The Left from the Right and The Right from the Left: an Analysis Through Political Memes

Chayapatr Archiwaranguprok

***Abstract*—This paper delves into the analysis of political memes within Reddit communities, specifically r/TheLeftCantMeme and r/TheRightCantMeme, employing topic modeling and corpus-based approaches. Noteworthy similarities in keywords hint at shared thematic elements, while nuanced differences emerge in the contextual usage of certain terms. Cross-corpora keyword extraction reveals linguistic disparities, and a resemblance between 'Top' and 'Controversial' posts.**

***Index Terms*—Internet Memes, Reddit, Corpus, Text Embedding, Data Clustering**

I. Introduction

A. Internet and Memes

Memes have emerged as powerful vehicles of cultural expression. Coined by evolutionary biologist Richard Dawkins, a meme, in its broadest sense, refers to an idea, behavior, or style that spreads from person to person within a culture [1]. In the age of the internet, memes evolve and take on a digital form, often combining images, text, and humor to convey a message. This evolution from cultural evolution to digital expression has given rise to the phenomenon of internet memes.

The evolutive and propagative nature of internet memes, makes them powerful mediums for communication in the online sphere [2]. In the political arena, memes have become instrumental in conveying complex ideologies, offering a simplified yet impactful means of expressing and critiquing political viewpoints [3].

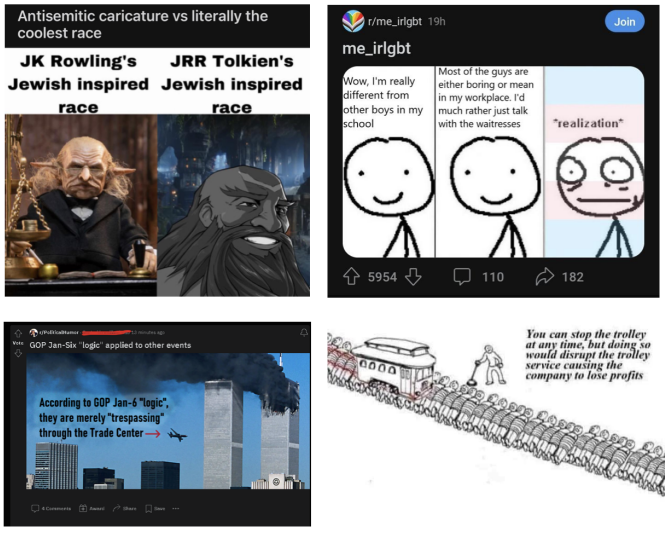

Fig. 1. Sample Internet Memes from r/TheLeftCantMeme

B. Reddit, r/TheRightCantMeme and r/TheLeftCantMeme

Reddit is a popular sprawling online community. Analytics by similar web shows that Reddit is the 17th most visited website globally with an average of approximately 1.9 billion visits monthly [4]. The platforms are organized into sub-rooms called "subreddits" that function as individual communities centered around specific topics. A subreddit, denoted by "r/" followed by its name, allows users to engage in discussions, share content, and form a community around a shared interest.

TABLE I
Brief statistics of the two subreddits (15 December 2023)

Subreddit	Subscribers	Rank on Reddit.com
r/TheLeftCantMeme	53,824	#7155
r/TheRightCantMeme	434,775	#1865

Two notable subreddits in the political meme landscape are r/TheLeftCantMeme [5] and r/TheRightCantMeme [6]. Created in 2017, they are dedicated to critiquing and satirizing memes from opposing political ideologies. As their names may suggest, r/TheLeftCantMeme and r/TheRightCantMeme criticize left-leaning and right-leaning ideologies respectively. By examining the content shared in these subreddits, we gain valuable insights into how each side perceives and responds to the other's use of memes.

Utilizing topic modeling and corpus-based analysis, we explore the interplay between political ideologies and meme culture aiming to uncover linguistic nuances within each subreddit. We cross-compare subreddits to identify shared topics of discussion and interests. This qualitative exploration provides valuable insights into the evolving landscape of digital communication and political meme culture.

C. Research Goal

Examining political discourse in the internet meme format within the r/TheLeftCantMeme and r/TheRightCantMeme subreddits, analyze similarities and differences in topics of discussion and notable emerging patterns.

II. Methodology

1) **Data Collection from Reddit** We collected data from relevant subreddits, focusing on r/TheLeftCantMeme and r/TheRightCantMeme, extracting posts and associated metadata for analysis.
2) **Preprocessing** Employing Optical Character Recognition (OCR) techniques, we converted image-based memes into text for comprehensive analysis.

