SUBREDDIT SENTIMENT & EMOTIONAL ANALYSIS REPORT

- Abhishek Raj Chowdary

-Damain James Moquin

- Krupa Minesh Shah

-Sunidhi Jain

AGENDA

- Problem Statement
- Data Preprocessing
- LDA
- Vader
- Empath
- Conclusion
- References

PROBLEM STATEMENT

The aim of our project, "Understanding the Impact of Social Media on Mental Health using Subreddit Sentiment and Emotional Analysis," is to explore how different discussion topics on social media platforms influence positive and negative emotions, as well as overall mental health. We plan to analyze top subreddits representing a wide array of topics, utilizing web-scraping techniques and various NLP models (like VADER, Empath, and LDA)

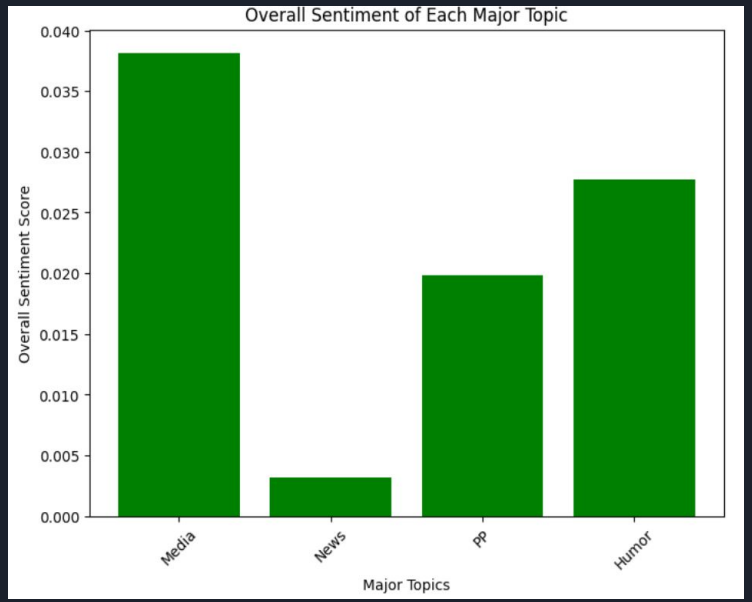
DATA SCRAPING & PREPROCESSING

- Used Python Reddit API Wrapper(PRAW) to extract the data.
- Categories:
  - Humor: Funny, Memes, Showerthoughts
  - Media: Gaming, Music, Movies
  - News: News, Upliftingnews, Worldnews
  - Politics/Philosophy: Politics, Philosophy, Unpopular opinions
- Dataset Structure: subreddit, post_title, post_score, post_url, post_num_comments, post_text_content, post_user_age_days, comment_text, comment_score, , comment_user_age_days
- Preprocessing: Tokenization, Removing Special characters & Stopwords, , Lemmatization

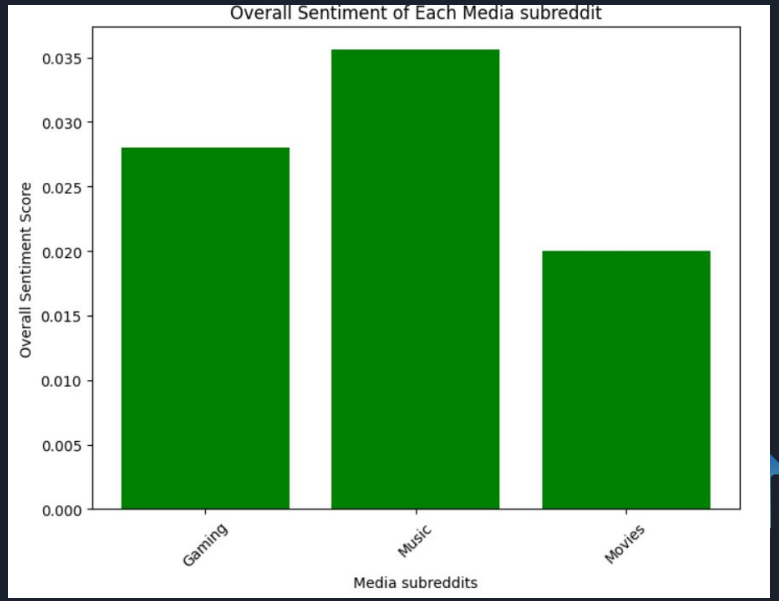
Latent Dirichlet Allocation(LDA)

- It is a powerful probabilistic model used for uncovering hidden topics within a collection of documents.
- LDA operates under the assumption that each document in a corpus is a mixture of various topics, and each topic is characterized by a distribution of words.
- This means that a document may discuss multiple topics simultaneously.
- Example:
  - "sports" might have a high probability for words like "game," "player," and "score,"
  - "politics" might have a high probability for words like "government," "election," and "policy."

