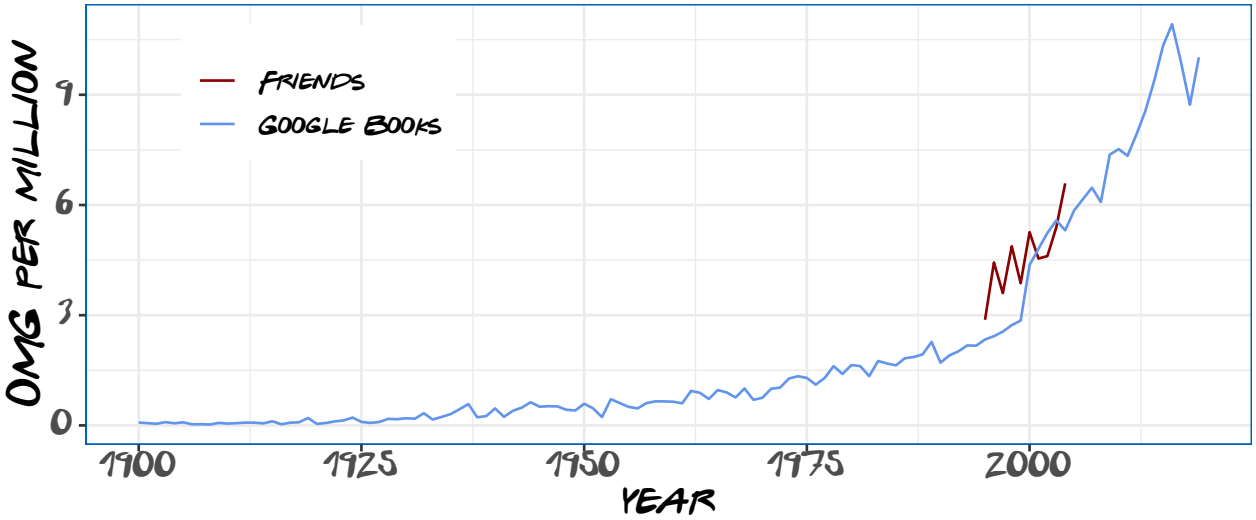


Figure 4: OMG per episode in Friends each season and yearly OMG per million trigrams in Google Books.

Figure 4 shows the frequency of OMG per million trigrams in Google Books with the number of OMG per episode of the ten seasons of friends superimposed. The frequency of OMG in books has consistently increased over more than a century. This rate apparently accelerated after Friends came out. Could Friends be responsible for additional growth in the frequency of OMG in books?

We tested the effect of Friends on the frequency of OMG in books with a causal inference design. First, we fit an ARIMA time series model of the log-transformed OMG frequency up to 1994, the year of the first season of Friends. We choose the model size by minimizing the BIC with the *auto.arima* function of the forecast R package (Hyndman et al. 2020; Hyndman and Khandakar 2008). The resulting model has an R^2 of 0.92 for log-transformed values and of 0.67 for raw frequencies, with a moving average term and a positive drift term that explains the exponentially increasing shape of Figure 4.

[1]We would say that our Google Books analysis an example of *culturomics*, but the term seems to be used nowadays more often to talk about gut bacteria than about culture product analysis

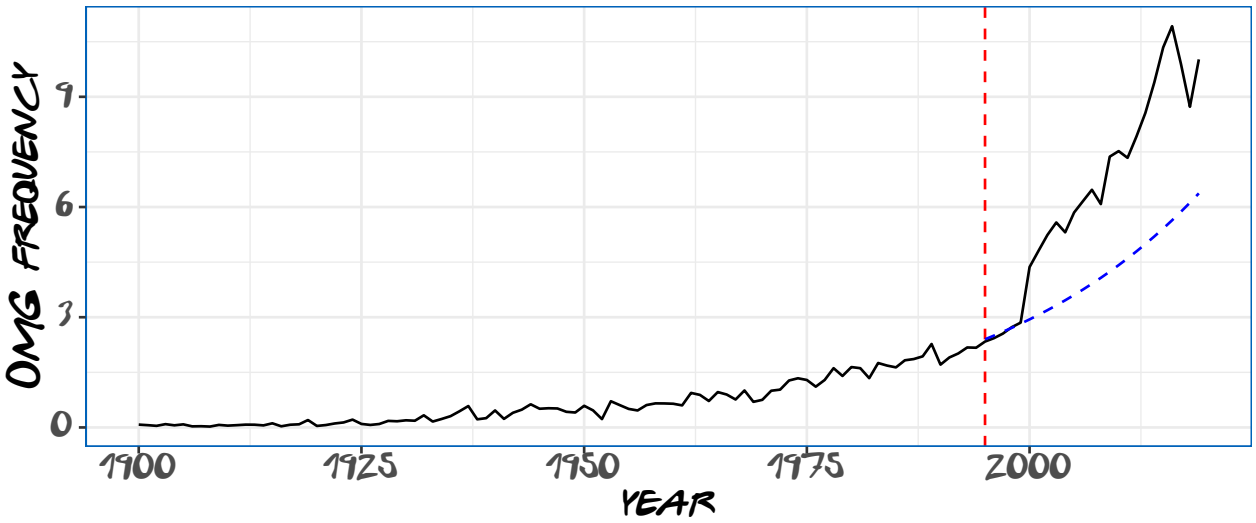


Figure 5: OMG per million trigrams in Google books (solid black) and time series model prediction (dashed blue). The vertical red line indicates the year of the first season.

