**PRODUCT SENTIMENT CLASSIFICATION**

**Abstract**

Sentiment Analysis also known as Opinion Mining refers to the use of natural language processing, text analysis to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine. In this project, we aim to perform Sentiment Analysis of product-based reviews. Data used in this project are online product reviews collected from “amazon.com”. We expect to do review-level categorization of review data with promising outcomes.

**Problem Statement & Description**

Analysing sentiments related to various products such as Tablet, Mobile and various other gizmos can be fun and difficult especially when collected across various demographics around the world. Analysing these sentiments will not only help us serve the customers better but can also reveal lot of customer traits present/hidden in the reviews.

**Dataset Description**

The unzipped folder will have the following files.

* **Train.csv – 6364** rows x **4** columns (Includes **Sentiment** Column as Target)
* **Test.csv – 2728**rowsx**3**columns
* **Sample Submission.csv –**sample format for submission file.

**Attribute Description**

* **Text\_ID –**Unique Identifier
* **Product\_Description –**Description of the product review by a user
* **Product\_Type –**Different types of product (**9** unique products)
* **Class –**Represents various sentiments
  + 0 – Cannot Say
  + 1 – Negative
  + 2 – Positive
  + 3 – No Sentiment

**Data Preprocessing**

This is a vital part of training the dataset. Here Words present in the file are accessed both as a solo word and also as pair of words. Because, for example the word “bad” means negative but when someone writes “not bad” it refers to as positive. In such cases considering single word for training data will work otherwise. So words in pairs are checked to find the occurrence to modifiers before 19 any adjective which if present which might provide a different meaning to the outlook.

**Training Data/ Evaluation**

The main chunk of code that does the whole evaluation of sentimental analysis based on the pre-processed data is a part of this. The following are the steps followed:

i) The Accuracy, Precision, Recall, and Evaluation time is calculated and displayed.

ii) Navie Bayes, Logistic Regression, Linear SVM and Random forest, XGB classifiers are applied on the dataset for evaluation of sentiments.

iii) Prediction of test data is done and Confusion Matrix of prediction is displayed.

iv) Total positive and negative reviews are counted.

v) A review like sentence is taken as input on the console and if positive the console gives 1 as output and 0 for negative input.

**Text Classification**

We are now ready to experiment with different machine learning models, evaluate their accuracy and find the source of any potential issues. We will benchmark the following five models:(Multinomial) Naive Bayes, Logistic Regression, Linear Support Vector Machine, Random Forest, XGB Classifier

**For evaluating our model’s performance, we must take into account the strong label imbalance; for this purpose, it’s better to choose an alternative metrics than standard accuracy:**

* Precision (tp / (tp + fp)) measures the ability of a classifier to identify only the correct instances for each class.

**Conclusion**

Sentiment analysis deals with the classification of texts based on the sentiments they contain. This article focuses on a typical sentiment analysis model consisting of three core steps, namely data preparation, review analysis and sentiment classification, and describes representative techniques involved in those steps. Sentiment analysis is an emerging research area in text mining and computational linguistics and has attracted considerable research attention in the past few years. Future research shall explore sophisticated methods for opinion and product feature extraction, as well as new classification models that can address the ordered labels property in rating inference. Applications that utilize results from sentiment analysis is also expected to emerge soon.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MODEL** | **F1 SCORE** | **ACCURACY** | **PRECISION** | **RECALL** |
| Multinomial NB | 67% | 66.7% | 88.5% | 67% |
| Logistic Regression | 73% | 68.3% | 80% | 68% |
| Linear SVC | 70% | 68% | 71% | 68% |
| Random Forest | 71% | 67.7% | 77.5% | 68% |
| XGB classifier | 73% | 65% | 88% | 66% |

By Observing all models, we got **Logistic Regression** - Accuracy: 68.3% - Weighted Precision: 80% as the highest.