Latent Dirichlet Allocation(LDA) Results

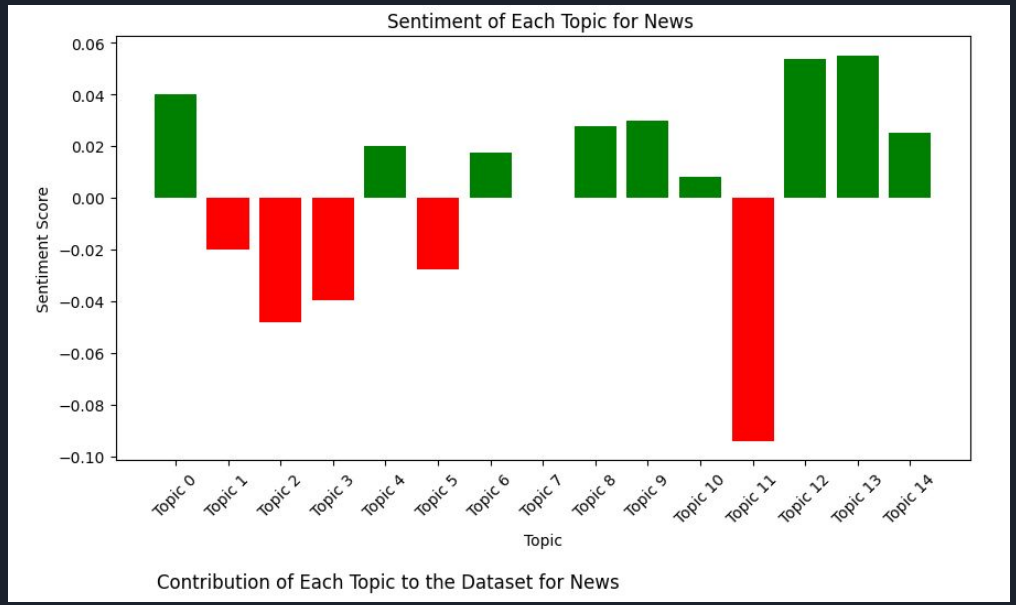


LDA Overall Sentimental Analysis



LDA Media Subcategories Sentimental Analysis

Latent Dirichlet Allocation(LDA) Results



LDA News Topics Modeling Analysis

Latent Dirichlet Allocation(LDA) Results

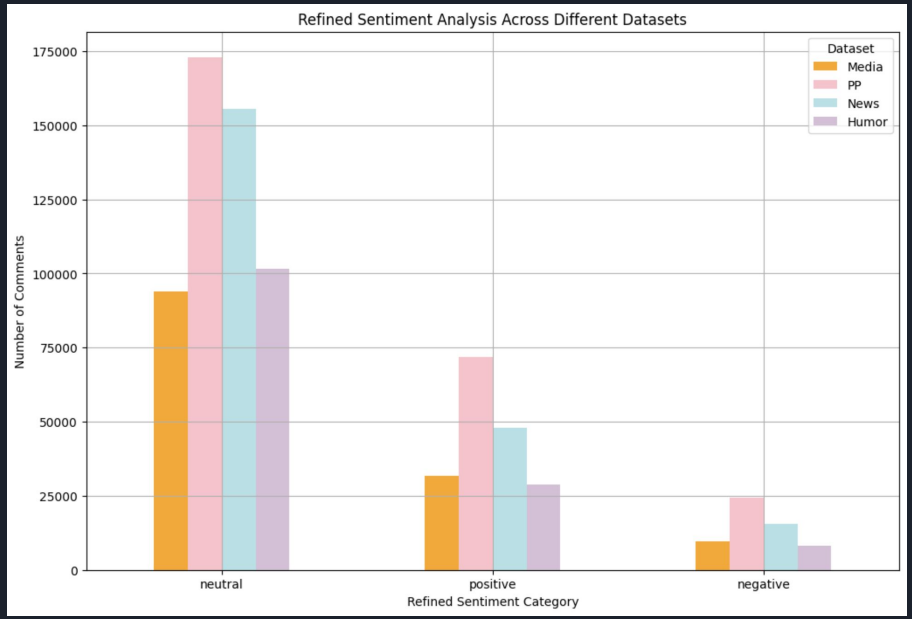
- People who tend to consume media(gaming,music,movie) tend to be happiest among the 4 major categories.
- Specifically, individuals who engage with music content within the media category tend to experience the highest levels of positive sentiment overall.
- Conversely, consumers of news content tend to have lower happiness levels compared to other major categories.
- Within the news category, individuals participating in controversial discourse and social critique display the most negative sentiment.

