Mid-Point Project Report

Jefford Secondes, C0865112

Jovi Fez Bartolata, C0869701

Keyvan Amini, C0866360

Lakpa Sherpa, C0874811

Maricris Resma, C0872252

Luz Zapanta, C0879190

Lucia Martinez Torres, C0863954

Pradeep Kumar Bhatt, C0875486

AML 2304: Natural Language Processing

Prof. Bhavik Gandhi

November 14, 2023

# 

# Abstract

This report presents a detailed sentiment analysis of Amazon product reviews, utilizing advanced natural language processing techniques to extract meaningful insights from customer feedback. The project began with a thorough data preprocessing phase, including text cleaning and lemmatization, to ensure the data's suitability for analysis. In the exploratory data analysis (EDA) stage, we conducted a comprehensive review of the data, examining word counts and employing word cloud visualizations to identify prevalent themes and expressions in customer reviews. A key focus of the project was the application of sophisticated feature extraction methods, notably Word2Vec and TF-IDF, which facilitated the transformation of text data into a numerical format for subsequent analysis. The core of our approach involved the use of Principal Component Analysis (PCA) for dimensionality reduction, followed by the development of a Logistic Regression model to classify reviews into various sentiment categories. The model demonstrated a notable accuracy, effectively categorizing reviews into negative, neutral, and positive sentiments. The evaluation of the model's performance was detailed through precision, recall, and F1-score metrics for each category, providing a comprehensive understanding of its efficacy in sentiment classification. The findings of this analysis not only underscore the power of machine learning in extracting customer sentiments but also offer valuable insights for businesses looking to enhance customer experience and service strategies.

## Project Overview:

### 1. Industry

The e-commerce industry, particularly the online retail sector, operates on the foundation of digital transactions and customer interactions. Sentiment analysis plays a crucial role in this domain, offering insights into customer preferences, satisfaction, and overall market trends. The industry's dynamics can be understood through several key aspects:Central Role of Customer Feedback: Customer reviews are vital in e-commerce. They influence purchasing decisions and provide businesses with insights for product and service improvement.

Data-Driven Decision Making: E-commerce relies on data analytics, with sentiment analysis forming a critical part of understanding customer preferences and experiences.

* Adaptation and Personalization: Rapid response to customer feedback is key. Sentiment analysis helps in personalizing offerings and improving user experience.

#### Simplified Flow of Sentiment Analysis in E-commerce:

* Collection: Gathering customer reviews from product pages, social media, and feedback forms.
* Preprocessing: Cleaning and normalizing data for analysis.
* Analysis: Employing NLP techniques like Word2Vec and TF-IDF to determine sentiments (positive, negative, neutral).
* Insight Generation: Producing actionable insights from analysis for strategic use.
* Business Actions: Implementing strategies based on insights for product enhancement and marketing.

### 2. Problem Statement

The project focuses on automating sentiment analysis of product reviews, a critical need in e-commerce for understanding and utilizing customer feedback effectively. Traditional manual methods of reviewing customer feedback are time-consuming and not scalable. This project introduces automated, objective, and scalable sentiment analysis using natural language processing, significantly enhancing efficiency and reducing subjectivity. By addressing these aspects, the project empowers companies to improve customer experience, guide product development, inform marketing strategies, and maintain a competitive edge by swiftly adapting to market and customer feedback trends.

### 3. Product

Our sentiment analysis project distinguishes itself from other market solutions through these key differentiators:

Advanced Natural Language Processing Techniques: We leverage cutting-edge NLP methods like Word2Vec and TF-IDF, combined with Logistic Regression, offering more nuanced sentiment analysis. This approach goes beyond basic sentiment detection, providing deeper insights into customer emotions and opinions.

Customization for E-commerce Reviews: Our tool is specifically tailored for the e-commerce domain, particularly Amazon reviews. This specialization ensures that the nuances of customer feedback in online retail are accurately captured and analyzed, which is often not the case with generic sentiment analysis tools.

User-Friendly Interface: A standout feature of our project is its easy-to-use and clear user interface, designed to make the analysis of complex sentiment data accessible and understandable even for users with minimal technical background. This ease of use ensures that businesses can quickly and effectively turn insights into action without the need for extensive training or technical expertise.

### 4. Project Board:

**A screenshot of a computer

Description automatically generated**

## 

## Data preprocessing

The initial step in the Natural Language Processing (NLP) project involves data preprocessing. This necessitates loading the data from the file into a data frame. The dataset comprises both constant words and comments, along with the associated values. Reading the file is essential, followed by the removal of constant words. Subsequently, the data is divided into two main sections: labels and comments.

A crucial action in this phase is the transformation of all data, encompassing both labels and comments. Labels are converted into two integer variables, 0 and 1, while all contents are transformed into strings.

In concluding this section, it is imperative to cleanse the data of any irregular words that could impact the performance of any model applied to the data. Given the time constraints of the project, a function is employed wherein regular expressions are defined separately. This allows other team members to easily add or remove them when fine-tuning the model. The cleaning process includes converting all text to lowercase, removing digits, handling double or more spaces, eliminating links (e.g., starting with http), handling punctuation, and excluding characters that do not contribute meaningful information to the analysis.

## EDA

Several important procedures were taken to analyze and understand the parts of the Amazon Customer Reviews dataset while it was being explored and summed up.

**Null Value Check** was done to make sure that the dataset was complete by checking for any missing values. This absence lowers the chance of biased or incomplete analysis, which makes the results more reliable.

**Average Word Count** calculation gave a number as a summary of the average length of a review in the dataset. This metric is a good starting point for getting a good idea of how long customer reviews are on average.

**Word Cloud Visualization and Word Frequency Count** were used to track how often each word was used in the reviews. This study finds patterns, trends, and common themes in customer feedback. It also gives us a better understanding of the language customers use and can be used as a base for other studies, like sentiment analysis.

**Word Count Distribution** shows how the word counts were spread out, which made it possible to find the average length of reviews. This method shows possible outliers or trends in the lengths of reviews, which leads to more research.

## 

## Feature extraction and word embedding

**Using Word2Vec**

The initial feature set was created using Word2Vec. It is a technique used to represent words as numerical vectors. For each token, the mean vector was computed. The output of Word2Vec has high dimension and cannot be visualized, therefore, the dimensionality of the data must be reduced. Using Principal Component Analysis, the first two principal components were obtained and used to plot the features. Based on the plot, class separation was not perfect and could still be improved because some classes overlapped in a given feature.

**Using TF-IDF**

The alternative method used for feature extraction uses TF-IDF. The resulting vector is the product of term frequency and inverse document frequency. It gives a higher score to the word that is rare in the corpus and common in a document. The series of clean text from the preprocessed data is considered as the corpus with each record pertaining to the document

TF-IDF is implemented using sklearn.feature\_extraction.text.TfidfVectorizer. Running this method is equivalent to running CountVectorizer which converts the corpus to a matrix of term counts and then running TfidfTransformer which transforms the matrix to a normalized tf-idf representation. Initially there are 239,919 term features when all parameters used were default. To improve it, parameter min\_df=.01 was applied in order to not include terms that appear in less than 1% of the dataset. The reasoning behind this is that if these terms appear that rare then it the more unlikely that they affect the sentiment of the whole document. After applying this, we end up with 649 features in the final vector to be used to fit the model. Based on the PCA plotting, visualization of the class separation shows it is also not separated very well due to overlapping. It is similar but slightly better than word2vec. We can therefore say that further improvement on data preprocessing is needed in the future.

## Model Building

The Logistic Regression approach was employed to train the model. Specifically, multiclass configuration was chosen since there were more than three variables in the target variable.

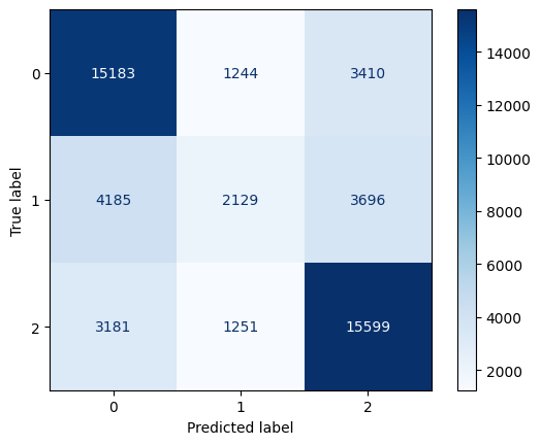
To avoid overfitting, regularization L2, or ridge, was applied.

This code is performing a logistic regression on 80% of the dataset that has been transformed using TF-IDF.

## 

## Model Evaluation:

The model was evaluated in terms of overall accuracy, precision, recall and F1 score. The Confusion Matrix below shows the resulting true and false predictions.



The table below gives an overview of the performance metrics:

| **Class​** | **Precision​** | **Recall​** | **F1-score​** |
| --- | --- | --- | --- |
| **Negative**​ | 67%​ | 77%​ | 72%​ |
| **Neutral**​ | 46%​ | 21%​ | 29%​ |
| **Positive**​ | **69%**​ | **78%**​ | 73%​ |

The classification report shows the Positive class having the highest precision of 69% and recall of 78%. With Negative class coming next with 67% precision and 77% Recall. Meanwhile, the model performed very badly on predictions for the Neutral class which showed lowest precision and recall. This implies that further data preprocessing and tuning is needed to improve our model.

​

# References

1. [Pang, B. and Lee, L. (2008) Opinion Mining and Sentiment Analysis.](https://doi.org/10.1561/1500000011)
2. [Principal Component Analysis - Ian Jolliffe](https://onlinelibrary.wiley.com/doi/10.1002/9781118445112.stat06472)
3. [Salton, G. and Buckley, C. (1988) Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management, 24, 513-523.](http://dx.doi.org/10.1016/0306-4573(88)90021-0)
4. <https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>