Sentiment Analysis for Amazon Reviews

Karthik Shankar
Department of Mathematical Sciences
Stevens Institute of Technology
Hoboken, United States of America
kshanka3@stevens.edu

Sambhav Chawla
Department of Mathematical Sciences
Stevens Institute of Technology
Hoboken, United States of America
schawla1@stevens.edu

***Abstract*—Because online marketplaces have grown in popularity over the last several decades, online vendors and merchants now invite their customers to give their thoughts on the things they've purchased. As a result, millions of evaluations are created every day, making it difficult for a potential customer to decide whether or not to purchase the goods. For product makers, analyzing such a large number of comments is difficult and time-consuming. This thesis investigates the topic of categorizing reviews based on their overall semantic content (positive or negative). On Amazon beauty items, two different supervised machine learning approaches, SVM and Nave Bayes, were used to conduct the study. Their precisions were then compared. When the data set is larger, the SVM technique beats the Naive Bayes strategy, according to the findings. Both algorithms, however, achieved promising accuracies of at least 80**

I. Introduction

Because online marketplaces have grown in popularity over the last several decades, online vendors and merchants now invite their customers to give their thoughts on the things they've purchased. Every day, millions of reviews are written all around the world. Different items, services, and locations may be found on the Internet. As a result, the most essential source of information and views on a product or service is the internet.

The overall semantic of customer reviews is determined by classifying them into positive and negative sentiment in this research, which uses supervised techniques to determine the overall semantic of customer reviews by classifying them into positive and negative sentiment.

The goal of sentiment categorization is to figure out what a written text's general objective is, whether it's praise or criticism. Machine learning methods like Naive Bayes and Support Vector Machine can help with this. As a result, the following is the problem that will be researched in the project: Which machine learning method is more accurate when it comes to Amazon product reviews?

People use Amazon every day for online shopping since it is one of the e-commerce behemoths that allows them to browse hundreds of evaluations left by other consumers about the things, they are interested in. These evaluations contain essential information about a product, such as its features, quality, and suggestions, allowing buyers to comprehend practically every aspect of the product. This is advantageous not just to customers, but also to vendors who manufacture their own items, since it allows them to better understand consumers and their demands.

The dataset comprises of 1128437 rows and 12 columns. The column includes Overall rating by the customer, Reviews and summary by the customer, the verification of the review, date of verification, reviewer id, reviewer name, asin, review time, style, vote and image of the product.

Our group plans to use Logistic Regression, Linear Support Vector Classifier, Decision Tree Regressor and Multinomial Naive Bayes to solve the problem as it suits our problem statement and needs.

II. Related Work

Sentiment analysis has gotten a lot of interest in recent years thanks to the explosion of online reviews. As a result, several studies have been conducted in this field[4]. Some of the most relevant research works are covered in this section.

For categorizing, Pang, Lee, and Vaithyanathan (2002) used supervised learning with the use of SVM and Nave Bayes and maximum entropy classification, movie reviews were divided into two categories: positive and negative. In terms of precision, all three approaches performed admirably[4]. In this work, they experimented with numerous features and discovered that when a bag of words was utilized as a feature in the classifiers, the machine learning algorithms performed better[4].

Three supervised machine learning algorithms, Nave Bayes, SVM, and N-gram model, were tried on internet evaluations about various tourism locations throughout the world in a recent survey done by Ye et al. (2009). They discovered in this study that properly trained machine learning algorithms perform very well for categorization of vacation destination reviews in terms of accuracy[4]. Furthermore, they revealed that the SVM and N-gram models produced superior results than the Nave Bayes technique[4]. However, increasing the quantity of training data sets lowered the gap between the algorithms dramatically[4].

Chaovalit and Zhou (2005) compared a supervised machine learning algorithm to an unsupervised approach to movie evaluation called Semantic orientation and concluded that the supervised technique was more trustworthy than the unstructured one[4].

Naive Bayes and SVM are two of the most widely utilized algorithms in sentiment classification issues, according to various studies (Joachims 1998; Pang et al. 2002; Ye et al. 2009)[4].

III. Our Solution

A. Description of Dataset

Data gathering for training and testing the classifiers is the initial stage. Because Amazon does not provide an API to obtain reviews like Twitter, the data is gathered through the SNAP data collection. The downloaded file was in JSON format, with one review per line. The file was converted to the Comma Separated Values (CSV) format, as this is a more convenient format for Python to handle. There are 1128437 reviews of various goods in the data set. Each review contains the following 12 features: Overall rating by the customer, Reviews and summary by the customer, the verification of the review, date of verification, reviewer id, reviewer name, asin, review time, style, vote and image of the product.

Our focus is on the columns- 'reviewText', 'summary' and 'overall'. After some pre-processing it was found that 'overall' column has an average score of 4.22 and almost all reviews are positive after the second quartile. There are a total of 500 null values in the dataset. As the size of the dataset is more than 100,000, it was decided to truncate the null values as ratio of null values to the total data is extremely low.

B. Machine Learning Algorithms

1. Multinomial Naïve Bayes- Multinomial naïve Bayes is a common algorithm which works well for data that can easily be turned into counts, such as word counts in text[1]. Naive Bayes is a simplified version of Bayes Theorem, where all features are assumed conditioned independent to each other (the classifiers), P(x—y) where x is the feature and y is the classifier[1]. Naïve Bayes uses probability theory and Bayes' Theorem to predict the tag of a text. It is planned to create a pipeline with steps-Count Vectorizer, TFIDF Transformer and Multinomial Naive Bayes, then fit the train set, predicted the test set and checked the accuracy score, mean squared error and f1 score of the model on the test set[1]. A confusion matrixis later created.

2. Logistic Regression- Logistic Regression is a classification algorithm which is used when the dependent variable is binary[1]. A logistic regression estimates a multiple linear regression function. Also, logistic regression is easier to implement and interpret[2]. A model in a pipeline with Count Vectorizer and Tfidf is implemented. Transformer and checked the accuracy score, mean squared error and f1 score and created the confusion matrix[2].

3. Linear Support Vector Classifier- SVC is a supervised convex optimisation problem which is relatively faster than non-linear classifier and fits N-models.The model in a pipeline is implemented[2]. The training set is fit to the model to help it perform better on the testing set[2]. We obtain the accuracy score,mean squared error, f1 score and the confusion matrix.

C. Implementation Details

To start with the project, the data was downloaded and extracted into Jupyter notebook. All the necessary libraries were imported, and the data was ingested into a pandas data frame. All the non-null values and datatype of the variables was checked for. To clean the data, the group checked for all the null values in each column and found that most of the null values were in the column- 'style', 'vote' and 'image'. These columns were dropped as they were non-significant. Entries with null values were truncated as the size of the dataset is massive and removing them wouldn't affect our model much. Attributes like 'reviewText' and 'summary' had comparatively fewer null values.

For analysing the data, the datetime format was changed. The group also checked the total number of reviews by unique products by applying group by function on the 'asin' column with 'overall' column and found that there are 48186 unique products. The percentage of positive, neutral, and negative sentiments were checked using 'overall' column. Post analysis, two columns (reviewText and summary) were concatenated as it would not complicate the training set and the new column was renamed as 'Reviews'.

All the stop words, numbers, url and punctuation from the 'Reviews' column were removed using Natural Processing Language as removing stop words and punctuation would make the content more efficient. It was ensured that all the words are lower case. After data preprocessing, an extra 'target' column was created and was filled with values according to the 'column and applied map function (if rating ¿3 it would return 2 which is positive, if rating overall' ¡3 it would return 0 which is negative and if rating=3 it would return 1 which is neutral). Basically, a classifier was built that can determine a review's sentiment. To start with machine learning algorithms, the data was split into train and test sets and their shape was checked to ensure uniformity.

Later a CountVectorizer was applied on the train set as it transforms each review into a numerical vector representation of words and n-grams. It is a highly flexible feature representation module for text as it also tokenize the sentences. Using TFIDF Transformer, the count matrix was transformed to a normalized TF or TF-IDF representation to measure weights as the word which has the highest weight provides more information.

A Logistic Regression model and Multinomial Naive Bayes Classifier model was implemeted. The performance metrics of the two models are stated below.

1) Multinomial Naive Bayes Classifier: i)Accuracy score on train set - 83.54ii)Accuracy score on test set - 83iii)Mean squared error - 44.87

2) Logistic Regression: i)Accuracy score on train set - 90ii)Accuracy score on test set - 89.63iii)Mean squared error - 20.19

It can clearly be inferred from the above results that Logistic Regression performs a little better as compared to Multinomial Naive Bayes Classifier. As for Support Vector Classifier and Decision Tree Regressor, they will be implemented and shared in the final project report.

References

[1]https://medium.com/@agarwalvibhor84/getting-started-with-machine-learning-using-sklearn-python-7d165618eddf
[2]https://datascience.stackexchange.com/questions/6987/can-you-explain-the-difference-between-svc-and-linearsvc-in-scikit-learn
[3]https://intellipaat.com/community/19783/which-one-is-better-linearsvc- or-svc
[4]https://www.divaportal.org/smash/get/diva2:1241547/FULLTEXT01.pdf