**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 helped refine the dataset, ensuring focus on the most relevant information for the study. Overall, ensuring the accuracy and applicability of our findings depended heavily on the dataset preparation process.

This section 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, data 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:**

A crucial stage in machine learning tasks is data annotation, which is labelling data with meaningful groups or classifications. Precise sentiment category labelling is required for sentiment analysis of reviews in order to build reliable predictive models. On the other hand, extensive manual data annotation can be expensive and time-consuming. This task was tackled using three different approaches: polarity score-based labelling for the sentiment of reviews in our dataset, emoji-based labelling, and active learning-based labelling.

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. Active learning-based labeling was also used, which involves iteratively choosing the most informative samples for labeling and including them in the labeled dataset, to overcome this constraint. With this strategy, the model can gain knowledge from the most pertinent samples and enhance its performance over time.

Lastly, an emoji-based categorization method was used, which entailed classifying assessments into five groups according to the presence of particular emoticons in the review text: Efficacy, Satisfaction, Uncertainty, Dissatisfaction, and Side Effects. This method has the benefit of being simple to understand and straightforward, but it also takes a lot of human labor to find the appropriate emoticons and categories the assessments.

In order to produce a more robust and trustworthy sentiment labelling, an ensemble learning technique was employed, combining the output of three labelling algorithms. This strategy can improve the accuracy and generalizability of the emotion labels by using the benefits of each unique methodology and compensating for its drawbacks.

**Polarity Score-Based Label:**

The collected reviews were initially labelled using a polarity score-based approach. The VADER (Valence Aware Dictionary and Sentiment Reasoner) tool was used to determine the polarity score for each review. Based on the polarity score, reviews were categorized into one of five groups: efficacy, contentment, uncertainty, dissatisfaction, or side effects. Reviews were categorized as effective or fulfilling based on their high polarity score, and unsatisfactory or having unwanted side effects based on their low polarity score. Reviews with a neutral polarity rating were classified as uncertain. This approach made it quick and simple to label the reviews, but it wasn't always able to properly capture their intricacies.

**Active Learning-Based Label:**

The data was split into labeled (star rating score) and unlabeled sets to facilitate this process. Logistic regression, random forest, and gradient boosting classifiers from the scikit-learn package were then utilized to train and test the models. More specifically, the labeled data was vectorized using TF-IDF vectorization, an active learner was initialized, and the most ambiguous samples were chosen for labeling using the uncertainty sampling technique. The selected samples were then labeled, added to the labeled set, and uncertainty scores for the selected samples were calculated. This procedure was repeated in other contexts. The model was then trained using this freshly labeled data after vectorization of the labeled data. 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 for labeling the reviews was emoji-based, where reviews were classified into five categories: Efficacy, Satisfaction, Uncertain, Dissatisfaction, and Side Effect. The process involved looping through the five classes and assigning the matching emoji as a label corresponding to its respective list. Many contradictory emojis were encountered, and decisions were made after carefully analyzing them against the review text. 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 objective of employing a combination approach was to overcome the limitations of individual labelling technique while improving the accuracy and reliability of the results. While polarity score-based labelling and active learning are effective methods, their use may be limited by the quantity and quality of the available data. Emoji labelling is a more modern method that has shown promising outcomes. One well-liked and proven method for combining many labels is the voting-based method. This method assigns a vote to each label; the label that receives the most votes becomes the final label. The voting-based technique has various advantages, such as simplicity, cheap computational cost, and ease of implementation—only the labels themselves are required.

In the current study, three labels—"Polarity Label," "Active Labels," and "Emoji Label"—were combined into one label using a voting-based method. This method was chosen due to its simplicity, convenience of use, and capacity to offer an effective way to combine labels without requiring extra resources. Voting-based label combination offers an easy and affordable way, and it may be applied to many different natural language processing and machine learning applications.