Valence Aware Dictionary and sEntiment Reasoner (VADER)

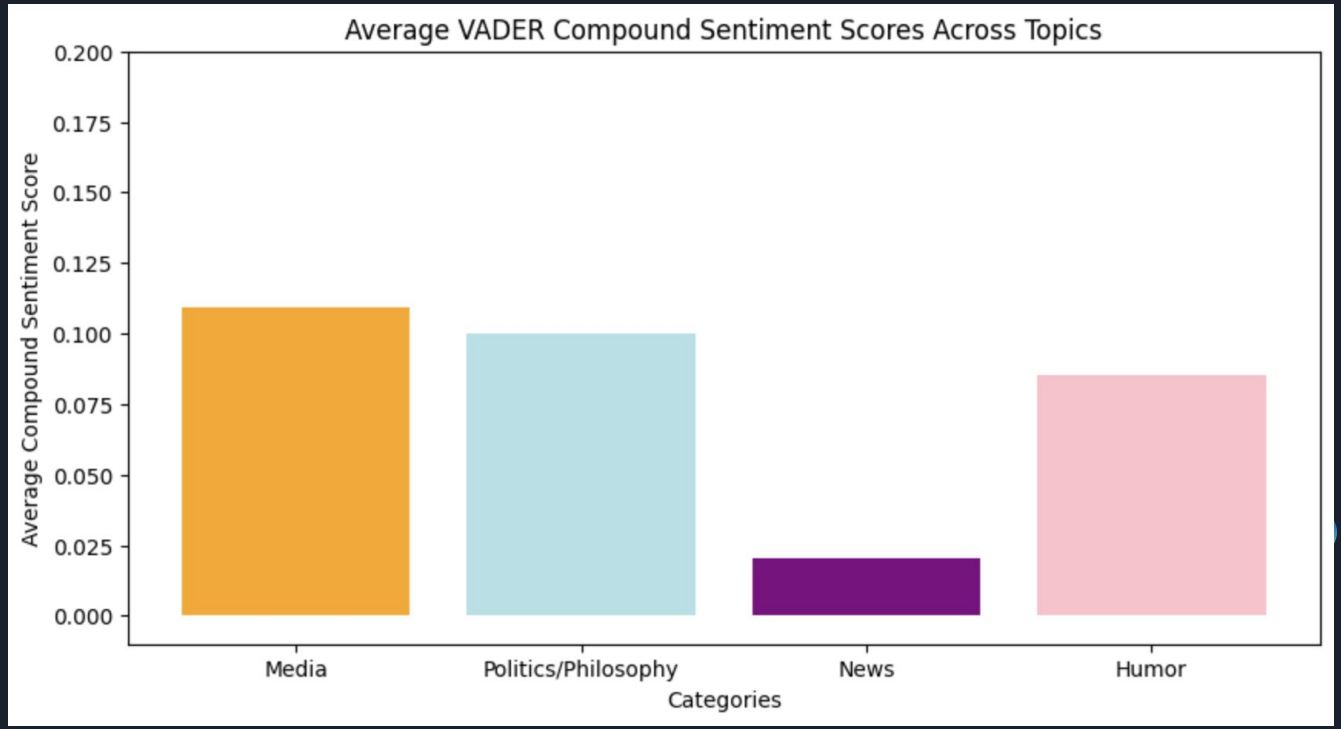
- VADER is a tool designed to analyze the sentiment of text, particularly focusing on texts from social media platforms, where expressions and slang are common
- It uses a lexicon to evaluate the sentiment of words in a text. Each word in the lexicon is scored for its positivity or negativity
- It is straightforward to implement and pre-built which can be used directly to assess sentiments
- It gives a compound score that combines the individual scores of words in the text, which indicates the overall sentiment (positive, neutral, or negative)

Valence Aware Dictionary and sEntiment Reasoner (VADER)

- A graph explaining individual Positive, negative and neutral values across all categories
- The result for average Vader Sentiment values across all topics



Valence Aware Dictionary and sEntiment Reasoner (VADER)

