Analyzing the Impact of High-Profile Events on Public Sentiment through Twitter

Team Name: SentiTweet
Logins: ymao36, clingzhi, eli55

Hypothesis

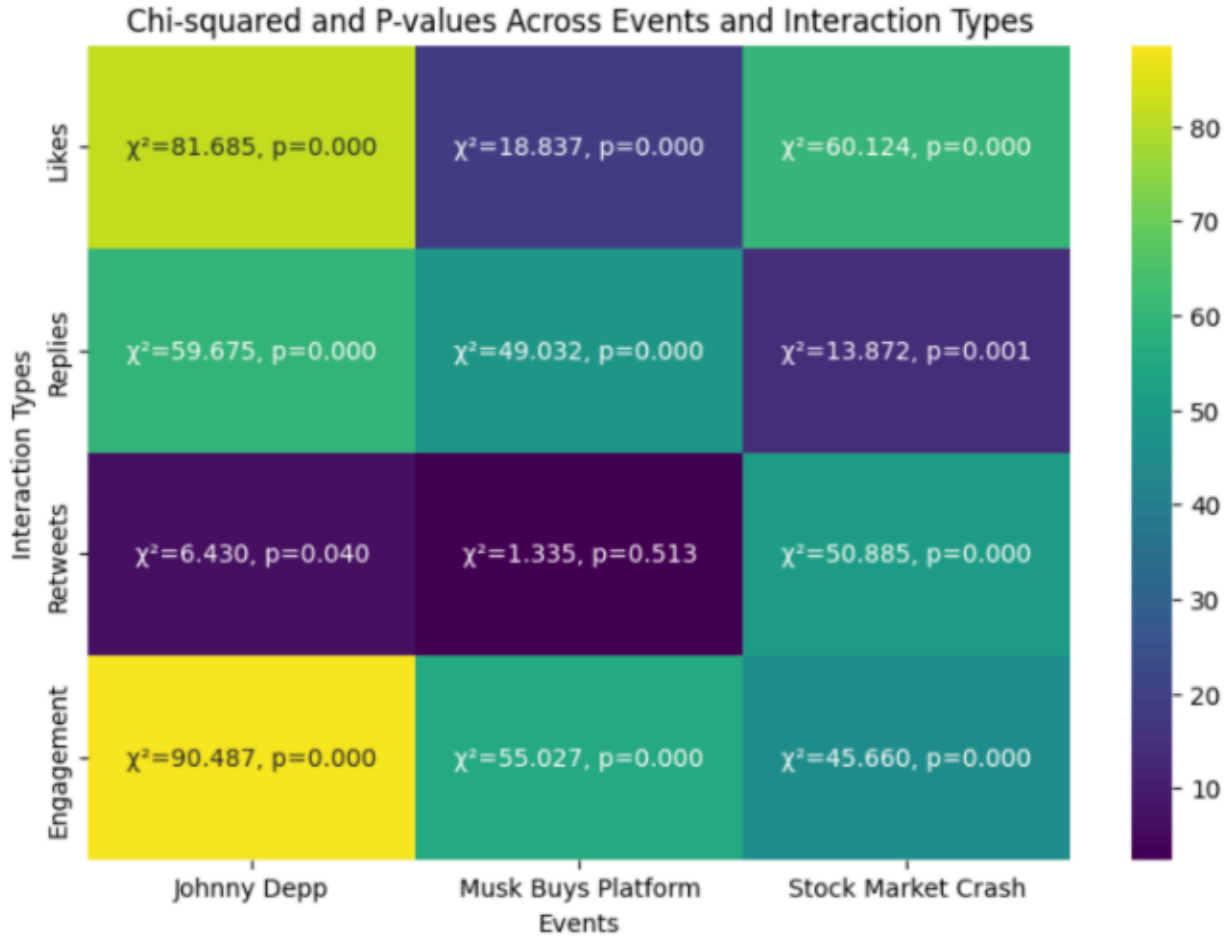
Social media platforms, like Twitter, reflect and can influence public sentiment. We hypothesize that engagement metrics (likes, comments, retweets) and user follower count positively correlate with the intensity of sentiment in social media comments. Posts with higher engagement and those from users with larger followings likely contain more pronounced sentiments. Understanding these relationships provides insights into the factors amplifying sentiment expression, allowing stakeholders to develop effective communication strategies, anticipate public reactions, and make informed decisions based on public opinion. Analyzing Twitter data is crucial for gauging public sentiment and its societal impact.

Data

Our dataset comprises three Kaggle Twitter datasets: "Users reaction to Musks' Takeover & Cancellation", "Twitter reacts to Johnny DEPP's win", and "Huge crash in the Stock market 2022". After preprocessing, we have 379,089 data points divided into three event-based groups with 50,760, 60,606, and 267,723 entries, respectively. The data was relatively clean, requiring only sentiment analysis using the open-source polyglot library and tokenization. The dataset provides a diverse sample of Twitter reactions to significant events, allowing for a comprehensive analysis of sentiment expression and its relationship with engagement metrics and user follower counts.

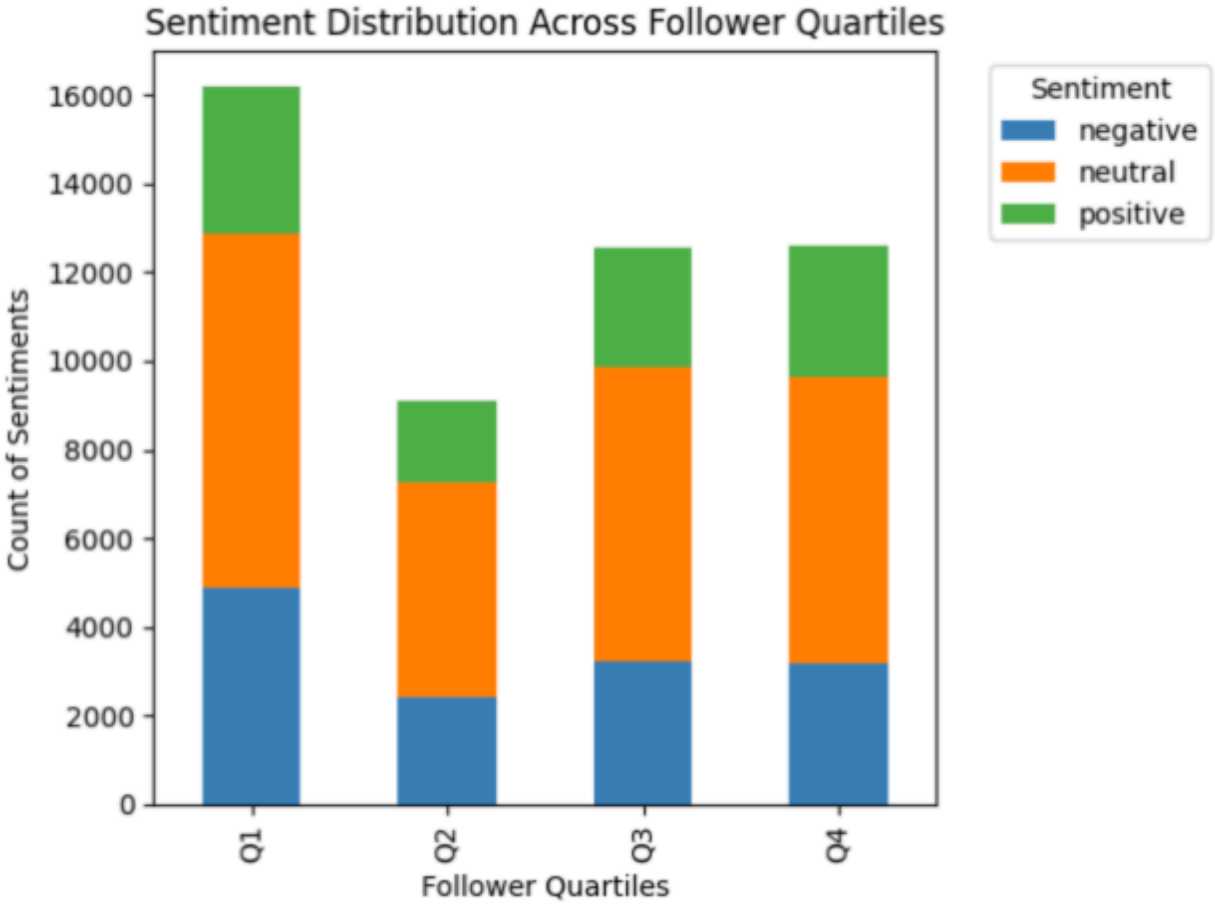
Findings

Claim #1: Engagement metrics such as likes, comments, and retweets have a correlation with more intense sentiment expressed in social media comments.



