Sentiment Analysis of Poverty Reddits

William W. Winters

Anderson College of Business and Computing

Regis University

MSDS 640: Ethics, Privacy, and Social Justice in Data Science

Dr. Ghulam Mujtaba

August 14, 2025

Sentiment Analysis of Poverty Reddits

Introduction

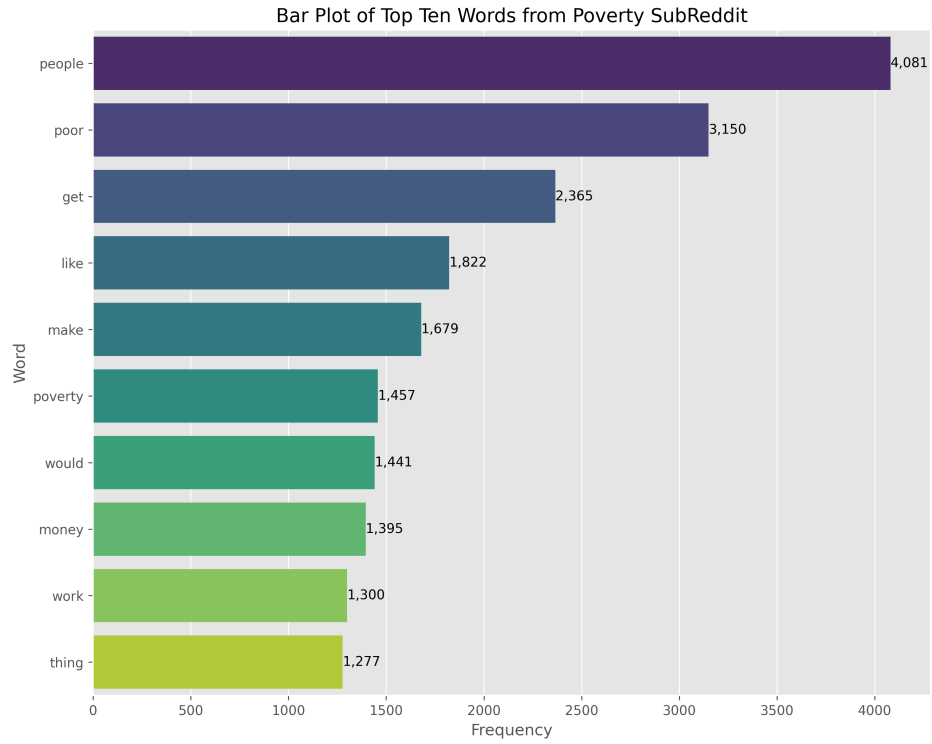
According to the U.S. Census Bureau (2025), there are 36.8 million people in the United States living at or below the poverty level. Poverty and homelessness are problems that require a comprehensive solution to solve. From a political perspective, some individuals believe that people who are poor or homeless are responsible for their circumstances, attributing their situation to poor decision-making and personal failures, rather than acknowledging potential external factors. This study will data mine posts from the poverty subReddit on the social media platform Reddit to determine what perceptions people have about poverty in the United States. These posts will be examined using natural language processing (NLP) techniques to determine each posts sentiment using the Valence Aware Dictionary and sEntiment Reasoner (VADER) algorithm. VADER not only provides a sentiment score of negative, neutral, and positive, it also indicates the intensity of the sentiment (nltk.org, 2024).

Method

Using the Reddit API, a dataset consisting of approximately 1000 posts and their comments were download into a sqlite3 database. The database was then queried to produce the dataframe used in this analysis. After data cleaning and other preprocessing steps, the final dataset consisted of 8,110 rows and 11 features. Text features related to the sentiment analysis being conducted were consolidated into a single feature and NLP text preparation techniques were applied. Exploratory Data Analysis (EDA) was conducted on the dataframe to include the top ten word counts and other processes to assess the usefulness of the data. See Figure 1 below.