Figure 5 shows the forecasted values of the model since 1995 versus the empirical frequency of OMG in Google Books. From 1995, there has been an additional 57.59% OMGs compare to what the null model predicts. This is an impressive surplus of OMG after Friends started, but we should not fool ourselves with such causal inference arguments. This could have been Friends or a coincidence with another influential source, for example South Park's "oh my God, they killed Kenny". What is clear is that Friends captured an increasing frequency of OMG and that, after the series, this frequency in books has grown even faster.

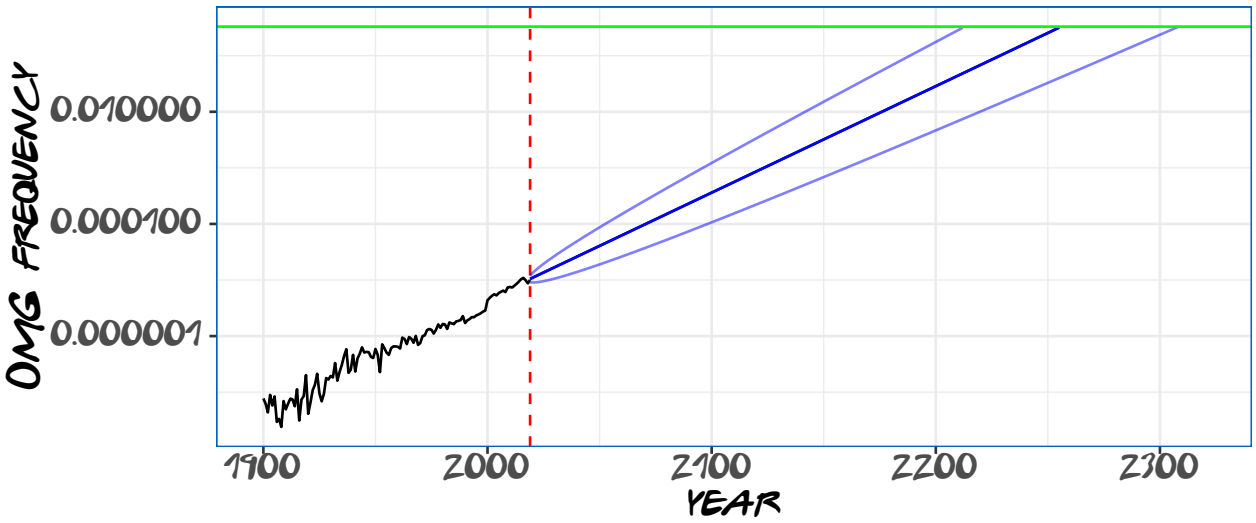


Figure 6: Model projection of OMG frequency per trigram from 2000. Light blue lines show 90% prediction intervals. The horizontal green line marks the 33% frequency line. The vertical dashed red line shows where predictions start.

What does this tell us about the future of the phrase "Oh my God"? To forecast trends, we apply the Complexity Science™ method of fitting straight lines to log scales. Figure 6 shows the empirical frequency in Google Books with a vertical logarithmic scale and the predictions of the model three centuries into the future after fitting it with the data since the year 2000. This suggest that approximately by the year 2256, all text in books will be composed of entirely "Oh my God". Our data-driven insights allow us to code a

model that writes English fiction literature like the one we will see in the 23rd century:

```
for (i in (seq(1,150)))
{
  cat("Oh my God")
  rnd <- runif(n=1, min=0, max=1)
  if (rnd <0.4)
    cat("! ")
  if ((rnd >=0.4) & (rnd <0.8))
    cat(". ")
  if (rnd >=0.8)
    cat("? ")
  }
```

An example of the output of the model can be found here:

Oh my God! Oh my God. Oh my God! Oh my God? Oh my God! Oh my God. Oh my God. Oh my God! Oh my God? Oh my God. Oh my God? Oh my God. Oh my God! Oh my God. Oh my God. Oh my God? Oh my God! Oh my God! Oh my God. Oh my God? Oh my God. Oh my God! Oh my God. Oh my God! Oh my God! Oh my God. Oh my God! Oh my God. Oh my God. Oh my God. Oh my God! Oh my God. Oh my God! Oh my God? Oh my God! Oh my God. Oh my God. Oh my God! Oh my God? Oh my God. Oh my God? Oh my God. Oh my God! Oh my God. Oh my God. Oh my God? Oh my God! Oh my God! Oh my God. Oh my God? Oh my God. Oh my God! Oh my God. Oh my God! Oh my God! Oh my God. Oh my God! Oh my God. Oh my God. Oh my God. Oh my God! Oh my God. Oh my God! Oh my God? Oh my God! Oh my God. Oh my God. Oh my God! Oh my God? Oh my God. Oh my God? Oh my God. Oh my God! Oh my God. Oh my God! Oh my God. Oh my God! Oh my God? Oh my God! Oh my God. Oh my God. Oh my God! Oh my God? Oh my God. Oh my God? Oh my God. Oh my God! Oh my God. Oh my God! Oh my God! Oh my God. Oh my God? Oh my God. Oh my God! Oh my God. Oh my God! Oh my God! Oh my God. Oh my God! Oh my God. Oh my God. Oh my God. Oh my God! Oh my God. Oh my God! Oh my God? Oh my God! Oh my God. Oh my God. Oh my God! Oh my God? Oh my God. Oh my God? Oh my God. Oh my God! Oh my God. Oh my God! Oh my God! Oh my God. Oh my God? Oh my God! Oh my God. Oh my God! Oh my God! Oh my God. Oh my God. Oh my God.

3 The one with the other languages

Given the international success of Friends, could this be happening in other languages too? We use the fact that Google Books covers eight languages to explore this possibility. We tried to find the closest equivalents of "my God", dropping the "oh" for better comparability across languages: "Dios mío" in Spanish, "mio Dio" in Italian, "mon Dieu" in French, and "mein Gott" in German. Adviced by native-speaking friends of us, we also included Chinese, Russian, and Hebrew versions, even though none of the authors speak these three languages[2].

