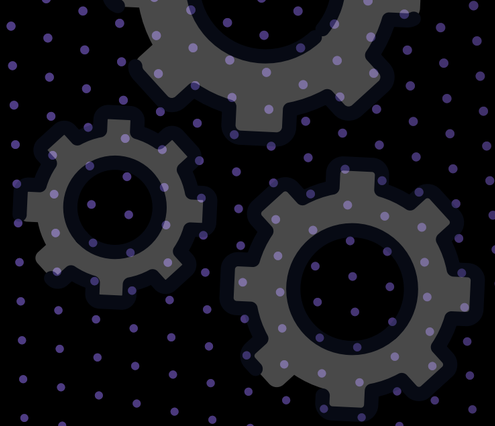


Sentiment Analysis using Machine Learning Algorithms

Sentiment Analysis : What You Need to Know

Introduction

Problem Statement

Existing body of work

Analysis

Our Approach

Final Results

INTRODUCTION

- Sentiments are expressed opinions or views, and they can be of various types like positive-negative, neutral, happy-sad, etc.
- Sentiment analysis is a tool used to analyze texts for polarity, i.e., positive to negative.
- Training the machines automatically to learn to detect sentiment without human input.
- These models can see beyond mere definitions like sarcasm, context, or misapplied words.
- Sentiment Analysis is a powerful approach that enables companies to understand the user emotions in their marketing campaigns.
- Understanding consumer psychology not only allows them to alter their product roadmap with more precision.

Timeline (After Mid-Semester)

Week 1

Used Count Vectorizer

Week 2

Fine tuning the vectorizer parameters and and use of unigrams and bigramas

Week 3

Applied Logistic Regreesion in 2-Class dataset

Week 4

Use GridSearchCV and RandomizedSearch CV to obtain optimal hyper-parameter.

Week 5

Visualization of Data and Results.

Problem Statement

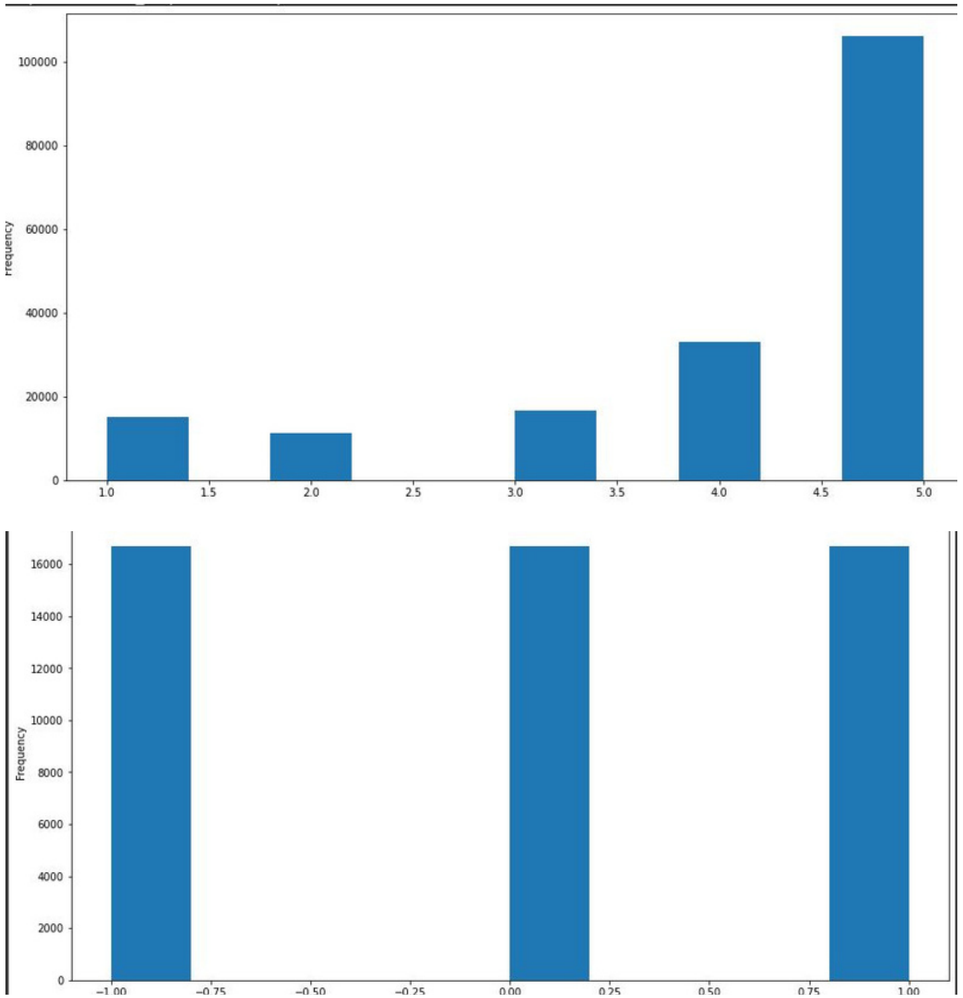
TO DETERMINE WHETHER THE INPUT TEXT CONTAINS NEGATIVE OR POSITIVE SENTIMENT.
DATASET USED: AMAZON BABY PRODUCTS.

