Suicidality on Twitter

Media Exposure, Contagion, and Representation

Em McGlone*

May 28, 2022

Abstract

Twitter, the world's largest microblogging platform, offers strong and available insights into public expressions of mental health. Google's BERT algorithm provides a general-purpose method for understanding short-form text. By applying BERT to tweets which do or do not express suicidal ideation, I create a predictive model for judging suicidal sentiment in tweets en masse, using a training data set of 1000 tweets.. The resulting model had an 87.6% classification accuracy. The model is then used to quantify changes in suicidal sentiment over time, due to exogenous shocks such as the release of the TV show *13 Reasons Why.* It found a -2.12% decrease in suicidal sentiment on Twitter over the media lifetime of *13 Reasons Why.*

*Department of Physics, Harvard University, Cambridge, Massachusetts, USA

1 Introduction

Suicide is the 10th leading cause of death in the United States, and the 2nd leading cause among those of ages 10-34 (Hedegaard H, 2021). Devastating, preventable, and always premature, suicidality remains hard to identify in individuals, hidden by stigma and consequent silence. Suicidality is defined by the presence of suicide-related behavior or thoughts, and can be expressed through communication, planning for suicide, or attempting suicide. This paper will focus on the intersection between the communication of suicidality on social media and the consumption of or exposure to suicide-related popular media.

Suicidality may be contagious; exposure to suicide increases the prevalence of suicidality among adolescents according to Swanson and Colman, 2013. *13 Reasons Why* was a massively popular Netflix show focusing on a high-schooler who commits suicide; the first season, released on March 31st, 2017, was at the time the second-most viewed Netflix season ever. Opinions on the show vary. Rotten Tomatoes gives it a critic score of 80%, and an audience score of 77%. On the other hand, many thought the show was triggering and overly romantic in its depiction of suicide, and would thus cause children to exhibit suicidal behavior and ideation. In any case, Netflix did edit at the suicide scene originally in the first season's finale, and add a content warning before each episode in response to these criticisms.

Between all the code, statistics, and opinions about a five-year-old Netflix show, it's important for both the author and audience to remember that the thing that's being discussed here is *human lives*. Suicide is a pervasive public health crisis, which touches many more lives than it ends. If there's nothing else to take away from this paper, remember the sensitivity of human life and mental health to changes in media and society.

2 Literature Review

Several attempts have been made before to create a machine learning model that predicts suicidality on Twitter. One example is O'Dea et al., 2015, who made a model to measure whether Twitter profiles were, in terms of suicidality, not concerning, slightly concerning, and very concerning. They created two models, one coded by humans, and one coded using machine learning. Their human model was 76% accurate, in comparison to the machine learning model, which was 80% accurate. Their findings confirmed that Twitter is used by individuals to express suicidality, but also note that the predictive power for actual suicidal behaviour and cause for intervention is not yet known.

Similarly, Braithwaite et al., 2016 compare participants' self-reported suicidality with predictions made by a machine learning model using each participant's Twitter data. They find that their model was able to identify clinically significant suicidality in 92% of cases, further validating that machine learning models are efficient for suicidality predictions.

As stated before, Swanson and Colman, 2013 show that exposure to suicide increases rates of suicidal ideation and attempts. They looked at children's exposure to suicide either through a classmate's suicide, or through personally knowing someone who committed suicide. They found that exposure to a classmates suicide increases the odds of suicidal ideation by two to four times, and that personally knowing someone who committed suicide increases the odds of suicidal ideation by three to five times.

On the other hand, Sinyor et al., 2021 study suicide contagion in Ontario due to Twitter events, and find no evidence of suicide contagion as a result. They identify suicide-related Twitter events with over 100 tweets, classify them from 1 to 5 as not concerning to very concerning, and compare these events to suicide trends in Ontario in the same time period. While the correlation is weak or non-existent, they stress that context and content are critical in evaluating the impact of media events, and assert that discourse about suicide on Twitter may not necessarily prompt suicide deaths.

3 Data

I collected data using the Twitter Developer API, which provides a convenient platform to collect individual tweets and their associated user information and metadata. Then, I used the rtweet package in R to connect to this API and download and clean tweet data. The main weakness of this method is that tweets must be searched either by user or by keywords. Keyword searches are problematic in this case because I used them to classify which tweets are suicidal or not based on basic phrases. As a result, I expect the BERT algorithm to associate these manually chosen keywords with its predictions. This effect is mitigated twice, however. First, because the Twitter API search generalizes the given keywords; that is, chosen keywords may be conjugated or separated, like if you Google search a phrase. Secondly, because the BERT algorithm further generalizes the meaning of these keywords, and is capable of making predictions that do not include the corresponding keywords.

