A Project Report

on

Amazon Reviews Sentiment Analysis

carried out as part of the course CC1634 Submitted by

Shardul Negi
169104101
6th Btech CCE
Saksham Agarwal
169104142
6th Btech CCE

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

In

Computer & Communication Engineering

MANIPAL UNIVERSITY
JAIPUR

Department of Computer & Communication Engineering,
School of Computing and IT,
Manipal University Jaipur,

April, 2019

CERTIFICATE

This is to certify that the project entitled "***Amazon Reviews Sentiment Analysis***" is a bonafide work carried out as part of the course ***Minor Project***, under my guidance by ***Shardul Negi***, student of ***Btech CCE*** at the Department of Computer & Communication Engineering, Manipal University Jaipur, during the academic semester ***6th***, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer & Communication Engineering, at MUJ, Jaipur.

Place:

Date:

Signature of the Instructor (s)

DECLARATION

I hereby declare that the project entitled "***Amazon Reviews Sentiment Analysis*** " submitted as part of the partial course requirements for the course **Minor Project,** for the award of the degree of Bachelor of Technology in Computer & Communication Engineering at Manipal University Jaipur during the ***(6th, Jan-May 2019)*** semester, has been carried out by me. I declare that the project has not formed the basis for the award of any degree, associate ship, fellowship or any other similar titles elsewhere.

Further, I declare that I will not share, re-submit or publish the code, idea, framework and/or any publication that may arise out of this work for academic or profit purposes without obtaining the prior written consent of the Course Faculty Mentor and Course Instructor.

Signature of the Student:

Place:

Date:

CERTIFICATE

This is to certify that the project entitled "***Amazon Reviews Sentiment Analysis***" is a bonafide work carried out as part of the course ***Minor Project***, under my guidance by ***Saksham Agarwal***, student of ***Btech CCE*** at the Department of Computer & Communication Engineering, Manipal University Jaipur, during the academic semester ***6th***, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer & Communication Engineering, at MUJ, Jaipur.

Place:

Date:

Signature of the Instructor (s)

DECLARATION

I hereby declare that the project entitled "***Amazon Reviews Sentiment Analysis***" submitted as part of the partial course requirements for the course **Minor Project,** for the award of the degree of Bachelor of Technology in Computer & Communication Engineering at Manipal University Jaipur during the ***(6th, Jan-May 2019)*** semester, has been carried out by me. I declare that the project has not formed the basis for the award of any degree, associate ship, fellowship or any other similar titles elsewhere.

Further, I declare that I will not share, re-submit or publish the code, idea, framework and/or any publication that may arise out of this work for academic or profit purposes without obtaining the prior written consent of the Course Faculty Mentor and Course Instructor.

Signature of the Student:

Place:

Date:

Table of Contents

INTRODUCTION

We have done amazon product review analysis based on the reviews provided by the consumers/customers.

Reviews are a very important aspect for any E-Commerce hub because reviews determine the product life, future demand and supply of the particular product and the related customer satisfaction level and hence studying them becomes important.

Statistics show that reading a text and understanding is much more tedious and time taking and in today's world with so much data around us, we cannot adopt the age old methodology of reading and understanding and when talking about big institutions, the problem grows exponentially and this is where the concept of star ratings /images comes to existence.

SCOPE OF THE WORK

Sentiment analysis is a uniquely powerful tool for businesses that are looking to measure attitudes, feelings and emotions regarding their brand. To date, the majority of sentiment analysis projects have been conducted almost exclusively by companies and brands through the use of social media data, survey responses and other hubs of user-generated content. By investigating and analyzing customer sentiments, these brands are able to get an inside look at consumer behaviors and, ultimately, better serve their audiences with the products, services and experiences they offer.

The future of sentiment analysis is going to continue to dig deeper, far past the surface. This forecast also predicts broader applications for sentiment analysis – brands will continue to leverage this tool, but so will individuals in the public eye, governments, nonprofits, education centers and many other organizations.

PRODUCT SCENARIOS

With the help of Natural Language Processing, machines are able easily to pick on what phrases and words are generally used by humans. The system finds what the user is actually searching for by using its understandings of the kind of language and the structure of sentence used. It detects patterns and creates links between the messages to derive the meanings of unstructured texts.

The performance of the NLP model is directly proportional to the amount of data and the quality of data that it is fed. The eCommerce sites have to consider the problem with slags and synonyms that works differently in different areas.

