**Analyzing User Behavior, Trends and Communities using Reddit Posts**

**Sentiment Analysis of Subreddits**

**Problem Statement:** Gather comments from different subreddits and perform sentiment analysis to assess whether certain topics tend to attract more positive or negative sentiment.

**Challenges:** Sarcasm and Irony, Jargon and Slang, Short and Informal Texts.

Diverse Topics and Communities, Multimodal Content.

User Anonymity, Irrelevant Comments, Temporal Dynamics and handling Visual data.

**2. Literature Review:**

**2.1 Sentiment Analysis of AMC, Clover, and GameStop Reddit Comments**[[1]](#83jesk9qmle6)**:**

***Summary:*** The research utilizes the Clov package in R Studio and the NRC lexicon for sentiment analysis on Reddit comments about AMC, Clover, and GameStop. Network diagrams and radar charts present nuanced insights into distinct sentiment trends for each product. The study combines network and sentiment analysis, offering a holistic view of user interactions.

***Drawbacks:*** However, limitations include reliance on a snapshot of data, potentially missing real-time sentiments, and insufficient exploration of scalability for larger datasets or multiple subreddits.

**2.2 Enhancing Tweet Sentiment Analysis through Feature Ensemble Modeling and CNN**[[2](#k7nxhog7qorz)]**:**

***Summary:*** The paper addresses limitations in existing sentiment analysis methods by introducing a novel feature ensemble model. This approach considers lexical, word-type, semantic, position, and sentiment polarity features in tweets containing fuzzy sentiment. Evaluation with real data, including DB1 and DB2 datasets, demonstrates enhanced sentiment analysis performance, notably in F1 scores.

***Drawbacks:*** the study acknowledges challenges related to the interpretability of deep learning models, especially CNNs, urging attention to the transparency of decision-making processes.

**2.3 Sentiment Analysis using RGWE embeddings with improved accuracy**[[3]](#44i78t23jlgh)**:**

***Summary:*** The paper introduces Refined Global Word Embeddings (RGWE) as a superior method for sentiment analysis, demonstrating its effectiveness across various datasets, experiment settings, and factors like sentiment concept weight and semantic similarity threshold. RGWE, incorporating position features and internal/external sentiment information, outperforms traditional embeddings in capturing sentiment nuances.

***Drawbacks:*** The study acknowledges limitations like optimization for specific datasets and a lack of discussion on the method's generalizability.  
**2.4 Sentiment analysis for measuring hope and fear from Reddit posts during the 2022 Russo-Ukrainian conflict**[[4]](#qdzvcpqmgl0a)**:**

***Summary :*** The paper introduces an unsupervised sentiment analysis method to measure “hope” and “fear” sentiments in the context of the 2022 Ukrainian-Russian Conflict. Reddit data from six relevant subreddits was extracted using a Python script. The data was analyzed to gauge interest in the conflict and count “hope” and “fear” word occurrences. Topic modeling was also performed on the text data, with optimal topics estimated and goodness-of-fit measures along with Word clouds This research provides a novel sentiment analysis approach for geopolitical conflicts.

**2.5 A survey on sentiment analysis methods, applications, and challenges**[**[**5]](#ep6todt2u1jl):

***Summary:*** The paper discusses sentiment analysis data collection methods, including web scraping, social media, news channels, E-commerce websites, forums, and other websites. It highlights the importance of feature identification for model development and discusses the use of “Uni-gram”, “Bi-gram”, and “Tri-gram” techniques. The paper also mentions pragmatic features that focus on word application. Feature extraction is emphasized as a crucial task in sentiment classification. The paper explores three sentiment analysis approaches: the Lexicon Based Approach, the Machine Learning Approach, and the Hybrid Approach. The authors also discuss a corpus-based method used by Park and Kim (2016), which uses linguistic constraints and connectives to determine a new token’s sentiment.