Data were collected from several sources. With a subscription to the Twitter API's full archive search capability, I first 3000 scraped tweets from 2017 to measure the change in suicidal sentiment due to the release of *13 Reasons Why*, and randomly sampled them down to 1000 tweets for runtime. The training data set (n = 1000) combined three individual searches (n $\approx$ 333 each) on January 1, 2017, three months before *13 Reasons Why* was released. These three searches used the following keywords, separated by ”OR”. Any individual tweet returned by the search must have at least one permutation of one keyword.

1. ”I'm suicidal OR I hurt myself OR hang myself OR hung myself OR kill myself OR I want to die OR I want death OR I want to be dead”

2. ”want OR don't OR do OR person OR will OR feel OR like OR gonna OR with OR was OR day OR can OR know OR make OR get”

3. ”died OR committed suicide OR overdosed OR passed away OR dying OR hospital OR hurts OR pain OR hate”.

Tweets from (1.) were marked as suicidal, and tweets from (2.) and (3.) were marked as not suicidal. (1.) collects obviously suicidal phrases, (2.) collects generic tweets using the most common (non-stopword[1]) English words, and (3.) collects tweets which are about death, but not suicidal in nature. The length of these queries were limited by the 128 character limit for searches in the Twitter Developer API. The goal of the split is to help the algorithm differentiate ”suicidality” from ”normalcy” and ”death.” The text of these 3000 tweets were embedded into numeric values, representing the semantic meaning of the tweets, using Google’s BERT. These word-embeddings were then fed into a random forest model, with a TRUE or FALSE for suicidal sentiment attached to each word-embedding as a response variable. The output is a model which can take in new word-embeddings (that is, the text of new tweets) and return a TRUE or FALSE for whether an individual tweet has suicidal sentiment. It may also be used to return a decimal probability between 0 and 1, which represents how much an individual tweet tends towards suicidal sentiment. For example, a result of 0.95 represents a tweet with a strong suicidal sentiment, a result of 0.05 represents a tweet with almost no suicidal sentiment, and a result of 0.60 represents a tweet with minor suicidal sentiment.

Testing data were collected from three days spanning the lifetime of *13 Reasons Why* media attention. All three searches use the following keywords:

- ”want OR don’t OR do OR person OR will OR feel OR like OR gonna OR with OR was OR day OR can OR know OR make OR get OR time OR make OR happy OR sad OR angry OR dog OR hate OR need OR over OR under OR about OR man OR woman OR they OR more OR less OR better OR worse OR middle OR right OR low OR high OR live OR take OR give”

Which represents a longer list of the most common words in English, now subject to Twitter Developer Full Archive Search Premium’s 1024 character limit for queries.

[1] Stopwords in this case are words which also act as code for searches. They cannot be used as search queries for this reason. This was a confounding element in designing search queries of common words, since many common words like ”for” and ”as” are stopwords.

The first data set (n = 500) was scraped from April 15th, 2017, the day *13 Reasons Why* media attention peaked according to Google trends. The second data set (n = 500) was scraped from March 1st, 2017, one and a half months before *13 Reasons Why* media attention peaked, when Google Trends frequency was basically zero. The third data set (n = 500) was scraped from June 1st, 2017, one and a half months after *13 Reasons Why* media attention peaked, when Google Trends frequency returned to a near-zero norm. The text of all these tweets were embedded into numeric values, and the random forest model was used to predict suicidal sentiment according to these word-embeddings.

4 Results

Using the text package in R, I ran a random forest on the training data, and trained and tested a model that classifies tweet text word embeddings as suicidal or not, as well as predicts a probabilistic suicidality score for each tweet. The model automatically makes several training-testing splits and tests on all the input data. Comparing the model's predictions to my keyword-based suicidality markers as indicated in Figure 1.

Of the 1000 training tweets, 27.4% were guessed correctly to be suicidal (True positive), 60.2% were guessed correctly to be non-suicidal (True negative), 6.1% were incorrectly predicted to be non-suicidal (False negative), and 6.3% were incorrectly predicted to be suicidal (False Positive). The cutoff between positive and negative predictions was optimized in the model for the highest accuracy, and correspondingly, the smallest ratio of true positive to false positive errors. This value was a suicidality probability score of 0.38. The total classification accuracy of this model was 87.6%. The Kappa statistic measures how well the model makes predictions above the accuracy random guesses; its value is 0.641, representing a significant agreement between the predicted and assigned suicidality values. The P-value of ~ 0 (actually $2.2 * 10^{-16}$) suggests that the results of the model are statistically significant, rejecting the null hypothesis. An RMSE of 0.599 for the suicidality score prediction puts a

