Analyzing Political Discourse on r/Politics: A Topic-Emotion Approach Using BERTopic and XLM-EMO

Giuseppe Bonsignore

giuseppe.bonsignore02@icatt.it

Abstract

This study explores political discourse on Reddit, specifically focusing on r/Politics subreddit, to investigate topic distribution and emotion trends. Using BERTopic for topic modeling and XLM-EMO for emotion classification, we analyze the most upvoted comments from highly engaged posts of r/Politics. The research aims to unveil the capabilities of large language models based on transformer architecture in providing valuable insights into emotional dynamics of political discussions, with potential applications in journalism, business intelligence, opinion monitoring and various other areas.

1 Introduction

The analysis of political discourse on social media platforms provides valuable insights into public opinion, sentiment trends and engagement patterns. This study focuses on Reddit, and specifically the r/Politics subreddit, a major hub for online discussion on contemporary political events in the United States. Reddit is a social news aggregation and discussion platform where users submit content (links, posts, images, and videos) which can be "upvoted" (voted up) or "downvoted" (voted down) by community members.[1]

To collect relevant posts and comments for our analysis, we utilize a library called Python Reddit API Wrapper (PRAW)[2]. This research investigates topic distribution and emotion trends within the most active ("hottest") posts on r/Politics fetched on February 2, 2025. We employ BERTopic for topic modeling, while emotion analysis is conducted using XLM-EMO.

Topic models are unsupervised machine learning models that learn on large datasets to induce groups of associated words from text that are very useful for discovering topical structure in documents. BERTopic is a topic modeling technique that leverages Hugging-Face Transformers and class-based TF-IDF to create dense clusters, allowing for great interpretability while preserving the most relevant words in the topic descriptors[3]. It is built upon a highly modular algorithm that consists of multiple customizable steps which may be tailored to different research goals.

More specifically, BERTopic is built upon Sentence-BERT[4] for the embeddings model, UMAP[5] for dimensionality reduction, HDBSCAN[6] for clustering, CountVectorizer[7] for tokenization, and c-TF-IDF[8] for weight normalization. On top of the last module, we also use KeyBERT[9] as our representation model, allowing for further fine-tuning of the topic representation through keyword extraction. The representation model is helpful to reduce noise

[1] https://en.wikipedia.org/wiki/Reddit
[2] https://praw.readthedocs.io/en/stable/
[3] https://maartengr.github.io/BERTopic/index.html
[4] https://sbert.net/
[5] https://umap-learn.readthedocs.io/en/latest/
[6] https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html
[7] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
[8] https://maartengr.github.io/BERTopic/api/ctfidf.html
[9] https://maartengr.github.io/KeyBERT/

generated by stop words, especially in short and noisy text from social media.

The other technique that is employed in this study is called sentiment analysis. Sentiment analysis is the field of study that analyzes people's opinions, sentiments and emotions towards entities such as products, services, organizations, individuals, issues, events, and topics. In this research, we specifically focus on the sub-task of sentiment analysis defined as emotion analysis, using XLM-EMO[10], a fine-tuned version of the XML-T model developed by the MilaNLP research team for emotion analysis, which achieved an F1 of 0.85 on the original test set.

By integrating BERTopic and XLM-EMO, this study aims to uncover themes, trends and recurring patterns of emotions shaping political discussions on r/Politics.

2 Data

The dataset used in this study consists of a collection of posts and comments from r/Politics. Data collection was conducted using the Python library PRAW via the Reddit API on February 2, 2025. More specifically, the 10 most upvoted comments were collected from the "hottest" reddit posts. Hot reddit posts are defined as those with higher levels of engagement, such as significant number of upvotes and comments.

This table presents a sample of the dataset structure:

Index	Post Title	Comment Text	Upvotes
30	Trudeau Tells Trump: Your Tariff War Will Shut American Factories	It's going to be a LONG 4 years 😞	575

Table 1: the r/Politics dataset

The dataset is stored in CSV format, consisting of 6057 rows, each of them corresponding to an individual comment. It contains four primary columns: Index, Post Title, Comment Text, and Upvotes. This dataset served as the point of departure for the analysis, and it was enriched by incorporating additional columns for topic and emotion classification.

3 Workflow

The workflow of this research was broken down into the following steps: 1) data collec-tion; 2) data normalization; 3) topic modeling; 4) emotion classification; 5) statistical analysis and data visualization.

For preprocessing, we applied minimal text normalization techniques, simply removing URLs and lowercasing every character, since a more intensive preprocessing, such as tokenization or stop word removal, might have caused interference with the models, which incorporate their own text processing mechanisms by default.