3) **Topic Modeling with K-means** Utilizing K-means clustering, we implemented topic modeling to identify distinct clusters of memes within each subreddit. Keywords of each cluster will be extracted and highlighted using TF-IDF.
4) **Corpus-based Analysis** Conducted a comprehensive corpus-based analysis to uncover nuanced information within the textual content of memes.

Details on the process in 2.1 and 2.2 will be discussed in Section 3, while the techniques and insights derived from 2.3 and 2.4 will be elaborated upon in Section 4 and 5 respectively.

III. Dataset and Preprocessing

A. Data Gathering

The dataset for this analysis was collected from the r/TheLeftCantMeme and r/TheRightCantMeme subreddits using the Python Reddit API Wrapper (PRAW) library in Python [7]. To ensure a balanced representation and capture a diverse range of content, the data was gathered over a one-year timeframe. This timeframe strikes a balance between avoiding overly broad or narrow datasets, allowing for the inclusion of memes with relevant and varied time-based contexts. The dataset was curated by employing two sorting methods: 'top' (sorted by most upvotes) and 'controversial' (posts with upvote-to-downvote ratios close to 0.5).

```
# Sample Gathered Data
{
    "post_id": "183woek",
    "title": "LOL those red-caps sure are racist!
        (This comic is a shame, I usually like this
        guy's stuff.)",
    "image": "t3_183woek.png" # file name
}
# image files are stored in a local folder
```

The data from the 'top' sorting method are used for the main analysis of this paper, while data from the 'controversial' sorting method are use as references and for additional comparison and analysis.

The data collection includes:

- **Title** Textual Header of Each Post
- **Image** For posts containing gifs and videos, they were converted to images using the first frame.

TABLE II
Gathered Data

	r/TheLeftCantMeme		r/TheRightCantMeme	
	Top	***Controversial***	***Top***	***Controversial***
Post Count	913	924	902	873
Word Count	28,520	28,679	22,781	22,281
Char Count	19,1246	192,186	153,175	149,880

For the counts in Table II, each post (title text and the result from OCR) is converted into a text file, then counted using `wc ./cluster/*.txt`.

B. Optical Character Recognition (OCR)

Optical Character Recognition (OCR) is utilized to convert textual data in the image into machine-readable text. Four OCR libraries were considered

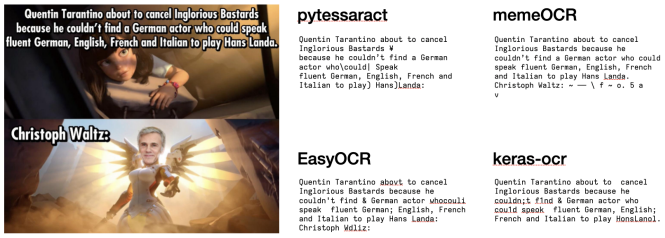

Fig. 2. Sample Result from the four OCR engines

- `pytesseract` by Google and Samuel Hoffstaetter (h)
- `EasyOCR` by Jaided AI
- `keras-ocr` by Fausto Morales (faustomorales)
- `Meme OCR` (a fine-tuned model of pytesseract) by John Lin (johnlinp)

To determine the most accurate OCR library, six images were randomly selected from the dataset, and the OCR results were compared against manually transcribed text. EasyOCR shows the most satisfying option due to its accuracy.

Subsequently, data from the four datasets (top and controversial from each subreddit) were subjected to OCR using EasyOCR, converting images to text for comprehensive analysis.

IV. Clustering Analysis

A. Data Cleaning

Given that many memes in the dataset are screenshots, several contain platform-specific interface data (e.g., Twitter retweets, Reddit post date, days since post, and other app interface elements). To refine the dataset, a two-step process was implemented:

- **Observation by Human** Initial observation by human annotators identified patterns indicative of platform-specific data.
- **Regular Expression (Regex) Substitution** Applying regex, platform-specific elements were systematically replaced with empty strings, effectively filtering out extraneous interface data.

This preprocessing step ensures a cleaner dataset, focused on the core content of the memes, facilitating a more accurate and insightful analysis.