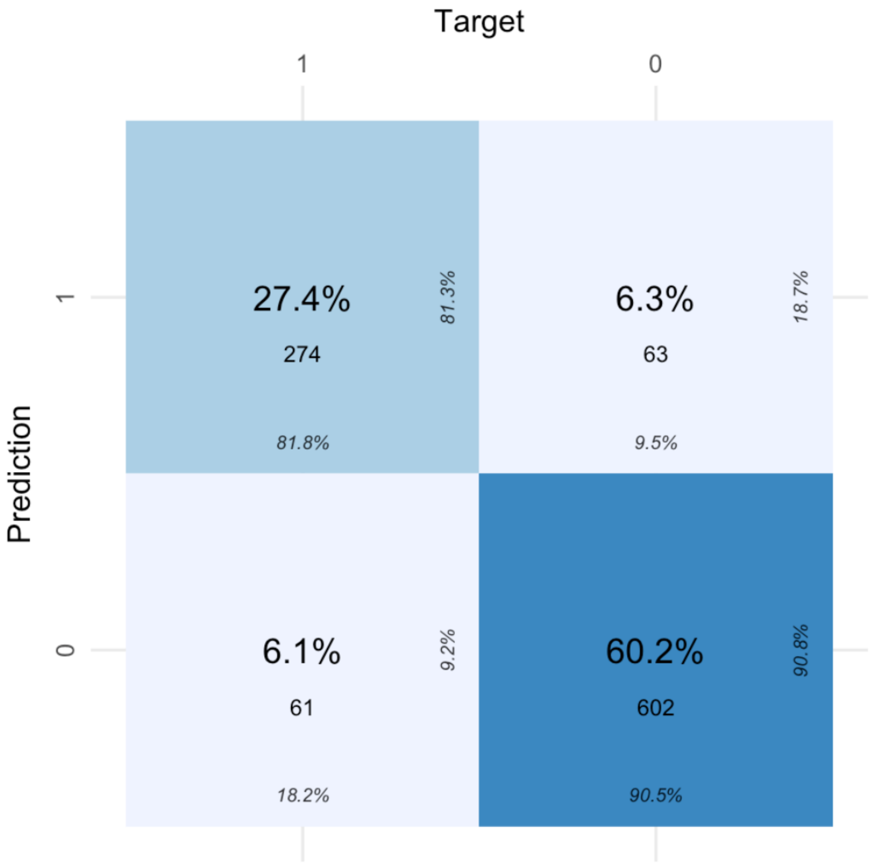


Figure 1: Confusion matrix for BERT Embeddings Random Forest Model. Target is 1 or 0 for if the tweet was marked suicidal or not based on my initial search. Prediction is 1 or 0 for if the BERT model guessed whether a tweet is suicidal or not. Percents along the edges represent the percentage of tweets in that target or prediction category which are in each cell. The total number of tweets in each cell is displayed as the number below the main percentage.

large error bar on results drawn from that model. All relevant accuracy statistics are listed in Table 1.

Figure 2 shows the Sensitivity-Specificity (ROC) curve for the BERT embeddings Random Forest model. Sensitivity represents the true positive rate; that is, how often the model predicts a tweet to be suicidal correctly. 1-Specificity represents the false positive rate; that is, how often the model predicts a tweet to be suicidal incorrectly. The area under the curve represents how much the true positive rate exceeds the false positive rate, making it a useful metric for evaluating the predictive power of this model in comparison to others. The value

BERT Word Embeddings Random Forest Model Accuracy Statistics							
Statistic	Accuracy	Precision	Sensitivity	Specificity	P-value	RMSE	Kappa
Value	87.6%	82.9%	98.5%	59.7%	∼ 0	0.599	0.641

Table 1: Relevant accuracy statistics. All statistics apply to the classification model accuracy, except for RMSE which applies to the suicidal probability score prediction.

of the area under the curve came out to be 0.93, out of a total of 1. This curve was useful in picking the optimal detection cutoff which maximized the ratio of true positive to false positive rates, which corresponded to a suicidality probability score of 0.38, a sensitivity of 0.985, and a specificity of 0.597.

