Sentiment Search: Make the Internet Your Focus Group

Justin Zandstra, Faris Durrani, RenChu Wang, Nemath Ahmed, Shuyan Lin, Laksmisree Iyengar | 2022

1. Introduction

In marketing, you need to be able to answer the question "How does the public view our product?". The primary method for doing this is focus groups, which are expensive and time consuming. We have created Sentiment Search, a tool which displays the general sentiment of the internet about a given topic by visualizing the average sentiments of multiple social media and online news sites, allowing marketers to explore the public's view of their product.

2. Data

To get a view of sentiments across the whole internet, we analyze text data from multiple platforms, including social media sites (Twitter, Facebook, and Reddit), and news sources (CNN, The New York Times, and The Guardian). Each platform had differing levels of security around their data, so for some site we had to collect data through web scraping, and others had large downloadable sets of data. We wanted to make sure that we collected the full text, the date published, and then computed its sentiment using a library. For Reddit and Twitter, we found large dowloadable datasets of posts. These datasets were too large to parse quickly, so we removed posts randomly to get a smaller sample. For the other platforms, we collected posts through scraping while limiting the amount collected. Once we had sources for our data, we collected the text, date, and platform of every post.

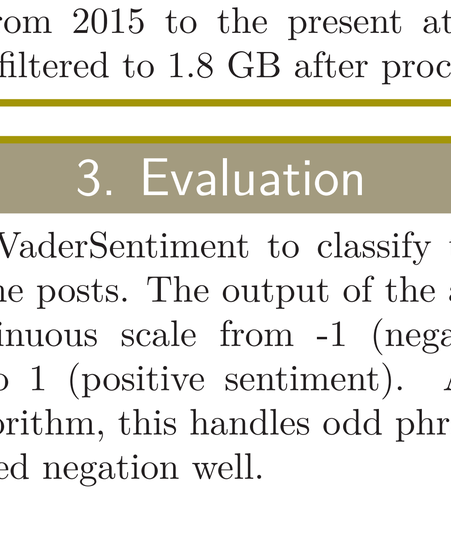
We also collected a set of significant events from the time period of the posts we collected to add some context to our visualization by scraping Wikipedia. We collected over 2.3 million posts ranging from 2015 to the present at 190 GB which we filtered to 1.8 GB after processing.

3. Evaluation

We used VaderSentiment to classify the sentiment of the posts. The output of the algorithm is a continuous scale from -1 (negative sentiment) to 1 (positive sentiment). As a rule based algorithm, this handles odd phrasing and complicated negation well.

We evaluated the results of the classification by testing sentences on the algorithm. To test the algorithm, we used Amazon reviews text and their rating as correct label. We found that Vader-sentiment can capture general sentiments well, and does a better job detecting strong emotions.

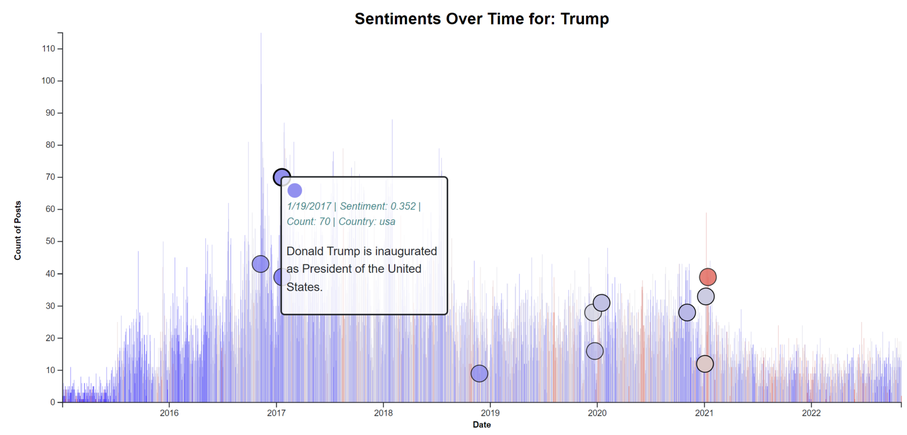
5. Platform Polarity



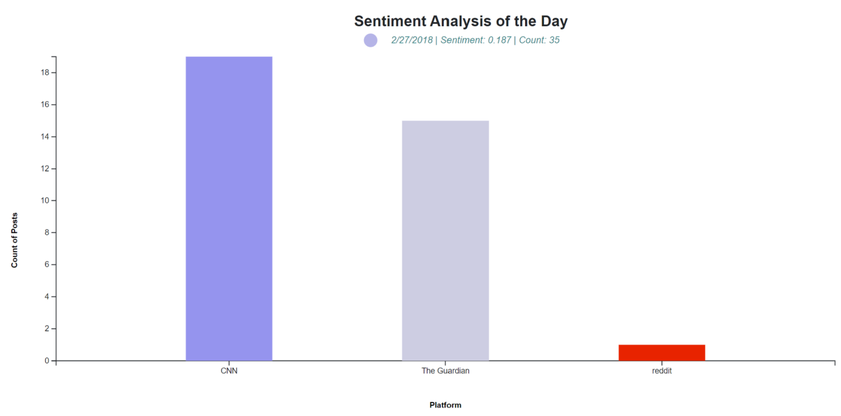
For a chosen media platform and timeframe, we created a pie chart to show the fraction of posts (that uses the selected number of top words) having positive/neutral/negative sentiment.

4. Approaches

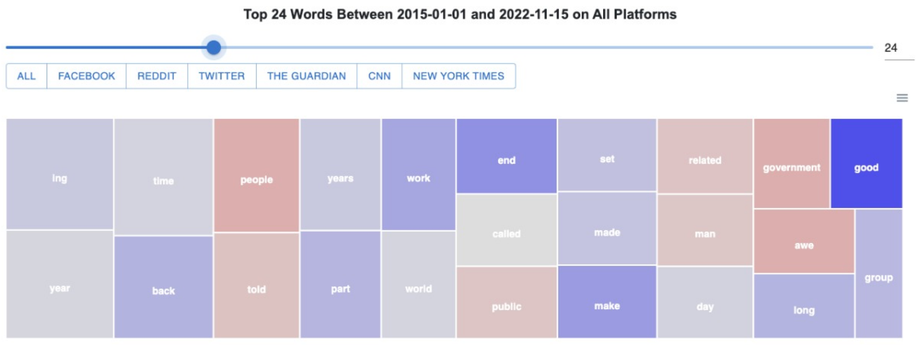
The user can select a timeframe and keywords to filter the data. When the user inputs filters, Sentiment Search quickly filters the data using caching and indexing techniques and displays four charts, implemented using D3.



The first shows the number of posts and the average sentiment of those posts over time. Other visualizations we found in this area did not show the number of posts, just the sentiment values. The number of posts is important for users to understand as a small sample may not be representative. Throughout the project we use a color scheme which represents high positive sentiment in blue and high negative sentiment in red, in order to keep the visualization as color-blind friendly as possible while keeping within norms associated with those colors. This chart also shows significant events on those days as bubbles showing a description when hovered over.



When the user hovers with their mouse over a bar in the main chart, a second chart shows the user a breakdown of the number of posts and mean sentiment by platform, something unique to Sentiment Search, so they can determine how different areas of the internet feel about the topic they are exploring.



Another piece of information that is useful for marketing and campaigning is which words are associated with your product. To make this accessible through our visualization, our tool gives users a way to see the top words people use in posts (above) about their product and the sentiment of the posts containing that word.

Users can control the number of top words shown, and we ensure that common words with no real value are ignored, for example "the", "and", and "a".

6. Conclusion

We have built a multidimensional visualization tool with dimensions across time and platform, which can be used to observe people's sentiments on various topics. This tool can be used to compare the sentiments expressed on different platforms and give insights on the most common words. We see future adaptations use real-time media posts to dynamically view the sentiment of the Internet.