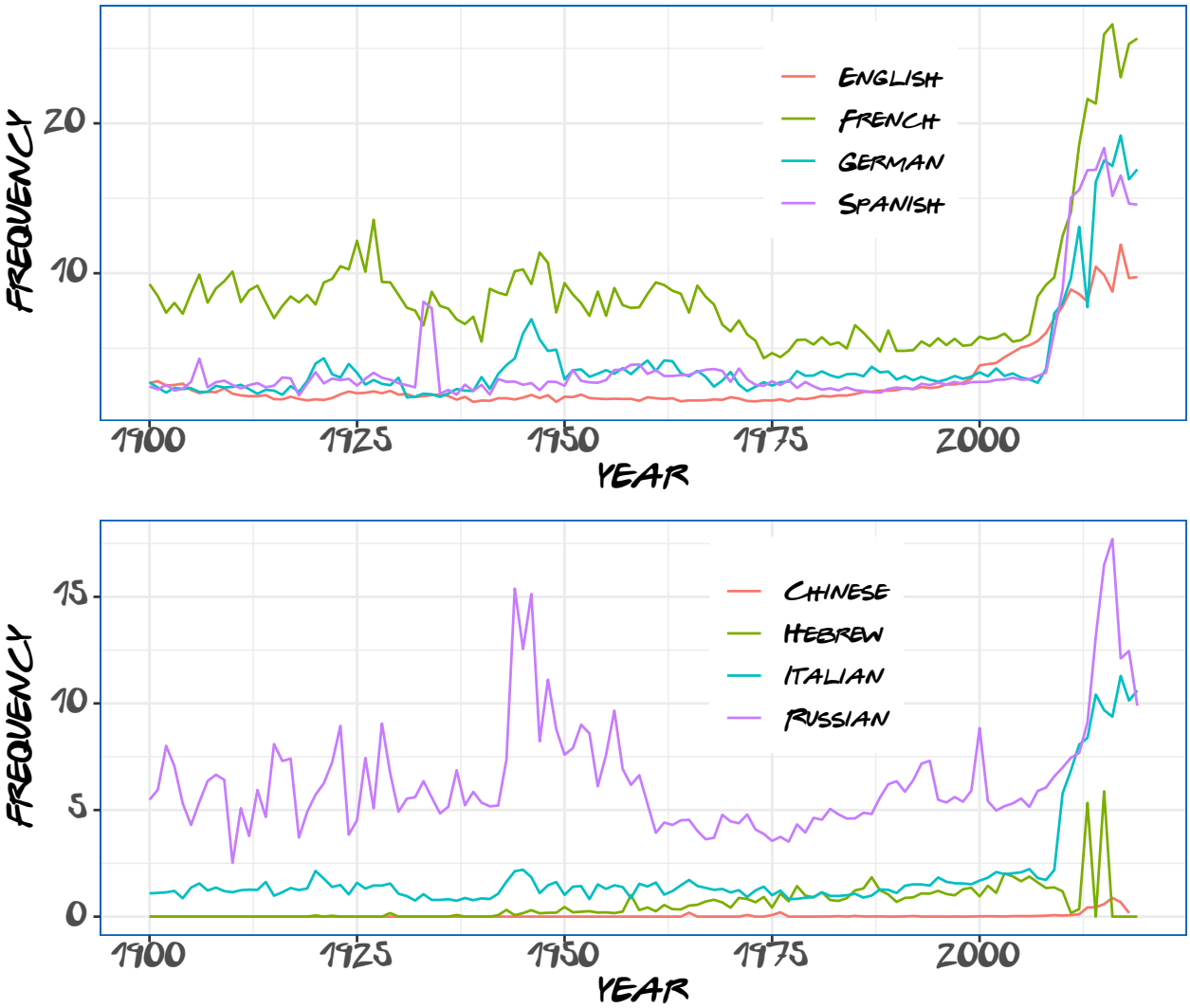


Figure 7: Frequency of terms equivalent to "my God" in the eight languages of Google Books.

Figure 7 shows the frequency of the equivalent of "my God" in English and the other seven languages. All languages except Chinese and Hebrew show sharp increases around 2008, coinciding with the growth of social media as a major communication mechanism. The pattern for English is softer, with a tendency to grow dating smoothly few decades back, and Russian also shows high historical frequency, especially around World War II. A common pattern in all languages is that their historical peak is in the last decade, showing that the overall tendency to use the phrase is growing beyond the English-speaking world.

[2]We thank Simon Schweighofer, Wenjuan Liang, Olga Antsiferova, and Amit Goldenberg for helping with our language handicaps

4 The one with the tweets

How is "oh my God" being used in social media nowadays? To explore this, we used our $14000/year subscription to Crimson Hexagon[3]. We analyzed tweets from 2010-09-01 to 2020-08-30 that were identified by Crimson Hexagon as written in English by users living in the US and with an identifiable gender. Our analysis includes a total of 180300519164 tweets with gender and among them 130796995 contain "oh my God" or "OMG".

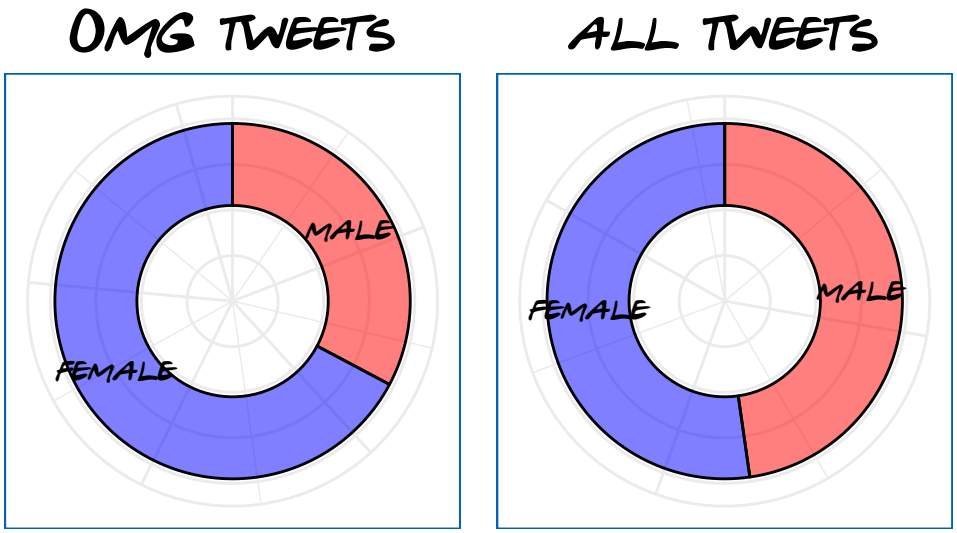


Figure 8: Donut plots of gender in tweets that contain OMG and in all tweets.

Figure 8 shows the fractions of tweets posted by users of each gender for tweets that include a form of OMG and for all tweets. 67.24% of OMG tweets are posted by women, while 52.26% of tweets in general are posted by women. This could mean that women use the phrase more often than men on Twitter or that the term is used as a predictor for gender by Crimson Hexagon. Distinguishing this would require other gender identification methods or self-reported data of Twitter users.

With respect to location, we can investigate the use of the phrase in the Los Angeles area by analyzing geolocated tweets:

The highest density of OMG tweets in Los Angeles is located around the areas of Santa Monica, Beverly Hills, and Hollywood. The San Fernando valley (top left) has also considerable density of OMG tweets, especially with respect to its overall lower tweet activity compared to the other regions. This shows how "Oh my God" is part of valleyspeak and in combination with the gender pattern, it explains the origin of the *valley girl* stereotype.

We can also use location information to see the regional distribution of OMG across the United States. Figure 10 shows the US colored by the frequency of OMG per million tweets. One can see a higher frequency of OMG in the West coast and the northeast, with a lower frequency along the Bible belt. The top states by OMG frequency are Oregon and Washington, with California surprisingly low in 9th position.

