# Sentiment Search: Make the Internet Your Focus Group

Faris Durrani, Nemath Ahmed, Justin Zandstra, RenChu Wang, Lakshmi Sree Lakshmanan, Shuyan Lin
Dec 02, 2022

## Abstract

*Sentiment Analysis has been used to identify changing moods of populations. In this project, we have analysed how sentiments revolve over topics from significant world events across social media platform (Facebook, Reddit, Twitter) and news sources (CNN, The New York Times, The Guardian). We have created an interactive visualization tool that allows to filter data on specific keywords and dates, and visualize time series sentiments along with the top words used in posts. This dashboard could potentially be used by businesses or political campaigns to analyze the effect of marketing strategies on public sentiment regarding their product, or to analyze the social climate surrounding certain ideas and issues on multiple platforms. Future steps could include a dynamic rendering of sentiments with new media posts using faster, more efficient algorithms.*

## 1. Introduction

Sentiment Analysis (SA) has been used to identify changing moods of populations and has been useful in cases like determining areas in need of disaster relief [4]. SA being a component of language processing aims to decode the underlying emotions given by a phrase which translate into ideas and insights. This can be useful in areas like marketing and politics where knowing what your customer / voter thinks given a certain idea can be crucial to planning your next steps. Traditional methods like surveys and polls can be expensive and difficult to obtain compared to automated methods like scouring over the user's activity online (already used in targeted online marketing).

Additionally, past research have shown multiple use cases for SA using a limited part of the Internet often closed to one or two social media platforms. In this project, "Sentiment Search: Make the Internet Your Focus Group", we aim to extend on the idea that it is possible to graph the general sentiment or feeling of the Internet over time given a certain keyword to show the affiliation and effect that keyword has to the general user. We further specify these analyses by breaking down these sentiments per platform and correlate general changes in sentiments with several significant events over time to add context. We put a lot of focus on keeping the tool dynamic and interactive to expand possibilities and insights in a concise manner. Another piece of information that is useful for marketing and campaigning is which words are associated with your product, which we showed by plotting the Top 20 or so words along with the mean sentiment of the posts containing those words.

## 2. Problem Definition

Sentiment change over-time, predominantly in the US, towards news and social media posts across the globe is tracked using the built map-based visualization tool. We have examined multiple sources including micro-blogs to news articles. Sources of opinions vary from individual platforms to news institutions and the goal of this interactive tool is to form interesting comparisons of opinions of a represented group versus opinions of aggregated individual toward a news topic.

## 3. Literature Survey

### 3.1. What Are We Trying to Achieve

In this section, we describe the definition and what people have tried to do with SA on social media. We focus on visualization on different social media and observe the behavior.

[5] examines the Social Helix as a way of visualizing two polarizing groups of people on social media. It mimics the 2D projection image of DNA. The strands of DNA are used to represent two different groups of people with polarizing opinions (e.g. republicans vs democrats). The Social Helix, however, only applies to two groups, and we have instead visualized overall sentiment.

To discuss one sentiment visualization method, [16] presented SentiView, an interactive visualization system that aims to analyze public sentiments for popular topics on the Internet. SentiView is designed to be flexible for varying tasks and is a good point for our platform to improve from. However, SentiView employs more difficult-to-obtain attributes like age and relationships, which we have limited data on.

[7] shows applications of sentiment analysis on social

media and provides a list of some insights social media has given us from 2015 to 2019 and the two methods for sentiment analysis- Lexicon-based and machine learning. It discusses possible problems we may encounter in the project while analysing social media posts other than twitter like Facebook posts with spelling errors.

### 3.2. SA Methodology (How Is It Done Today)

Most work is done on one social media platform but fails to look at bigger picture-Internet as a whole. Different platform has different length of blogs which makes it difficult to choose one algorithm.