**3. Methodology:**

**3.1 Datasets:**

3.1.1 *Tensorflow dataset*[[6]](#3r5nxpwyhhdw): This dataset comprises processed posts sourced from Reddit, containing a total of 3,848,330 entries.

3.1.2. *Pushshift Reddit Datase*t[[7]](#5zwfewzx5ir): Pushshift Reddit Dataset for research, focusing on two popular subreddits: r/AskReddit and r/WritingPrompts. The authors note that Reddit’s structure often favors entertaining content over informative responses.

3.1.3 *Sarcastic Comments - REDDIT Dataset*[[8]](#7fode4x0mvrv): Sarcastic comments generally mean the opposite of what the text suggests, so we will consider the probability of a text to be sarcastic.

3.1.4 *PRAW (Python Reddit API Wrapper)*: Use PRAW to retrieve data from Reddit. Depending on our requirements, we will fetch information about comments, or other Reddit entities.

3.1.5 *Pushshift API*[*[9*]](#ve03qnxvlbeo): This API allows to retrieve 20000 comments from Reddit per API request

**3.2 Approach:**

**3.2.1 Data Acquisition:**

3.2.1.1 *Selecting Existing Datasets:* We have chosen existing datasets that align with our goal. Considered the topics, domains, and types of data available. We ensure that the existing datasets have the necessary information(comments) required for analysis. The topics we have finalized are Movies, Sports, Politics, Education, Science and Technology, Literature, Agriculture, etc.

3.2.1.2 *Define Data Schema*: Extracted the data using PRAW so that it has a similar or compatible schema with the existing datasets. This is crucial for merging and combining datasets effectively. Our dataset contains comments and other important details from various subreddits.

3.2.1.3 *Reddit Data Extraction:* Used PRAW to extract data from Reddit on the decided subreddits, types of posts, and other parameters that are relevant to our goal of mining emotions. Extract the necessary information such as post title, content, comments, scores, etc.

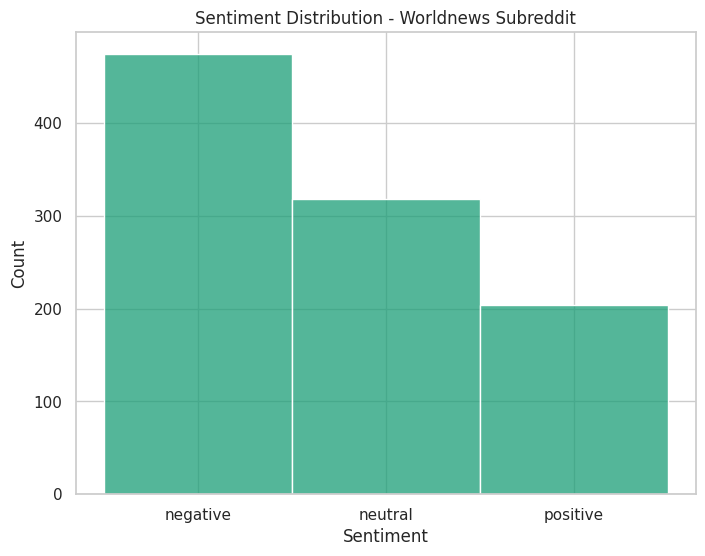
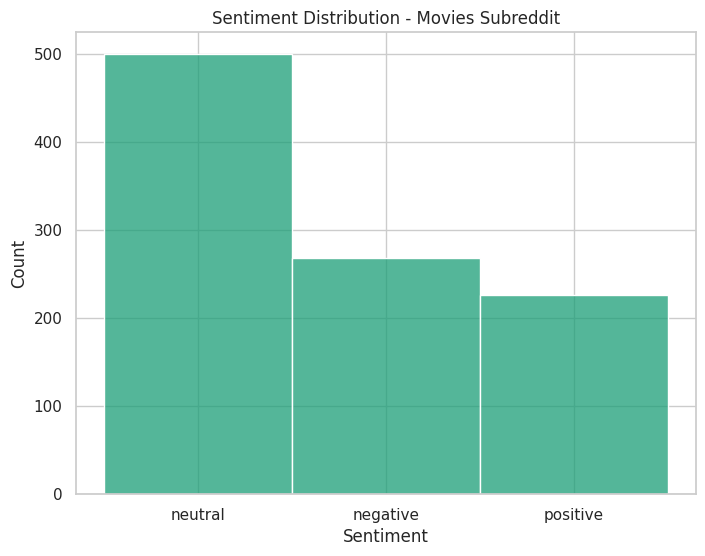
3.2.1.4 *Merge Datasets:* Merge the datasets obtained using PRAW based on common identifiers and make sure data has a unified schema.