[3]Gently payed by other serious research projects

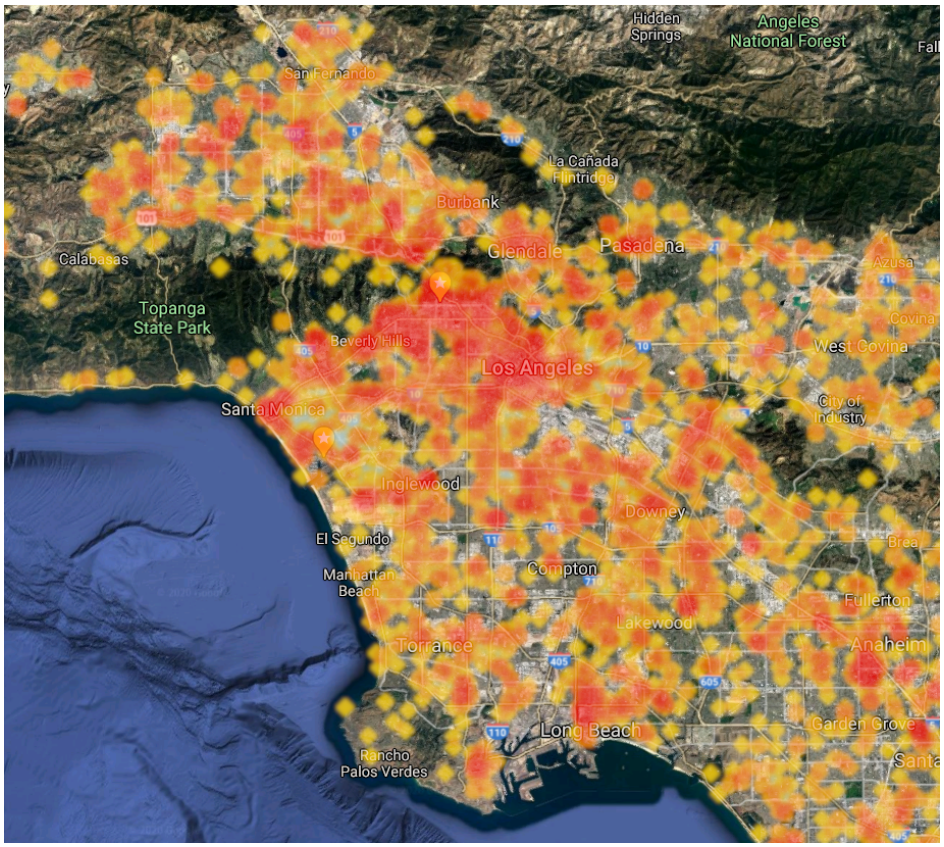


Figure 9: Heatmap of OMG tweets in the Los Angeles area.

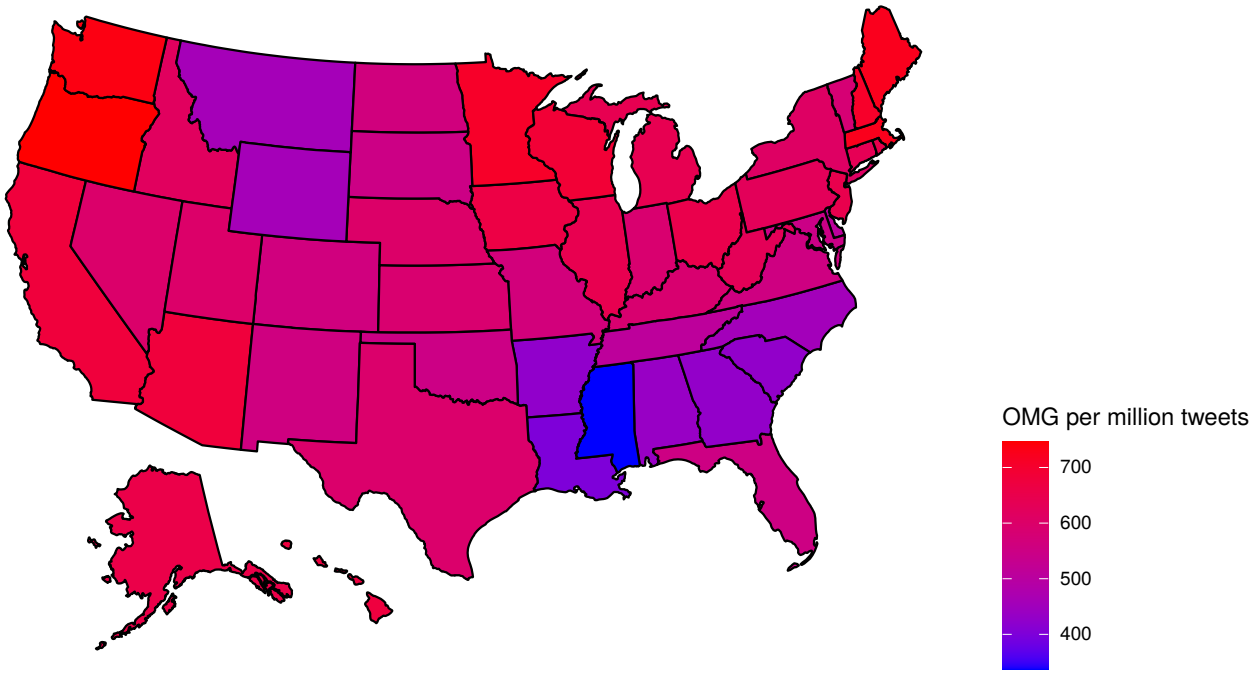


Figure 10: Map of the United States with states colored according to their frequency of OMG per million tweets.

