**Project 1**

**STQD6114 – Unstructured Data Analytics**

**P152419 – Hazim Fitri Bin Ahmad Faudzi**

**Part 1 – Task 3**

**INTRODUCTION**

A publicly available reviews dataset that has been chosen for this task is the Amazon Fine Food Reviews from Kaggle. This dataset consists of 500,000 consumer reviews of food products purchased on Amazon.

The Amazon Fine Food Reviews dataset comprises more than five hundred thousand genuine customer comments on food items purchased through the Amazon platform. Each entry records the written feedback, a star rating from one to five, the product’s name, and the date of submission. Such a wealth of unstructured textual data offers a window into consumer attitudes and experiences. To translate raw text into measurable sentiment, three established lexicons were employed. The first, AFINN, assigns each word an integer value ranging from negative five for very negative sentiment up to positive five for very positive sentiment. The second, Bing Liu’s lexicon, classifies each term simply as either positive or negative. The third resource, the NRC emotion lexicon, associates words with eight basic feelings—anger, anticipation, disgust, fear, joy, sadness, surprise and trust—as well as with overall positive or negative polarity.

**DISCUSSION**

Prior to analysis, all review texts underwent rigorous preprocessing. Each comment was tokenized into individual words, converted to lowercase to ensure consistent matching with lexicon entries, and stripped of common words such as “the,” “and” or “to,” which carry minimal emotional weight. Numbers and punctuation marks were also removed so that only meaningful vocabulary remained. This cleaning ensured that subsequent sentiment scoring would reflect genuine expressions of opinion rather than artifacts of formatting or grammar.

The Bing Liu lexicon was first applied to gauge overall positive and negative tone. Positive words were counted across the entire corpus and compared to the total of negative words. Results showed that expressions of approval outnumbered expressions of disapproval by nearly three to one, indicating a strong predominance of favorable language. Such a distribution suggests that most customers tend to leave positive feedback when satisfied with a purchase.

To capture the intensity of sentiment, the AFINN lexicon was next employed. Scores for each word in a review were summed to produce a single value per comment, which could range anywhere from very negative to very positive on the negative five to positive five scale. The average score across all reviews settled just below positive two, reflecting a generally mild to moderate level of positivity. The inclusion of numerical grading allowed for the identification of outlier reviews exhibiting exceptionally strong sentiment, whether adverse or laudatory.

Further insight was obtained by applying the NRC emotion lexicon. Counts of words linked to each emotion revealed that trust-related words appeared most frequently, followed closely by words of joy and words expressing anticipation. Emotions such as sadness, fear, anger and disgust were far less common. When these raw counts were converted into percentage terms, trust accounted for roughly twenty percent of all emotion-bearing words, joy for about nineteen percent and anticipation for around seventeen percent. Negative emotions each comprised less than seven percent of the total.

These findings align with expectations for consumer feedback on a commercial platform. Shoppers who feel confident in the products they purchase often describe their experiences with language of trust and satisfaction. Expressions of joy appear when products exceed expectations. Fewer customers use forceful negative language unless a significant issue has been encountered. Nevertheless, the identification of reviews featuring strong negative scores or frequent appearances of anger and disgust can highlight serious problems requiring immediate attention.

Each lexicon offers distinct advantages. The simple positive versus negative classification delivers a rapid overview of general mood. Numerical scoring provides an estimated strength of sentiment, enabling prioritization of the most extreme feedback. Emotion tagging reveals the precise nature of feelings, which supports targeted responses such as reinforcing trust or addressing customer concerns. Together, these complementary methods create a comprehensive framework for understanding customer sentiment.

A graph with different colored bars

AI-generated content may be incorrect.

**CONCLUSION**

In conclusion, the comparative analysis of the Amazon Fine Food Reviews dataset through multiple lexicons demonstrates the value of a multi-faceted approach. By combining count-based polarity measures, intensity scores and nuanced emotion categories, it becomes possible to derive actionable insights. Companies may thereby focus efforts on improving products that inspire fear or disgust, while celebrating and leveraging feedback that fosters trust and joy. Such balanced analysis enhances decision making and ultimately supports better customer satisfaction and loyalty.