**Name:** Sri Varsha Naharajan
**Semester:** Fall 2024

**Sentiment Analysis of Popular Diets in Health-Related Reddit Communities (1470 words)**

## 1. Introduction:

The popularity of diets like keto, vegan, and paleo has led to widespread discussions on social media regarding their health benefits, sustainability, and overall lifestyle impacts, particularly within Reddit's health and nutrition communities. These communities offer a space where people share stories, trade advice, and debate the benefits and challenges of different dietary choices. Social media has become a vital source of public sentiment on health topics, with sentiment analysis helping to reveal the feelings people attach to various diets.Farzindar and Inkpen (2015) explore how sentiment and emotion analysis techniques on social media reveal public perspectives on food and health choices. Similarly, Tan, Lee, and Lim (2023) review sentiment analysis methods, such as machine learning and rule-based approaches, showing how they capture attitudes toward lifestyle trends in digital spaces. Choudhury and De (2014) show that Reddit's health communities provide rich, firsthand insights into dietary habits and health experiences, often linking diet discussions to broader themes like lifestyle and identity—emphasizing the importance of analyzing sentiment in these spaces. While these studies have broadly explored social media sentiment or specific health-related discussions, little attention has been paid to how specific communities discuss and perceive these dietary approaches. This study will explore how Reddit users feel about keto, vegan, and paleo diets, giving insight into community values and attitudes surrounding popular diets.

## 2. Research Question:

How do Reddit users within health-focused communities express their sentiments toward popular diets, specifically keto, vegan, and paleo?

## 3. Method:

### 3.1 Data

The dataset for this study consists of 522 comments collected from four relevant subreddits (r/keto, r/vegan, r/paleo, and r/nutrition), spanning 10 posts to ensure coverage of discussions around the keto, vegan, and paleo diets. All posts were selected from

within the past two years to maintain relevance. For the keto diet, two posts from r/keto were chosen using the keyword "keto": "Do you actually enjoy keto?" (68 comments) and "Why does keto get so much shit? I've never felt better" (81 comments). These posts were selected due to their high engagement, each having over 300 comments. For the vegan diet, two posts from r/vegan were selected using the keyword "vegan": "Being like 90% Vegan is embarrassingly easy, what gives?" (67 comments) and "Please stop mocking new vegans or vegan-curious people on here." (60 comments), both featuring over 200 comments and 500+ upvotes. For the paleo diet, three posts from r/paleo were identified using the keyword "paleo": "Are you still paleo?" (29 comments), "Paleo in 2023" (31 comments), and "My thoughts on Paleo" (22 comments), all fitting within the timeframe and deemed relevant. Recognizing the potential for bias inherent in dedicated subreddits, additional comments were collected from r/nutrition to provide a more balanced perspective. In r/nutrition, three posts were selected using the keywords "keto," "vegan," and "paleo": "Being Keto and then reading a book by Michael Greger has me confused on what to believe" (77 comments), "Does going vegan help your body?" (76 comments), and "Will Paleo be considered mostly plant-based eventually?" (12 comments). Keywords were not case-sensitive, and data was collected using the Python Reddit API Wrapper (PRAW), focusing on top-level comments. Deleted comments and those from deleted users were excluded to ensure data quality. The difference in comment volume for keto, vegan, and paleo reflects the varying popularity and engagement levels of these diets within Reddit communities. This approach ensures a comprehensive and balanced dataset for sentiment analysis across the three diets.

### 3.2 Analysis

This study used VADER from the vaderSentiment library for sentiment analysis and Latent Dirichlet Allocation (LDA) for topic modeling. VADER effectively categorized comments as positive, negative, or neutral, providing sentiment scores from -1 (most negative) to 1 (most positive), offering a nuanced measure of sentiment intensity, while LDA identified recurring themes in the discussions.

**Data Preprocessing and Preparation**: To ensure consistency and efficiency, comprehensive text preprocessing was performed at the beginning of the analysis using the pandas and nltk libraries. The dataset included key columns: 'Date,' 'Text,' and 'Keyword,' which were essential for both sentiment and topic analysis. Text preprocessing involved converting the 'Text' column to lowercase, removing punctuation, and splitting the text into individual words (tokens). Non-informative stopwords, as well as any non-alphabetic tokens (e.g., numbers and symbols), were removed. The cleaned tokens

were then joined back into a single string, resulting in a uniform text format suitable for analysis. This preprocessing step ensured that the text data was suitable for both sentiment analysis and topic modeling.

**Sentiment Analysis:** The VADER Sentiment Intensity Analyzer was then applied to the cleaned text data. Each comment received a compound score, which was used to classify the sentiment as "Positive" (score $\geq 0.05$), "Negative" (score $\leq -0.05$), or "Neutral" (score between -0.05 and 0.05). These thresholds were chosen based on VADER's validated methodology, which is optimized for distinguishing subtle sentiments in social media content. The results were recorded in new columns: 'sentiment_score' for the numerical sentiment score and 'sentiment' for the sentiment category. The sentiment column was added to the DataFrame, and a grouped bar chart was used to visualize the overall sentiment distribution across different diet keywords, effectively illustrating the prevalence of positive, negative, and neutral sentiments for each diet category.

