Intersect, Vol. 17, No. 2 (2024)

Impact of Traumatic News on Social Media

Nidhi Vadlamudi

Abstract

The spread of traumatic and triggering content on social media harms many users on a daily basis, as people cannot always control the content they view. Those who have been affected by past trauma may be triggered seeing social media posts that contain news of a similar nature. In this study, we collect tweets from two different news sources' Twitter accounts. We use both regression and neural network models to classify these tweets as traumatic or non-traumatic based on a dataset of 600 events rated by trauma level. We then use various tweet engagement metrics including retweets and replies to identify the reach and impact these potentially traumatic or triggering tweets have on Twitter users.

Introduction

As social media rises in popularity to become a primary source for sharing information, concerns have risen about how people are influenced by content they see online. It is difficult to control the content that comes up on a social media feed, since social media platforms often suggest and promote new content beyond the content a user actively follows. With the amount of content people consume on social media, news sources often turn to social accounts to share news with a wider audience, despite some of this news containing potentially disturbing content. Since information spreads so quickly and easily through social media, and with little oversight or control, the impact of traumatic or triggering content can cause significant harm to many groups of people. One such group is those affected by Post Traumatic Stress Disorder (PTSD). People with PTSD can be triggered by violent news events, which are often discussed on social media and are difficult to avoid hearing about. Because of this, the concept of trigger warnings, disclaimers shared before providing potentially traumatic or disturbing content, have become more widely used on social media. Trigger warnings, however, have been shown to have a negative impact on social media users. Our study aims to identify the impact disturbing or traumatic content has on social media users as well as develop methods of identifying this content in order to potentially curtail its spread across social media.

There has been extensive research conducted on identifying hate speech, offensive language, and their spread on social media, but there is very little identifying and analyzing the impact of triggering social media posts. Identifying traumatic content on social media and analyzing its reach and impact on social media users may help us better understand how people are influenced by triggering media and combat this spread.

In this study, we determine the reach of traumatic news compared to non-traumatic news and compare their impact on Twitter users. We use a dataset of 600 events [1] rated by participants on a scale from “not at all traumatic” to “extremely traumatic.” We pull tweets from various news sources’ Twitter accounts and use word2vec vectorization to represent the tweets and events as vectors. We then use multiple models to classify these tweets as traumatic or non-traumatic. We count the likes, retweets, and comments on each tweet. We compare the reach of the tweets identified as containing potentially traumatic content and the tweets identified as non-traumatic. We then analyze the sentiment of the replies to each tweet in order to determine how learning about traumatic and non-traumatic news events impact the general population.

Background

Psychological Response to Traumatic News

Numerous studies [2, 3] have shown that consuming violent or otherwise traumatic news through media can trigger trauma responses, especially in children and teenagers. Many teenagers use social media excessively, which has been linked to depression, anxiety, and other mental health disorders. These issues are then further exacerbated by exposure to disturbing or violent content on social media. Learning about distressing news events can also trigger post-traumatic stress, as people with PTSD can be reminded of past trauma.

Spread of News Across Social Media

There has been controversy surrounding social media algorithms and what type of content is provided further reach. At their base, social media algorithms are dependent on user interaction; the more you engage with a post, the more of those types of posts you will see in your feed. However, because these algorithms are developed by humans, they can still contain biases [4] that affect what type of content appears in a user’s social media feed. With the rise of social media activism, many people engage with posts discussing current events in order to spread awareness on issues they care about. These posts, often discussing controversial social and political topics, may contain information that is distressing to read about. However, higher engagement through social media activism causes these posts to have a further reach, which can lead to unintended triggering of trauma responses in viewers with little control over the content they see.

Terms

Psychological Trauma: A psychological response to an event that a person finds distressing [5].

Trigger: Defined by Psych Central [6] as “a stimulus that causes a painful memory to resurface.”

Traumatic Content: We define traumatic content as potentially triggering or traumatic news. News is classified by this study as triggering or traumatic if it discusses or relates to events rated as traumatic in the dataset Psychological Response Data on the Traumatic Nature of 600 Written Events [1]. We identify traumatic and triggering content using machine learning algorithms trained on this dataset.

