Sentiment Analysis of Major News

Elijah Kim (cys8qu)
12/17/2025

Motivation

During major events, shifts in sentiment often correlate with changes in public opinion

- Elections, crises, wars

Organizations want early signals

- Journalists, public agencies care about “how are people reacting now”

Goal:

- When engagement spikes, does sentiment shift? (pre → during → post)
- Do different communities show different sentiment direction?

Background

Why Reddit?

- Communities are separated
- More than short-form text
  - Timestamps, upvotes, number of comments
- Threaded structure
- Large online presence

A lot of mixed sentiment

- Sarcasm and internet shorthand
- Posts with neutral titles with polarized community responses

Related Work

“Reddit Sentiment Analysis on the Impact of AI Using VADER, TextBlob, and BERT”

- Kristi Pham, Krishna Chaitanya Rao Kathala, Shashank Palakurthi (May 2025)

“Sentiment Analysis and Topic Modeling of Reddit Data”

- Drashya Nileshbhai Babariya, Gauri Prakash Dhanawade, Elamathy Sivakumar, Vamshi Krishna Dara (January 2025)

Claim & Target Task

Target Task

- Build a workflow that takes subreddit text data and outputs:
  - Sentiment per post
  - Detect sentiment spikes over weeks (identify spikes in engagement)
    - Includes shifts in sentiment
  - Comparison between different subreddits

Claim

- Spikes in engagement can serve as an indicator for major events, and combining spike detection with sentiment scoring yields interpretable event-centric insights

Proposed Solution

Sentiment scoring

- Assign each post title a sentiment label using a tuned classical model

Event discovery

- Identify “events” as time windows with unusually high engagement within each subreddit
  - Engagement is measured with interaction signals, such as total number of comments, average number of comments, number of upvotes

Event analysis

- Quantify how sentiment changes from before, during, and after the spike in engagement, then compare changes across subreddits

Implementation

Data ingestion + Cleaning

- Load CSVs (from data_sets/*.csv)
- Standardize columns
- Drop missing/empty titles; invalid dates; megathreads

Models used

- SVM training
  - Tuned both
    - Word TF-IDF + Linear SVM
      - Tokenizes text into words
    - Character TF-IDF + Linear SVM
      - Tokenizes text into character n-grams
  - Performed best on labeled dataset + is fast
  - Tuned using GridSearch
- RoBERTa sentiment model
  - Used for comparison as a “second opinion” on hard cases
  - Map model outputs to [-1, 0, 1] for comparison
  - Uncertainty gate (SVM margin)
    - SVM margin = top class score - runner-up score
      - Large margin → confident prediction
      - Low margin → ambiguous post

Data Summary

Training dataset

- ~37,000 labeled samples
- Label distribution shown (class imbalance handled via class_weight=”balanced”)

Reddit datasets

- 6 subreddits with ~6,000 total posts
- Features include:
  - Title text, time created, number of comments, number of upvotes

Known constraint

- Dataset is “top posts,” so results reflect highly visible content

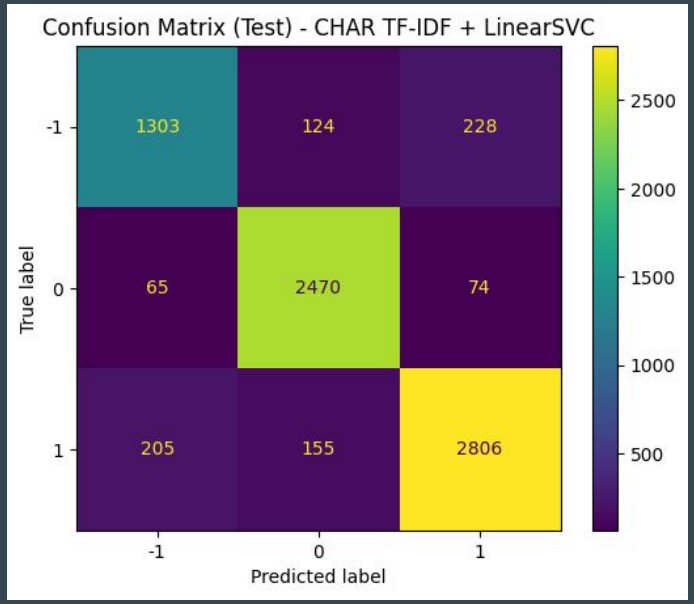
Results (Model Quality)

Test performances:

- Word TF-IDF
  - CV macro-F1: ~0.821
- Character TF-IDF (best model)
  - CV macro-F1: ~0.865
  - Accuracy on test data: ~0.885

Confusion matrix highlights:

- Neutral class is easiest
- Hardest cases are short/ambiguous phrases, sarcasm, negation



Results cont.