Figure 1

Top Ten Words from Poverty SubReddit

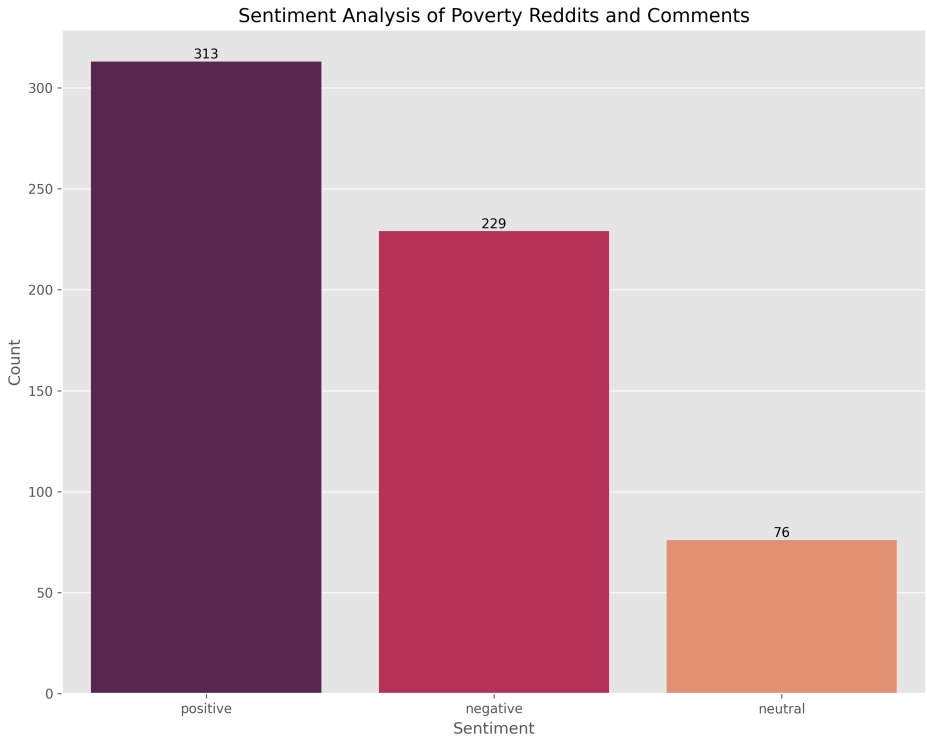


In some instances rows were dropped and in others columns were dropped or combined. After EDA and some basic data cleaning, the text to be analyzed was converted to lower case, numbers removed, punctuation removed, and extra spaces were handled. All words except English stop words were lemmatized to prepare it for sentiment analysis.

VADER was applied to the processed text and a sentiment score for reach post was obtained. The sentiment scores were limited to negative, neutral, and positive, although VADER can also provide the intensity of the sentiment. The positive sentiment was the majority class, followed by negative and neutral.

Figure 2

Sentiment Analysis Counts



Discussion

Many communities are affected by poverty and it is a complex issue with no easy answers. Generally, it is believed that the effects of poverty could be reduced or minimized; yet, there is debate on whether it can be totally eliminated. There are many causes of poverty in the U.S., but many agree that lower levels of education, lower levels of health, lack of access to affordable healthy food, and living in high-crime areas all contribute to individuals becoming impoverished. While poverty is a complex issue with several causes, the only clear method of exiting it, is a quality education according to Tackie (2021).