REQUIREMENT ANALYSIS

1. Functional Requirements

i. Acquiring a product reviews dataset which is big in terms of number of reviews and reviewers so as to obtain better results and accuracy.
ii. Cleaning the dataset obtained so as to remove the deduplication, html tags, stop words etc.
iii. Using various text featurisation techniques such as bag of words, n-gramming, tf-idf, word to vector etc.
iv. It should be able to handle english text.

2. Non-functional Requirements

i. The syntax rules of Python allowed us to express concepts without writing additional comments. At the same time, Python, unlike other programming languages, emphasized on code readability, and allowed us to use English keywords. Hence, we used Python to build our model . The readable and clean code base helped us to maintain and update the model without putting extra time and effort.

ii. The Jupyter Notebook is an open-source web application that allowed us to create and share documents that contain live code, equations, visualizations and explanatory text. It's Uses include: data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more.

3. Use Case Scenarios

i. Retailers who are new to the business can predict and study consumer behaviour, consumer expectation and launch the products with the required specifications.
ii. Developing Recommender Systems and recommending products to the consumers.
iii. E-commerce hubs can get an insight into the advertisement of brands (What, When and How a product needs to be advertised).
iv. Competitive monitoring between brands can be done.

4. Other Methodologies

i. Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

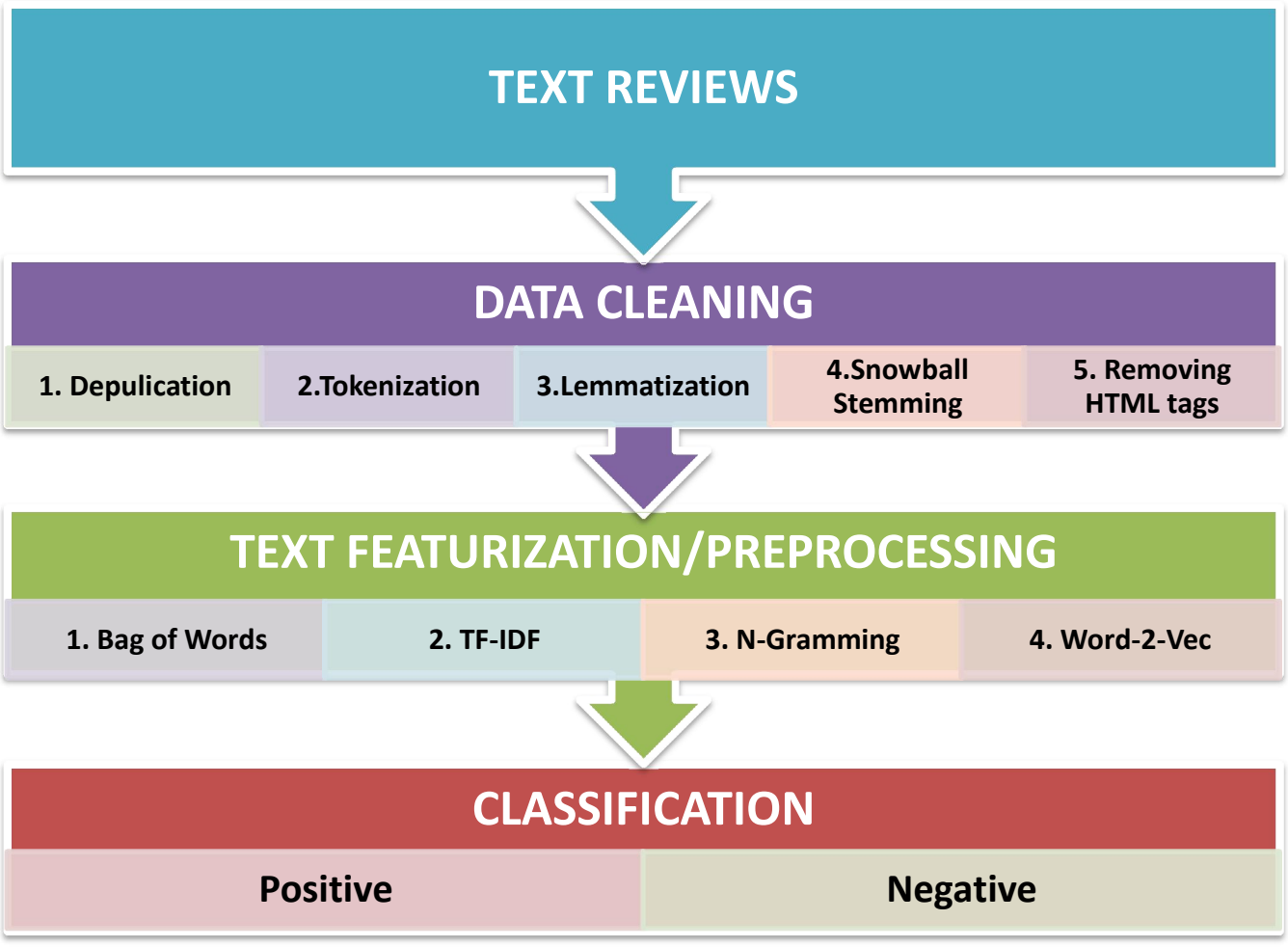
ii. NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

System Design

1. Design Goals

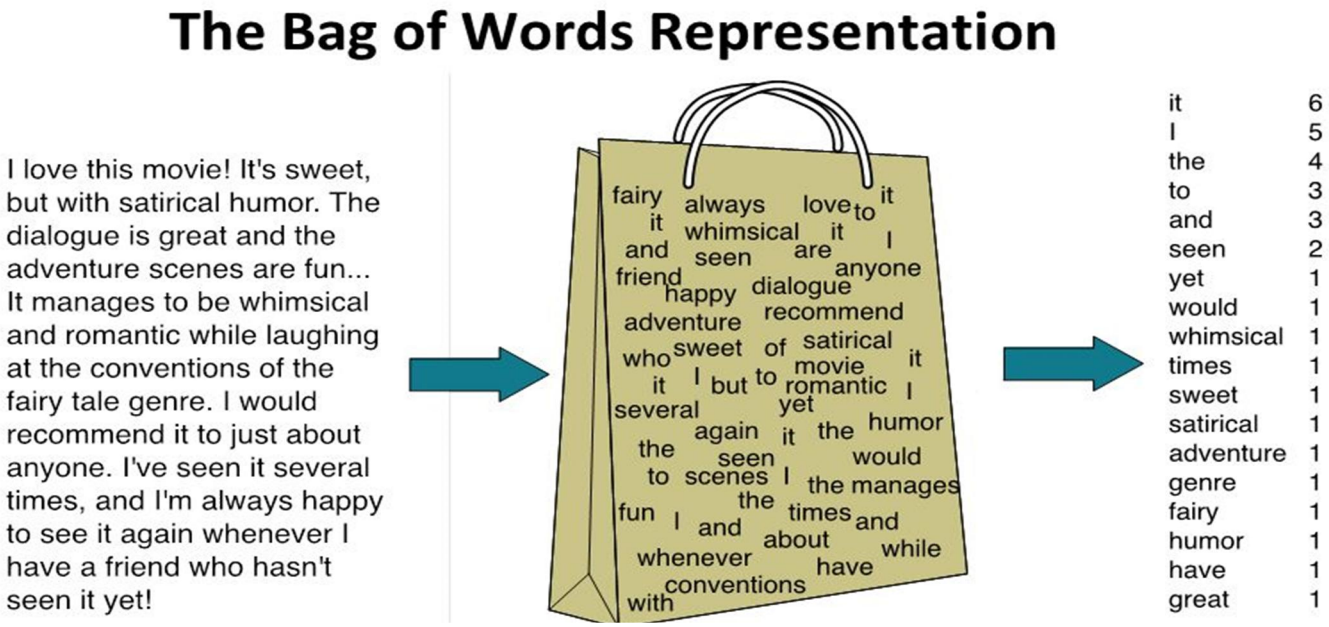
Our main objective was to use a text data-set which consists of reviews from people around the world for varied food product range and uses the text of the reviews and its summary so as to build a machine learning model that is successfully able to classify the reviews based on their sentiment polarity and weight that has been obtained using the techniques such as bag of words and term frequency-inverse document frequency, as positive and negative and assign them binary values 0 and 1 and also classify the new input text based on its learning from the reviews.