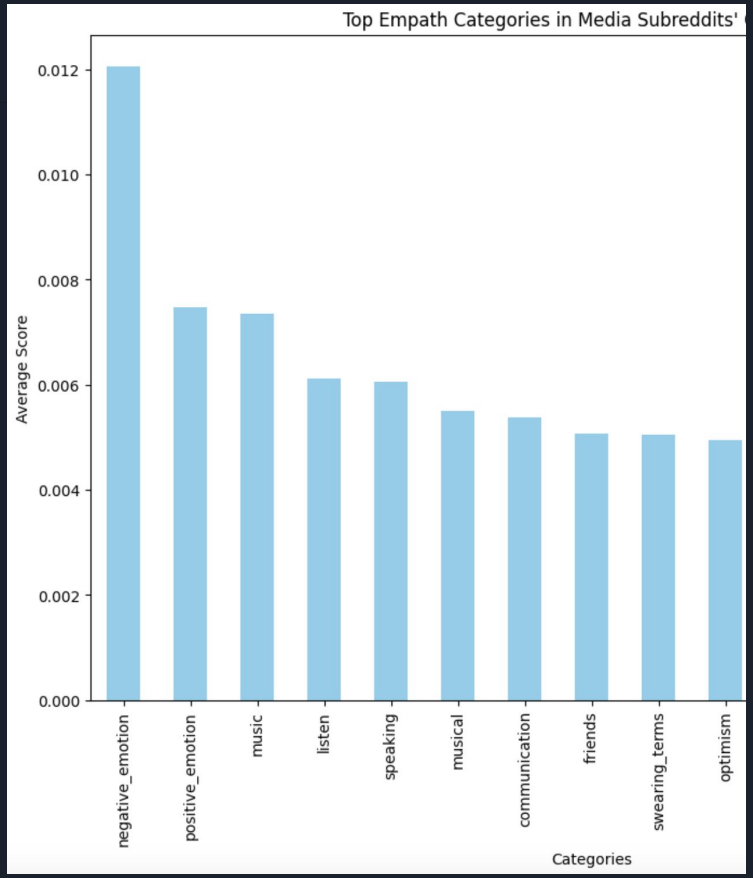


EMPATH

- “Tool that generates and validates lexical categories from a set of seed terms”
- 200 categories for topics and emotions, a mix of topic modeling and emotional categorization
- Similar to Linguistic Inquiry and Word Count (LIWC)
  - LIWC has hand crafted proprietary dictionaries for categories
  - Empath uses deep neural embeddings for categories along with crowdsourced validation
- Subreddits put into subcategories (Media, News, Politics/Philosophy, Humor), Empath algorithm run on each subcategory
- Plotted top 15 mean category values across posts in each subcategory

EMPATH Results

- “negative_emotion” category had the highest score for each reddit subcategory
- The News subcategory had the largest difference between the “negative_emotion” and “positive_emotion” categories
  - Most “negative” subcategory
- Humor had the smallest difference, making it the post “positive” subcategory
- Every subcategory had high rankings for categories with negative connotation (“violence”, “swearing_terms”, etc.)



CONCLUSION

- Exact sentiment/emotional values vary based on the method used (LDA, VADER, EMPATH)
- The News subcategory is the most negative compared to Media, News, or Politics/Philosophy
  - News possibly fosters negative discussion because the most common topics tend to be ‘bad news’ due to reporting bias and popularity
- Media tends to be the most positive subcategory
  - People tend to enjoy talking about the media (movies, games, music) they like

REFERENCES

H. U. Abro, Z. S. Shah and H. Abbasi, "Analysis Of COVID-19 Effects On Wellbeing - Study Of Reddit Posts Using Natural Language Processing Techniques," 2022 International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE), Lahore, Pakistan, 2022, pp. 1-7, doi: 10.1109/ETECTE55893.2022.10007300.

N. S. Kamarudin, G. Beigi and H. Liu, "A Study on Mental Health Discussion through Reddit," 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), Pekan, Malaysia, 2021, pp. 637-643, doi: 10.1109/ICSECS52883.2021.00122.

T. Nandurkar, S. Nagare, S. Hake and K. Chinnaiah, "Sentiment Analysis Towards Russia - Ukrainian Conflict: Analysis of Comments on Reddit," 2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP), Nagpur, India, 2023, pp. 1-6, doi: 10.1109/ICETET-SIP58143.2023.10151571.

S. Albota, "Linguistic and Psychological Features of the Reddit News Post," 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT), Zbarazh, Ukraine, 2020, pp. 295-299, doi: 10.1109/CSIT49958.2020.9321991.

M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Detection of Depression-Related Posts in Reddit Social Media Forum," in IEEE Access, vol. 7, pp. 44883-44893, 2019, doi: 10.1109/ACCESS.2019.2909180.

S. Thukral et al., "Analyzing Behavioral Trends in Community Driven Discussion Platforms Like Reddit," 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 2018, pp. 662-669, doi: 10.1109/ASONAM.2018.8508687.

E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding Topic Signals in Large-Scale Text," in Proc. 2016 CHI Conf. Hum. Factors Comput. Syst., May 2016, doi: 10.1145/2858036.2858535

THANK YOU!