**Topic Modeling:** Following sentiment analysis, Latent Dirichlet Allocation (LDA) was applied to identify recurring themes in the discussions. Using the sklearn library, the cleaned text data was vectorized with bi-grams using CountVectorizer, allowing the model to capture both individual words and meaningful word pairs. LDA then extracted key topics within each diet-related discussion, revealing themes such as health benefits, challenges, and lifestyle impacts. By applying LDA to each diet keyword separately, the analysis provided an understanding of diet-specific themes.

This structured approach, combining efficient text preprocessing, sentiment analysis, visualization, and topic modeling, ensured a thorough and insightful examination of how diets are discussed and perceived on social media.

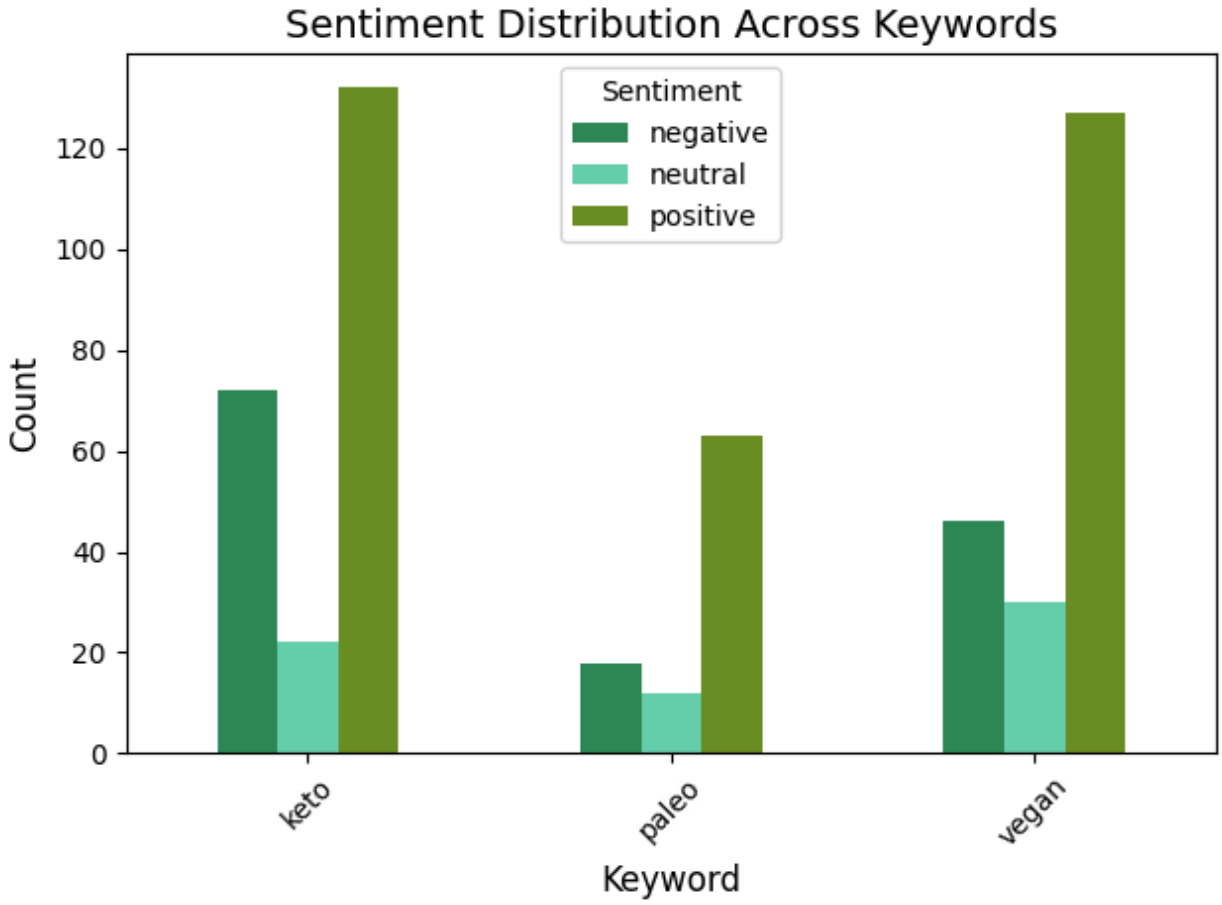
## 4. Results:

The sentiment analysis results are summarized in Table 1. Among the diets analyzed, the keto diet generated the highest number of comments, with 58.4% expressing positive sentiment, 31.9% negative, and 9.7% neutral. This indicates a mix of enthusiasm for benefits like weight loss and health improvements, alongside concerns about challenges. Paleo discussions had the most positive sentiment distribution (67.7%), with only 19.4% negative comments, suggesting high satisfaction with its ancestral health approach. Vegan discussions also leaned positive (62.6%) but had a higher proportion of negative sentiment (22.7%), reflecting challenges like meeting nutritional needs.

