**Dataset Preparation**

Any research project involving any task involving natural language processing must begin with the preparation of the dataset. The dataset used in this study was collected from the Amazon Product Reviews dataset available on Amazon Web Services (AWS) Public Dataset [9]. The dataset contains product reviews from multiple categories; four categories were selected for this study: Health and Personal Care, Personal Appliances, Gift Cards, and Beauty. To narrow down the scope of this study, only the reviews that contained emoticons were considered. This decision was made to explore the impact of the use of emoticons in product reviews on sentiment analysis and emotion detection.This process assisted us in honing our dataset and ensuring that we were looking at the most pertinent information for our study. Overall, ensuring the accuracy and applicability of our findings depended heavily on the dataset preparation process.

In this section, we describe the important steps taken during the data preparation phase, which includes data collection, data preprocessing, data annotation, and generating final labels using a combination approach.

**Data Preprocessing:**

In natural language processing activities like sentiment analysis, text preparation is a vital stage. In this work, we cleaned and normalized the raw text data gathered from the Amazon Product Reviews dataset using a number of text preparation approaches. Python was used to implement the preprocessing processes, together with its NLTK, spaCy, emoji, and scikit-learn packages.

The text has been lemmatized, user mentions and URLs have been removed, emojis have been compressed to a single word, punctuation and digits have been eliminated, all text has been converted to lowercase, the HTML tag has been removed, stop words have been eliminated, and the text has been removed from punctuation. The final product is a cleaned-up version of the original text, which may be utilized as input into a text classification model for data annotation.

**Data Annotation:**

The process of data annotation is a crucial step in machine learning tasks, as it involves the labeling of data with meaningful categories or classes. In the context of sentiment analysis of reviews, accurate labeling of sentiment categories is essential to build effective predictive models. However, the task of manual annotation of large amounts of data can be expensive and time-consuming. To address this challenge, we explored and applied three different approaches for labeling the sentiment of reviews in our dataset: polarity score-based label, active learning-based label, and emoji-based label.

A typical strategy is the polarity score-based label approach, which determines the sentiment of a text based on the polarity score of the words it includes. Although this method has the advantage of being simple and automated, it might not always be able to capture the subtleties and context-dependent aspects of sentiment expression. We also used active learning-based labelling, which includes iteratively choosing the most informative samples for labelling and included them in the labelled dataset, to overcome this constraint. With this strategy, the model can gain knowledge from the most pertinent samples and enhance its performance over time.

Finally, we used an emoji-based categorization strategy, which entailed categorizing reviews into five groups based on the presence of certain emoticons in the review text: Efficacy, Satisfaction, Uncertainty, Dissatisfaction, and Side Effects. This method has the benefit of being straightforward and simple to understand, but it also necessitates a substantial amount of human work to find the pertinent emoticons and categorize the reviews appropriately.

We used an ensemble learning strategy that integrates the results of the three labelling algorithms to provide a more robust and reliable sentiment labelling. By taking use of the advantages of each distinct methodology and making up for its shortcomings, this method has the potential to increase the precision and generalizability of the emotion labels.

**Polarity Score-Based Label:**

We first employed a polarity score-based method for labelling the gathered reviews. We calculated the polarity score for each review using the VADER (Valence Aware Dictionary and Sentiment Reasoner) tool. We divided the reviews into one of five categories based on the polarity score: efficacy, contentment, uncertainty, dissatisfaction, or side effects. Reviews with a high polarity score were classified as efficacious or satisfying, whereas those with a low polarity score were classified as unsatisfactory or having undesirable side effects. Uncertain was the designation given to reviews with a neutral polarity rating. The reviews could be labelled quickly and easily using this method, but it wasn't always able to fully capture the subtleties of the reviews. Consequently, we used additional techniques for further

**Active Learning-Based Label:**

We split the data into labelled (star rating score) and unlabeled sets in order to do this, and we then utilized logistic regression, random forest, and gradient boosting classifiers from the scikit-learn package to train and test the models. To be more specific, we vectorized the labelled data using TF-IDF vectorization, initialized an active learner, and chose the most ambiguous samples to label using the uncertainty sampling technique. We then labelled the selected samples, added them to the labelled set, and calculated the uncertainty scores for the selected samples. We repeated this procedure in other contexts. The model was then trained using this freshly labelled data after the labelled data had been vectorized. The test data and the remaining unlabeled data had their labels predicted using the trained model. The summary of accuracy for different settings is shown in table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Setting** | **Model** | **Iterations** | **Instances** | **Accuracy** |
| 1 | Logistic Regression | 20 | 50 | 81% |
| Random Forest | 82% |
| Gradient Boosting | 82% |
| 2 | Logistic Regression | 40 | 100 | 82% |
| Random Forest | 83% |
| Gradient Boosting | 84% |
| 3 | Logistic Regression | 100 | 160 | 83% |
| Random Forest | 83% |
| Gradient Boosting | 85% |

Table 2. Different Settings for Active Learning Method

**Emoji-Based Labels:**

The third approach we used for labeling the reviews was emoji-based. We classified the reviews into five categories: Efficacy, Satisfaction, Uncertain, Dissatisfaction, and Side Effect. We looped the five classes and the matching emoji assigned as label corresponding to its respective list. We encountered many contradictory emojis, after carefully analyzing against review text decision were made. Few examples are: “🌵” assigned to efficacy because corresponding reviews are very positive, “👀”, “💵”, “🙀”, “💧”,” 🏽”, “🙉”, “⁉️”, “💦”, “🐒”, “💈”, “💰”, “😮”, “😷” has large number of satisfaction reviews, “😼” has large number of dissatisfaction reviews. The table 2 shows common emojis against respective categories.

|  |  |
| --- | --- |
| **Classes** | **Emojis** |
| Efficacy | 💗, 😍, ✅, 💛. 😘, 💜, 😊, 💕, ⭐, 💚 |
| Satisfaction | 🙌, 🌹, 👀, 👌, 🆗, 🌼, 👍, 😃, 💪, 🙏 |
| Uncertain | 😑, 🐝, 💻, 🔱, 🚩, ➕, 🍭, 💇, 🤡, 🔹, 💳 |
| Dissatisfaction | 😪, 😕, 😰, 😬, 😶, 😨, 👎, 💥, 😭, ⚠ |
| Side Effect | 💩, 🌚, 😡, 🤢, 💔, ✖, ❎, 👿, ⛔, 💀, ❌, 😡 |

Table 3 Most Common Emojis Against Each Label

**Final Labels by Using Combination Method:**

The goal of utilizing a combined strategy was to increase the accuracy and dependability of the results while overcoming the drawbacks of individual labelling technique. Active learning and polarity score-based labelling are efficient techniques, but they can be constrained by the caliber and volume of the data. A more recent technique that has had encouraging results is labelling using emojis. A popular and well-established technique for merging numerous labels is the voting-based approach. With this method, each label is given a vote, and the label with the most votes is chosen as the final label. Since only the labels themselves are needed, the voting-based approach has several benefits, including simplicity, ease of implementation, and low computational cost.

In the context of the present study, a voting-based approach was employed to combine three labels - 'Polarity Label', 'Active Labels', and 'Emoji Label' - into a single label. This approach was selected for its simplicity and ease of implementation, as well as its ability to provide an efficient means of label combination without the need for additional resources. The voting-based approach provides a simple and cost-effective means of label combination, and is suitable for a wide range of applications in natural language processing and machine learning.