Navigating Depressive Discourse: NLP Analysis of Reddit and Twitter

Annapia Borraccino

CUNY Graduate Center

aborraccino@gradcenter.cuny.edu

Abstract

This project explores how users express feelings of depression on social media platforms like Twitter and Reddit. We leverage Log Odds Ratio to identify linguistic variations in both content and non-content words. By training and evaluating models based on these features, we aim to achieve better performance compared to a baseline in classifying user content across three classification tasks.

Keywords:[1] NLP, corpus linguistics, depression detection, social media analysis

1. Introduction

Social media offers a unique window into users psychology, as people often express themselves more freely online, where at times anonymity is guaranteed. Depression's significant impact is undeniable. Ranked as the fifth leading cause of disease burden in the US [9], it demands innovative solutions. Social media platforms like Twitter and Reddit, coupled with powerful tools like Natural Language Processing (NLP) and Machine Learning (ML), offer immense potential and provide us with tools for monitoring depression-related linguistic trends over time, ultimately leading to a deeper understanding of the disorder.

Social media platforms shape online discourse in unique ways. Twitter's character limit (280) encourages concise messages, while Reddit's higher limit (8,000) allows for more in-depth discussion. Additionally, Reddit's looser age verification practices might attract a younger demographic compared to Twitter, though the minimum age is 13 for both platforms. Recent data [3] shows Twitter's strongest user base in the US is young adults (42% aged 18-29), followed by those aged 30-49 (27%). Understanding these platform differences is crucial because age can significantly influence the language, the register, and topics users discuss within the corpora.

2. Literature Review

Previous literature has focused on mental health detection, each to a different degree. Fine [5] highlighted how speech patterns can be a window into the mind, and this insight has been leveraged by psychiatrists in diagnosing mental health conditions. In Jain et al. [7] they apply ML and NLP techniques to analyze language related to depression and suicidal ideation using data from two subreddits. The goal was to distinguish between language indicating clinical depression versus language indicating high risk of suicide ideation. They concluded that the 'r/SuicideWatch' subreddit provides valuable signals for identifying suicidal behavior online, and their models can aid mental health professionals in detecting high-risk individuals based on the language used. The core contribution is applying text classification methods on Reddit data to distinguish depression versus suicidal ideation language.

Fine et al. [6] explored the potential of using NLP on public social media data to estimate population-level mental health in real-time. They focused on anxiety, depression, and suicide risk among Twitter users in the United States. Data was collected from both healthcare professionals and a general population sample over a six-month period in 2020. The researchers employed pre-trained models to analyze each tweet and assess the likelihood of the author experiencing these mental health concerns. By comparing trends in these scores before and after significant events like COVID-19 lockdowns and the death of George Floyd, they observed a rise in distress levels across both healthcare workers and the general public. This study highlights the potential of social media data as a valuable tool for rapidly measuring the mental health impact of crises and informing public health interventions. In Parkar et al. [10] the authors propose a novel approach for depression detection leveraging NLP on video search data from YouTube. Microblogging platforms like

[1] github:annapiab/NLP_depression: Navigating Depressive Discourse: NLP Analysis of Reddit and Twitter (github.com)

YouTube offer users a space to express themselves, potentially revealing behavioral patterns and personality traits. This study proposes a system for detecting depression based on the titles and keywords users search for within their application. By integrating the YouTube API, the system can recommend related videos based on user searches. Additionally, the authors suggest employing NLP algorithms to analyze user search history and categorize users into depressed or non-depressed groups.

All of these studies have pre-processed their corpus to focus on content words only. Nevertheless, non-content words, often overlooked, have previously provided valuable insights into users' engagement. For instance, a study by Preoţiuc-Pietro et al. [12] highlighted distinct patterns: content words often revolved around 'illness management' on Twitter, while non-content words reflected an 'increased focus on the self'. Moreover, while Twitter data has undergone extensive analysis over the past decade, research on Reddit corpora – focusing on depression only – remains limited but is steadily expanding. Following the suggestion in Rajput & Ahmed [13], this project aims to juxtapose the insights gained from a recent Twitter corpus (2020) with those documented in previous literature.

3. Data

Both corpora used in this study can be found on Kaggle. The Twitter corpus [14] contains 10.314 tweets in English extracted using the Twitter API. Each one has a label: 0 = not_depressed, 1= depressed. This makes the dataset a popular and valuable source, often used for sentiment analysis. The Reddit corpus [2] contains circa 7.000 text entries also purely in English language. The raw data was collected through web scraping depression-related Subreddits and was provided after being cleaned and labeled (1 = is_depressed, 0 = is_not_depressed). Both datasets were resized to roughly the same size.