[14] proposed another idea of evaluate sentiments. The familiar category for SA is positive, negative and neutral. [11] introduces a new hybrid method for sentiment analysis that uses text mining, a CNN classifier with word to vector, and an RNN classifier with Long Short Term Memory and word to vector that performs better than regular ML classifiers. [17] discussed the traditional lexicon method for text mining and the R package, Twitter that helps to preprocess the tweet to structured data. [13] analyzes the timeframe in which social media responds to events, and how quickly indicators of sentiment. They introduced 3 parameters to analyze the problems: history window size, prediction bandwidth and response time. Using SVM, they achieve 85 percent accuracy with broad categories. [1] mentions a two-step process. First, feature selection based on forward selection and backward elimination steps coupled with recursive feature exclusion. Second, SVM classifier with radial basis kernel is used along with optimal hyperparameters. [12] process both the local and global context of words in Twitter, MySpace, and Digg to improve sentiment analysis accuracy. [18] is a general survey of the state of SA in 2015. Stating the existence of sentiment aware platforms, and how sentiment can be seen as emotion connected with action. [3] preprocesses tweet data by replacing words with tags denoting their sentiment implication and uses both unigrams and bigrams to improve sentiment classification. This helped us towards improving the quality of our data. [2] focuses on the reasons for sentiment change. It presents a framework (Filtered-LDA) that outperformed existing methods of interpreting sentiment variations on Twitter. However, when the classifier was tested on other platforms, its accuracy fell. [15] focuses on aspect-level sentiment analysis, where the goal is to find and aggregate sentiment on entities mentioned within documents or aspects of them. The issues in Comparative Opinions and Conditional Sentences still remain even with recent techniques. [9] utilized the following techniques to analyze sentiments across different time frames:Sentiment velocity, Frequency component, outliers from their predictions.

### 3.3. Who cares

Social media such as Twitter and Facebook are heavily used in today's world by people to express their opinions/feeling on a current/personal issue. A lot of them shown from the research papers are for political purposes.

[4] research using sentiment analysis for disaster relief by determining where is especially affected, which can be used to improve response times. [9] aims to find out the relation between sentiment present in social media and mass media, and how they change depending on each other. [6] glossed over some election prediction approaches including the volumetric approach, sentiment analysis approach, and social network analysis approach. On the same note of predicting elections, [10] compared predictive models for election results using social media in three developing Asian nations using volumetric, sentiment, and social media influence approaches. [8] is a direct related survey relating to twitter SA and the techniques they used to such as emotion detection, irony detection, tracking sentiment overtime and sentiment quantification.

## 4. Proposed Method

### 4.1. Intuition

Sentiment analysis refers to the method to extract subjectivity and polarity from the text and semantic orientation refers to the polarity and strength of words, phrases, and texts. In marketing, it is important to understand the public's view of your product, which can be difficult to do and results can be difficult to interpret. More formally, we are attempting to solve the problem of the anecdotal and imprecise nature of determining the public view of various topics by harnessing the power of sentiment analysis.

Other sentiment visualizations do not have much interact-ability to break down data. They do not span multiple platforms, which means a simultaneous comparison cannot be made. We created an interactive tool that allows users to visualize the average sentiment over multiple platforms related to a specific topic over a chosen timeframe. The tool, while spanning the filtered data over multiple platforms and average sentiment, also provides a detailed breakdown of sentiment distribution on each one of them. This allows the user to focus on sentiment overall as well as for specific platform, which is not provided by tools available today. Moreover, we help users visualize what specific frequently occurring words contribute to positive/negative/neutral sentiments and their corresponding fraction of posts over a pie chart. This again, is better than state-of-the-art techniques for the same, as the tool lets the user choose the number of top words.

### 4.2. Data Extraction

1. We have extracted data from five different sources geared towards this single purpose of analyzing polarization. Previous works around similar topics have focused on 2-3 data sources on average.

2. We believe including more datasets will give us a more comprehensive view of the internet rather than focusing on one or two sources which might ultimately have some kind of platform-specific bias, and we focus our dataset timeframe from January 2015 to November 2022.

3. We focus on both Individual User Articles/Posts (Social Media - Twitter, Reddit, Facebook) and Curated Articles (CNN, The New York Times, The Guardian)

4. We have also extracted data that represents the list of important events from Wikipedia to correlate it with patterns observed in sentiments of content from five of our data sources.