Support for Claim #1:
The chi-square analysis shows significant differences in sentiment distribution across interaction types for most events. For the "Johnny Depp" event, likes (χ^2=81.685), replies (χ^2=59.675), and engagement (χ^2=90.487) show highly significant differences (p=0.000), while retweets (χ^2=6.430, p=0.040) have a less significant impact. The "Stock Market Crash" event shows significant differences (p≤0.001) for all interaction types, with chi-square values ranging from 13.872 to 60.124. However, the "Musk Buys Platform" event has mixed results, with retweets (χ^2=1.335, p=0.513) showing no significant impact on sentiment distribution, while other interaction types demonstrate significant differences (χ^2=18.837 to 55.027, p=0.000).

Claim #2: Users with higher follower counts tend to express more pronounced sentiments, whether positive or negative, possibly due to their broader reach and potential sense of accountability to a larger audience.



Support for Claim #2:
The chi-square analysis shows significant differences in sentiment distribution across follower quartiles (χ^2=144.37, p=0.000). Quartile 1 (lowest followers) has the highest proportion of negative (4884) and neutral (7983) sentiments, while Quartile 4 (highest followers) has the highest proportion of positive sentiments (2953). Quartiles 2 and 3 show a similar pattern, with neutral sentiments being the most common (4843 and 6620, respectively), followed by negative (2417 and 3233) and positive (1858 and 2696) sentiments. These results suggest that follower count significantly impacts sentiment expression across different social media user groups.

Socio-historical Context and Impact Report

socio-historical context

Describe a few of the broader societal issues and their relationship to your data, prediction goal, and/or hypothesis.

Our project analyzes the sentiment of tweets posted by people for different social events and makes predictions based on existing data. However, different events occur at different times, and different times have different social contexts. In our analysis and prediction, we don't take those social contexts into account too much, and this could affect our source datasets, hypothesis, and prediction accuracy. For example, the severity of the pandemic increases the negative sentiments of the general public to varying degrees, and the current economic situation can also affect the sentiments of the general public. And these are invisible factors that can change people's sentiments to social events.

Who are the major stakeholders in this project? What is your relationship to these stakeholders?

The major stakeholders in our project are the general public (including us), the media or business people who want to keep tabs on public opinion, and the people involved in the specific social event itself. If someone tries to use the results of our research to make a profit by publishing a misleading statement or tries to manipulate public opinion, they will benefit and the general public will be potentially harmed.

Summarize the most relevant technical or non-technical research that has already been conducted about your project topic.

Research that is most relevant to our project topic is sentiment analysis itself. Sentiment analysis is common, and it is often applied to analyze social media reactions toward new products, helping businesses better understand user feedback and detect potential business opportunities. It digitizes sentiment and is an important tool in business.

ethical considerations

What biases might exist in your interpretation of the data?

For the selection of data sources, we started with the idea of choosing relatively neutral events (not overly positive or negative) and would subconsciously want to try to avoid very sensitive events that would make the viewer with a special identity uncomfortable. This idea may influence the development of our hypotheses.

How could an individual or particular community's privacy be affected by the aggregation or analysis of your data? / Is data being used in a manner agreed to by the individuals who provided the data?

Consider a situation: a user posts a tweet (irrationally and negatively), then quickly regrets and deletes it, but someone has already recorded it in a dataset and made it available for others to analyze. For this tweet, the user possibly didn't want it to be permanently recorded or even used for sentiment analysis. And if someone had recorded it and used it for analysis, the user wouldn't have known about it. So an individual's privacy could be affected by the aggregation of the data, and when we used the dataset, we might not use the data in a manner agreed to by the individuals who provided the data.

What are possible misinterpretations or misuses of your project results and what can be done to prevent them?

Just as previous discussion, the possible possible misinterpretations or misuses include if someone tries to use the results of our research to make a profit by publishing a misleading statement or tries to manipulate public opinion.
To prevent this, first of all, because we didn't analyze recent events but events from 2022, it is difficult to use the results for profit. Unable to make a profit means it will just be used primarily for research and misuse will be reduced. Other things can be done to prevent possible misinterpretations or misuses included, we will hide relatively sensitive content and blur precise numbers if we release the project publicly. And we will also remind users of privacy and ethical considerations by featuring them in the public release.

Citations

Talevi D, Socci V, Carai M, Carnaghi G, Faleri S, Trebbi E, di Bernardo A, Capelli F, Pacitti F. Mental health outcomes of the CoViD-19 pandemic. Riv Psichiatr 2020;55(3):137-144. doi 10.1708/3382.33569

Mathew, W. (2024, March 4). The what, why and how of sentiment analysis. Meltwater. https://www.meltwater.com/en/blog/analyse-sentiment-with-media-intelligence

Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of big data privacy. IEEE access, 4, 1821-1834.