	Twitter	**Reddit**
Corpus size	**3804**	**3197**
Depressed	1902	1400
Non-depressed	1902	1797

Table 1: Corpora size and label numbers

During the process, all personally identifiable information (e.g., user IDs, usernames) was removed from the data to ensure privacy and anonymity. Consequently, the resulting corpora contain only the text data and the corresponding labels. The labels originate from external annotators. For the purpose of this study, the label assignments have undergone a simple validation process based on personal judgment to ensure accuracy and reliability.

3.1. Data pre-processing

Numerous data preprocessing techniques were employed to adapt the data to the project goals. The Log Odds Ratio computation involved filtering the text data for content words initially, followed by filtering for non-content words. Subsequently, these words, along with their frequencies, were saved in TSV files.

In the training of the SVM model, preprocessing followed the classic steps: tokenization, case-folding, removal of stopwords, and lemmatization. As far as tokenization is concerned, it is important to highlight that due to the prevalence of nonstandard English or Internet-like language in many essays, the reliability of tokenizer or tagger performance may not be as consistent as it would be with more conventional language usage, i.e., data coming from sources like official documents, research papers, etc.

4. Methods

4.1 Log Odds Ratio

The first question that the study aims to answer is: ‘What are the linguistic differences in expressions of depression between Twitter and Reddit? And what do these suggest, for both content and non-content words?’. One of the approaches to comparing corpora can be computing the Log Odds Ratio between them [8]. This can be exemplified by:

$$lor(w) = log\left(\frac{f^i(w)}{n^i - f^i(w)}\right) - log\left(\frac{f^j(w)}{n^j - f^j(w)}\right)$$

The equation provided offers a way to estimate whether a word *(w)* has greater likelihood in corpus *i* or corpus *j*. It calculates this by dividing the frequency count of *(w)* in corpus *i*, denoted as $f^i(w)$, by the difference between this count and the total

number of words in corpus *i*, n^i. This process is mirrored for corpus *j*, and the resulting ratio can be any real number. Positive ratios suggest higher odds in corpus *i,* while negative ratios indicate higher odds in corpus *j*.
This computation was executed four times:

- between content words between the Twitter corpus and the Reddit one;
- between non-content words between the Twitter corpus and the Reddit one;
- between depressed and non-depressed content words in the Twitter corpus;
- between depressed and non-depressed content words in the Reddit corpus;

4.2 Classification Tasks

The second question driving this study centers on the efficacy of linguistic features from diverse social media platforms in identifying posts suggestive of depression. To address this, a series of Support Vector Machine (SVM) models were trained utilizing various features extracted from the text data.
The classification task revolves around binary classification, aiming to discern the presence or absence of symptoms associated with depression within the textual content of social media posts. This method involves using linguistic clues found in the text to predict an individual's mental health status.
By deploying SVM models trained on distinct feature sets, this research seeks to highlight the discriminative power of linguistic features in detecting signs of depression within the digital discourse landscape.

4.3 Feature Selection

Two distinct sets of features were extracted for the classification task, each offering unique insights into the textual content:

1) **TF-IDF Features:** leveraging the TfidfVectorizer from the scikit-learn library, this feature set quantifies the significance of each term within the text corpus. By assigning weights based on term frequency-inverse document frequency, we manage to capture the relevance of terms in distinguishing patterns associated with depression.
2) **Empath Categories:** While traditionally, the LIWC categories have been the golden standard [11], logistical constraints prevented their utilization in this study. Instead, we turned to the Empath categories [4] as a viable alternative. By leveraging the Empath library, this feature set analyzes the emotional and psychological subtleties woven within the text. Through categorizing text into empathic themes like emotions, relationships, and mental states, including depression, this methodology offers a comprehensive understanding of the prevailing sentiments and themes indicative of depression in textual content.

4.4 Classification Model

The dataset was divided into training and testing subsets, with an 80-20 split ratio maintained throughout. To establish a performance benchmark, a dummy classifier was trained. Its predictions were then compared against those generated by the SVM classifier. Using the designated feature set, SVM classifiers with a linear kernel were trained on the training data. Comprehensive evaluation was conducted, encompassing a range of metrics including accuracy, precision, recall, and F1-score. Moreover, cross-validation techniques were employed to evaluate the model's ability to generalize and to mitigate potential overfitting. Additionally, word clouds were created to visually represent the most influential words.