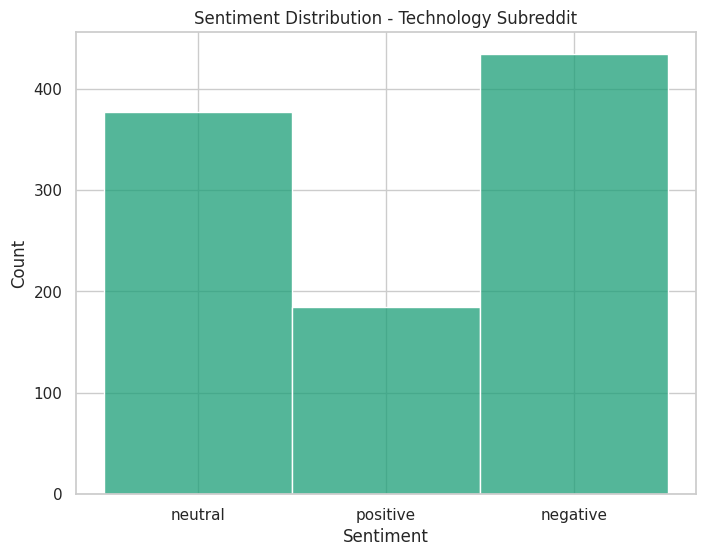
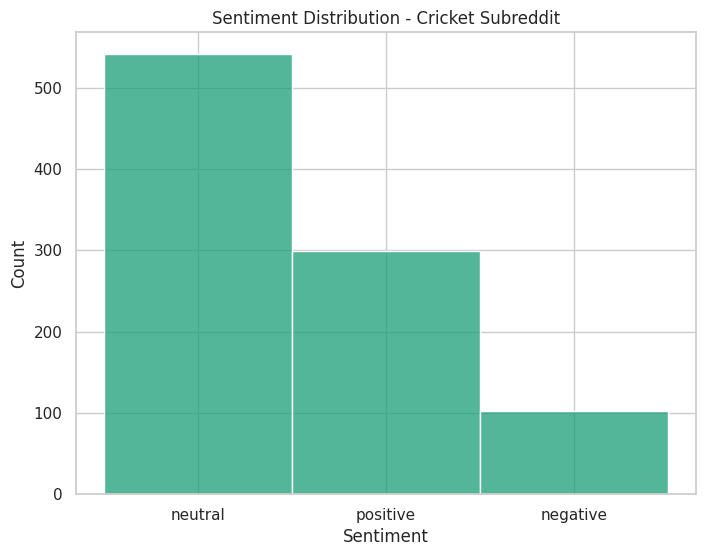
**3.2.2 Exploratory Data Analysis:**

For the initial phase of our analysis, we focused on extracting insights from headlines across four distinct subreddits: Cricket, Movies, Worldnews, and Technology. Our primary goal was to understand the sentiment dynamics within each subreddit by performing a basic sentiment analysis on the post headlines. We utilized PRAW (Python Reddit API Wrapper) to extract headline data from the selected subreddits to extract information including post titles and relevant metadata. Later, employed a sentiment analysis tool (VADER) to categorize headlines into Positive, Negative, and Neutral sentiments and created a sentiment distribution for each subreddit based on the analyzed headlines. Finally, generated separate graphs for each subreddit, illustrating the distribution of positive, negative, and neutral sentiments.The graphical representation confirmed observable differences in sentiment patterns among the chosen subreddits.