Reach: We define the “reach” of a tweet as the extent of its spread over social media. We calculate the reach of a tweet by determining its engagement based on the number of likes, retweets, quote tweets, and replies.

Impact In this study, we define the impact of a tweet as Twitter users’ reactions to a tweet. We determine the impact of a tweet by analyzing the general sentiment (positive or negative) of the replies.

Sentiment Analysis Analyzing text to determine positive or negative sentiment. Sentiment analysis can be used to determine attitudes toward a piece of media. In this study, we use sentiment analysis to determine the traumatic nature of news tweets.

Word2Vec A method of word vectorization which takes into account the context of the word and is used to train a machine learning algorithm to understand the meaning of words in context.

TF-IDF Vectorization TF-IDF stands for Term Frequency - Inverse Document Frequency. This method of vectorization takes into account the frequency of a term within a document as well as its frequency across documents to determine its overall relevance.

Regression Model The regression model we use in this study is the *SGDClassifier*, or Stochastic Gradient Descent Classifier. This model takes training data and fits it to a linear model, allowing it to predict new values.

Neural Network We use *Keras*, which is a high level deep learning API developed by Google for implementing neural networks.

Related Work

In this section, we discuss various methods used by other studies for detecting hate speech and offensive language on social media, as well as the potential impact of traumatic content on social media users.

Detecting Hate Speech and Offensive Language

Davidson et al. [7] discuss the drawbacks of using bag-of-words techniques to identify hate speech, as the context of the words being used must often also be taken into consideration. Supervised approaches to hate speech detection often confuse hate speech with offensive language. The study uses a lexicon of hate speech words to search for tweets containing these terms, but when users were asked to determine which of the tweets actually contained hate speech, only 5% of the tweets were identified as containing hate speech. The study creates unigrams, bigrams, and trigrams, using TF-IDF vectorization to weight terms considered most important to identifying hate speech. They test various models on the data, finding that logistic regression and linear SVM performed the best.

Spread of Hate Speech on Social Media

In the study "Characterizing and Detecting Hateful Users on Twitter," Ribeiro et al. [8] use a directed graph of retweets, where an edge from one user to another indicates a retweet. The study identifies hateful users based on a sample of 200 of their tweets and a lexicon of hateful words.

In order to examine the spread of hate speech on the social media platform Gab, Mathew et al. [9] use a lexicon of unigrams and bigrams commonly associated with hate speech to filter potential hate speech. They use a similar method to identify hateful users, flagging a user as hateful based on the number of posts containing phrases from the lexicon.

Both studies examine how users marked as hateful are connected between one another as well as users not flagged as hateful. They find that hateful users are more connected amongst one another (i.e. reposting one another's posts, following each other, etc.)

Trigger Warnings and Traumatic Content

Trigger warnings are becoming more prevalent on social media and act as a warning for any user who may be triggered by the content in the post due to past trauma. Jones et al. [10] discover through their study that trigger warnings on social media posts generally do not help in reducing trauma responses, and in fact could marginally increase anxiety in a person who has experienced trauma related to the trigger warning.

Classification of Trigger Warnings on Social Media

Sekerka-Bajbus [11] uses both regression models and neural networks to classify Reddit posts by trigger warning based on the subreddit in which they were posted. The project extracts posts from nine subreddits associated with certain triggers, such as anxiety, depression, and abuse, splitting the data and using four different classification methods to identify the triggers associated with each post.

Case Study

In one case study [12], a woman continuously received Facebook ads surrounding health conditions in children and adults. Constantly seeing these ads caused her anxiety, as she had lost a parent to cancer and was the mother of a young child. In addition, she was unable to control what ads she saw, as when she requested to view fewer ads about parenting or health, similar ads would eventually reappear under new categories. In this case, she had no control over what she was seeing on social media, and it was adversely affecting her mental health, bringing up past trauma and causing anxiety.

Our objective in this study is to determine the impact such content has on social media users. We attempt to identify traumatic content in the news on Twitter and analyze the impact it has on Twitter users. Our study is a step towards combating the reach of traumatic content and reducing its negative impact on users.