position	state	perMillion	position	state	perMillion
1	Oregon	746.5444	41	Delaware	506.6569
2	Washington	734.2142	42	Montana	457.7706
3	Massachusetts	724.0687	43	Wyoming	457.6181
4	Maine	723.2998	44	North Carolina	454.8653
5	Minnesota	706.8594	45	Alabama	435.6764
6	New Hampshire	706.3069	46	Georgia	428.4706
7	Wisconsin	688.0421	47	South Carolina	426.1360
8	Arizona	677.9887	48	Arkansas	424.6931
9	California	677.7345	49	Louisiana	400.4717
10	Hawaii	672.7159	50	Mississippi	337.1873

Table 1: Top and bottom 10 states by frequency of OMG in tweets.

The bottom states by frequency of OMG in tweets are southern states like Mississippi and Louisiana. Since "Oh my God" can be considered blasphemy under traditional Christian values, one can expect the frequency of this expression to be lower in more conservative states. To test this idea, we collected the US 2020 presidential election results from https://cookpolitical.com/2020-national-popular-vote-tracker.

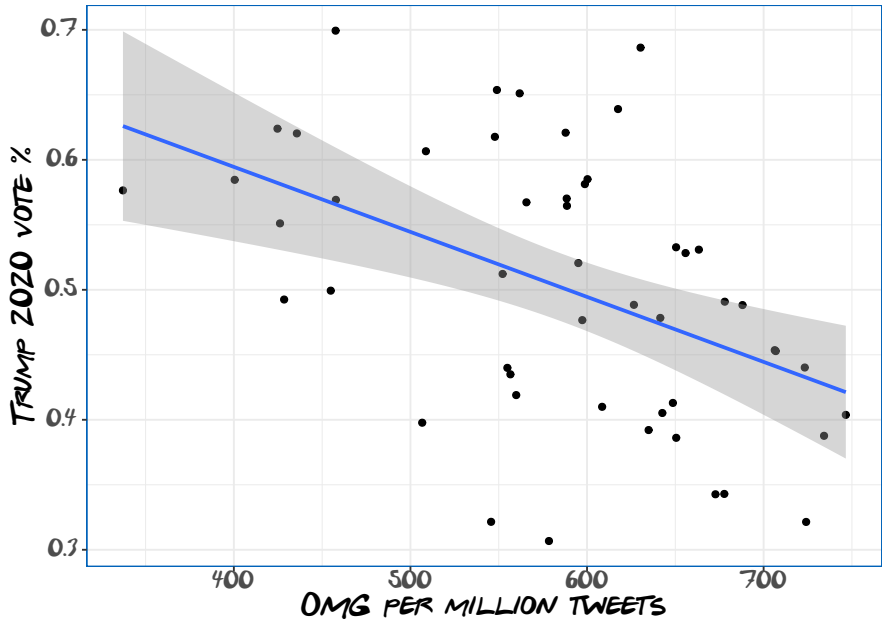


Figure 11: Scatter plot of Donald Trump vote percentage in 2020 versus OMG per million tweets across US states. The line and shaded areas show the results of a linear fit.

Figure 11 shows the association between the percentage of votes for Trump in 2002 and the frequency of OMG per million tweets. Over states, the percentage of votes to Donald Trump is negatively correlated with the frequency of OMG tweets (Pearson's correlation coefficent of -0.47 with p-value 0.0006). This significant correlation satisfies the necessary condition for any paper or blog post to refer to Donald Trump or to COVID-19 to get any attention on Twitter.

5 The one with Artificial Intelligence

To understand better the meaning of "Oh my God", we use everyone's favorite technology nowadays: *"Artificial Intelligence"*. More precisely, we inspect three deep language models trained against large-scale data: BERT (Devlin et al. 2018), GPT-2 (Radford et al. 2019), and RoBERTa (Liu et al. 2019). We do this through the models available in Huggingface and the transformers python package.

BERT allows us to predict a word based on the context in which it appears. We calculated the distribution of word probabilities to fill the sentence "Oh my _____" to see if God seems to be the most frequent word. The top 10 words by probability are shown below:

word	score
god	0.9291785
goodness	0.0245932
gods	0.0092625
lord	0.0071211
goddess	0.0059075
word	0.0016963
zeus	0.0013340
go	0.0009128
my	0.0008139
head	0.0007671

Table 2: Top words with BERT scores for presence after "Oh my".