*Fig: Sample dataset obtained using PRAW on Cricket subreddit*



*Fig: Comparison of sentiments for 4 different subreddits*

**Progress:**

**Data Acquisition and Visualization:** Identify data sources, and hot topics on Reddit. Formed extensive dataset using API’s and existing data. Did basic visualization and analysis.

**Next Steps:**

**1. Data Cleaning and Preprocessing:**

Perform data cleaning and preprocessing on both the existing datasets and the data obtained from Reddit. This includes handling missing values, standardizing formats, and addressing spelling errors. There are abbreviations, internet slang, sarcastic comments, emojis, etc used by people these days and it is important to consider its effect on data interpretation.

Explore Top Keywords: Identify the most common keywords in the comments of each subreddit.

**2. Sentiment Analysis on Comments:** Extend analysis to include sentiment analysis on comments. This would be a baseline model.

**3. Time Series Analysis:** Consider performing sentiment analysis over time to see if there are temporal trends or patterns in sentiment within each subreddit.

**4. User Engagement Metrics:** Explore metrics related to user engagement, such as the number of upvotes, downvotes, and the length of discussions in comments. Additionally create Word Clouds, look into User Demographics to gain additional context.

**5. Sentiment Analysis Models:** Once we are done with Data collection, preprocessing, exploration, and feature engineering, we use a popular sentiment analysis model that suits the data from- VADER, TextBlob, BERT, GPT (Generative Pre-trained Transformer), LSTM Networks, etc. Finalize a function to deal with GIFs(Graphics Interchange Format).

**References:**

[1] Abiona A.A. et al, International Journal of Computer Science and Mobile Computing, Vol.11 Issue.9, September- 2022, pg. 34-40

[2] H. T. Phan, V. C. Tran, N. T. Nguyen and D. Hwang, "Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model," in IEEE Access, vol. 8, pp.14630-14641, 2020, doi: 10.1109/ACCESS.2019.2963702

[3] Y. Wang, G. Huang, J. Li, H. Li, Y. Zhou and H. Jiang, "Refined Global Word Embeddings Based onSentiment Concept for Sentiment Analysis," in IEEE Access, vol. 9, pp. 37075-37085, 2021, doi:10.1109/ACCESS.2021.3062654

[4] Guerra A, Karakuş O. Sentiment analysis for measuring hope and fear from Reddit posts during the 2022 Russo-Ukrainian conflict. Front Artif Intell. 2023 Apr 5;6:1163577. doi: 10.3389/frai.2023.1163577. PMID: 37091300; PMCID: PMC10113549.

[5] Wankhade, M., Rao, A.C.S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif Intell Rev* 55, 5731–5780 (2022). <https://doi.org/10.1007/s10462-022-10144-1>.

[6] M. Völske, M. Potthast, S. Syed, and B. Stein, "TL;DR: Mining Reddit to Learn Automatic Summarization," in Proceedings of the Workshop on New Frontiers in Summarization, Copenhagen, Denmark, Sep. 2017, pp. 59-63, doi: 10.18653/v1/W17-4508

[7]Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, Omer Levy, “LIMA: Less Is More for Alignment”, arXiv:2305.11206v1 [cs.CL] 18 May 2023.

[8]Mikhail Khodak and Nikunj Saunshi and Kiran Vodrahalli, “A Large Self-Annotated Corpus for Sarcasm”, [https://arxiv.org/abs/1704.05579](https://arxiv.org/abs/1704.05579%7D), year=2017

[9]Lindskog, Sebastian and Serur, Juan Andrés, Reddit Sentiment Analysis (August 15, 2020).