Dataset

We used the Twitter API to extract the 400 most recent tweets at the time of data collection from the *Fox News* and *The New York Times* Twitter accounts. Additionally, we extracted the likes, retweets, quote tweets, and replies. To train our sentiment analysis models, we utilize two different datasets. The dataset Psychological Response Data on the Traumatic Nature of 600 Written Events is a dataset rated by participants on a scale of 1 ("not at all traumatic") to 7 ("extremely traumatic"). We also use the Sentiment140 dataset, containing 1.6 million tweets marked as containing positive, negative, or neutral sentiment. We cleaned the data by removing mentions, links, and images contained within the tweets.

Approach/Methodology

Vectorization

We used the GloVe Twitter [13] pre-trained word vectors to represent our data. We created sentence vectors by averaging the word vectors of every word in the sentence. This method was used for both the extracted tweets and the datasets we used to train our sentiment analysis models.

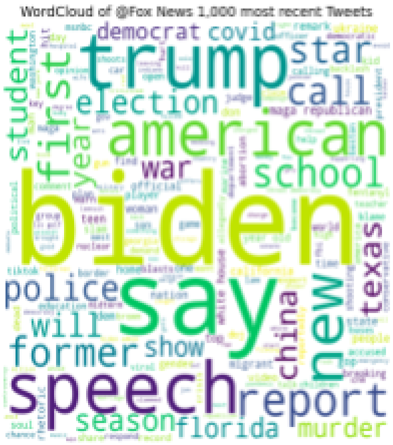
Training Models

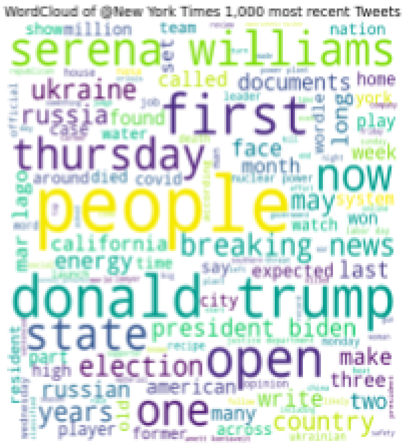
Our objective was to classify the tweets as traumatic or non-traumatic. To achieve this, we trained two different models on both the dataset of 600 traumatic events and the Sentiment140 dataset, creating four different classifications. We utilized both regression and neural network methods.

We used both the *SGDClassifier* regression model from the *scikit-learn* library [14] and the *Keras* neural network from the *TensorFlow* library [15]. We first trained both models on the dataset of 600 traumatic events, using an 80-20 train-test split. The regression model produced an accuracy of around 80.83%, while the neural network achieved an accuracy of around 84.16%. We then trained both models using the Sentiment140 dataset. When using this dataset, the regression model had an accuracy of 73.03%, and the neural network achieved 75.63% accuracy.

Implementing Models

Our two best-performing models were the regression and neural network trained on the dataset of 600 events, so we used these to classify tweets from *Fox News* and *The New York Times* as traumatic or non-traumatic. The *SGDClassifier* regression model identified 43.78% of the *New York Times* tweets and 45.00% of the *Fox News* tweets as traumatic, while the *Keras* neural network identified 44.04% of the *New York Times* and 37.00% of the *Fox News* tweets as traumatic.





Determining Reach

Twitter defines engagement with a tweet as the "total number of times a user interacted with a Tweet." The more engagement a tweet gets, the more it is promoted on the platform. For our purposes, we examine four types of tweet interactions: likes, replies, retweets, and quote tweets. Retweets and quote tweets increase the reach of a tweet because retweets and quote tweets are likely to be seen by the followers of the retweeter or quote tweeter as well as the followers of the original author of the tweet, increasing engagement with the original tweet. Additionally, when a user replies to a tweet, this reply appears on the user's profile, increasing the likelihood of the user's followers seeing the reply and the original tweet. Accordingly, we weight retweets and quote tweets the highest, with replies weighted higher than likes. We calculate reach based on the following metric:

$$Reach = Retweets + Quote\ Tweets + Replies + Likes$$

Analyzing Impact

We determined the impact (positive or negative) of a tweet by analyzing the sentiment of the tweet's replies. In order to achieve this, we used the VADER (Valence Aware Dictionary for Sentiment Reasoning) sentiment analyzer from the *Natural Language Toolkit (NLTK).* The VADER model accounts for polarity (positive and negative) as well as intensity (strength of emotion).