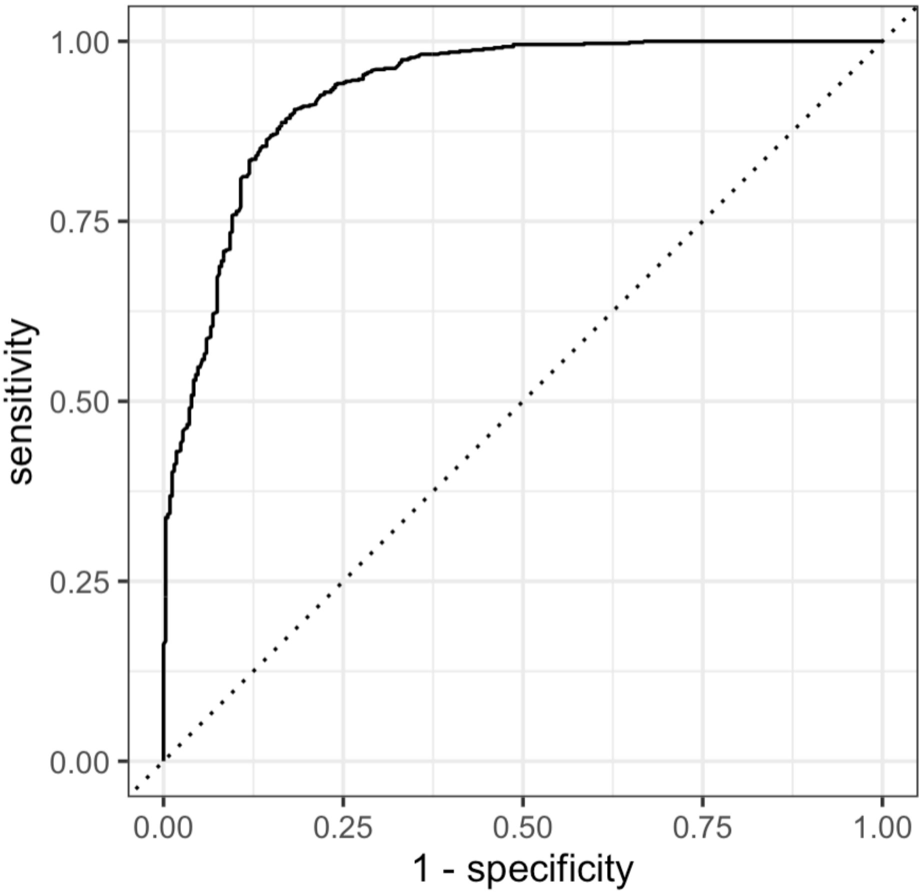


Figure 2: Sensitivity-Specificity curve for the BERT embeddings Random Forest classification model.

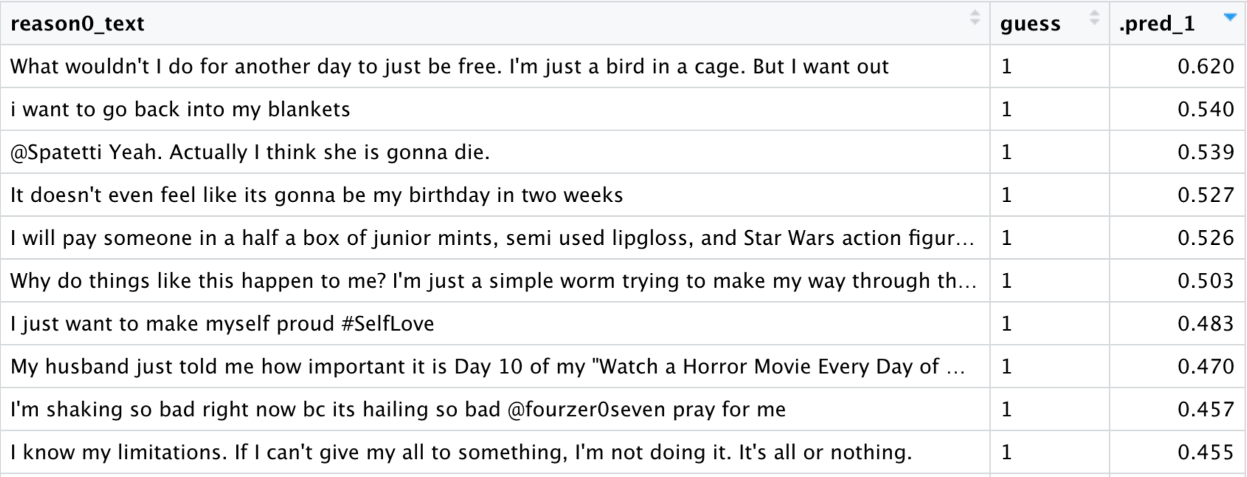
After scraping and cleaning the three testing data sets straddling the release of *13 Reasons Why* (n = 1000 each), I ran the model on those tweets and predicted a suicidality classification (1 for suicidal or 0 for non-suicidal) and a suicidality probability score (between 1 and 0, weighting towards either suicidal or non-suicidal). Then, I summed the number of tweets classified as suicidal in each sample, and summed the suicidality probability scores for all the tweets, to simulate a total amount of suicidal sentiment in the sample. These values, and their corresponding dates, are listed in Table 2. A percent change between the values on March 1, 2017 and June 1, 2017 is also listed.

