Sentiment Analysis on Amazon Reviews

September 6, 2023

Abstract

Reviews can significantly impact a company's reputation in the market, potentially influencing its overall business outcomes, either positively or negatively. This is especially crucial for companies that operate primarily through e-commerce platforms. Hence, it is vital for companies to pay close attention to customer reviews.Sentiment Analysis, often referred to as "opinion mining," is a significant procedure in Natural Language Processing (NLP) which serves the purpose of ascertaining the emotional tone of a provided text and categorizing it into positive, negative, or neutral perspectives.In this paper, sentiment analysis methodology is presented for classifying amazon reviews which utilizes a large dataset of reviews and employs Multinomial Naive Bayesian (MNB), Support Vector Machine (SVM), Maximum Entropy (ME), and Logistic Regression as the primary classifiers by the authors. With the aid of machine learning, we employed a supervised learning approach to an extensive Amazon dataset in order to categorize it based on sentiment polarity, achieving a high level of accuracy for the results.Here, we utilized the Kaggle dataset that includes a substantial volume of reviews and associated metadata which comprises customer review and ratings on amazon products.

Index terms: Sentiment Analysis, Machine learning, Natural Language Processing, Naive Bayesian (MNB), Support Vector Machine (SVM), Maximum Entropy (ME), Logistic Regression,feature extraction, text classification.

1 Introduction

The long-term viability of businesses like amazon is heavily contingent on their capacity to effectively fulfill customer requirements.Countless individuals share their opinions about various services or prod- ucts through different platforms like social networking sites, blogs, or popular review sites that will comes up just with a google search. So,in recent times it has become a common practice to search for reviews before making a purchase decision.Hence, disseminating customer reviews and feedback about online products or services can significantly impact the perceptions of new customers regarding the organization. Additionally, Amazon can examine these reviews to determine their authenticity and assess whether they might be part of a competitor's scheme to manipulate perceptions.Also, con- ducting a thorough analysis of customer sentiments empowers Amazon's business enthusiasts to gain deeper insights into the market, enabling them to make informed decisions that preemptively address customer needs and concerns, like betterment or elimination of any existing product, increasing or decreasing monetary value for marketing reach etc. Sentiment analysis leverages natural language pro- cessing techniques and text analysis to delve into what customers are conveying, how they articulate their thoughts, and the underlying meaning behind their expressions.The primary objectives of this paper is to extract the sentiments conveyed in customer reviews and analyze these senti-

various classifiers were employed iteratively we utilized the labeled datasets to conduct further processing while the extracted features were classified through different classifiers.We utilized a combination of two feature extraction methods: the bag of words approach and the tf-idf and Chi-square approach to enhance the accuracy of our results

2 Literature Review

With the help of machine learning algorithms, classification of textual data for sentiment analysis is done easily for the detection of positive and negative reviews about the products on amazon. By using machine learning algorithms, we concentrated more on the preciseness of the classification of the reviews. So, we extracted data from the reviews on amazon site and created a business model by surveying the result of the reviews. Based on the data, we detected positive and negative reviews of the customer by using Multinomial Naive Bayesian(MNB), Support Vector Machine(SVM), Maximum Entropy(ME) and Logistic Regression as the main classifier. By adapting a supervised machine learning algorithm, we predicted the ratings of textual reviews on a given numerical scale. In the paper we used natural language processing for classifying the textual reviews. To label the reviews given by the customers whether it is positive or negative, Naive Bayesian and Decision list is used as classifiers. The main objective is to create a system that portrays the sentiment of the reviews in the chart. Maximum Entropy came under the category of probabilistic classifier and it did not take into account the elements independent of each other. Also added that Maximum Entropy solved a huge amount of text classification problems. So, it is very popular in the field of sentiment analysis. Naive Bayesian constructed classifiers that assigned a class label to the problem examples. Depending on the value of the features being defined, Naive Bayesian expressed a vector form where labels came from the finite sets.Despite not being a standard algorithm, Naive Bayesian is dependent on the principles being used. For classification problems, another algorithm called Support Vector Machine is used widely. Besides, they are popular for being especially effective at separating data points that belong to various classes using the ideal hyperplane, making them particularly well-suited for classifi- cation issues. They distinguished the classes very effectively. On the other hand Logistic Regression known as the statistical approach and machine learning algorithm on the idea of probability is used to solve classification problems. It is applicable if the dependent variable is categorical. To determine whether the output of the reviews are positive, neutral or negative, we used logistic regression. As a result to get more accurate results for our research paper, we used logistic regression, Support Vector Machine, Naive Bayesian and Maximum Entropy for the prediction of amazon reviews.

3 Methodology

3.1 Collected Data

In our effort to conduct sentiment analysis on Amazon reviews, we made use of a carefully curated dataset obtained from Kaggle, a well-known community for data science enthusiasts. We used the Kaggle dataset for our investigation, which contains a sizable number of reviews and related metadata, to build a strong sentiment analysis framework. 21 columns and 34,660 rows make up this dataset, which depicts various facets of the review data in each row and column. We concentrated on using 10 essential features from this large dataset for our particular investigation.Our sentiment analysis targets extracting valuable insights from the Kaggle dataset, illuminating the sentiments expressed within Amazon reviews by focusing on these 10 carefully selected features. This group of features provides a targeted and effective way to analyse consumer sentiment on Amazon and discover the variables affecting how they feel about specific products.

3.2 Proposed Methodology

One of Amazon's most valuable and established features are customer reviews. For the purpose of determining the polarity of a review, we have used the reviews.text and reviews.rating features from our dataset.We have preprocessed our data by removing null values from our data and replaced them with spaces, as well as any extraneous features, since If they are not handled properly, missing values could introduce bias and errors into the analysis. Missing values have the potential to skew statistical metrics like means, medians, and correlations. There are a finite number of categorical variables, most of which take the form of "strings" or "categories."Data preparation is made more challenging by categorical values' lack of numeric values.Algorithms can find relationships and patterns in the data by converting category values into numerical representations, which are widely used to describe qualitative aspects.. Additionally, categorical variables like reviews.text are included in our project. The code in our project converts the "reviews.text" column from a string to an integer type. CountVectorizer and TfidfTransformer are used to convert the text data (reviews) into numerical characteristics that machine learning models can use. The dataset was separated into training and training dataset 70 and 30 percent, respectively, and scrambled at random. The accuracy was then assessed using machine learning approaches such as SVM, logistic regression, Decision tree, naive bayes, and random forest algorithm.

4 Conclusion

Ultimately, our exploration of sentiment analysis in Amazon reviews has demonstrated the enormous importance of comprehending customer comments. We've opened the way for businesses to obtain profound insights into the thoughts of their customers by harnessing the power of machine learning techniques like

Multinomial Naive Bayesian, Support Vector Machine,Decision Tree, and Naive Bayes. Now it is possible to categorize customer reviews, which are sometimes buried in unstructured data, as either good, negative, or neutral. We were able to manage a sizable dataset using our methodology, which was informed by data mining and computational linguistics, and we were able to derive meaning from these reviews. To increase the accuracy of our research, we have expanded upon and modified previous findings. Businesses may better meet client wants, develop their goods and services, and improve their market placement by undertaking detailed sentiment analysis. In the end, sentiment analysis proves to be a crucial tool for businesses functioning in the digital age to flourish, change, and be successful.