Results/Analysis

Reach

We calculated the reach scores of each tweet and found the mean and median of scores among tweets classified as traumatic and non-traumatic.

The New York Times Reach Scores	Median	Mean
SGDClassifier Traumatic	298.00	974.41
SGDClassifier Non Traumatic	298.50	814.10
Keras Traumatic	285.00	902.41
Keras Non Traumatic	319.00	871.35

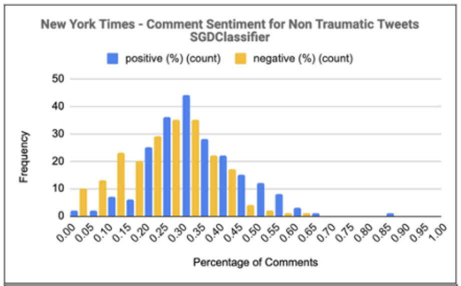
Among the *New York Times* tweets, the median reach score appears to be similar across all classifications. However, the mean reach score appears to be higher for events classified as traumatic both by the *SGDClassifier* regression model and the *Keras* neural network, but especially by the *SGDClassifier*. This indicates that though the distributions of reach scores for tweets classified as both traumatic and non-traumatic are positively skewed, the *New York Times* tweets classified as traumatic tend to have much higher reach scores than non-traumatic tweets on the high end. These results suggest that among *New York Times* tweets, news classified as traumatic by our models tends to have a further reach than news classified as non-traumatic.

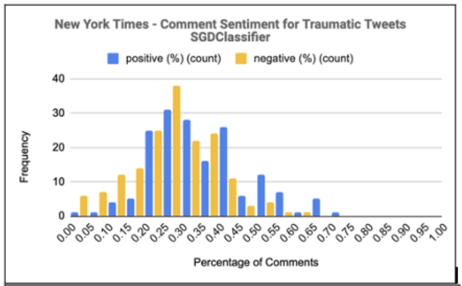
Fox News Reach Scores	Median	Mean
SGDClassifier - Traumatic	166.50	345.64
SGDClassifier - Non Traumatic	196.50	438.74
Keras - Traumatic	170.50	377.55
Keras - Non Traumatic	182.00	407.43

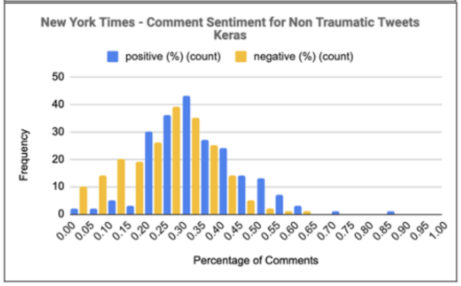
Among the *Fox News* tweets, on the other hand, the mean and median reach scores appear to be similar across classifications, with the tweets classified as non-traumatic actually having higher mean reach scores than the tweets classified as traumatic. This data suggests that among *Fox News* tweets, the traumatic nature of the news as identified by our models has little effect on the reach of the tweets.

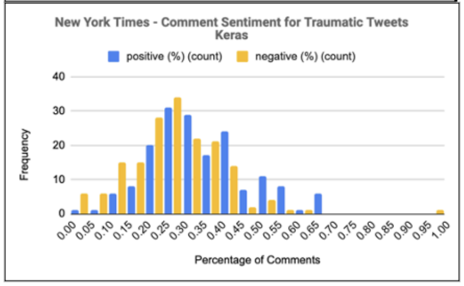
Impact

Using histograms, we plotted the distribution of percentage of positive and negative replies among tweets classified as traumatic and non-traumatic. We excluded replies identified as having neutral sentiment, since they provided no additional information about the impact of the tweets on Twitter users.