13 Reasons Why: Suicidality Effects on Twitter		
Date of Tweets	Number of Suicidal Tweets (n = 1000)	Sum of Suicidality Predictions
March 1, 2017 (Before *13 Reasons Why*)	35	102.922
April 15, 2017 (During *13 Reasons Why*)	33	101.095
June 1, 2017 (After *13 Reasons Why*)	33	100.744
Percent Change (from March 1 to June 1)	-5.71%	-2.12%

Table 2: Results of BERT Embeddings Random Forest model on tweets spanning the media timeline of *13 Reasons Why*. Number of Suicidal Tweets corresponds to the number of tweets which the BERT model classified as suicidal. Sum of Suicidality Predictions is the sum of suicidality probability scores between 1 and 0 that the BERT model predicts for all tweets. A higher number thus represents more suicidal sentiment in the dataset as a whole.

The number of tweets classified as suicidal decreased from 35 out of 1000, to 33 out of 1000, a percent change of -5.71%. The sum of suicidality probability predictions decreased by -2.12%. Therefore, suicidal sentiment on Twitter actually decreases during the media lifetime of *13 Reasons Why*.

Just looking at the tweets from around the media timeline of *13 Reasons Why* and their corresponding predictions made by the model, it is clear that about only half of the tweets marked as suicidal have clear suicidal ideation in their texts. It does seem that the model divides up text easily into some positive- and negative- sentiment split, but whether this truly represents ”suicidality” en masse is unclear. Examples of high-suicidality predictions made by the BERT embeddings Random Forest model on tweet text from March 1, 2017 are shown in Figure 3. Examples of low-suicidality predictions from tweets on March 1, 2017 are shown in Figure 4; these feature many ads, tweets not in English, discussion of Donald Trump, and many links and @s.



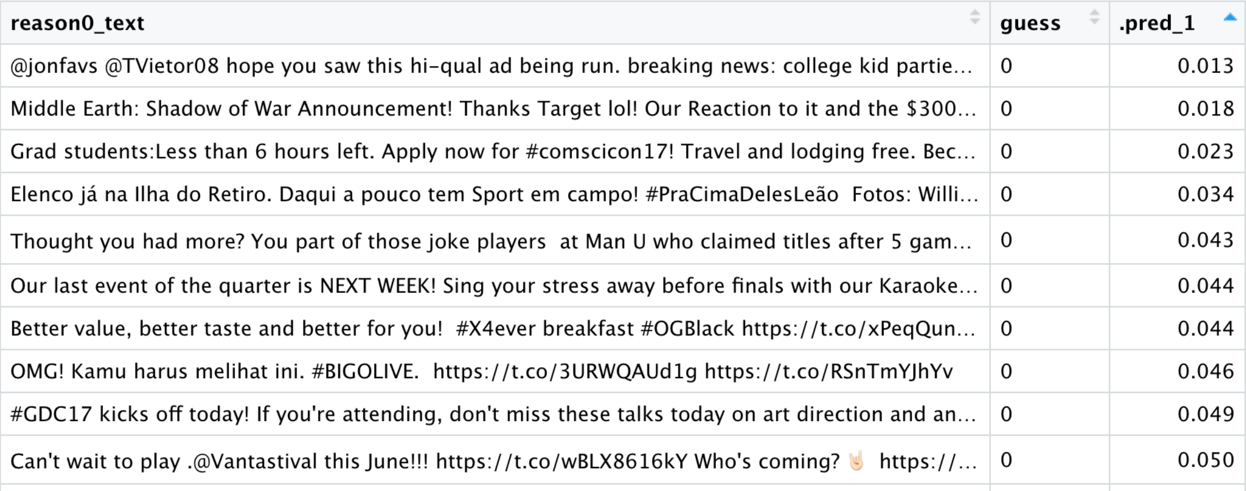
reason0_text	guess	.pred_1
What wouldn't I do for another day to just be free. I'm just a bird in a cage. But I want out	1	0.620
i want to go back into my blankets	1	0.540
@Spatetti Yeah. Actually I think she is gonna die.	1	0.539
It doesn't even feel like its gonna be my birthday in two weeks	1	0.527
I will pay someone in a half a box of junior mints, semi used lipgloss, and Star Wars action figur…	1	0.526
Why do things like this happen to me? I'm just a simple worm trying to make my way through th…	1	0.503
I just want to make myself proud #SelfLove	1	0.483
My husband just told me how important it is Day 10 of my "Watch a Horror Movie Every Day of …	1	0.470
I'm shaking so bad right now bc its hailing so bad @fourzer0seven pray for me	1	0.457
I know my limitations. If I can't give my all to something, I'm not doing it. It's all or nothing.	1	0.455

Figure 3: Tweets with high-suicidality predictions made by the BERT model. reason0 text represents a tweet on March 1, 2017. guess represents a 1 for if the tweet is classified as suicidal, and 0 for if a tweet is not. .pred 1 is the suicidality probability score; a higher number represents a tweet more likely to be suicidal.