5. For most platforms have used both Native and Third-Party APIs to scrape the data and intermediate data cleansing and transformation steps such as language translation, and data formatting has been performed for sentiments to be properly extracted.

6. For some platforms, we were able to find large sets of posts with all of the information needed for our purposes. In fact, some of these sets were so large that they were difficult to work with, so we used OpenRefine to remove impurities in their data and select the most important posts by using the number of votes in Reddit and the number of Retweets on Twitter to create cutoffs.

7. Many prebuilt datasets have limited information on each post/article. On the other hand, we collect info on timestamps, post content, location, author, likes/comments, and a URL source, allowing us to have high customization of information.

### 4.3. Detailed Description of Approach

From a large amount of structured data from different platforms, the objective was to serve them through a consistent API. The RESTful API was used to achieve modularity and platform independence.

Due to the large size of the total data which we collected (over 190 GB), our original SQLite database loaded extremely slowly on initialization and in querying, so we pivoted to saving our data in CSV files and querying the data in memory through a pandas dataframe. We also improved the speed some by indexing the dataframe and caching data from previous queries some. This got our response time down to around one second for simple queries.

However, after splitting indexes, we still weren't satisfied with the performance. We also noticed there's a huge imbalance between different platforms (some platforms like Facebook have less than 1% of data of other platforms). Hence, to do that, we cleaned the data further month by month and randomly discarded unfit data to drive down the amount of data we have. We were able to serve the entire database of around 2.3M posts directly in memory with a pandas dataframe which only used 1.8 GB of storage.

In order to avoid mischaracterizing words in the keyword search, for example, a post containing "trumpet" being included when searching from "Trump", we tokenized the words in the text portion of the data. To improve the quality, in addition to simply using text search, we also utilized production quality tokenizers (from huggingface) to first tokenize the texts, then perform the filters. However, since this step is computationally too expensive, so we pre-compute the answers for all our entries and was able to achieve a 5-10 times speed up.

MapReduce algorithm was used for processing. We utilized word tagging and third-party machine learning framework for sentiment-generating purposes and utilized sorting/interval trees for indexing purposes to speed up serving the data.

### 4.4. Sentiment Analysis Algorithm

To evaluate sentiment, we utilized the VaderSentiment analysis tool (https://github.com/cjhutto/vaderSentiment). The processing time for sentiment is quick, and the time complexity of the algorithm is O(N). The algorithm is lexicon and ruled-based. It's attuned to social media sentiments. For example, punctuation will increase the intensity of emotion such as "this is a good movie!!!". This compares to a typical machine learning algorithm which cleans the data by removing punctuation and maintaining case uniformity even though casing and punctuation can give away a lot about sentiments. It can translate emojis to quantify sentiments. It can understand slang words and take consideration of all caps. The advantage of this rule-based algorithm is it's very flexible. A lot of machine learning algorithms, it has a vocabulary bank. Usually, slang vocabulary is not included because most of the algorithm is trained on Wikipedia page. It will not able to read slang words and acronyms. VaderSentiment first tokenize the article
How good the algorithm relies on how good the sentiment lexicon is. A sentiment lexicon is a collection of words associated with their sentiment score. The sentence to be evaluated is tokenized and matches the vocabulary. It can be one of the disadvantages of the algorithm. One challenge of using both lexicon and machine-learning-based algorithms is their accuracy gets affected by spelling. It's a challenge for

any algorithm. Essentially lexicon rule-based is similar to the machine learning algorithm besides sentiments probabilities are determined by humans or by algorithms. One benefit of the Vader algorithm over deep learning is that it incorporates more heuristics of word order. For example, degree modifiers will have an impact on sentiments.
To conclude, VaderSentiment uses a rule-based which can give more transparency to the decision, while the deep learning model's decision is not interpretable.

### 4.5. Visualizations

The interactive tool located here produces four visualizations:

1. Sentiment Color:
   To add a third dimension in our visualization, we used a linear gradient color scheme going from red (very negative sentiment or -1 as outputted by the VaderSentiment library) to blue (very positive sentiment or +1) with gray denoting neutrality (0). This simple legend shows up before the graphs, denoting that bars and areas colored red have negative sentiment posts, i.e., posts that are unpleasant, depressing, or harmful, and vice versa for blue.

