Name: Sri Varsha Naharajan
Semester: Fall 2024

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

Research Question:

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

Background:

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. Although previous studies have looked broadly at social media attitudes toward diets, 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.

Data:

For this study, I will collect a dataset of user comments from Reddit, specifically focusing on discussions about keto, vegan, and paleo diets within health-related subreddits, such as r/nutrition, r/keto, r/vegan, r/Paleo, r/loseit, and r/health. Through the Python Reddit API wrapper (PRAW) , I will gather at least 500 comments, aiming for a roughly equal distribution across each diet category to ensure balanced representation. Comments will be retrieved from posts containing keywords like "keto," "ketogenic," "low-carb," "vegan," "plant-based," "paleo," and "whole foods," with keywords not case-sensitive to capture relevant discussions across each diet type. Only top-level comments and direct first-level replies that respond directly to the respective posts will be included, while any deleted or removed comments will be excluded. The dataset will focus on posts from within the past year to ensure that the sentiment analysis reflects current perspectives on these diets within Reddit's health-focused communities.

Method:

To analyze the Reddit comments, I will first preprocess the data by removing special characters, URLs, emojis, and stop words to ensure only relevant text remains. I will then use VADER sentiment analysis, a lexicon-based tool designed for social media, to assign each comment a compound score from -1 to +1, categorizing it as "Positive" (≥ 0.05), "Negative" (≤ -0.05), or "Neutral" (between -0.05 and 0.05). This categorization enables a comparative analysis of sentiment across keto, vegan, and paleo diet

discussions. Following this, I will apply Latent Dirichlet Allocation (LDA) to identify recurring themes within each diet's discussions, revealing topics like health benefits, challenges, or lifestyle impacts. Together, VADER and LDA will uncover both the emotional tone and thematic content of discussions within Reddit's health-focused communities.

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

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