5 Discussions

When I first saw that suicidality on Twitter decreased, I accidentally interpreted it backwards and thought suicidality increased, likely as a result of the *13 Reasons Why* being a trigger for contagious suicidal behavior. This was my original thesis, so I ran with the inverse result for a while, in an almost-catastrophic fit of confirmation bias.

This decrease in suicidality conflicts with the result of Swanson and Colman, 2013, suggesting that suicide contagion may be limited to cases of personal exposure to suicide rather than media exposure to suicide. It also replicates the result of Sinyor et al., 2021, showing that media depictions of suicide have limited effects on real-world suicidality. This model achieves a higher accuracy than O'Dea et al., 2015, likely thanks to the sheer intelligence of Google's BERT above the Support Vector Machines and Logisitic Regression models that



reason0_text	guess	.pred_1
@jonfavs @TVietor08 hope you saw this hi–qual ad being run. breaking news: college kid partie…	0	0.013
Middle Earth: Shadow of War Announcement! Thanks Target lol! Our Reaction to it and the $300…	0	0.018
Grad students:Less than 6 hours left. Apply now for #comscicon17! Travel and lodging free. Bec…	0	0.023
Elenco já na Ilha do Retiro. Daqui a pouco tem Sport em campo! #PraCimaDelesLeão Fotos: Willi…	0	0.034
Thought you had more? You part of those joke players at Man U who claimed titles after 5 gam…	0	0.043
Our last event of the quarter is NEXT WEEK! Sing your stress away before finals with our Karaoke…	0	0.044
Better value, better taste and better for you! #X4ever breakfast #OGBlack https://t.co/xPeqQun…	0	0.044
OMG! Kamu harus melihat ini. #BIGOLIVE. https://t.co/3URWQAUd1g https://t.co/RSnTmYJhYv	0	0.046
#GDC17 kicks off today! If you're attending, don't miss these talks today on art direction and an…	0	0.049
Can't wait to play .@Vantastival this June!!! https://t.co/wBLX8616kY Who's coming? 🤘 https://…	0	0.050

Figure 4: Tweets with low-suicidality predictions made by the BERT model. reason0 text represents a tweet on March 1, 2017. guess represents a 1 for if the tweet is classified as suicidal, and 0 for if a tweet is not. .pred 1 is the suicidality probability score; a lower number represents a tweet less likely to be suicidal.

they use. This model achieves a lower accuracy than Braithwaite et al., 2016, however this is likely not a useful comparison, since they had access to actual identified clinical diagnoses of suicidality, rather than basing suicidality in tweets on a keyword search.

Interpreting this result is difficult since it does not track individuals and their suicidal ideation, but rather gives a snapshot in time across twitter of suicidal sentiment. So, I can say that suicidal sentiment on Twitter does decrease from March 1, 2017 to June 1, 2017. Whether this is caused by *13 Reasons Why* is unclear; a correlation exists, but no causal link. As a result, other exogenous effects could be causing this change, like events in Donald Trump's presidency, or other media. Google Trends does indicate that *13 Reasons Why* was probably the most popular thing on the internet on April 15th, 2017, giving it a relative frequency of 100 on that day, with other high frequencies in the neighboring weeks.

Finally, a litany of model improvements:

- The limitations of using keywords to specify suicidality, as discussed in Section 3.
- If I had more search requests, I would scrape more data to train and test the model. I would also make sure all searches use the full query length in a more systematic

manner.

- "Apples to Oranges" - the fact that I used different keyword searches for the training and *13 Reasons Why* data decreases validity, because they do not represent similar samples. The model's predictions do weight well, predicting suicidality to be in 0.3% of tweets in the wild, compared to 33% in the training data.
- The model does not handle comedy and irony well. More well chosen keywords, or another model, may be able to account for this. However, whether or not tweets that express suicidal ideation in a comedic or ironic manner may be considered to be suicidal is an interesting question.
- The model as it stands uses solely the text of tweets. It would be a relatively simple addition to make a more complex metadata model, which embeds not only the text of tweets, but the also the tweeter's name, screen name, and profile description. It would also include metadata of the individual tweets, such as number of likes and the number of followers of the tweeter. More complex additions to this model involve the ability to open and understand links, videos, follower networks, and profile images.
- The model should account for outside trends in suicidality. It's hard to get specific enough data for this, but as it stands, any other exogenous events from March to June 2017 could be causing the change in suicidality on twitter.
- In any case, if I had more requests, it would be great to re-scrape new tweets in the same time period and see if the trend holds up.

6 Conclusion

For one, it may be the case that *13 Reasons Why* handled suicidal ideation in a productive and helpful manner, which reduced suicidality in its viewers. If Rotten Tomatoes is to be believed, then *13 Reasons Why* is in fact a mature and age-appropriate depiction of suicidality, and perhaps it ought to be credited. Whether the question of what constitutes a helpful depiction of suicide on TV is better left to critics or psychologists or clinicians or researchers is an open question. Regardless, in this case, it seems Good Depictions of Suicide in Media May Reduce Suicidality.