5. Results & Discussion

5.1 Log Odd-Ratio – results

The initial Log Odds Ratio was calculated for the content words found in both the Twitter and Reddit corpora, without distinguishing whether they were indicative of depressive discourse or not. The results can be observed in *Figure 1*.
The darker bars represent words most likely associated with Twitter, while the lighter bars signify those likely originating from Reddit. Although no distinction was made between 'depressed' and 'non-depressed' content words, the results offer insights into potential topics and associated emotions. For instance, words like *"tire"*, *"depressed"*, *"annoy"*, *"bore"*, and *"disappoint"* convey the prevailing sentiments among Twitter users. Conversely, words

like *"milk"* or *"prettier"* could indicate underlying reasons for these feelings; Interestingly, "milk" seems to be associated with eating disorders and consequently with feelings of depression upon examination of the text data.
In contrast, Reddit words seem to center around topics that may evoke distress and concern—illustrated by terms such as *"job"*, *"college"*, *"money"*, and *"health"* in the chart. Reddit's content words appear to offer deeper insights into these topics, possibly reflecting the platform's propensity for more extensive and freer discussions.

Concerning non-content words, it is observed that those with the highest negative log-odds ratios (e.g., *"you," "yours," "be," "yourself"*) likely denote terms frequently utilized in Twitter discussions, possibly reflecting self-referential language and the dispensation of advice using the indefinite second person. Conversely, non-content words with the highest positive log-odds ratios (e.g., *"any", "few", "him", "his"*) may typify terms more prevalent in Reddit conversations, potentially indicating a communication style focused on outward expression or the dissemination of information, along with soliciting advice, in contrast to the discourse style observed on the other platform (see *Figure 2*).
To go deeper into the language used to describe depressive discourse on each platform, we conducted a Log Odds Ratio analysis on content words extracted from posts labeled as depressed compared to those labeled as non-depressed. In *Figure 3*, the blue bars represent content words associated with depression in Reddit posts. Here, words such as *"anxiety", "mental", "health", "therapist", "advice", "symptom"*, or *"struggle"* indicate that Reddit users utilize the platform not only to express their emotions but also to seek guidance. This may involve seeking advice on coping with depression, whether through discussions about consulting a therapist or managing somatic symptoms of depression, a space that users have not been given on Twitter. In fact, looking at this platform's results, the chart leads to a similar conclusion obtained at the beginning. Users seem to just be able to vent on the platform, hence words like *"mood", "tire", "idk", "mentally", "suicide", "anymore"* (see *Figure 4*).

5.2 Classification Tasks – results

The initial model employed a Support Vector Machine (SVM) utilizing TF-IDF features. It underwent training and testing (80-20 split) on a combined corpus, encompassing both Reddit and Twitter labeled datasets. Subsequently, a similar SVM model was exclusively trained and tested on Twitter data, followed by another one solely focusing on Reddit data. For each model, we established a baseline using the DummyClassifier from sklearn, and notably, all models performed significantly better than this baseline. Specifically, the accuracy scores were .62 for the mixed corpus, .49 for Twitter, and .76 for Reddit. These results are presented in *Table 2*.

SVM + TF-IDF

	Mixed Corpora	**Twitter**	**Reddit**
	F-1	**F-1**	**F-1**
0 = non_depressed	.97	.95	.99
1 = depressed	.95	.95	.98
accuracy	.97	.95	.99

Table 2. Results of SVM + TF-IDF[2]

Following that, we proceeded to train the SVM using TF-IDF features on the Reddit dataset and evaluated its performance on the Twitter corpus through again an 80-20 split. In this scenario, the baseline accuracy stood at .49. Overall, the model exhibited better performance compared to the baseline; however, it fell short of the performance achieved by the previous models. This outcome was anticipated due to the previously mentioned differences between the Reddit and Twitter datasets, like distinct user bases, posting styles, and topics of discussion, which result in variations in language use. Therefore, a model trained solely on Reddit data may not generalize as effectively to the Twitter dataset, leading to a comparatively lower performance.

[2] These tables only report F-1 scores, though recall and precision are available on the Notebook.