B. Keywords for Substitution

- **Watermarks** made with mematic, imgflip.com, tenor
- **Interface Data** retweets, quotes, likes, views, comments, bookmarks, `(join)?\d+\s?h\s?reddit` (Reddit is post time data), `\d+\s?k` (Reddit is post upvote information)

C. Text Embedding

Given the inherently brief nature of memes, the risk of context loss and the limitations of comparing text alone arise. Notably, variations in linguistic forms expressing similar meanings (e.g., ”USA,” ”United States,” ”America”) demand a more nuanced approach. Traditional methods like Word2Vec [8] are deemed insufficient due to their reliance on older datasets (e.g., fasttext-wiki-news-subwords-300 from 2017 and glove-wiki-gigaword-200 from 2014). These methods often prioritize frequent words, potentially overlooking the meaning of new terms such as ”COVID-19” (emerging in 2019) and ”TikTok” (founded in 2018), as well as internet slang neologisms like ”skibidi,” ”fr fr,” and ”touch grass.”

To address these challenges, the text-embedding-ada-002 transformer-based embedding engine by OpenAI [9] was chosen. This selection serves a dual purpose: firstly, the dataset used for training is more recent, originating from 2021, capturing contemporary linguistic nuances. Secondly, the transformer-based approach is adept at understanding sentence context, accommodating the peculiarities of internet slang and new terminologies. This strategic choice aims to enhance the contextual understanding of meme content for a more robust analysis.

```
# Get Text Embedding from OpenAI API
def get_openai_embedding(text,
        model="text-embedding-ada-002"):
    response = openai.embeddings.create(
        input=[text], model=model)
    return response.data[0].embedding
```

For every post, the textual content from the post title and the textual information extracted from the image are combined into a unified string. This composite string is then embedded using the mentioned engine, producing a result represented as a 1,536-dimensional array of floats. The result embedding of each of the two clusters (i.e. r/TheLeftCantMeme and r/TheRightCantMeme, each sorted by ’top’) are then stored locally.

D. Clustering

The embeddings of each cluster undergo a clustering process utilizing the k-means clustering method. This iterative process runs from a range of clusters (k) starting from 1 and extending up to 30. The optimal number of clusters is algorithmically determined through the application of the elbow method, ensuring the identification of the most meaningful and distinctive groupings within the dataset. The process are performed utilizing functions from Python’s scikit-learn library [10]

The clustering process applied to the embeddings from r/TheLeftCantMeme results in 9 distinct clusters, whereas r/TheRightCantMeme yields 12 clusters.

E. Topic Modeling

As embeddings of the posts are clustered, TF-IDF is applied to the original text within each cluster to extract important keywords from each group. The process yield the following results:

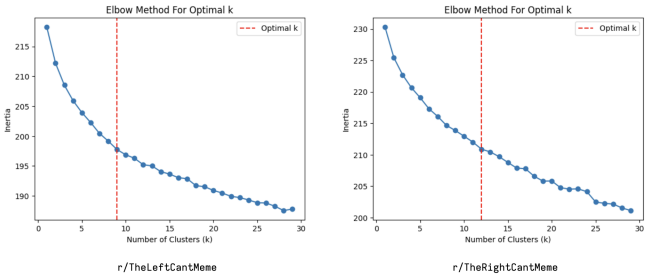

Fig. 3. Clusters of r/TheLeftCantMeme and r/TheRightCantMeme

1) Top Keywords from r/TheRightCantMeme Clusters:

- **Cluster 1** trump donald biden million america
- **Cluster 2** climate science think years electric
- **Cluster 3** babies fetus child life abortion
- **Cluster 4** kids trans children meme sex
- **Cluster 5** movie man like religion jewish
- **Cluster 6** dont slavery people im settle
- **Cluster 7** white black woke racist people
- **Cluster 8** people lgbtq gay month pride
- **Cluster 9** capitalism communist workers socialist korea
- **Cluster 10** right left wing conservative leftists
- **Cluster 11** trans gender pronouns men girl
- **Cluster 12** elon musk views woke like

2) Top Keywords from r/TheLeftCantMeme Clusters:

- **Cluster 1** com dont just did black
- **Cluster 2** trump president political republican humor
- **Cluster 3** hogwarts game legacy gaming circle
- **Cluster 4** trans women rights men people
- **Cluster 5** share reddit memes join rl
- **Cluster 6** kids gay people lgbt trans
- **Cluster 7** meme right left lgbt reddit
- **Cluster 8** communism capitalism communist work people
- **Cluster 9** people white right conservative nazi

The topic modeling analysis of each cluster reveals noteworthy similarities between the two subreddits, shedding light on shared thematic elements. Notably, both r/TheRightCantMeme and r/TheLeftCantMeme exhibit discussions around race and religion (evident in r/TheRightCantMeme Clusters 4, 6, 7 and r/TheLeftCantMeme Clusters 1, 9), politics and political ideologies (r/TheRightCantMeme Clusters 1, 9, 10 and r/TheLeftCantMeme Clusters 2, 9), and sex and gender (r/TheRightCantMeme Clusters 4, 8, 11 and r/TheLeftCantMeme Clusters 4, 6, 7).

