**Academia International College**

Tribhuvan University

Institute of Science and Technology



**Project Proposal Report**

**On**

**“Sentiment Analysis on Daraz for Product Review using Naive Bayes”**

**Submitted To:**

Department of Computer Science and Information Technology

Academia International College

In partial fulfillment of the requirement for Bachelor Degree in Computer Science and Information Technology

**Submitted By:**

Ashish Adhikari (10811/073)

Bhuwan Shahi (10816/073)

Yagya Raj Bhatta (10849/073)

**Date:2076/12/5**

# Abstract

Sentiment analysis or opinion mining is one of the major tasks of NLP (Natural Language Processing). Sentiment analysis has gained much attention in recent years. In this paper, we aim to tackle the problem of sentiment polarity categorization, which is one of the fundamental problems of sentiment analysis. A general process for sentiment polarity categorization is proposed with detailed process descriptions. Data used in this study are online product reviews collected from Daraz.com. Experiments for both sentence-level categorization and review-level categorization are performed with promising outcomes.Keywords: Natural Language Processing, Sentiment Analysis, Machine Learning, Naïve Bayes, Confusion Matrix, Feature extraction.

Table of Contents

[1. Introduction 1](#_Toc35422350)

[1.1 Sentiment Analysis 1](#_Toc35422351)

[1.2 Problem Definition 1](#_Toc35422352)

[1.3 Objective 2](#_Toc35422354)

[1.4 Challenges of sentiment analysis 2](#_Toc35422355)

[2. Literature Review 3](#_Toc35422356)

[3. Methodology 4](#_Toc35422357)

[3.1 System Analysis 4](#_Toc35422358)

[3.1.1 Data Collection 4](#_Toc35422359)

[3.2 System Design 5](#_Toc35422360)

[3.3 Proposed System 6](#_Toc35422361)

[3.4 Naive Bayes 9](#_Toc35422362)

[4. Development Model 11](#_Toc35422363)

[4.1 Development Tools 11](#_Toc35422364)

[4.2 Requirement Analysis 11](#_Toc35422365)

[4.2.1 Functional Requirement 11](#_Toc35422366)

[4.2.2 Non-Functional Requirement 11](#_Toc35422367)

[4.3 System Flowchart 12](#_Toc35422368)

[5. Working Schedule 13](#_Toc35422369)

[6. Expected Result 13](#_Toc35422370)

[7.Conclusion 14](#_Toc35422371)

[References 14](#_Toc35422372)

[Figure 1-Flow diagram for Data Filtration 5](#_Toc35418021)

[Figure 2-System Architecture of Product Review 6](#_Toc35418022)

[Figure 3-Illustration of proposed system using sentiment analysis 7](#_Toc35418023)

[Figure 4-Diagram for testing proposed model on dataset 8](#_Toc35418024)

[Figure 5-Flowchart for Machine Learning Implementation 10](#_Toc35418025)

[Figure 6-Gantt Chart 13](#_Toc35418026)

# Introduction

## Sentiment Analysis

Here we propose an advanced Sentiment Analysis for Product Rating system that detects hidden sentiments in comments and rates the product accordingly. The system uses sentiment analysis methodology in order to achieve desired functionality. This project is an E-Commerce web application where the registered user will view the product and product features and will comment about the product. System will analyze the comments of various users and will rank product. We use a database of sentiment-based keywords along with positivity or negativity weight in database and then based on these sentiment keywords mined in user comment is ranked. Comment will be analyzed by comparing the comment with the keywords stored in database. The System takes comments of various users, based on the comment, system will specify whether the product is good, bad, or worst. Once user login the system he can view the product and product features. After viewing product user can comment about the product. User can also view comment of other users. The role of the admin is to add product to the system and to add keywords in database. User can easily find out correct product for his usage. This application also works as an advertisement which makes many people aware about the product. This system is also useful for the users who need review about a product.

## Problem Definition

In this world of digital marketing,there are lots of product in ecommerce website for selling and its very difficult to trust which produt are reliable for customer.Product rating help customer to gain trust on the product which they are seeking for.There is comment section on the product purchasing section where user can gice review about the product which helps other people to rely on it.The customer who purchases product from daraz are complaining that they didnit get the product they want.So this system helps people to choose the actual product what they are looking for.