The probablity of "god" is by far the highest, close to 0.93. This is followed by "goodness", which fits as a similar exclamation but without having religious connotations. Beyond that, all probabilities are below 0.01, showing how prevalent is the word "god" in this context.

With GPT-2, we can generate a longer text following "Oh my God!" to have an idea of the context of the phrase. First, we produce the most likely text following the sentence:

Oh my God! I'm so sorry. I'm so sorry. I'm so sorry. I'm so sorry. I'm so sorry. I'm so sorry. I'm so sorry. I'm so sorry. I'm so sorry. I'm so sorry. I'm so sorry. I'm so sorry. I'm so sorry.

Interestingly, the model gets stuck in a loop of "I'm so sorry". We have run the generation up to very long texts and the result was the same, an endless stream of "I'm so sorry" is the most likely text to follow "Oh my God!" in GPT-2.

We generated 10.000 random texts with GPT-2 to have an idea of the contexts that can follow "Oh my God!". Below we show the first four as examples:

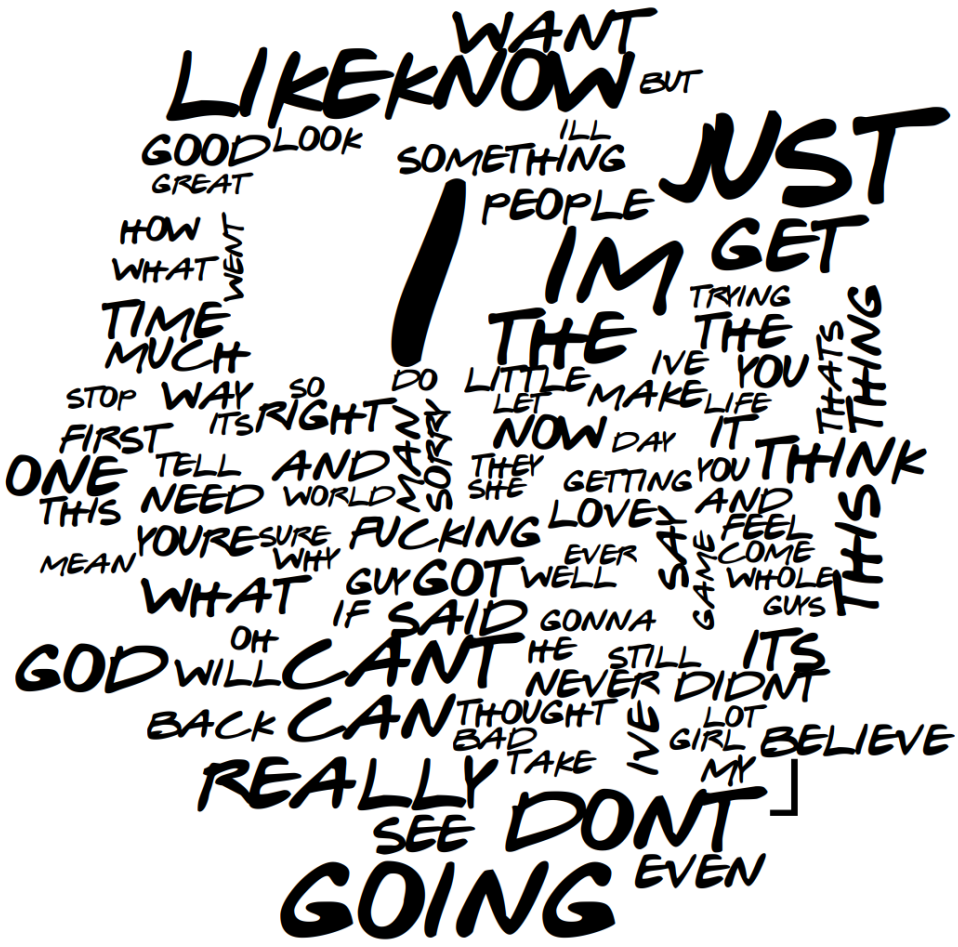
[1] "*Oh my God! Seriously, that's unbelievable. And to think that we are so ignorant as to not care about*"

[1] "*Oh my God! I just can't believe that this is that. I mean I*"

[1] "*Oh my God! I'm going to be dead now and we're going to be dead for years to*"

[1] "*Oh my God! It is so incredibly cute and very, very hot! You know,*"

We see that the phrase can lead to both negative statements, for example the first, but also to postive statements, like the last one. A common factor is the general expression of surprise and of first person singular sentences that highlight the expression of emotions following the sentence. This becomes clearer if we generate a word cloud[4] of the generated texts:



The word "I" is by far the most frequent, followed by verbs and that express personal experiences and feelings. The word "like" is also visible, highlighting this way another of the language patterns of *valleyspeak*.