After dataset creation, comments were clus-tered into topics using BERTopic. We decided to keep customization to a minimum when training the model on our data. With respect to the default settings of the model, we only replaced the default embedding model with all-MiniLM-L6-v2 sentence transformer model[11]. For additional coherence and informativity of keywords, we also used KeyBERT as the representation model. Once trained, the model assigned topic labels to each comment.

After topic modeling, emotion analysis was performed using XLM-EMO, classifying each comment across four dimensions: anger, joy, fear and sadness. XLM-EMO was used with its default configuration.

Once all data were assigned topic clusters and emotion labels, we conducted statistical analysis to investigate topic-emotion distribution, and we visualized the results to enhance interpretability.

[10] https://huggingface.co/MilaNLProc/xlm-emo-t

[11] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

4 Results and Discussion

BERTopic classified the comments into 112 clusters, each corresponding to a distinct topic. Each topic corresponds to a probability over a set of words. According to this approach, words receiving the highest probability score in each topic are considered as the most representative words of the specific topic to which they are assigned. A label can be assigned to each topic by extracting the top-n most probable words from each topical cluster. In this study, we set n=3, allowing for a concise but interpretable topic representation. For example, the following comment was classified under the topic "miseducation – schooling - indoctrinating":

"As an American, I really wish we invested in education over corporations. Really sad time for any US citizen that's actually paying attention."

The example shows how BERTopic assigns meaningful labels based on word co-occurrence patterns.

The following bar chart shows the top 20 most commented topics in r/politics according to our topic model:

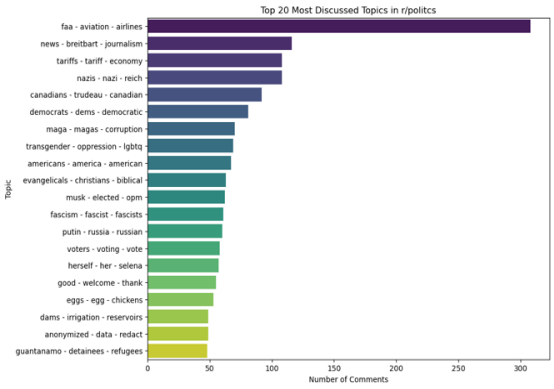

Figure 1: the 20 most discussed topics on r/Politics

It is important to highlight that the most frequently occurring topic, labelled as "tariffs – Trump – federal", was excluded from the chart, as it was identified by the topic model as an outlier based on its significantly higher frequency in comparison to the other topics. Nevertheless, we decided to keep the topic in the dataset, as it may provide meaningful insights, for example into the recent introduction of tariffs on Canadian, Mexican and Chinese imported goods, which has recently emerged as one of the most widely discussed debates in latest global politics.

The data suggest strong engagement with aviation policies, followed by discourse about journalism, media and economy. Additionally, themes related to Nazism and Fascism suggest ongoing discourse polarized around political ideologies which might represent a gateway into the study of current historical parallels that people are making.

Finally, themes including "Transgender-Oppression-LGBTQ" and "Evangelicals-Christian-Biblical", show engagement with social, cultural, and religious debates.

Overall, the chart reflects a mix of policy, ideology, leadership and cultural discussion dominating political discourse on the subreddit under analysis. However, we maintain that it is not appropriate to delve into the discussion, as we are not experts in the field. Therefore, in this research we mostly limit ourselves to presenting the results obtained, without engaging into the more political aspects.

Using BERTopic built-in visualization methods, we can plot the clusters on a two-dimensional map:

Figure 2: topic clusters

The map visualizes topic clusters, with each tiny dot representing a single comment. The colored dots indicate comments that were classified by our model as representative of

a specific category. The map is also interactive, allowing users to inspect the content of each individual comment by clicking on it, which increases the overall interpretability of each topic and allows for deeper investigations of topic clusters. The following charts, instead, show the 5 words that received the top probability scores over the 10 most discussed topics:

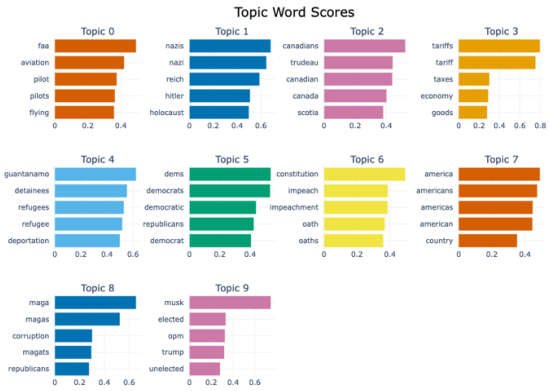

Figure 3: topic word scores

Emotion analysis was conducted using XLM-EMO, through which comments were distributed across four emotional dimensions: anger, joy, sadness and fear. The emotion distribution in the dataset is highly skewed towards anger:

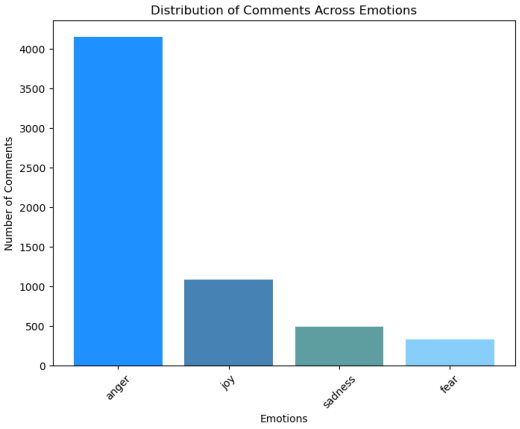

Figure 4: comment distribution across emotions

Again, since we are not domain experts, we do not claim to make definitive inferences of any sort based uniquely on the data at our disposal. However, the prevalence of anger in the dataset might be an indicator of a general sense of frustration with government policies. More generally, the analysis indicates that online political discussions on r/Politics tend to be strongly polarized and emotionally charged, with fear and rage creating higher engagement than positive sentiments, as it is evidenced by the distribution of average upvotes across emotions:

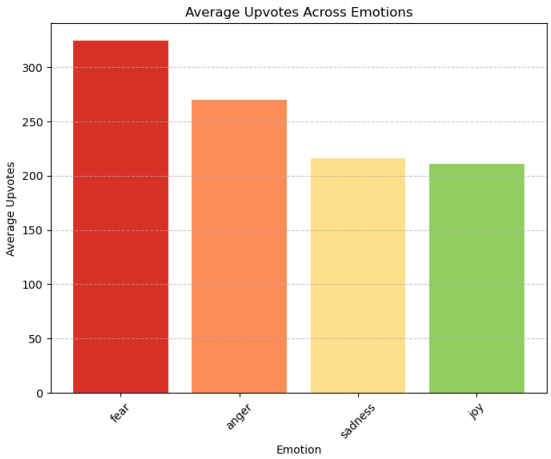

Figure 5: average upvotes across emotions

Examining the emotion distribution across the 100 most upvoted comments, we observe similar trends, with 73% anger, 13% joy, 8% fear and 6% sadness.

Furthermore, one useful aspect that can be analyzed is the emotion distribution across specific topics. For instance, the topic labeled as “stock – market” shows a distribution of 34,2% fear, 26,3% joy, 26,3% anger, and 13.1% sadness, while the topic labelled as “fentanyl – cartels – smuggling”, which seems to point very explicitly to drug trafficking, shows a distribution of 87,5% anger, 6,25% sadness, and 6,25% joy.

Overall, these types of investigation might offer deep insights into the emotional dynamics underlying online political discussion, with many possible applications such as journalism, business intelligence, stock market prediction, and various other fields, where sentiment analysis can provide a better understanding of public opinion and engagement, with an even finer

grained classification if combined with the capabilities of topic models.

5 Future work

We are aware that collecting data from a single platform and specifically from one subreddit introduces potential sampling biases. Thus, the discussion and the emotions that emerge in this subreddit may not be representative of the entire United States political scenario. Factors such as demographics can influence the nature of the data that we observe.

Future research might address these limitations by incorporating data from multiple platforms and more politically diverse online community, allowing for the construction of a more representative and balanced analysis of political discussion.

Besides cross-platform analysis, temporal analysis would be an interesting research path, especially considering the capabilities offered by BERTopic in dynamic modeling.

This research line could focus on investigating how emotions and topic distribution change over time, with the aim of identifying how certain political events affect the intensity and nature of online political discussion and seeking to uncover deeper correlations in political events and their emotional perception.

References

D. Jurafsky & J.H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition with Language Models, 3r edition, pages 56; 103-104; 111-114; 203; 223; Online manuscript released January 12, 2025.

M. Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084

B. Liu. 2012. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

F. Bianchi, D. Nozza & D. Hovy. 2022. XLM-EMO: Multilingual Emotion Prediction in Social Media Text. In Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 95-203, Dublin, Ireland. Association for Computational Linguistics.

Hugging Face. (n.d.). Introduction to natural lan-guage processing with Hugging Face (Chapter 1). https://huggingface.co/learn/nlp-course/chapter1/1

The White House, February 1, 2025, President Don-ald J. Trump Imposes Tariffs on Imports from Canada, Mexico and China [Fact Sheet]. https://www.whitehouse.gov/fact-sheets/2025/02/fact-sheet-president-donald-j-trump-imposes-tariffs-on-imports-from-canada-mexico-and-china/