Spike Scores

- Robust z-score computed based on weekly total comments for each Reddit post
  - Can now compare between different subreddits

In table below, can see how the model detects certain keywords and sentiment values

event_id	subreddit	time_bin	spike_score	total_comments	post_count	mean_sent_during	top_keywords	top_titles
worldnews__2022-02-28	worldnews	2022-02-28 00:00:00+00:00	14.224169	153143.0	29	-0.068966	ukraine, russia, russian, war, zelensky, sanct...	Vladimir Putin says Russia Has "no ill Intenti...
worldnews__2022-02-21	worldnews	2022-02-21 00:00:00+00:00	12.252073	133577.0	23	-0.130435	russia, ukraine, russian, putin, kyiv, swift, ...	Putin puts Russia's nuclear deterrent forces o...
worldnews__2020-01-06	worldnews	2020-01-06 00:00:00+00:00	10.851870	119685.0	12	-0.166667	iran, shot, multiple, crashes, commit, commit ...	Multiple rockets hit Taji base in Iraq | Penta...
news__2020-06-01	news	2020-06-01 00:00:00+00:00	8.284456	110553.0	23	-0.304348	police, protest, officers, buffalo police, for...	Active duty troops deploying to Washington DC ...

Results cont.

Event windows

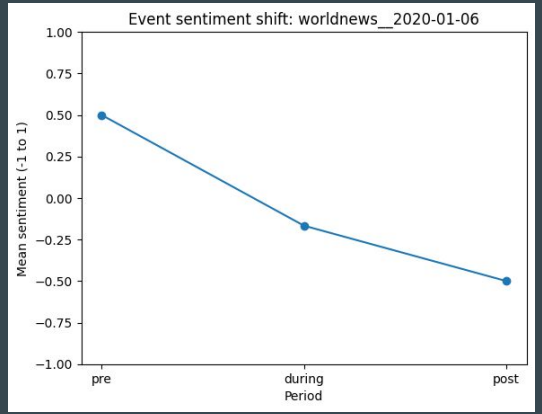
- For each detected spike week, build a 3 period window plotting against the sentiment in that week
  - Pre: 1 week before
  - During: spike week
  - Post: 1 week after

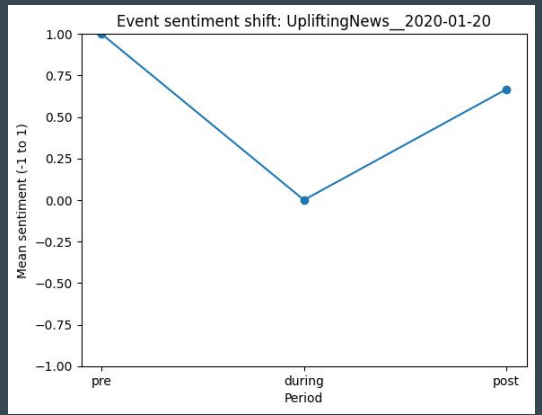
Top image (worldnews_2020-01-06)

- Downward trend
- Keywords: Iran, rockets, Ukrainian plane shot down

Bottom image (UpliftingNews_2020-01-20)

- Downward spike
- Keywords: Policy actions, health-cost relief





Analysis & Conclusions

Model Error Pattern (Confusion Matrix)

- Neutral (0) is easiest; More polarized between positive (1) and negative (-1)

Confidence profile (SVM margin): Median margin ~0.84, meaning ~45% of all posts fell below 0.75

- A large chunk of post titles were borderline/ambiguous
- Could mean there are large domain differences between the training of the transformer and our SVM model

When large events happen, sentiment shifts sharply

- Change in sentiment from pre- to during- can be strongly positive or negative
- Sentiment also differs greatly by community
    - Saw r/UpliftingNews stay neutral to positive in contrast to r/worldnews

Character TF-IDF + LinearSVC

- Strong baseline for noisy, short text

Future Work

Future work

- Add a larger sentiment scale
- Incorporate comment text (better reflection of discussion between users)
- Directly compare sentiment of the same event in different subreddits
- Utilize more up-to-date news
    - Get access to Reddit's Developer API

References

Drashya Nileshbhai Babariya, Gauri Prakash Dhanawade, Sivakumar, E., Dara, V. K., & Jha, N. (2025, January 11). Sentiment Analysis and Topic Modeling of Reddit Data. https://doi.org/10.13140/RG.2.2.10004.21125

GeeksforGeeks. (2021, January 20). Understanding TFIDF (Term FrequencyInverse Document Frequency). GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/understanding-tf-idf-term-frequency-inverse-document-frequency/

Jain, A. (2024, February 4). TF-IDF in NLP (Term Frequency Inverse Document Frequency). Medium. https://medium.com/@abhishekjainindore24/tf-idf-in-nlp-term-frequency-inverse-document-frequency-e05b65932f1d

Pham, K., Chaitanya, K., & Shashank Palakurthi. (2025). Reddit Sentiment Analysis on the Impact of AI Using VADER, TextBlob, and BERT. Procedia Computer Science, 258, 886–892. https://doi.org/10.1016/j.procs.2025.04.326

RoBERTa. (n.d.). Huggingface.co. https://huggingface.co/docs/transformers/en/model_doc/roberta

scikit-learn. (2019). sklearn.metrics.f1_score — scikit-learn 0.21.2 documentation. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html