[4] Just for illustration purposes

To have an idea of the affective meanings of the phrase, we use a RoBERTa model tuned for NLI. We set up a zero-shot classification task as a Natural Language Inference task that classifies the phrase "Oh my God!" using the hypothesis "This sentence expresses {}.", where the brackets are filled with a single word that will serve as the candidate label. As candidate labels, we use the 24 emotion words of the GRID project (Fontaine, Scherer, and Soriano 2013). The words with the highest score in this task are the following:

word	score
surprise	0.0982359
disappointment	0.0909134
disgust	0.0818704
irritation	0.0763046
interest	0.0741176
despair	0.0681486
being hurt	0.0652222
sadness	0.0578116
anger	0.0518757
contempt	0.0506069

Table 3: Top words by score in the zero-shot emotion classification task.

As expected, "surprise" is the closest emotion, followed by a series of negative emotions including "disappointment" and "disgust". In the top 10, the only two non-negative emotions are "surprise" and "interest", suggesting that, in terms of emotion classification, the phrase "Oh my God" is first related to surprise and then to negative emotions, with positive emotions being lower in the ranking.

6 The one with the conclusions

To summarize, we have found the following:

- The frequency of OMG per episode in Friends increased over its run, as it did in contemporary TV and movies. However, Friends had more than four times the frequency of the phrase as other TV and movie content.
- The frequency of OMG in books was exponentially increasing before Friends was released. This growth rate accelerated after Friends became popular.
- This growth pattern appears in othe languages too, having a pronounced acceleration at the beggining of the 2010s decade.
- Tweets that contain OMG in the US are more likely to be posted by women and regions with higher frequency of OMG tweets tend to have lower vote shares for Donald Trump in the 2020 election.
- "God" is the word that fits best after "Oh my" and the text following the phrase can be both positive and negative, but in general subjective.
- Large language models associate the phrase with surprise and then with negative emotions.

Thanks for reading and all the best for 2021!

References

Albright, Alex. 2015. "The One with All the Quantifiable Friendships the Little Dataset." *The Little Dataset That Could*. https://thelittledataset.com/2015/01/20/the-one-with-all-the-quantifiable-friendships/.

Bhattacharyya, Shilpi. 2019. "Sentiment Analysis of the Lead Characters on F.R.I.E.N.D.S." *Medium*. https://towardsdatascience.com/sentiment-analysis-of-the-lead-characters-on-f-r-i-e-n-d-s-51aa5abf1fa6.

Carmody, Sean. 2020. *Ngramr: Retrieve and Plot Google N-Gram Data*. https://github.com/seancarmody/ngramr.

Davies, Mark. 2010. "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English." *Literary and Linguistic Computing* 25 (4). Oxford University Press: 447–64.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv Preprint arXiv:1810.04805*.

Feinerer, Ingo, and Kurt Hornik. 2020. *Tm: Text Mining Package*. https://CRAN.R-project.org/package=tm.

Feinerer, Ingo, Kurt Hornik, and David Meyer. 2008. "Text Mining Infrastructure in R." *Journal of Statistical Software* 25 (5): 1–54. https://www.jstatsoft.org/v25/i05/.

Fontaine, Johnny RJ, Klaus R Scherer, and Cristina Soriano. 2013. *Components of Emotional Meaning: A Sourcebook*. Oxford University Press.

Hyndman, Rob, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. 2020. *forecast: Forecasting Functions for Time Series and Linear Models*. https://pkg.robjhyndman.com/forecast/.

Hyndman, Rob J, and Yeasmin Khandakar. 2008. "Automatic Time Series Forecasting: The Forecast Package for R." *Journal of Statistical Software* 26 (3): 1–22. https://www.jstatsoft.org/article/view/v027i03.

Liu, Yinhan, 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*.

Loscalzo, Michael. 2018. "Oh, My God!" *Medium*. https://towardsdatascience.com/oh-my-god-cb69dd74839c.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331 (6014). American Association for the Advancement of Science: 176–82.

Pechenick, Eitan Adam, Christopher M Danforth, and Peter Sheridan Dodds. 2015. "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution." *PloS One* 10 (10). Public Library of Science: e0137041.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models Are Unsupervised Multitask Learners." *OpenAI Blog* 1 (8): 9.

Sahakyan, Elizabeth Ter. 2019. "The One with the Data Scientist: A Closer Look at the Friends of Friends." *Medium*. https://medium.com/@liztersahakyan/the-one-with-the-data-scientist-a-closer-look-at-the-friends-of-friends-d3530d1902af.

Sohoye, Yusuf. 2019. "The One with All the FRIENDS Analysis." *Medium*. https://towardsdatascience.com/the-one-with-all-the-friends-analysis-59dafcec19c5.

The Economist. 2019. "Why 'Friends' Is Still the World's Favourite Sitcom, 25 Years on," September. https://www.economist.com/prospero/2019/09/20/why-friends-is-still-the-worlds-favourite-sitcom-25-years-on.

Wikipedia. 2020. "Valleyspeak." https://en.wikipedia.org/w/index.php?title=Valleyspeak&oldid=992610782.