For another, it may be the case that independent of the quality of *13 Reasons Why*, the mere Discussion or Depiction of Suicide in Media May Reduce Suicidality. This could be true for many reasons. Perhaps just discussing suicide in a public forum reduces stigma, or perhaps depictions of suicide are real enough to prevent people from ever trying it. In this case, let's talk a lot more about suicide.

Either of these cases, of course, need more testing around other similar exogenous shocks in culture and media to be further validated and compared. However, the sheer lack of suicide-focused content in media makes this difficult. Many films and TV shows feature suicide-related content, but not many focus on it as much, nor were as large as *13 Reasons Why*. A good start would be to repeat the test on *13 Reasons Why*.

As it stands, this model is not particularly useful for identifying targets for intervention, since it only takes in the text of individual tweets, without considering profile characteristics. A full metadata model, as discussed in Section 5, would be a more powerful tool for judging the suicidality of an individual, rather than of Twitter as a whole.

The idea that media can literally save lives is a strong claim which is not contradicted by my intuition. People form strong and meaningful attachments to media. Representation matters, and helps to change narratives and reduce stigma. So, again, let's talk about suicide.

References

Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., & Hanson, C. L. (2016). Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR Mental Health*, *3*(2), 21. https://doi.org/10.2196/mental.4822

Hedegaard H, W. M., Curtin SC. (2021). Suicide mortality in the united states, 1999–2019. *NCHS Data Brief*, (398). https://doi.org/https://dx.doi.org/10.15620/cdc:101761.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on twitter. *Internet Interventions*, *2*(2), 183–188. https://doi.org/https://doi.org/10.1016/j.invent.2015.03.005

Sinyor, M., Williams, M., Zaheer, R., Loureiro, R., Pirkis, J., Heisel, M. J., Schaffer, A., Cheung, A. H., Redelmeier, D. A., & Niederkrotenthaler, T. (2021). The relationship between suicide-related twitter events and suicides in ontario from 2015 to 2016 [PMID: 32366171]. *Crisis*, *42*(1), 40–47. https://doi.org/10.1027/0227-5910/a000684

Swanson, S. A., & Colman, I. (2013). Association between exposure to suicide and suicidality outcomes in youth. *CMAJ*, *185*(10), 870–877. https://doi.org/10.1503/cmaj.121377