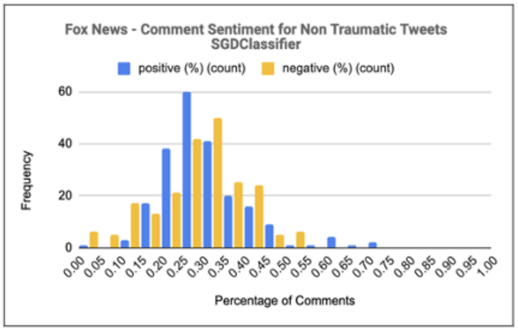


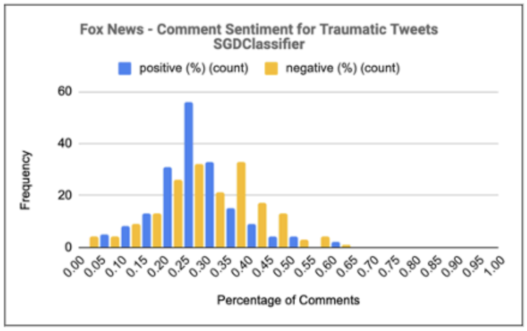


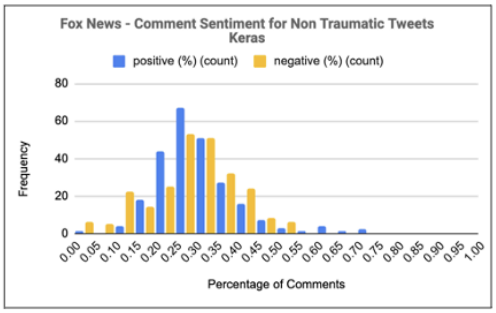


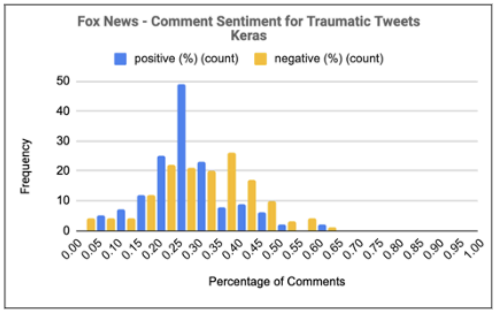


Among *New York Times* tweets, the distributions of positive and negative replies appear to be around the same for tweets classified as both traumatic and non-traumatic. The centers of both distributions appear to be around the same across all classifications, indicating that among the *New York Times* tweets, the percentage of positive replies and the percentage of negative replies are around the same regardless of whether the tweet was classified as traumatic or non-traumatic. This suggests that the traumatic nature of New York Times tweets had little impact on Twitter users, given that the general sentiment of their replies was relatively unchanged.









Among *Fox News* tweets, the distributions of positive and negative replies appear to be similar across tweets classified as non-traumatic. Across tweets classified as traumatic, however, the centers of the distributions of negative replies appear to be slightly higher than the centers of the distributions of positive replies, indicating that among the *Fox News* tweets, tweets classified as traumatic tended to have a higher percentage of negative replies than tweets classified as non-traumatic. This suggests that the traumatic nature of *Fox News* tweets did have an impact on Twitter users, since their replies tended to be generally more negative among tweets classified as traumatic than among tweets classified as non-traumatic.

Limitations

The study was limited by the small size of the traumatic events dataset. In further work, we aim to expand the data by collecting tweets from other social media accounts and potentially other social media platforms or news articles. We also plan to collect further data on replies and retweets, including collecting data on the followers of the users who retweet news tweets, as this will give us a better metric for determining the true reach of potentially triggering or traumatic tweets. We will consider incorporating a dictionary of words commonly associated with traumatic events to improve identification of tweets containing traumatic content. In addition, manual classification of tweets as traumatic or non-traumatic may be useful in accurately evaluating the efficacy of our models in identifying traumatic content on social media.

Conclusion

We used two different methods of classification to identify traumatic news on social media. We chose to collect tweets from both *The New York Times* and *Fox News* because of the vastly differing political leanings of the two news sources, which may have influenced the reach and impact of the tweets. In the case of *The New York Times*, content identified as traumatic appears to have a farther reach but little difference in the sentiment of the replies from content identified as non-traumatic. On the other hand, tweets from *Fox News* appear to have similar reach regardless of their traumatic nature but have a higher percentage of negative replies on tweets identified as traumatic by our models.