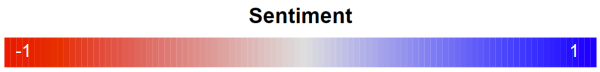


Figure 1. Sentiments being color-coded

2. Sentiments on a Timeframe:
   The main graph [Fig. 2] is the Timeline, which shows the count of posts per day over time, and the color of the bars denoting the sentiments. The user can hover over each bar to find out the sentiment and count of posts of that date.

   We also show several "significant events" (of our own choosing from Wikipedia) that contain those keywords in circles. These circles have the same height and color as the bar on that date. This allows the user to correlate certain significant change in sentiment with a significant event.

   As an example, 1/12/2021 has a great deal of fairly negative posts which correlates with President Trump's second impeachment.

3. Frequency Chart:
   As the user hovers over the Timeline above, this frequency chart also updates to show the breakdown of posts on that date. [Fig. 3]

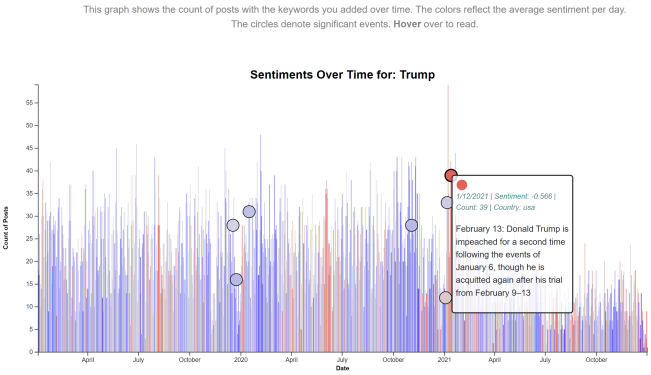


Figure 2. Sentiments on a Timeframe

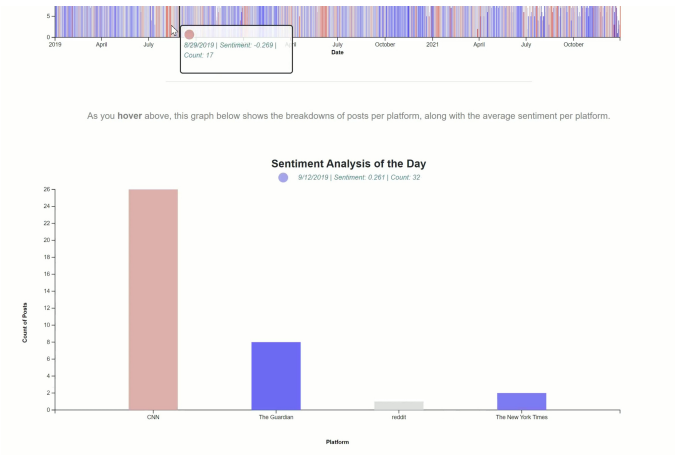


Figure 3. Sentiments breakdown for a given day

4. Visualizing bag of words:
   For every selected media and timeframe, the top words (specified number by user) are displayed. The words are color-coded based on the sentiment intensity. The Pie chart denotes the fraction of posts with the selected number of top words that are positive/neutral/negative. [Fig. 4]

## 5. Experiments

### 5.1. Questions Answered by Experiments

The following questions are to be answered by the experiments:

1. Is the sentiment analysis algorithm working fine on posts?
2. Are the significant events correctly mapped on the timeline?
3. Do the keyword/ timeframe correctly filter the data?
4. Are the bar graph and pie-chart visualizations correct?