## Objective

* Product reviews are used on shopping sites to give customers an opportunity to rate and comment on products they have purchased. Other consumers can read these when making a purchase decision.
* They help build trust and loyalty, and typically describe what sets your products apart from others and provides recommendations to the user of similar products that they actually seeking for.
* Understand what your customers like and dislike about your product.
* Compare your product reviews with those of your competitors.
* Save hundreds of hours of manual data processing.

## Challenges of sentiment analysis

Some of the major challenges in sentiment analysis:

* The comments given by user for a product is considered positive at one situation and negative at another situation.
* Some people don’t express opinions in the same way. Most reviews will have both positive and negative comments, which somewhat manageable by analyzing sentences one at a time.
* Sometimes people may give fake comments about the product, which gives the bad review about the product.

# Literature Review

They analyzed the Naive Bayes, Max Entropy and Support Vector Machine models for sentiment analysis on unigrams and bigrams of data. In their experiment, SVM paired with unigram feature extraction produced the best results. They reported a result of 82.9% accuracy. [1]

The only study which mentions the mismatch between rating and text. They ignore reviewer ratings and employ a new set of ratings for the training and testing of their system. From their work, however, it is neither clear to what extent the new ratings differ from the user ratings as they do not report inter-annotator agreement scores nor what the effect is of the different ratings on classifier performance. [2]

The task of sentiment analysis on product reviews was originally performed to extract overall sentiment from the target texts. However, the difficulty shown in the experiments, the whole sentiment of a document is not necessarily the sum of its parts. Then there came up with research works shifting focus from overall document sentiment to sentiment analysis based on product attributes. [3]

The product features commented by the customer in the review are mined. Natural language processing and Data mining techniques are used for mining. The opinions in the review are identified and the opinions are classified as positive or negative. Set of adjectives words called opinion words are identified and semantic orientation of the opinion words is determined. WordNet can be used to identify the semantic orientation and the opinion orientation of each sentence is decided. [4]

# Methodology

## System Analysis

## Data Collection

Datasets **acquiring:** We acquire datasets from several repositories.

Data pre-processing: To pre-process the raw data to convert into usable format.

Dataset as input: Taking input as our dataset through the R code.

Reviews filtering: Filtering the reviews to remove stop words and get adjectives.

Classifying the data: Classifying the data depending on the sentiment present in them.

Output graph: Displaying output as a graph.

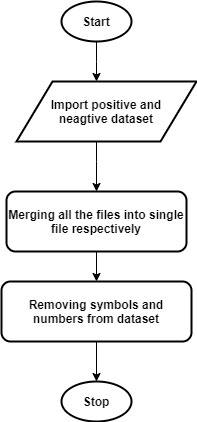


Figure 1-Flow diagram for Data Filtration

Data filtering is the task of reducing the content of noise or errors from measured process data. It is an important task because measurement noise masks the important features in the data and limits their usefulness in practice. Various techniques have been developed to filter process data, and include model-free techniques, model-based techniques, and techniques based on empirical models. The data sets are refined into simply what a user (or set of users) needs, without including other data that can be repetitive, irrelevant or even sensitive.

## System Design

Feature Extraction:

The dimensionally reduction process of extracting informative and non-redundant values from a given dataset is called Feature Extraction. This module extracts numerical

features from the given movie or product reviews which are in

text format in the following way:

* Each string is converted into a unique ‘token’.
* Frequency of occurrence of each of these tokens is calculated.
* Tokens are organized based on the frequency of occurrences.

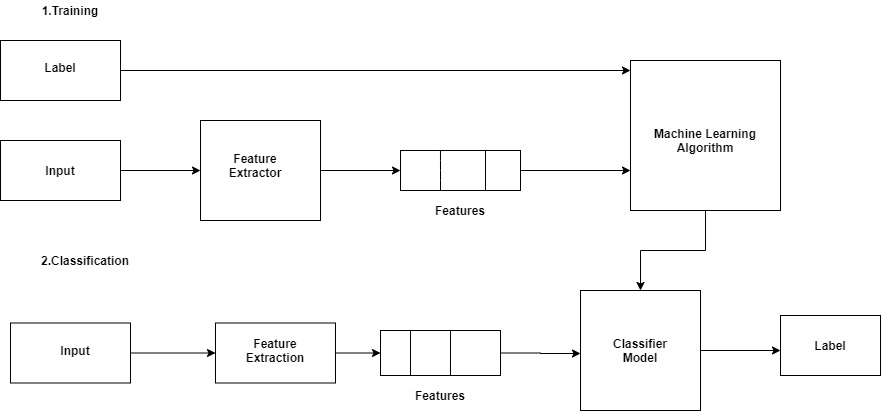


Figure 2-System Architecture of Product Review

## Proposed System

The below mentioned figure illustrates the proposed algorithm for sentiment analysis. This proposed algorithm is divided into three phases as:

* Data filtration
* Training model
* Testing model

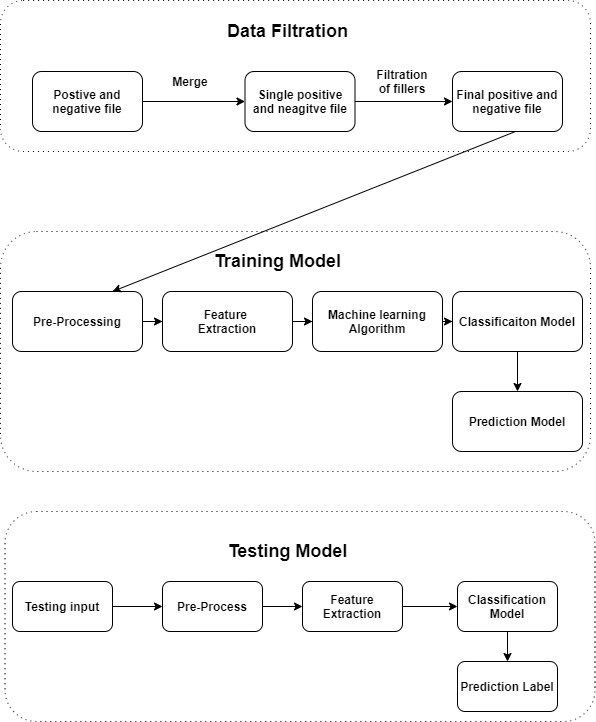


Figure 3-Illustration of proposed system using sentiment analysis

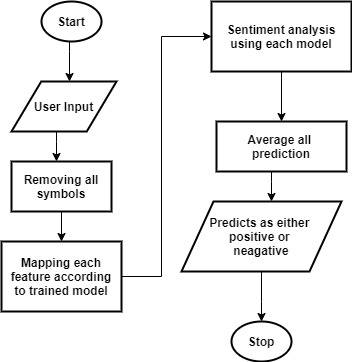


Figure 4-Diagram for testing proposed model on dataset

In this testing model on dataset, user can test and analysis the respective model by performing preprocessing over the given input. We preprocessed the dataset by removing the symbol and number. The dataset which are already trained is used for prediction. Then the model accuracy is tested.

Initially, we researched about the model feature, its broader scope, implementation advantages, and difficulties via internet. This helps us to be more familiar with the product review model.

While involve in the research with the product review model, by knowing their online opinions, we make the system which helps to give rating of product by understanding their comments on online product. The benefits of developing this model is to understand peoples how good, average and bad product who are shopping products through online.

We research many algorithms and we find Naïve Bayes which is more efficient and suitable on review type model. The benefits of using Naïve Bayes algorithm is that it can handle both continuous and discrete data, it can make probabilistic predictions, it requires less training data. Some disadvantages using this algorithm is data scarcity, chances of loss of accuracy.

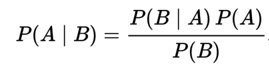
## 3.4 Naive Bayes

Naive Bayes is one of the most commonly used algorithms in machine learning. This is based on the collection of classification algorithms based on Bayes Theorem. Naive Bayes is not just a one algorithm, but it is the mix of algorithms that all share a common principle, that every feature being classified is independent of the value of the any other feature. Let us take an example of a watermelon, watermelon has features like it is green in color, diameter of about 6”, oval in shape. If we consider a Naïve Bayes classifier to identify this fruit then this Naïve Bayes classifier will treat each of these features to contribute independently to the probability that the fruit is a watermelon, regardless of any correlation between the features.  Since features are not always independent this is the shortcoming of Naïve Bayes algorithm and so it is labelled as “naive”.