While the size of the data collected may have impacted our results, the results of our study demonstrate that traumatic content on social media does have a far-reaching impact on social media users. Further work on this topic, including expanding our dataset of tweets as well as identifying other methods of classifying traumatic content, could provide more information on the impact traumatic and triggering content has on social media users. These findings can help us understand how to mitigate the negative impact of traumatic news—what type of language elicits negative reactions from users, and whether there are ways of rewording information to soften the impact of such news on readers.

Though trigger warnings are one method of self-regulating what content one consumes on social media, they have been extensively researched and proven to negatively affect social media users. Detecting traumatic content using machine learning methods rather than manually applied trigger warnings will allow us to understand the specifics of how certain language and content affects users and can help in preventing this content from harming social media users.

References

1: Jones, P. J., Bellet, B. W., Levari, D. E., & McNally, R. J. (2021). Psychological Response Data on the Traumatic Nature of 600 Written Events. Journal of Open Psychology Data, 9(1), 3. DOI: http://doi.org/10.5334/jopd.46

2: McHugh, B.C., Wisniewski, P., Rosson, M.B. and Carroll, J.M. (2018), "When social media traumatizes teens: The roles of online risk exposure, coping, and post-traumatic stress", Internet Research, Vol. 28 No. 5, pp. 1169-1188. https://doi.org/10.1108/IntR-02-2017-0077

3: Abdalla, S. M., Cohen, G. H., Tamrakar, S., Koya, S. F. & Galea, S. (2021). Media Exposure and the Risk of Post-Traumatic Stress Disorder Following a Mass traumatic Event: An In-silico Experiment. Front.Psychiatry12:674263.doi:https://doi.org/10.3389/fpsyt.2021.674263

4: Kim, S. A. (2020, August 27). *Social Media Algorithms: Why you see what you see*. Georgetown Law Technology Review. Retrieved September 6, 2022, from https://georgetownlawtechreview.org/social-media-algorithms-why-you-see-what-you-se e/GLTR-12-2017/

5: Leonard, J. (2020, June 3). *What is trauma? types, symptoms, and treatments*. Medical News Today. Retrieved August 20, 2022, from https://www.medicalnewstoday.com/articles/ trauma

6: Pedersen, T. (2022, April 28). *Triggers: What they are, how they form, and what to do*. Psych Central. Retrieved August 20, 2022, from https://psychcentral.com/lib/what-is-a-trigge r#what-is-a-trigger

7: Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). In International AAAI Conference on Web and Social Media. Retrieved from https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665/14843

8: Ribeiro, M., Calais, P., Santos, Y., Almeida, V., & Meira Jr., W. (2018). Characterizing and Detecting Hateful Users on Twitter. Proceedings of the International AAAI Conference on Web and Social Media, 12(1). Retrieved from https://ojs.aaai.org/index.php/ICWSM/articl e/view/15057

9: Mathew, B., Dutt, R., Goyal, P. & Mukherjee, A. (2019). Spread of Hate Speech in Online Social Media. 173-182. 10.1145/3292522.3326034. https://doi.org/10.1145/3292522.3326034

10: Jones, P. J., Bellet, B. W., & McNally, R. J. (2020). Helping or Harming? The Effect of Trigger Warnings on Individuals With Trauma Histories. Clinical Psychological Science, 8(5),905–917. https://doi.org/10.1177/2167702620921341

11: Sekerka-Bajbus, K. (2021). Classification of 'Triggering' Content on Social Media. GitHub. Retrieved September 6, 2022, from https://github.com/ksek87/trigger-warning-classification

12: *Algorithms of trauma: New case study shows that Facebook doesn't give users real control over disturbing surveillance ads*. Algorithms of trauma: new case study shows that Facebook doesn't give users real control over disturbing surveillance ads | Panoptykon Foundation. (2021, September 28). Retrieved September 6, 2022, from https://en.panoptykon.org/algorithms-of-trau ma

13: Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. Retrieved September 6, 2022, from https://nlp.stanford.edu/projects/glove/

14: Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Retrieved from https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

15: Chollet, F., et al. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras