Sentiment Asymmetry in Social Media: How Negative Political Content Drives Disproportionate Engagement

Alex Alvarez Mannan Mishra Margaret Elizabeth Royal Fuling Wang

2025-04-21

0.1 Introduction

Social media has proven itself a vital source of information for many in today's society. Breaking news is announced through tweets from journalists, and news agencies. Tweets covering corporate announcements have an effect on the stock market's volatility. Many rely on social media in their everyday lives for communication with friends and family. Ultimately, we aim to understand if there exists a relationship between polarized sentiment from tweets and their engagement generated? Posing the question can help expose patterns in user's behavior, have algorithmic implications for social media platforms incentivized to generate engagement, and shape perceptions on the internet possible through forms of mistrust and misinformation. Conducting a descriptive analysis on the relationship between polarized sentiment and engagement helps lay the groundwork on more complex studies like causal and explanatory analysis.

0.2 Description of the Data Source

Sourcing the data to conduct a descriptive analysis came directly from the user Enryu on the platform, Hugging Face (Enryu. 2023). An open source repository providing pre-trained models & datasets for NLP tasks. We took two large datasets each containing 100 million rows of records. One dataset containing tweets and their relevant attributes such as number of replies, retweets, and likes (Enryu. 2024). The other dataset focused on Twitter users and their receptive attributes such as follower count (Enryu. 2021). We joined the two dataset by username for our analysis. The large scale of records posed an issue running the data in-memory on our devices. As a solution we sourced the collection of the data through a randomization method called reservoir sampling (Konstantinovsky. 2024). Reservoir sampling is an algorithm that works by allowing us to randomly sample records from a stream of data without needing to specify the size of the stream. For instance, we select a random sample of the entire list or stream of data with equal chance of being selected and collected in our reservoir (sample) or skipped. This method was selected to ensure efficiently selecting with equal probability a random distinct sample of tweets from the large dataset. We chose to pull tweets from March 2023 only since the data was pulled from April 2023 we wanted the follower counts to be closely representative of the followers at the time of the tweet. Overall, the number of units selected totaled 500,000 tweets.

0.3 Data Wrangling

In the tweets dataset, we generated new columns leveraging langdetect, a LLM trained on identifying languages (Danilk, 2016). The final dataset will only incorporate tweets that were tagged as English and discarded other languages and unidentified languages (URLs and emoji only tweets). Additionally utilizing the LLM model that was pre trained on evaluating topics for each tweet with categories: News & Social Concern (politics) and made all other topics (Diaries & daily life, Sports, film tv & video, music, etc.) as Other (Tweet-Topic-Base-Multilingual. 2024). The topic model also came with a score for confidence the model is in that topic. We opted to take tweets with low confidence in the topic and move them to the other category.

We can define engagement as the interactions generated from other users when a tweet was posted. This can take the form of other users liking, retweeting, replying, and/or quote tweeting a posted tweet. Engagement is the numeric score derived from quantifying & summing a tweet's number of likes, retweets, replies, and quotes into an aggregated metric score. This measurement is a variation found in a similar online site qualifying engagement (Twitter Engagement Calculator, n.d.).

We then used a language model called roBERTa, which is based on BERT but specifically fine-tuned to classify Twitter sentiment (twitter-roberta. 2021). We can define sentiment as the emotions conveyed from a tweet's content. This can take the form of sentiment conveying negative, neutral, and positive emotions. It produces corresponding sentiment scores (0 to 1) for negative, neutral, and positive emotions. To categorize a tweet by sentiment, we assigned a tweet by the dominant sentiment score and its respective sentiment label.

To better capture the relationship between sentiment strength and engagement we used the 2-dimensional approach for negative and positive sentiment, since a strongly neutral sentiment doesn't convey meaningful information about sentiment intensity.

0.4 Operationalization

In order to prepare the intended dataset, a number of considerations and filtering criteria were applied. We established a data pipeline that had a filtering process for the tweets sampled. The criteria include the exclusion of tweets with video/links since we wanted to focus our analysis on text, also seen in a similar study looking at Twitter data on the 2020 presidential election (Ondocin. 2020).

The removal of tweets with zero engagements to ensure that the sample would be representative of twitter at large. The removal of tweets that were direct replies to other tweets sampled. This scenario poses a risk for the independently identically distributed (i.i.d) assumption for our model. Since tweets and their respective replies may have an influence on the engagement generated on replies there exists a dependence on the engagement generated violating i.i.d (Schöne et al., 2021).

We then had to exclude Twitter users that had less than 100 followers and exclude those with a large following in the top 5th percentile since they are mostly commercial accounts. Commercial accounts and celebrities' behavior has been known to differ from normal users (Brady et al., 2017). Establishing a filtering criteria for tweets that were only 1-2 words and avoiding tweets with excessively long word count, >75 words. Tweets with only 1–2 words, such as 'yes' or 'yes sir,' provide little information. At the upper end, 99% of tweets have 75 words or fewer. Since tweets have a character limit, a tweet with more than 75 words is unlikely to contain normal text. By removing these outliers, we ensure that the word count variable is more meaningful for analysis. Overall the number of observations pulled from the 100 million were 500,000 and as part of the filtering process totalled 384,742 records removed leaving the 115,258 tweets. According to prior literature the requisite number of tweets needed for a study of this nature is 100,000 (Brady et al., 2017). In our final single data set there is no clustering on users since each user represents less than 0.1% of the observations in our filtered dataset.

0.5 Data Visualized

Before building our model we examined the relationship between sentiment score and engagement. The best way to visualize this relationship is by making negative sentiment negative and to color by topic. Figure 1 includes the full zoomed out version. We can see that there are a lot of outliers in our dataset that make it hard to see visually what this relationship is. As we zoom in to a smoothing fit of the data we can see that on the positive side there is little effect of the strength of sentiment on engagement. For negative sentiment, strength of sentiment has a greater effect on engagement for politics than for other topics. In general politics gets more engagement than the other topics we can see here a higher intercept.

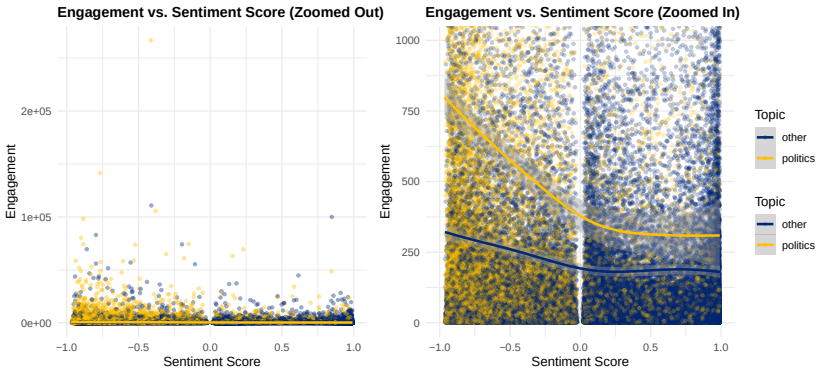

Figure 1: Engagement vs Sentiment Score

0.6 Model Specifications

To examine how sentiment strength relates to engagement for both positive and negative tweets, we tested several control variables — follower count, topic, word count, hashtag indicators, readability, mentions and average user engagement. After evaluating various model specifications – including using each variable individually, summing selected variables, and including interaction terms (see Model Specifications in appendix) – we retained follower count, topic, and word count as controls.

We also tested various transformations for both X and Y variables and applied a log transformation to the engagement score to address its heavy right skew. This helps compress the wide range of engagement scores and better align it with the scale of X variables. The plots below show the distribution of engagement scores before and after the transformation.

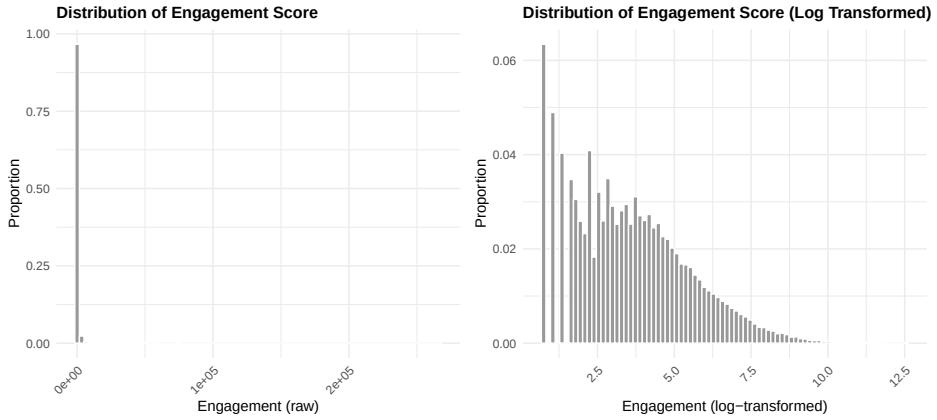

Figure 2: Engagement Score Distributions

0.7 Model Assumptions/Limitations

Our first assumption is that model residuals are independent and identically distributed (i.i.d.). To evaluate independence, we conducted a permutation-based Ljung-Box test. The histogram below shows that the p-values from permuted residuals are uniformly distributed between 0 and 1, as expected under the null hypothesis of independence. However, our observed p-value lies on the far left of this distribution, suggesting there is a strong correlation in our residuals. While this violates the independence assumption of linear regression, we verified that there is no apparent clustering in the data—such as repeated tweets from the same user or time-based patterns. Given our large sample size (>100k) and exogenous regressors, the OLS estimates should remain unbiased and consistent. However, standard inference (e.g., p-values and confidence intervals) may be invalid without adjustments for autocorrelation.

Our second key assumption is the existence of a unique Best Linear Predictor (BLP), which requires no perfect collinearity among predictors and finite residual variance. We assessed multicollinearity using the Variance Inflation Factor (VIF), and found that all variables had VIF values below the conventional threshold of 10, suggesting no problematic collinearity. Additionally, the residual histogram appears approximately normal, without extreme skew or heavy tails, supporting the assumption of finite variance. Taken together, these diagnostics suggest that a unique and stable BLP is likely in our model.

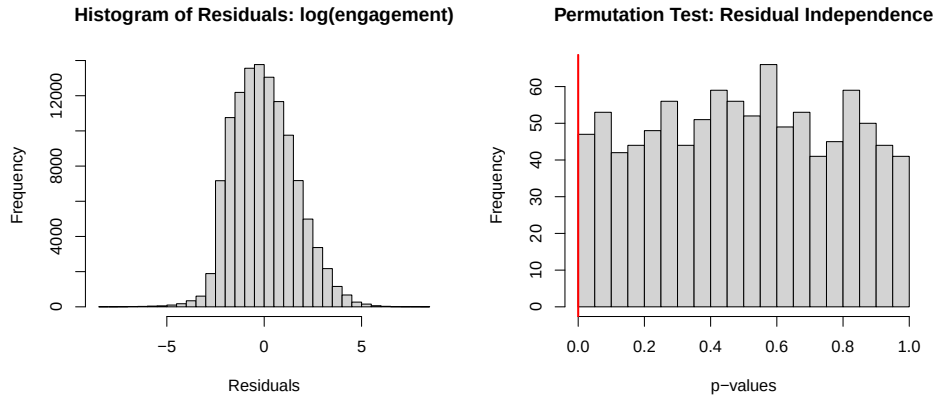

Figure 3: Residual Diagnostics

0.8 Model Results and Interpretation

The goal of the model is to understand the relationship between the strength of sentiment and engagement for both positive and negative tweets. Additionally we are interested in how this varies between political tweets and non political tweets. To do this we needed to find the coefficients (or slopes) of the relationship between sentiment score and engagement for four categories: negative politics, positive politics, negative other, and positive other. There were two ways to model these in our OLS regression. One was to interact with the 3 terms: sentiment score, sentiment label (positive/negative) and topic (politics/other). This would give us the relationships between sentiment and engagement for each of the categories as a relationship to one another (Table 2 in Appendix). This analysis shows that there is a statistically significant difference in the relationship between sentiment score and engagement for these 4 categories, but makes the results difficult to interpret because they are all in reference to one another. To help you better understand the relationships, we will focus on the four-category model in this report. This model is slightly different from the one model version because our control variables of follower and word count are also able to vary for the four categories.

Table 1: Engagement vs. Sentiment Score Regression Models

	Output Variable: Engagement			
	(1)	(2)	(3)	(4)
Sentiment Score	0.920***	0.237***	0.316***	0.041***
	(−0.001)	(−0.001)	(−0.001)	(−0.0002)
Topic: Politics	0.00001***	0.00001***	0.00001***	0.00001***
	(−0.000)	(−0.000)	(−0.000)	(−0.000)
Topic: Other	0.011***	0.009***	0.012***	0.012***
	(−0.00002)	(−0.00004)	(−0.00001)	(−0.00001)
Constant	2.439***	2.547***	2.421***	2.479***
	(0.001)	(0.002)	(0.001)	(0.0003)
Observations	25,698	9,009	32,426	48,125
R^2	0.331	0.258	0.286	0.235
Adjusted R^2	0.331	0.257	0.286	0.235
Residual Std. Error	1.685 (df = 25694)	1.612 (df = 9005)	1.563 (df = 32422)	1.552 (df = 48121)
F Statistic	4,233.243*** (df = 3; 25694)	1,041.280*** (df = 3; 9005)	4,326.079*** (df = 3; 32422)	4,938.584*** (df = 3; 48121)
Note:	*p<0.05; **p<0.01; ***p<0.001 HC_1 robust standard errors in parentheses. Politics is the reference category.			

Table 1 shows a larger coefficient for the relationship between sentiment score and log(engagement) for negative politics than for any of the other categories. There is statistical significance for all of the features in all of the models. To make these results more interpretable we have plotted them on top of our data below,

using the average followers and average word count. To adjust for the bias caused by the log transformation we added +(s2/2) to the output, where s is the sample standard deviation (Giles, 2016). The practical significance of the modeling results is that for an average follower count and word count there is a 176 increase in engagement as you go from neutral to negative for politics, but practically no increase for other, and for positive politics. As a tweet increases in its negativity we expect a proportional increase in engagement. These results vary from our original smoothing fit in Figure 2, due to the log transformation, which is more resistant to the effects outliers than the smoothing fit, which uses an average. In the appendix, Figure 5, you can see what a median smoothing curve looks like and it is more similar to our model results. Figure 6, in the appendix shows that there is some correlation between the fitted values and the residuals, meaning that there is some variance that is not explained by the model. There are many factors of what makes a tweet something users want to engage with that we did not take into consideration, and further analysis could improve this.

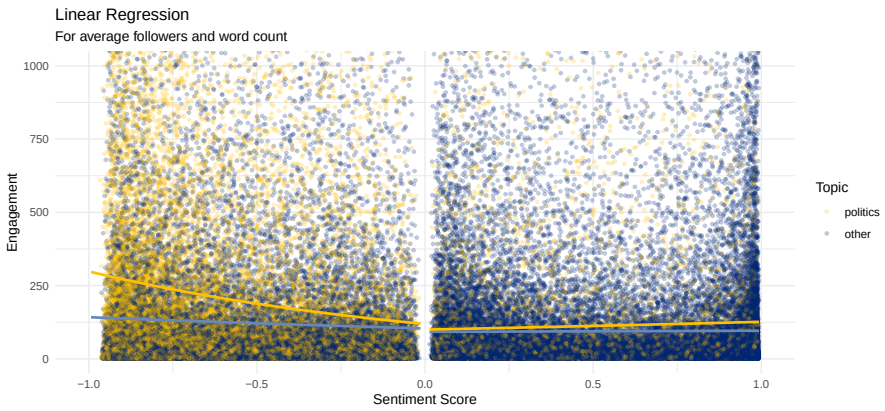

Figure 4: Linear Regression: Predicted Engagement vs Sentiment Score

0.9 Overall Effect

Our analysis reveals a significant pattern in social media engagement: negative sentiment, particularly in political content, generates substantially higher engagement compared to neutral or positive sentiment. For political tweets, there is an increase of approximately 176 engagement units when moving from neutral to negative sentiment, while positive political content and non-political content show negligible increases. This pattern creates concerning incentives within social media ecosystems, potentially rewarding content creators for producing increasingly negative and polarizing content to gain visibility.

These findings suggest emotionally charged negative information is more likely to achieve viral status, particularly regarding politically sensitive topics, potentially outpacing fact-checking mechanisms. This could contribute to information pollution and fuel a self-reinforcing cycle of negativity across platforms, including hate speech, as users adapt their posting behavior to mirror what receives attention. For platform designers and policymakers, our research highlights the tension between engagement-maximizing behavior and healthy public discourse, suggesting the need for interventions that maintain engagement without incentivizing harmful content patterns.

0.10 Appendix

0.10.1 Datasets:

- Enryu. (2024, August 31). twitter100m_tweets. Huggingface.co. https://huggingface.co/datasets/enryu43/twitter100m_tweets
- Enryu. (2021). twitter100m_users. Huggingface.co. https://huggingface.co/datasets/enryu43/twitter100m_users

0.10.2 Model Specifications

- Experimented with additional variables and didn't find they improved the model: hashtag bool, hashtag count, readability score, mention count, average user engagement.
- Experimented with several topic categories, but found politics the most interesting.
- One-dimensional sentiment encoding explains more variance alone, but adds no additional value when combined with sentiment labels. As our focus is on sentiment strength rather than polarity, we chose the two-dimensional encoding.
- Compared likes, retweets, replies and quotes and found that retweets increased the most when comparing neutral to negative for political tweets.
- Tried removing outliers and found it did not make the residuals more normal. The log transformation alone was sufficient for handling outliers.

Table 2: Engagement Regression Model with Sentiment and Topic Interactions

	Output Variable: Engagement (Natural Log)
Sentiment Score	0.904^{***}
	(0.040)
Sentiment Label: Positive	-0.079^{*}
	(0.039)
Topic: Politics	0.049
	(0.032)
Topic: Other	0.00001^{***}
	(0.00000)
Followers	0.012^{***}
	(0.0003)
Word Count	-0.334^{***}
	(0.014)
Hashtag Bool	-0.633^{***}
	(0.065)
Constant	-0.638^{***}
	(0.051)
sentiment_labelpositive:topic_other	0.004
	(0.046)
sentiment_score:sentiment_labelpositive:topic_other	0.410^{***}
	(0.075)
Constant	2.506^{***}
	(0.028)
Observations	115,258
R^2	0.302
Adjusted R^2	0.302
Residual Std. Error	1.589 (df = 115247)
F Statistic	$4{,}977.253^{***}$ (df = 10; 115247)
Note:	*p<0.05; **p<0.01; ***p<0.001
	HC_1 robust standard errors in parentheses. Politics and Negative Sentiment are the reference categories.

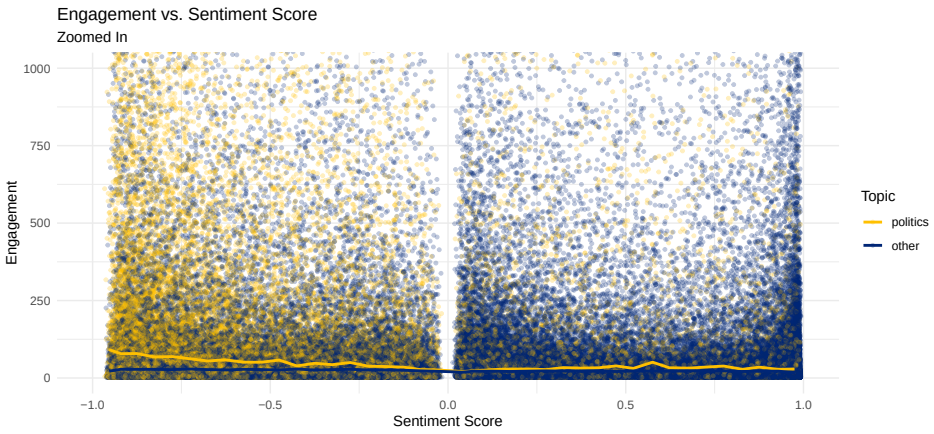

Figure 5: Figure B: Smoothed Median Trend - Engagement vs Sentiment Score

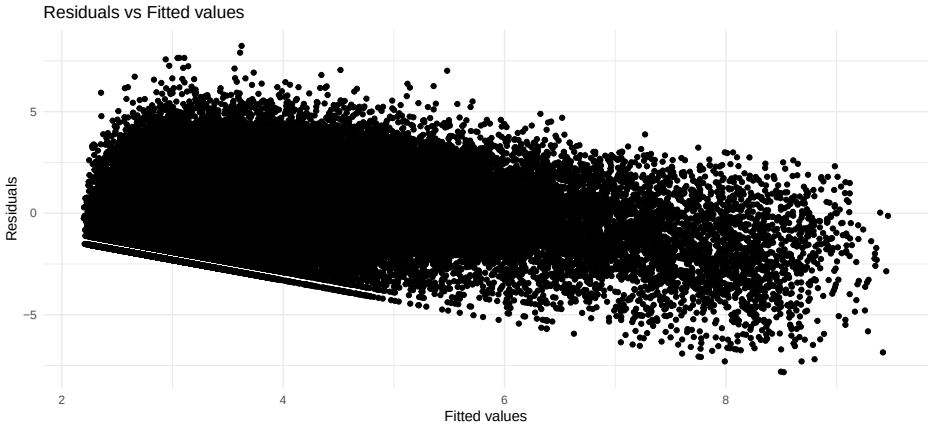

Figure 6: Figure C: Residuals vs Fitted Values - Residuals vs Fitted Values

1 References

Antypas, D., Preece, A., & Camacho-Collados, J. (2023). Negativity spreads faster: A large-scale multilingual twitter analysis on the role of sentiment in political communication. Online Social Networks and Media, 33, 100242. https://doi.org/10.1016/j.osnem.2023.100242

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. Review of General Psychology, 5(4), 323–370. https://doi.org/10.1037//1089-2680.5.4.323. PDF:https://assets.csom.umn.edu/assets/71516.pdf

Blau, Francine D, and Lawrence M Kahn. 2017. "The Gender Wage Gap: Extent, Trends, and Explanations." Journal of Economic Literature 55 (3): 789–865.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences, 114(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Bronnenberg, Bart J, Sanjay K Dhar, and Jean-Pierre H Dubé. 2009. "Brand History, Geography, and the Persistence of Brand Shares." Journal of Political Economy 117 (1): 87–115.

Cardiffnlp/tweet-topic-base-multilingual · Hugging Face. (2024, November 25). Huggingface.co. https://huggingface.co/cardiffnlp/tweet-topic-base-multilingual

Cardiffnlp/twitter-roberta-base-2021-124m · Hugging Face. (n.d.). Huggingface.co. https://huggingface.co/cardiffnlp/twitter-roberta-base-2021-124m

Conover, Pamela Johnston, Stanley Feldman, and Kathleen Knight. 1987. "The Personal and Political Underpinnings of Economic Forecasts." American Journal of Political Science, 559–83.

Danilk, M. (2016, October 3). langdetect. PyPI. https://pypi.org/project/langdetect/

Enryu. (2023, May 2). Fun with large-scale tweet analysis - Enryu - Medium. Medium. https://medium.com/@enryu9000/fun-with-large-scale-tweet-analysis-783c96b45df4

Giles, D. (2016). More on Prediction From Log-Linear Regressions. Blogspot.com. https://davegiles.blogspot.com/2014/12/s.html

Hagemann, L., & Abramova, O. (2023). Sentiment, we-talk and engagement on social media: insights from Twitter data mining on the US presidential elections 2020. Internet Research. https://www.researchgate.net/publication/367205481_Sentiment_we-talk_and_engagement_on_social_media_insights_from_Twitter_data_mining_on_the_US_presidential_elections_2020

Konstantinovsky, T. (2024, August 29). Dipping into Data Streams: The Magic of Reservoir Sampling. Medium; The Pythoneers. https://medium.com/pythoneers/dipping-into-data-streams-the-magic-of-reservoir-sampling-762f41b78781

linusha. (2020). GitHub - linusha/twitter-sentiment-2020-election: Code for data collection, processing and analysis as well as the data-set used for the research paper "Crafting Audience Engagement in Social Media Conversations: Evidence from the U.S. 2020 Presidential Elections" presented at HICSS 2022. An extended version has been published in "Internet Research". GitHub. https://github.com/linusha/twitter-sentiment-2020-election

Ondocin, R. J., Cao, J., Datar, P. P., & Shripad Laddah. (2020, December 15). Sentimental Analysis of Twitter Data from the US Presidential Election of 2020. https://doi.org/10.13140/RG.2.2.14910.92488

Schöne, J. P., Parkinson, B., & Goldenberg, A. (2021). Negativity Spreads More than Positivity on Twitter After Both Positive and Negative Political Situations. Affective Science, 2, 379–390. https://pmc.ncbi.nlm.nih.gov/articles/PMC9383030/

Twitter Engagement Calculator (Free Tool). (n.d.). Mention.com. https://mention.com/en/twitter-engagement-calculator/