Although it is a simple idea, Naive Bayes can work more efficiently than other sophisticated algorithms and so it is used in many real-world applications now-a-days. For spam detection and document classification also, this can be used. Also, we are used naïve Bayes in our project for product review.

**Bayes’ Theorem**

This lets us examine the probability of an event based on the prior knowledge of any event that related to the former event. So, for example, the probability that price of a house is high, can be better assessed if we know the facilities around it, compared to the assessment made without the knowledge of location of the house. Bayes’ theorem does exactly that.



**Above equation gives the basic representation of the Bayes’ theorem. Here A and B are two events** and,

P(**A|B**): the conditional probability that event **A** occurs, given that **B** has occurred. This is also known as the posterior probability.

P(**A**) and P**(B**): probability of **A** and **B** without regard of each other.

P(**B|A):** the conditional probability that event **B** occurs, given that **A** has occurred.

# 4. Development Model

## 4.1 Development Tools

* Programming Language: Python 3
* IDE: PyCharm
* Web Scraping tools: PySpider, Beautiful soup
* Framework: Django for Web Development
* Libraries: Pandas, NLTK, String, Collection
* Algorithm: Naive Bayes, Random Forest., K-means cluster

## Requirement Analysis

## 4.2.1 Functional Requirement

The functional requirements of the Sentiment Analysis System for Product Review are as follows:

* The system can perform the classifier training process and display the model in the form of feature sets of the term data from the training data.
* The system can display the test data result and display confusion matrix generated from the classifier testing.
* The system can display a set of movie review dataset terms derived from tokenizing, filtering, and stemming processes.
* The system can display sentiment analysis result derived from reviews submitted by users.

## 4.2.2 Non-Functional Requirement

Meanwhile, the non-functional requirements of the Sentiment Analysis System for Product Review are as follows:

* The system can run in various web browsers which support the system environment.
* The system gives a fast response.
* The system has a user-friendly interface design.

## 4.3 System Flowchart

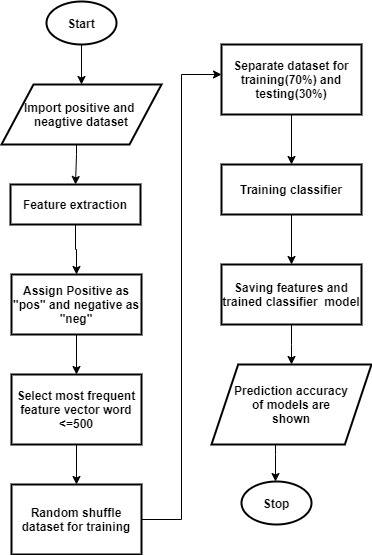


Figure 5-Flowchart for Machine Learning Implementation

# 5. Working Schedule

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Start Date | End Date | Duration |
| Requirement gathering | 20-Feb-2020 | 27-Feb-2020 | 7 |
| Coding | 28-Feb-2020 | 30-Mar-2020 | 32 |
| Unit testing | 1-Apr -2020 | 8-Apr-2020 | 8 |
| System Testing | 9-Apr-2020 | 17-Apr-2020 | 8 |
| Bug Fixes | 18-Apr-2020 | 10-May-2020 | 12 |
| Final Testing | 3-May-2020 | 20-May-2020 | 17 |

Table 1-Working Schedule

Figure 6-Gantt Chart

# 6. Expected Result

We will retrieve the data from different websites and processed it for its sentiment. The retrieved data will be the review towards some products. The sentiment of that data will help to find how the customer thinks about those products. From this project the customer can find the product review/rating which was analyzed by the review of another customer. The new customer can find how the old user of those product think about the products which will help them to find whether to buy it or not.

# References

|  |  |
| --- | --- |
| [1] | L. a. V. Pang, Movie reviews, Philadelphia,PA, 2002, pp. 79-86. |
| [2] | C. d. Albornoz, A joint model for feature mining and sentiment analysis for product review rating, Dublin,Ireland: Procedings of ECIR-2011, 2011. |
| [3] | P. D. Turney, Semantic orientation applied to unsupervised classification of reviews, Philadelphia,USA: Association for Computational Linguistic(ACL'02), 2002. |
| [4] | A.-M. P. a. O. Etzioni, Extracting product features and opinions from reviews, Vancouver,Canada: Human Language Technology Conference and NLP Conference(HLT/EMNLP'05), 2005. |