While the topic modeling analysis highlights shared thematic elements between r/TheRightCantMeme and’r/TheLeftCantMeme, it is crucial to acknowledge that the presence of similar keywords doesn’t necessarily imply unanimous agreement or alignment in opinions. For instance, discussions around a shared keyword like ”Donald Trump” may encompass diverse perspectives, ranging from praise to critique.

Despite these shared topics, certain clusters illuminate distinct focuses that potentially underscore stereotypical issues and topics of concern within each ideological wing. For instance, r/TheRightCantMeme Cluster 2 brings attention to climate change and science, indicative of concerns often associated with right-wing ideologies. Meanwhile, Cluster 3 emphasizes abortion, aligning with conservative viewpoints. On the other hand, r/TheLeftCantMeme Cluster 3 revolves around the controversies surrounding the game "Hogwarts Legacy," released in February 2023. This controversy emerged from J.K. Rowling's series of statements, considered by some as transphobic, leading to a left-wing boycott of the game.

F. Visualization

The clusters derived from the topic modeling process are visualized using 2d t-distributed stochastic neighbor embedding (t-SNE), a technique that transforms high-dimensional data into a lower-dimensional space, emphasizing the preservation of local structures.

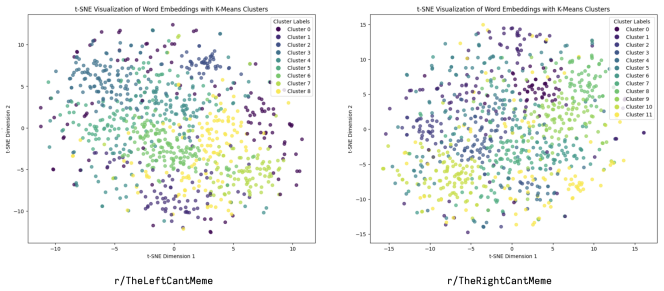

Fig. 4. 2d t-SNE visualization of the clusters

However, it is noteworthy that the resulting clusters may not exhibit distinct separation due to the close proximity of some clusters and the presence of sparse datapoints in certain clusters (e.g., Cluster 11 of r/TheRightCantMeme). This observation suggests that while posts can be effectively clustered, the inherently sparse and decentralized nature of Reddit content and the internet itself may contribute to a less clearly delineated clustering structure.

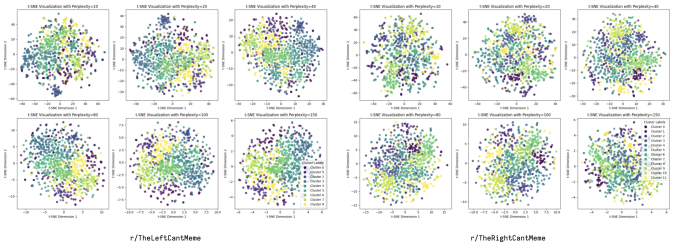

Fig. 5. 2d t-SNE visualization with a variety of perplexity constants

V. Corpus-based Analysis

To further analyze the data, we employ AntConc to delve deeper into the textual content of each cluster. Each of the posts are converted into a separate text file, organized into folders by subreddits and sorting methods ('top' and 'controversial'). They are then imported to AntConc as four corpus.

A. Lemmatization

Unlike the topic modeling process, watermarks and platform-based features are retained here to preserve contextual information while converting data into a .txt file.

However, the textual data are instead lemmatized using the NLTK library's WordNetLemmatizer [11]. This approach maintains contextual richness while reducing words to their base form, producing more concise information for the text.

```
# Lemmatization (sentences are splited into tokens)
tokens = [lemmatizer.lemmatize(token)
    for token in tokens]
```

B. Keyword Analysis

Fig. 6. Keywords of the Data Using AmE06 as a Reference Corpus

Keywords are extracted from the two corpora, utilizing AmE06 as a reference corpus. Notably, the words "right" and "people" emerge as significant keywords in both r/TheRightCantMeme and r/TheLeftCantMeme. Intriguingly, the collocates for the word "people" exhibit similarities across both subreddits, emphasizing themes related to race, with terms like "black," "white," and "victim" prevailing.