Appendix A. Donald Trump's Twitter

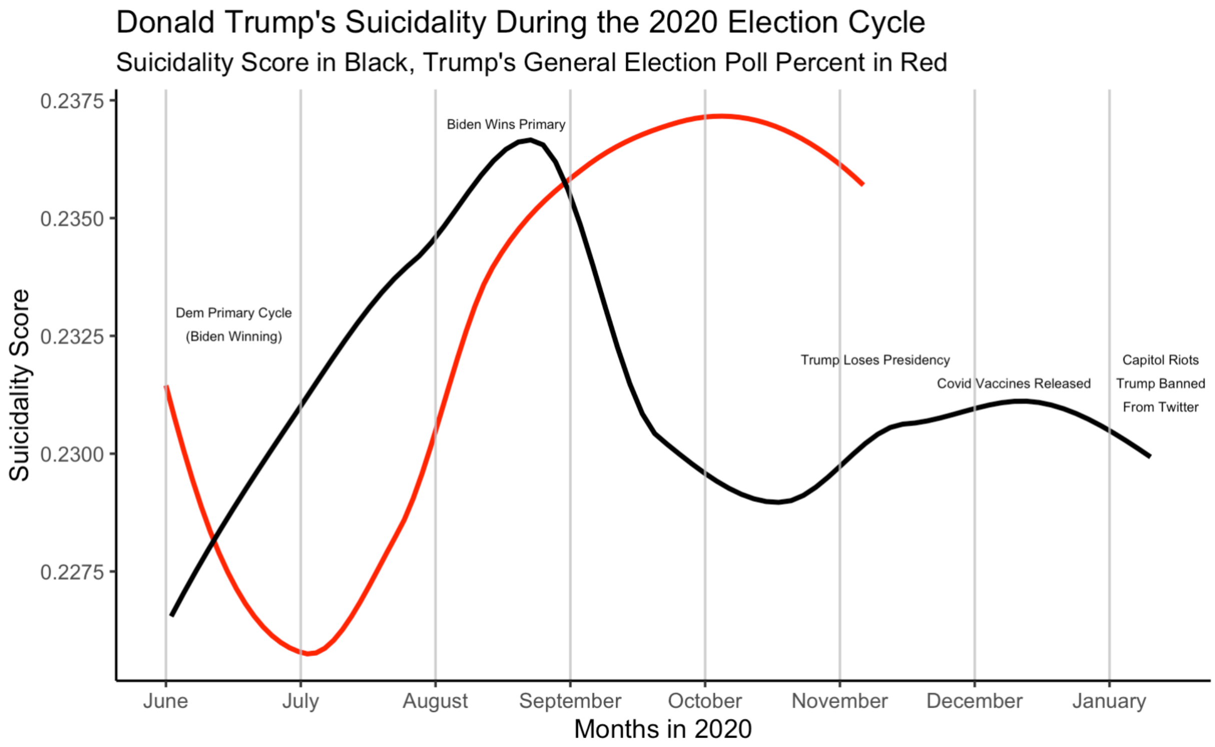
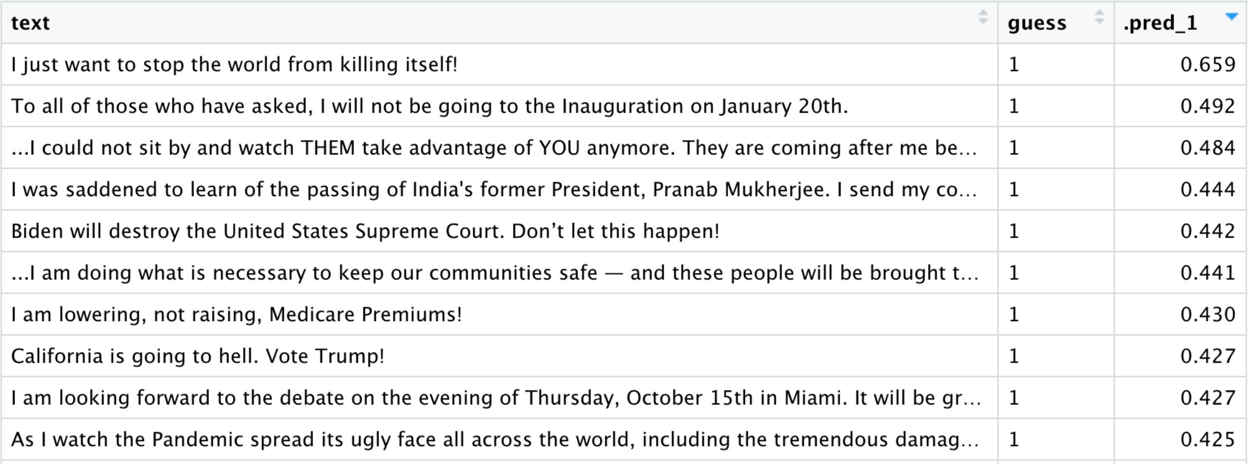


Figure 5: Bert Embeddings Random Forest Model used on Donald Trump's last 1880 tweets. Trump's suicidality score increases after a lag when his poll numbers decrease, and decreases after a lag when his poll numbers increase. Suicidality scores increase as Biden wins the Democratic Primary, and when Biden wins the presidency.

I scraped Donald Trumps last 1880 tweets before his Twitter ban, starting on June 1st, 2020, and ending on January 6th 2021. Then, I applied my suicidality probability prediction model to measure his suicidality during the 2020 election cycle. This must assume that 2017 suicidal tweets apply to Donald Trump in 2020. These scores are graphed in black in Figure 5, along with his relative polling percentages in red. Both are smoothed for clarity. It is important to take this with a grain of salt, since I chose which election events to annotate with, and how to display the polling scores. The polling scores do tell a story, though, where Trump's suicidality trends up with a lag when his polls go down, and vice versa.

Trump's tweets which were predicted to have the highest suicidality score are displayed in Figure 6. It seems none of these tweets clearly show ideation, although the first is interesting.



text	guess	.pred_1
I just want to stop the world from killing itself!	1	0.659
To all of those who have asked, I will not be going to the Inauguration on January 20th.	1	0.492
...I could not sit by and watch THEM take advantage of YOU anymore. They are coming after me be…	1	0.484
I was saddened to learn of the passing of India's former President, Pranab Mukherjee. I send my co…	1	0.444
Biden will destroy the United States Supreme Court. Don’t let this happen!	1	0.442
...I am doing what is necessary to keep our communities safe — and these people will be brought t…	1	0.441
I am lowering, not raising, Medicare Premiums!	1	0.430
California is going to hell. Vote Trump!	1	0.427
I am looking forward to the debate on the evening of Thursday, October 15th in Miami. It will be gr…	1	0.427
As I watch the Pandemic spread its ugly face all across the world, including the tremendous damag…	1	0.425

Figure 6: Tweets with high-suicidality predictions made by the BERT model. reason0 text represents a tweet on Donald Trump’s most recent weets. guess represents a 1 for if the tweet is classified as suicidal, and 0 for if a tweet is not. .pred 1 is the suicidality probability score; a higher number represents a tweet more likely to be suicidal.