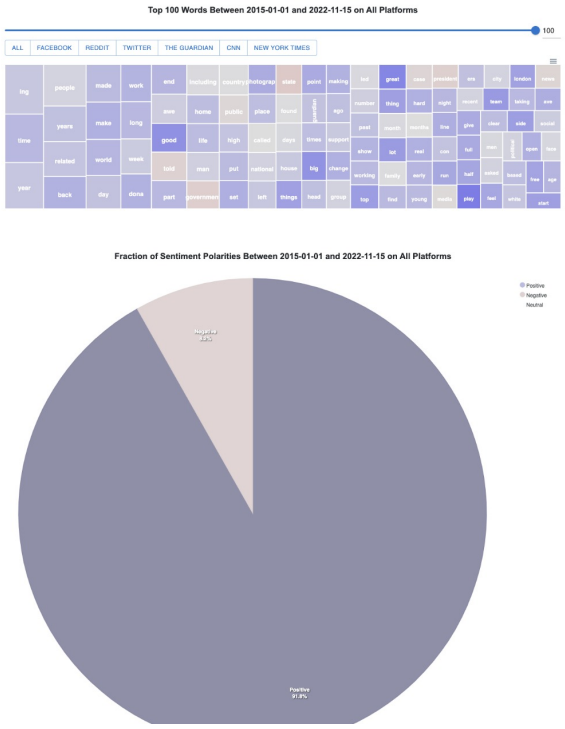


Figure 4. Bag of top words with pie chart showing fraction of positive and negative sentiment

### 5.2. Evaluation

To evaluate we tested the sentiment analysis algorithm on Amazon product reviews. The table shows some examples that we tested. It can be seen that if the writer has a strong sentiment, the algorithm does a good job predicting the sentiments. If the writer is calm but articulating the defects, the VaderSentiment cannot capture them. It's understandable because not even machine learning can completely understand the syntax. This is under an acceptable margin of error.

| Text review | rating | Sentiment result |
|---|---|---|
| Just finished the first three chapters of this book and you can really feel the enthusiasm of the author. He put so much effort in making the book comprehensible. | 5 star | 0.6986 |
| I have used L-tyrosine off and on for years to help with attention problems. This product works well and has a good value for the money. If I could change one thing, I would add B-12 to the supplement as a few other companies do. The body uses B-12 to convert L-tyrosine into dopamine, and having them in one capsule is highly convenient. Aside from that, it is a great product. | 4 star | 0.8885 |
| Set up was easy, however it doesn't serve much of a purpose if features don't actually work. I researched to make sure everything I wanted to add to and use Alexa for were possible on this device, and supposedly they were. However, my Blink systems does NOT work on this device. | 2 star | 0.6369 |
| I have bought many crocs and have never had this happen. My grandson would wear my red crocs all the time… so I bought him a pair in red. Mind you, my crocs are worn in the water, the beach, camping, etc, and they are left out all the time in the sun. My grandsons were left out and they shrunk and do not fit his feet anymore!! I tried them on and one is actually now alittle smaller than the other!! I just bought these end of July! These are not cheap and he has only worn them for a month! Like I said I have many pairs and mine have never done this especially my red ones! I am so disappointed.. I these are defected!! My grandson is so sad that he cannot wear them! What the heck Crocs!! | 1 star | -0.9419 |
| I bought these shoes in Aug in preparation for winter. As an older person, I need shoes to STAY on my feet, so that's why I got the clogs with the straps. However, within approximately two weeks of actually wearing them regularly, the strap broke on one, rendering it absolutely worthless. And, of course, I can no longer return them and get my money back. I would have considered buying another pair, however...not now! | 1 start | -0.8169 |

Upon testing, the significant events are correctly mapped on the timeline for search using the keyword 'Trump' and specified timeline. We also see that the data is filtered correctly. We could use human judgment to evaluate the visualizations (whether or not the model is making good extractions) by the relatively small, labeled datasets focused on the sentiments of each post we found on Kaggle. If the model performs well on those small datasets, we believe that we have a reasonable model for visualization.

## 6. Conclusion and Discussion

### 6.1. Effort Distribution

Everyone on our team contributed equally to this project.

### 6.2. Conclusion

We have built a comprehensive multidimensional interactive visualization tool with dimensions across time and platform, which can be used to observe people's sentiments on various topics. Users can select different time frames and keywords to compare views in many areas. This tool can be used to compare the sentiments expressed on different platforms. The tool also gives users insight on the words most commonly associated with the topic of their choice, and how those change from platform to platform. This tool is especially applicable to marketing of both products and people.