SVM + TF-IDF

	Train: Reddit Test: Twitter
	F-1
0 = non_depressed	.78
1 = depressed	.62
accuracy	.72

Table 3. Results of SVM + TF-IDF, trained on Reddit, tested on Twitter

Lastly, we trained the SVM model using Empath feature extraction, employing it across the mixed corpora, Twitter, and Reddit datasets. Notably, while the model's performance remained above baseline, it lagged behind that of models utilizing TF-IDF features. Given Empath's capability to extract numerous categories, future research might benefit from refining hyperparameters or strategically selecting features to potentially enhance performance even further.

SVM + Empath

	Mixed Corpora	Twitter	Reddit
	F-1	**F-1**	**F-1**
0 = non_depressed	.86	.75	.95
1 = depressed	.67	.66	.80
accuracy	.81	.72	.92

Table 4. Results of SVM + Empath

5.4 Limitations

One limitation of this study is the relatively small size of the datasets utilized which may affect the robustness of machine learning models. Another constraint is that only text data is being considered, potentially overlooking speech data that may also be indicative of depression. Additionally, while Empath offers a rich set of categories for feature extraction, its potential may not have been fully realized in this study. Future research could explore more sophisticated methods for leveraging Empath features and enhance classification performance.

5.5 Ethical considerations

There is a risk of stigmatization and discrimination associated with labeling individuals as "depressed" based solely on their online activity. This could perpetuate harmful stereotypes and lead to social marginalization of individuals struggling with mental health issues.

Furthermore, there is a need to consider the potential impact of the research findings at scale. Highlighting differences in language use between depressed and non-depressed individuals could inadvertently reinforce negative perceptions or exacerbate existing disparities in access to mental health care.

6. Conclusion

In conclusion, this study offered valuable insights into how users express depressive discourse, the topics they discuss, and the unique roles that Twitter and Reddit play in these conversations. We evaluated the effectiveness of SVM models using TF-IDF and Empath features to classify depressive discourse, finding that while both methods outperformed baseline models, TF-IDF features provided superior performance. Identified limitations highlight areas for future improvement.

Although this study has not uncovered groundbreaking findings, the goal is to identify language indicative of depression online and ultimately promote more positive and hopeful messages on affected individuals' pages. Additionally, we envision this data to be used to train a therapy chatbot, for example, for deployment in countries where mental health is still not widely recognized as a legitimate concern.

References

[1] Alduailaj, A. M., & Belghith, A. (2023). Detecting Arabic Cyberbullying Tweets Using Machine Learning. *Machine Learning and Knowledge Extraction, 5*(1), 29-42. https://doi.org/10.3390/make5010003

[2] *Depression: Reddit Dataset (Cleaned)*. (2022, August 10). Retrieved from https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned

[3] Dixon, S. J. (2024). Distribution of X (formerly Twitter) users worldwide as of April 2024, by age group. In *Statista*. Retrieved May 12, 2024, from https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/

[4] Fast, E., Chen, B., & Bernstein, M. S. (2016). *Empath: Understanding topic signals in large-scale text.* Stanford University. Retrieved from https://hci.stanford.edu/publications/2016/ethan/empath-chi-2016.pdf

[5] Fine, J. (2006). *Language in psychiatry: A handbook of clinical practice.* Equinox London.

[6] Fine, A., Crutchley, P., Blase, J., Carroll, J., & Coppersmith, G. (2020). Assessing population-level symptoms of anxiety, depression, and suicide risk in real time using NLP applied to social media data. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science* (pp. 50-54). Online: Association for Computational Linguistics.

[7] Jain, P., Srinivas, K. R., & Vichare, A. (2022). Depression and suicide analysis using machine learning and NLP. *Journal of Physics. Conference Series, 2161(1),* 012034. https://doi.org/10.1088/1742-6596/2161/1/012034

[8] Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2009). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis, 16*(4), 372-403.

[9] Murray, C. J., Atkinson, C., Bhalla, K., Birbeck, G., Burstein, R., Chou, D., Foreman, Lopez, Murray, Dahodwala, Jarlais, Fahami, Murray, Jarlais, Foreman, Lopez, Murray, & US Burden of Disease Collaborators. (2013). The state of US health, 1990-2010: burden of diseases, injuries, and risk factors. *Journal of the American Medical Association*, 310(6), 591-608. https://doi.org/10.1001/jama.2013.13805

[10] Parkar, B., Lanjekar, S., Mulla, A., & Patil, V. (2021). Depression Detection Using NLP Algorithm On YouTube Data. *International Research Journal Of Engineering And Technology*.