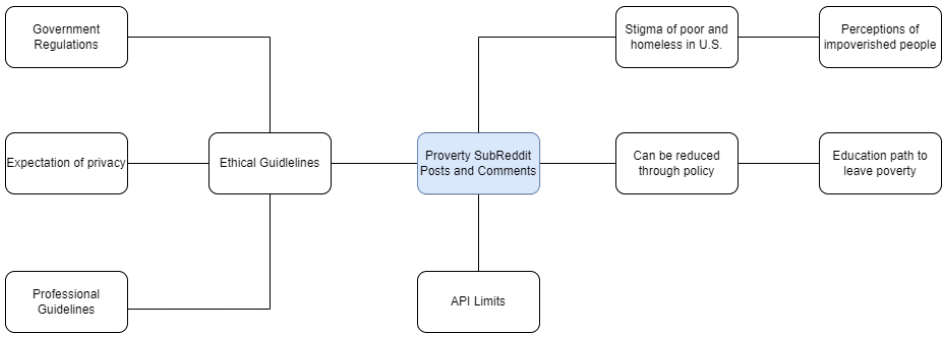
One of the problems with poverty in the U.S. is people have different perceptions of the issue. There is a large amount of stigma attached to being poor or destitute and this makes it difficult to look at it as a social problem. The extent to which the impoverished population is stigmatized reflects on the assumptions about the causes and possible policy

solutions to the problem (Bowen & Capozziello, 2024). This leads me to the main objective of my analysis: to apply sentiment analysis to Reddit posts about poverty in order to determine whether the Reddit community tends to view poverty in a negative or positive light.

As I engage in data mining of social media, I must consider the ethical implications of collecting and analyzing data from individuals who participate in online activities in publicly accessible spaces. One question I must ask myself is: do individuals who share information in these "public" online spaces have a reasonable expectation of privacy? This concern has sparked a growing debate among researchers, prompting various professional organizations to establish ethical guidelines for the use of social media data in research. While regulations vary across countries, some common trends have emerged, including the development of regulations governing internet-based research, aimed at balancing the requirement for data-driven insights with the need to protect individuals' privacy and rights (Fiesler et al., 2024). The mind map below ties all of these concepts together.

Figure 3

Poverty SubReddit Mind Map



Findings and Conclusions

The imbalance between positive, negative, and neutral sentiment classifications on the text data indicates poverty is a polarizing issue and many individuals in the Reddit community either post positively or negatively on the subject with not many being neutral. In addition, the frequency of each sentiment was charted with its associated Reddit score

range. When a Reddit is posted, the community will either up-vote or down-vote it. These votes are used to give the post a Reddit score. The goal is to visualize any correlation between a Reddit score and the sentiments in the post. The visualization was inconclusive and no meaningful correlation was found; however, the Reddits that scored between 0 and 500, contained the largest numbers of comments. Although the sentiment groups are imbalanced, most of the responses are positive.

The literature review indicates poverty is a world-wide problem and there is no clear solution on how to solve it. Tackie (2021) listed a quality education as a clear path out of poverty, but since it is a world-wide issue, education may not be the only answer. Furthermore, ethical questions about the expectation of privacy were brought up and discussed by Fiesler et al. (2024). According to his paper, this is an ethical dilemma with no clear answer at this time.

References

Bowen, E., & Capozziello, N. (2024, May). Faceless, Nameless, Invisible: A Visual Content Analysis of Photographs in U.S. Media Coverage about Homelessness. *Housing Studies*(3), 746-765. doi: 10.1080/02673037.2022.2084044

Fiesler, C., Zimmer, M., Proferes, N., Gilbert, S., & Jones, N. (2024, Jan). Remember the Human: A Systematic Review of Ethical Considerations in Reddit Research. *ACM*, *8*(GROUP). doi: 10.1145/3633070

nltk.org. (2024, Aug). nltk.sentiment.vader module. Retrieved from `https://www.nltk.org/api/nltk.sentiment.vader.html`

Tackie, D. (2021). An Examination of Poverty: Dimensions, Causes, and Solutions. *Journal of Rural Sciences*, *36*(2), 1-25. Retrieved from `https://egrove.olemiss.edu/jrss/vol36/iss2/2/`

U.S. Census Bureau. (2025, Jan). National Proverty in America Awareness Month: January 2025. Retrieved from `https://www.census.gov/newsroom/stories/poverty-awareness-month.html`