### 6.3. Future Steps

Our project could be improved in the future by making it update in real time as people post to the sites that we have analyzed. This is not something we could not do now due to our low server space and the immense storage that would be required to do this. Another limitation to achieving this is the speed at which we wanted the tool to update. We

already had to significantly reduce the amount of data used compared to what we collected because the tool simply took too long to load all of the data and update when filters were changed, despite the methods we used to speed it up. Another area for improvement is that we could include information on the locations that posts are coming from. We could not find a reliable way to get location data for many of our platforms in the time we had to complete the project, so we did not display information about sentiment location, but we had originally planned to do so using a heat map or a choropleth map. It could also be improved by providing a more holistic view of the changing sentiment by platform, perhaps by creating a multi-line chart.

## References


[1] Mohammed H Abd El-Jawad, Rania Hodhod, and Yasser MK Omar. Sentiment analysis of social media networks using machine learning. In *2018 14th international computer engineering conference (ICENCO)*, pages 174–176. IEEE, 2018. 2

[2] Fuad Alattar and Khaled Shaalan. Using artificial intelligence to understand what causes sentiment changes on social media. *IEEE Access*, 9:61756–61767, 2021. 2

[3] Alexandra Balahur. Sentiment analysis in social media texts. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 120–128, 2013. 2

[4] Ghazaleh Beigi, Xia Hu, Ross Maciejewski, and Huan Liu. An overview of sentiment analysis in social media and its applications in disaster relief. *Sentiment analysis and ontology engineering*, pages 313–340, 2016. 1, 2

[5] Nan Cao, Lu Lu, Yu-Ru Lin, Fei Wang, and Zhen Wen. Socialhelix: visual analysis of sentiment divergence in social media. *Journal of visualization*, 18(2):221–235, 2015. 1

[6] Priyavrat Chauhan, Nonita Sharma, and Geeta Sikka. The emergence of social media data and sentiment analysis in election prediction. *Journal of Ambient Intelligence and Humanized Computing*, 12(2):2601–2627, 2021. 2

[7] Zulfadzli Drus and Haliyana Khalid. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161:707–714, 2019. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia. 1

[8] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):1–41, 2016. 2

[9] Anastasia Giachanou and Fabio Crestani. Tracking sentiment by time series analysis. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 1037–1040, 2016. 2

[10] Kokil Jaidka, Saifuddin Ahmed, Marko Skoric, and Martin Hilbert. Predicting elections from social media: a three-country, three-method comparative study. *Asian Journal of Communication*, 29(3):252–273, 2019. 2

[11] Abhishek Kumar, Vishal Dutt, Vicente García-Díaz, and Sushil Kumar Narang. Twitter sentimental analysis from time series facts: the implementation of enhanced support vector machine. *Bulletin of Electrical Engineering and Informatics*, 10(5):2845–2856, 2021. 2

[12] Aminu Muhammad, Nirmalie Wiratunga, and Robert Lothian. Contextual sentiment analysis for social media genres. *Knowledge-Based Systems*, 108:92–101, 2016. New Avenues in Knowledge Bases for Natural Language Processing. 2

[13] Le T Nguyen, Pang Wu, William Chan, Wei Peng, and Ying Zhang. Predicting collective sentiment dynamics from time-series social media. In *Proceedings of the first international workshop on issues of sentiment discovery and opinion mining*, pages 1–8, 2012. 2

[14] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*, 2005. 2

[15] Kim Schouten and Flavius Frasincar. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830, 2016. 2

[16] Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang. Sentiview: Sentiment analysis and visualization for internet popular topics. *IEEE Transactions on Human-Machine Systems*, 43(6):620–630, 2013. 1

[17] Eman MG Younis. Sentiment analysis and text mining for social media microblogs using open source tools: an empirical study. *International Journal of Computer Applications*, 112(5), 2015. 2

[18] Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60(2):617–663, 2019. 2