However, a nuanced distinction surfaces when exploring the collocates for the word "right." In r/TheRightCantMeme, the top three collocates—"wing," "left," and "far"—suggest a discourse primarily centered around political ideologies. In contrast, r/TheLeftCantMeme presents a different narrative with the top collocates being "cant," "meme," and "h." Further inspection reveals that "h" likely originates from the interface of Reddit, specifically in reference to the screenshot of r/TheLeftCantMeme, implying a critique of content within that subreddit. The presence of "h" as a collocate also aligns with Reddit's notation (`(\d)+h` indicating hours from the post time), adding a layer of platform-specific context to the analysis.

For the less significant terms in both corpora reveals intriguing parallels, with words like "trans" and "memes" sharing comparable collocates and contextual significance in both subreddits. This convergence in language suggests a shared vocabulary within the distinct political spheres, hinting at recurring themes that bridge the observed ideological divide in more pivotal keywords similar to analysis in Section IV.

C. Cross-Corpora Keywords Extraction

Keywords are then extracted from r/TheLeftCantMeme using r/TheRightCantMeme as a reference corpus. The keywords extracted are illustrated in a wordcloud below. Meanwhile, the reverse exploration from the reference corpus to r/TheRightCantMeme yields no hits.

Fig. 7. Keywords of r/TheLeftCantMeme Using r/TheRightCantMeme as a Reference Corpus (vice versa return no hits)

VI. Comparison between 'Top' and 'Controversial'

Considering the comparison between 'Top' and 'Controversial' posts within both r/TheLeftCantMeme and r/TheRightCantMeme. Keywords from the 'Controversial' category are extracted following the procedure outlined in Section IV. Surprisingly, the results exhibit a striking similarity to the keywords extracted from the 'Top' posts.

Further investigation is conducted using AntConc, comparing 'Controversial' and 'Top' posts from both subreddits. However, this analysis, conducted by treating one category as the main corpus and the other as a reference corpus, yields no hits.

This suggests a potential convergence in thematic content across these post categories within each subreddit. Alternatively, distinctions between these post types may manifest in other dimensions, such as variations in image types or arguments presented in the comment sections.

VII. Discussion

The striking similarity in keywords between r/TheLeftCantMeme and r/TheRightCantMeme illuminates shared topics of concern or contention between the two ideological camps. Notably, this convergence doesn't imply agreement in sentiments; rather, it unveils the focal points of heated arguments and discussions. For instance, the nuanced examination of keywords reveals variations in their usage contexts, such as some keywords originating from interface data. This suggests potential screenshot actions, hinting at one side scrutinizing the other. Moreover, it unveils preferences for specific platforms and other contextual information that influence the meme-sharing process.

Intriguingly, the absence of certain keywords in one dataset, like "Hogwart Legacy," suggests a deliberate imposition of discourse or stereotype by one ideological group onto another. This observation raises questions about the deliberate construction of narratives and the portrayal of specific themes to reinforce ideological perspectives. It underscores the nuanced ways in which memes serve not only as expressions of opinions but also as tools for shaping narratives and reinforcing stereotypes.

The unexpected lack of obvious textual differences between 'Top' and 'Controversial' posts within each subreddit prompts a closer examination of other dimensions where distinctions may lie. This absence of divergence in keywords between these post categories hints at a potential convergence in thematic content. However, the true distinctions might be embedded in the type of images shared or the nature of arguments unfolding within the comment sections. This has a potential to be further investigated in the future works.

The source code for this paper is accessible via https://www.github.com/chayapatr/corpus-memes

References

[1] Dawkins, R. (2006). The Selfish Gene. Oxford University Press.

[2] Savvas Zannettou, Tristan Caul!eld, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the Origins of Memes by Means of Fringe Web Communities. In Proceedings of IMC '18. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3278532.3278550

[3] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. 2019. A Characterization of Political Communities on Reddit. In 30th ACM Conference on Hypertext and Social Media (HT '19), September 17–20, 2019, Hof, Germany. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3342220.3343662

[4] Reddit.com Traffic Analytics, ranking stats, tech stack — similarweb. (n.d.). https://www.similarweb.com/website/reddit.com/

[5] r/TheLeftCantMeme. Reddit. (2017). https://www.reddit.com/r/TheLeftCantMeme/

[6] They just can't... — r/TheRightCantMeme. Reddit. (2017). https://www.reddit.com/r/TheRightCantMeme/

[7] praw-dev. PRAW: The Python Reddit API Wrapper. Github. (2023). https://github.com/praw-dev/praw

[8] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/1301.3781

[9] Embeddings - OpenAI API. OpenAI. (2023). https://platform.openai.com/docs/guides/embeddings

[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[11] Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc.