NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM,

APPROVED BY AICTE & GOVT.OF KARNATAKA

Data Mining Project Proposal

On Sentimental Analysis

Submitted in the Partial fulfillment of the requirements of Semester 5 of

BACHELOR OF TECHNOLOGY IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

Darshan MR	JVSS Pavan Kumar	Saiel Gaonkar
1NT19CS063	1NT19CS087	1NT19CS166

Under the Guidance of

Dr. Vani V

Dept. of CSE

Introduction

Everyday we come across various products in our lives,on the digital medium we swipe across hundreds of product choices under one category. It will be tedious for the customer to make selection.Here comes 'reviews' where customers who have already got that product leave a rating after using them and brief their experience by giving reviews. As we know ratings can be easily sorted and judged whether a product is good or bad. But when it comes to sentence reviews we need to read through every line to make sure the review conveys a positive or negative sense.

Sentiment Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral.Understanding people's emotions is essential for businesses since customers are able to express their thoughts and feelings more openly than ever before.It is quite hard for a human to go through each single line and identify the emotion being the user experience.Now with technology, we can automatically analyzing customer feedback, from survey responses to social media conversations, brands are able to listen attentively to their customers, and tailor products and services to meet their needs.

Python sentiment analysis is a methodology for analyzing a piece of text to discover the sentiment hidden within it. It accomplishes this by combining machine learning and natural language processing (NLP). Sentiment analysis allows you to examine the feelings expressed in a piece of text.

Data Mining task

- **Data preprocessing :** Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.
- **Binarization :** Binarization is the process of transforming data features of any entity into vectors of binary numbers to make classifier algorithms more efficient. In a simple example, transforming an image's gray-scale from the 0-255 spectrum to a 0-1 spectrum is binarization.
- **Data Selection :** It is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection.
- **Data transformation :** Data transformation is the process of changing the format, structure, or values of data. For data analytics projects, data may be transformed at two stages of the data pipeline.

Data set

This file has reviewer ID , User ID, Reviewer Name, Reviewer text, helpful,
Summary(obtained from Reviewer text),Overall Rating on a scale 5, Review time

Description of columns in the file:

reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B

asin - ID of the product, e.g. 0000013714

reviewerName - name of the reviewer

helpful - helpfulness rating of the review, e.g. 2/3

reviewText - text of the review

overall - rating of the product

summary - summary of the review

unixReviewTime - time of the review (unix time)

reviewTime - time of the review (raw)

Methods And Models :

- **Normalization:** Database normalization is the process of structuring a database, usually a relational database,in accordance with a series of so-called normal forms to reduce data redundancy and improve data integrity.In machine learning and data mining, data normalization is used to make model training less sensitive to feature scale. As a result, our model can converge to better weights, resulting in a more accurate model. It is generally useful for classification algorithms.

- **Associative Rule Mining:** Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an itemset occurs in a transaction.By using regression we predict whether the review is postove or negative.

- **Model Building:** The resultant output of the normalization process will be taken as the input for the model building.Since the primary goal of the project is to predict a attribute., we will have to apply one of the regression algorithms to obtain the desired result.First select the best peforming model by using cross validaton. Let's consider all the classification algorithm and perform the model selection process.We go with with logistic regression with hyperparameter tuning.

Logistic Regression with Hyperparameter tuning :We use regularization parameter and penality for parameter tuning.

Classification metrics :Here we plot the confusion matrix with ROC and check our f1 score

Assesments:

As a classification problem, Sentiment Analysis uses the evaluation metrics of Precision, Recall, F-score, and Accuracy. Also, average measures like macro, micro, and weighted F1-scores are useful for multi-class problems. Depending on the balance of classes of the dataset the most appropriate metric should be used.

- **Precision** (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved. Both precision and recall are therefore based on relevance.
- **Recall** is the ratio of correctly predicted positive observations to the all observations in actual class
- **F1 Score** is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.
- **Accuracy** is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

Presentation and Visualization :

- **Sentiments vs Helpful rate:** We look whether there any relationship between sentiment of review and helpfulness of it.
- **Year vs Sentiment count :**n this block we will see how many reviews were posted based on sentiments in each year from 2004 to 2014
- **Day of month vs Reviews count :** Here we if check there are any relationship between reviews and day of month
- **N-gram analysis :** Here we will be using ngrams to analyse the text, based on it's sentimen.
- **Monogram analysis** : Here we will plot most frequent one word in reviews based on sentiments

- **Bigram analysis** : Here we will plot most frequent two words in reviews based on sentiments
- **Trigram analysis :** Here we will plot most frequent three words in reviews based on sentiments
- **Wordcloud-Positive reviews :** Here we look at the word cloud of positive reviews
- **Wordcloud-Neutral reviews :** Here we look at the word cloud of neutral reviews
- **Wordcloud-Negative reviews** : Here we look at the word cloud of negative reviews

Roles

The roles of all the team members are as follows:

- **Darshan MR:** Drafting the project proposal. Work on the initial data understanding,preparation, selection of the regression algorithm, and assist with training of the model.Focusing on dividing the dataset for training & testing.
- **JVSS Pavan Kumar:** Drafting the project proposal implementing the algorithm, and training the model, working on the aggregation of the results for the other stated goals of the project, and working on the data visualization.
- **Saiel Gaonkar:** Working on the data visualization,presentation, assessment of the model trained, and identifying possible ways to improve the model. Training the model, Drafting the project report.

Schedule :

Date	Tasks
06/01/2022	Data Preprocessing
09/01/2022	Loading dataset and performing data mining tasks
12/01/2022	Data Visualization
15/01/2022	Model Evaluation
17/01/2022	Project Report

Bibiliography

- **https://techvidvan.com/tutorials/python-sentiment-analysis/**
- **https://www.analyticsvidhya.com/blog/2021/06/rule-based-sentiment-analysis-in-python/**
- **https://www.digitalocean.com/community/tutorials/how-to-perform-sentiment-analysis-in-python-3-using-the-natural-language-toolkit-nltk**
- **https://www.kaggle.com/benroshan/sentiment-analysis-amazon-reviews#Table-of-Contents**
- **https://www.kaggle.com/sid321axn/natural-language-processing-sentiment-analysis**