**Table 1: Sentiment Analysis Results**

| Keyword | Negative | Neutral | Positive | Total |
|---|---|---|---|---|
| Keto | 72 (31.9%) | 22 (9.7%) | 132 (58.4%) | 226 |
| Paleo | 18 (19.4%) | 12 (12.9%) | 63 (67.7%) | 93 |
| Vegan | 46 (22.7%) | 30 (14.8%) | 127(62.6%) | 203 |

**Figure 1: Sentiment Distribution Across Keywords**




The LDA topic modeling identified recurring themes across the diets (Tables 2, 3, and 4). Keto discussions highlighted scientific aspects, including carbohydrate management, insulin response, and sugar reduction, alongside lifestyle adjustments like reducing processed foods and managing dietary fat. Vegan discussions focused on plant-based eating, ethical concerns, and protein adequacy, with practical themes such as the availability of vegan alternatives and the ease or difficulty of adopting the diet. Paleo

discussions emphasized ancestral eating habits, community experiences, and dietary restrictions like avoiding grains and gluten, with debates on the scientific validity of the diet and its comparison to keto.

**Table 2: Topics for Keto Discussions**

| Topic | Keywords |
|---|---|
| 1 | weight, good, meat, fat, healthy, carbs, keto |
| 2 | way, carb, think, fat, foods, don’t, keto |
| 3 | foods, better, health, carbs, feel, sugar, keto |
| 4 | glucose, think, weight, high, health, good, keto |
| 5 | diets, lot, processed, don’t, foods, keto, carbs |

**Table 3: Topics for Vegan Discussions**

| Topic | Keywords |
|---|---|
| 1 | going, lot, plant-based, plant, meat, based, vegan |
| 2 | foods, easy, healthy, way, want, don’t, vegan |
| 3 | protein, animal products, time, foods, products, vegan |
| 4 | lot, meat, diets, foods, easy, vegans, vegan |
| 5 | got, animal, meat, don’t, make, vegans, vegan |

**Table 4: Topics for Paleo Discussions**

| Topic | Keywords |
|---|---|
| 1 | potatoes, paleo years, crossfit, right, foods, paleo |
| 2 | need, AIP, want, foods, I’m, keto, paleo |
| 3 | way, feel, meat, paleolithic, strict, ancestors, paleo |
| 4 | know, post, I’ve, I’m, vegan, paleo, years |

| | |
|---|---|
| 5 | strict, I’ve, keto, lot, I’m, think, paleo |

These results provide a nuanced understanding of how Reddit users discuss and perceive these diets, revealing common themes around health, ethics, lifestyle challenges, and community engagement for each dietary approach.

## 5. Conclusion and Limitations

The analysis revealed that Reddit users in health-focused communities express varied sentiments toward the keto, vegan, and paleo diets, with distinct themes emerging for each. The sentiment analysis showed that discussions around all three diets are largely positive, with keto and vegan diets having some notable negative feedback related to dietary challenges. Keto discussions often focus on scientific aspects, like carbohydrate management and the physical effects of the diet, reflecting a strong interest in health benefits and the complexities of maintaining a low-carb lifestyle. Vegan discussions emphasize ethical considerations and nutritional concerns, particularly around protein adequacy, while also addressing the practicality of a plant-based diet. Paleo discussions are characterized by personal experiences and an emphasis on ancestral eating habits, with users frequently sharing tips and debating the diet’s scientific foundations. These findings help answer the research question by providing a nuanced view of how users express their attitudes and concerns, shaped by both health motivations and lifestyle impacts.

Despite the insights gained, this study has several limitations. The analysis was restricted to Reddit, which may not represent broader public opinions on these diets. Additionally, the use of VADER for sentiment analysis, while effective for social media text, may not capture more complex or nuanced expressions of sentiment. The topic modeling approach, while useful for identifying common themes, may overlook less frequent but potentially significant topics. Future research could explore other social media platforms or employ more sophisticated sentiment analysis techniques to gain a more comprehensive understanding of public perceptions.

## 6. References:

Farzindar, A., & Inkpen, D. (2015). The use of sentiment and emotion analysis and data science to assess the language of nutrition-, food-, and cooking-related content on social media: A systematic scoping review. *Nutrition Research Reviews*, 28(1), 1–20. https://doi.org/10.1017/S095442241500017X

Choudhury, M. D., & De, S. (2014). Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM)*, 71–80. https://ojs.aaai.org/index.php/ICWSM/article/view/14526

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences, 13*(7), 4550. https://doi.org/10.3390/app13074550