Amazon/Flipkart customer sentiment prediction using ML and DL.

Team Members :

Nitesh Marchande
Anurag Mhatre
Harsh Shah
Maheshwar Reddy

Abstract

Opinion information is very important for businesses and manufacturers. They often want to know in time what consumers and the public think of their products and services. However, it is not realistic to manually read every post on the website and extract useful viewpoint information from it. If you do it manually, there is too much data. Sentiment analysis allows large-scale processing of data in an efficient and cost-effective manner. Application of sentiment analysis on business is to basically understand its strengths and limitations. This project deals with Amazon dataset which builds a model to predict the sentiment of the comment given the customers by using Python and machine learning algorithms.

Keywords—Sentiment Classification, Machine Learning, EDA, NLP, Counter Vectorizer, Logistic Regression

Table of Contents

Introduction

As the commercial site of the world is almost fully undergone in online platform people is trading products through different e-commerce website. And for that reason reviewing products before buying is also a common scenario. Also now a days, customers are more inclined towards the reviews to buy a product. So analyzing the data from those customer reviews to make the data more dynamic is an essential field nowadays. In this age of increasing machine learning based algorithms reading thousands of reviews to understand a product is rather time consuming where we can polarize a review on particular category to understand its popularity among the buyers all over the world.

The objective of this project is to categorize the positive and negative feedbacks of the customers over different products and build a supervised learning model to polarize large amount of reviews. A study on amazon last year revealed over 88% of online shoppers trust reviews as much as personal recommendations. Any online item with large amount of positive reviews provides a powerful comment of the legitimacy of the item. Conversely, books, or any other online item, without reviews puts potential prospects in a state of distrust. Quite simply, more reviews look more convincing. People value the consent and experience of others and the review on a material is the only way to understand others impression on the product.

Problem Statement

An application that collects reviews from the users about a certain product and analyzes them. It would segregate the reviews into positive and negative reviews. The negative reviews will be helpful to the companies to further enhance their product based on the user's' feedback. The application will further provide reports about the sentiment analysis performed on the products.

Dataset And Data Collection

For Amazon sentiment prediction , we utilize the dataset of Amazon dedicated to segregate the reviews. Data set contains various garbage values and attributes like product_title, customer_id, product_category, etc.
To train the model properly, a large dataset of 1000 rows is used to gain accurate result and accuracy.

MODEL AND METHODOLOGIES

a. NLP- Spacy

Natural Language Processing is one of the principal areas of Artificial Intelligence. NLP plays a critical role in many intelligent applications such as automated chat bots, article summarizers, multi-lingual translation and opinion identification from data. Every industry which exploits NLP to make sense of unstructured text data, not just demands accuracy, but also swiftness in obtaining results. Natural Language Processing is a capacious field, some of the tasks in nlp are – text classification, entity detection, machine translation, question answering, and concept identification.

spaCy

Spacy is written in cython language, (C extension of Python designed to give C like performance to the python program). Hence is a quite fast library. spaCy provides a concise API to access its methods and properties governed by trained machine (and deep) learning models

Benefits of NLP:

- Better data analysis. Unstructured data such as documents, emails, and research results are difficult for computers to process.
- Streamlined processes.
- Improved customer experience

b. Counter Vectorizer

In order to use textual data for predictive modeling, the text must be parsed to remove certain words – this process is called tokenization**.** These words need to then be encoded as integers, or floating-point values, for use as inputs in machine learning algorithms. This process is called feature extraction (or vectorization)**.**

Scikit-learn's CountVectorizer is used to convert a collection of text documents to a vector of term/token counts. It also enables the pre-processing of text data prior to generating the vector representation. This functionality makes it a highly flexible feature representation module for text.

```
Data = ['The', 'quick', 'brown', 'fox', 'jumps', 'over', ' the', 'lazy', 'dog']
```

	The	quick	brown	fox	jumps	over	lazy	dog
Data	2	1	1	1	1	1	1	1

a. Logistic Regression

In statistics the **logistic model** (or **logit model**) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.

Logistic regression uses an equation as the representation, very much like linear regression.

Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.

Below is an example logistic regression equation:

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*

Approach

To train the model we have input data and the expected output data. This is a case for Supervised Learning.

To create such model, we have to go through the following phases:

- model construction
- model training
- model testing
- model evaluation

Model Construction

For construction of the model, the null values need to be removed or replaced.

Shuffling of the dataset needs to take place as only a certain part of the dataset will be used to train the model. Logistic regression model is used for training as well as testing the dataset to analyze sentiment prediction of the customers.

Model Training

Once the Model Layers are finalized we can proceed with Training

First we perform the train test Split to get our training and testing datasets.

Then we can train the model with the training data.

Model Testing

Model Testing can be performed on the split testing dataset. In this case we also have a test dataset separately available that can be used for validation.

Model Evaluation

For evaluation purposes we need to select a baseline model. For this purpose we can use a simple NLP model to get a baseline score. After the implementation of the model the desired accuracy can be achieved.

Drawbacks

The accuracy of the whole system majorly depends on the dataset used. The larger the dataset the better the accuracy. Null and recurring values in the dataset can prove to be disadvantageous.