[11] Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count: LIWC* [Computer software]. Austin, TX: liwc.net.

[12] Preoţiuc-Pietro, D., Sap, M., Schwartz, H. A., & Ungar, L. (2015). Mental Illness Detection at the World Well-Being Project for the CLPsych 2015 Shared Task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 40-45). Denver, Colorado: Association for Computational Linguistics.

[13] Rajput, A., & Ahmed, S. (2019). *Making a Case for Social Media Corpus for Detecting Depression.* International Journal of Advanced Computer Science and Applications, 10(4), 407-412.

[14] Ravehgillmore. (2021, July 14). *Predicting depression from tweets using BERT.* Retrieved from https://www.kaggle.com/code/ravehgillmore/predicting-depression-from-tweets-using-bert/notebook

APPENDICES

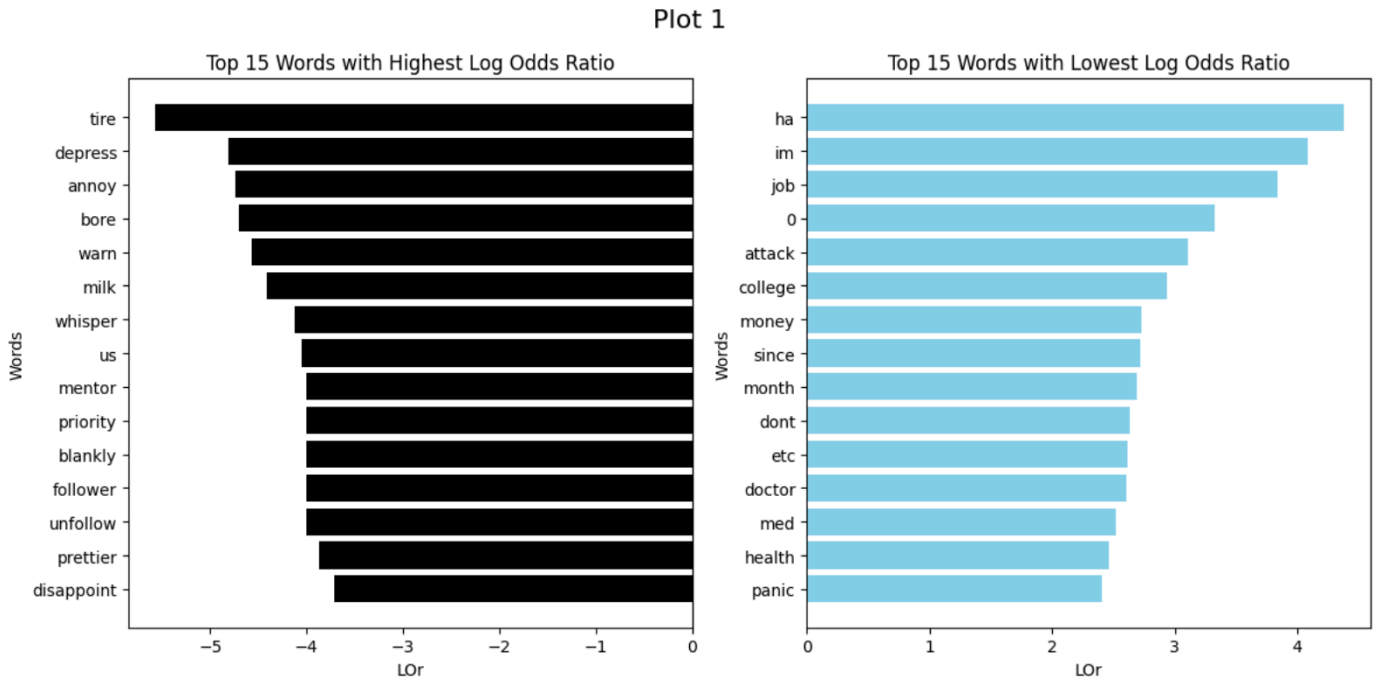

Figure 1. LOr of content words Twitter – Reddit

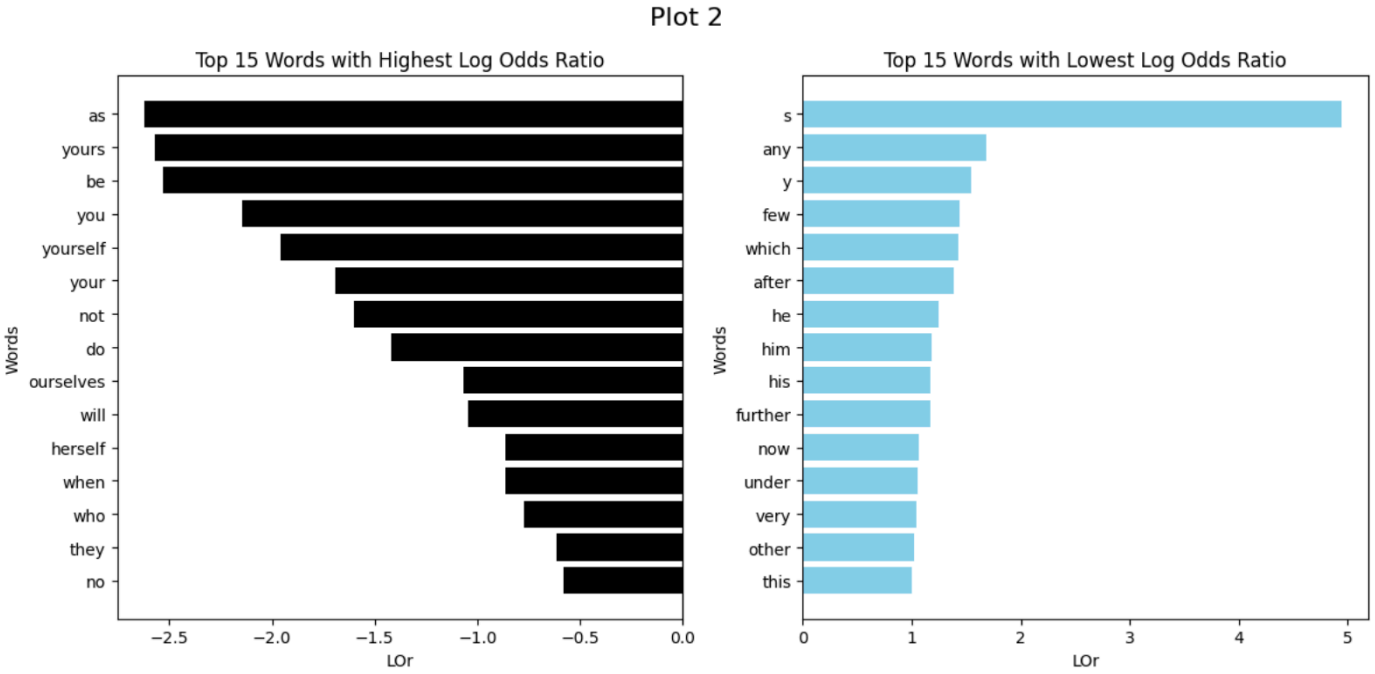

Figure 2. LOr of non-content words Twitter – Reddit

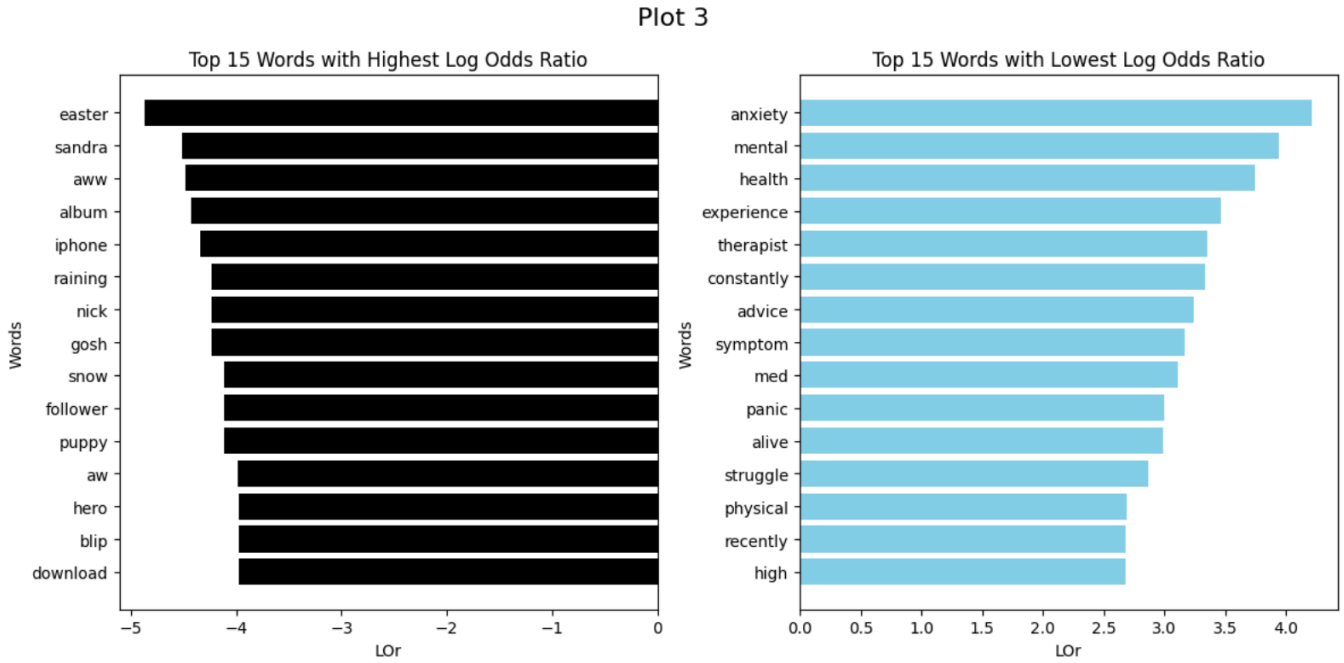

Figure 3. LOr of content words non-depressed vs depressed – Reddit

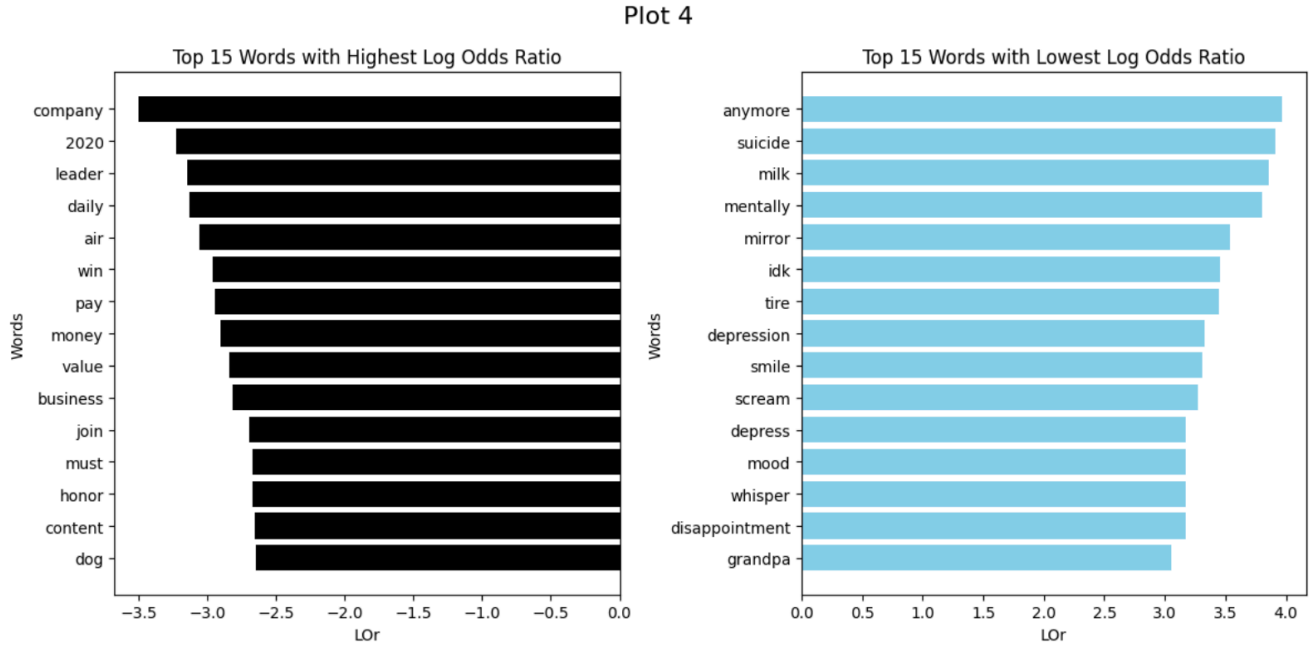

Figure 4. LOr of content words non-depressed vs depressed – Twitter