Existing Body of Work

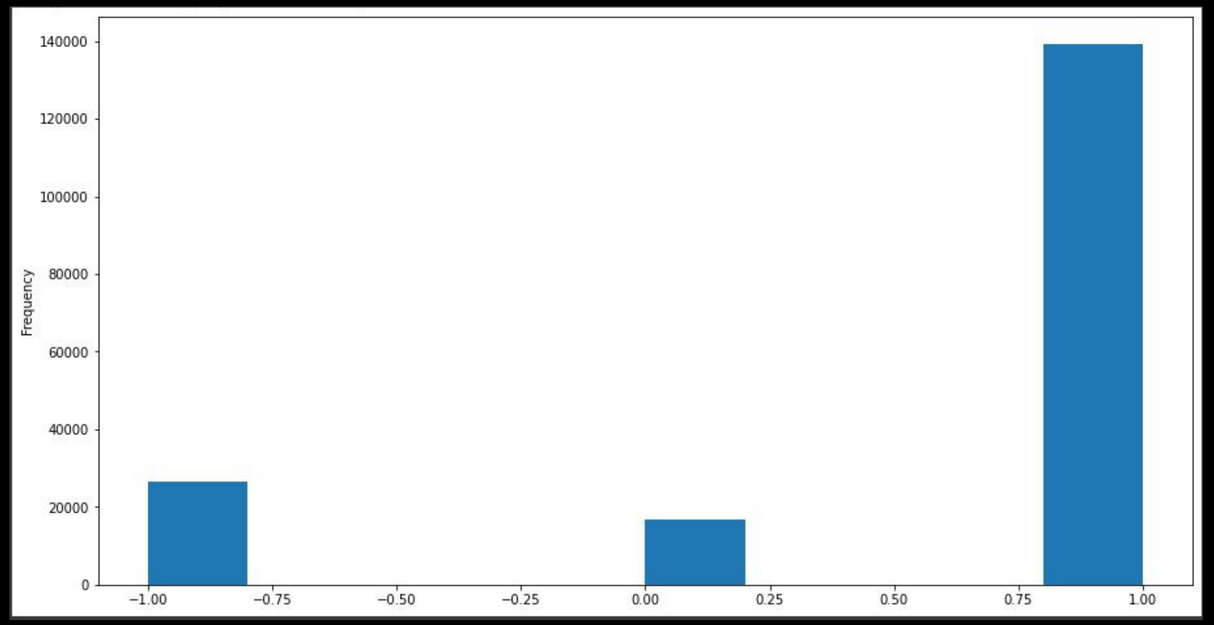
- The paper [3] proposes using maximum entropy techniques for text classification. They conducted their research in 1999 and used the Maximum Entropy Algorithm, through which they achieved an average accuracy of 72% in different datasets.
- In one of the articles, the author performed sentiment analysis with a Naive Bayes Classifier. The author divided the dataset into positive, negative, and neutral values by rating them larger than 1, lower than 1, and 1, respectively.[3]
- In this blog, the author has explained the NaiveBayes algorithm in detail and explained how it works with examples. the author has selected google play reviews as their data set and achieve an accuracy of 85% in the dataset.[2]

Data Analysis

- Selected Amazon Baby dataset for the models.
- The data was positively skewed with more than 77% of data having a rating of more than 3 i.e. positive hence we made them equal.
- As we only required positive, negative, and neutral sentiment, we found out for which sentiment we had the lowest data and then decreased the size of other sentiments to make divide the dataset among all classes equally.



- Initially the model had 180,000 rows of data of which 140,000 we having positive sentiment.
- Neutral review were the least of all of them so we reduced the no. of rows from Positive and Negative sentiments to make it equal.
- For 2 Class classification we dropped all the neutral values and reduced the positive rows to make it equal to negative rows.



Feature Engineering

We have used unigrams and bigrams using Count Vectorizer to further increase the precision of the model and make it more accurate for real-world data.

Lemmatization of the reviews so that the no. of words in the vocabulary could be reduced. For effective Lemmatization, we used POS tagging, which tags each word as a noun, verb, adjective, etc.

Converted the content to lowercase, then removed all the stop words with the help of the stop words parameter in the vectorizer.

Initial Approach

Pre-processed the data and fine tuned it for the models.

Vectorized the data using tfidf vectorizer.

Only applied Multinomial Naive Bayes.

Current Approach

Feature Engineering aspects were added to the model such as the use of unigrams and bigrams.

Divided Data into two different Datasets

Applied the Decision Tree Classification

Used Logistic Regression

Used GridSearchCV method for obtaining optimal hyperparametres

Final Results

After fine-tuning the data and pre-processing it various times in different ways, the results we received are:

Multinomial Naive Bayes

(3-Class) An accuracy of 65.87%.
(2-Class) An accuracy of 86.91%.

Decision Tree Classifier

(3-Class) An accuracy of 42.39%.
(2-Class) An accuracy of 86.91%.

Logistic Regression

An accuracy of 89%

Conclusion

The models worked very well on the real-world data, which was not the case before.

All our models performed well on the dataset except the decision tree model

the user had compared the product, so comparative words like “Better,” which convey a positive sentiment, may show the sentiment is negative.

Role of each team member

Kalp Ranpura
1920134

Extraction of dataset and preprocessing it for model fitting. Created PPT and Report

Dhrumil Mistry
1920184

Preprocessing the data and implementing different models. Created PPT and Report

Rashika Jain
2020178

Contributed to the Report and visualized the data and results.

References

[1] Sentiment Analysis & Machine Learning. MonkeyLearn Blog. (2020, April 20). Retrieved March 20, 2022, from https://monkeylearn.com/blog/sentiment-analysis-machine-learning/#:~:text=Sentiment%20analysis%20is%20a%20machine,detect%20sentiment%20without%20human%20input

[2] Performing sentiment analysis with naive Bayes classifier! Analytics Vidhya. (2021, July 13). Retrieved March 20, 2022, from https://www.analyticsvidhya.com/blog/2021/07/performing-sentiment-analysis-with-naive-bayes-classifier/

[3] Foy, P. (2021, July 26). Naive Bayes for sentiment analysis & Natural Language Processing (NLP). MLQ.ai. Retrforved March 20, 2022, from https://www.mlq.ai/sentiment-analysis-with-naive-bayes/

[4] Kuzminykh, N. (2020, October 23). Sentiment Analysis in python with textblob. Stack Abuse. Retrieved March 20, 2022, from https://stackabuse.com/sentiment-analysis-in-python-with-textblob/

Timeline (After Mid-Semester)

Week 1

Used Count Vectorizer

Week 2

Fine tuning the vectorizer parameters and and use of unigrams and bigramas

Week 3

Applied Logistic Regreesion in 2-Class dataset

Week 4

Use GridSearchCV and RandomizedSearch CV to obtain optimal hyper-parameter.

Week 5

Visualization of Data and Results.