2. System Architecture



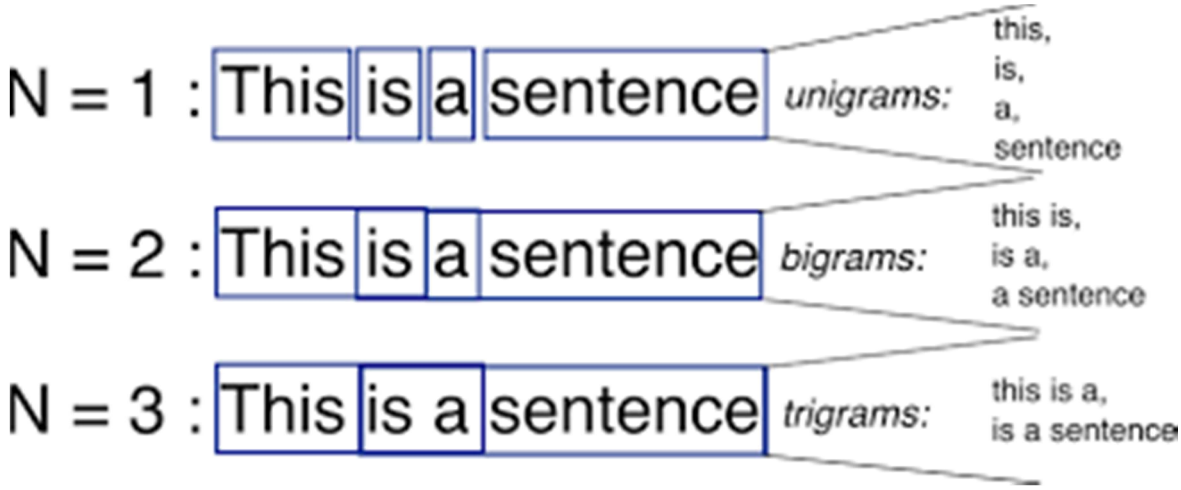
3. Detailed Design Methodologies

I. BAG OF WORDS



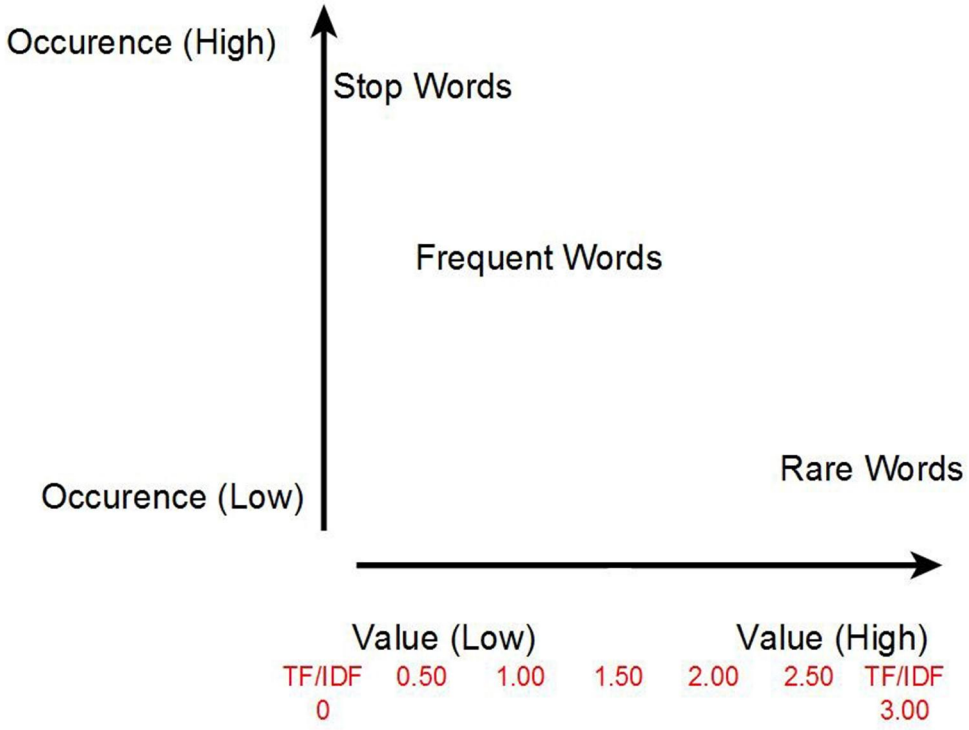
II. BI-GRAM AND N-GRAM

N-gram model predicts the occurrence of a word based on the occurrence of its N – 1 previous words. So here we are answering the question – how far back in the history of a sequence of words should we go to predict the next word? For instance, a bigram model (N = 2) predicts the occurrence of a word given only its previous word (as N – 1 = 1 in this case). Similarly, a trigram model (N = 3) predicts the occurrence of a word based on its previous two words (as N – 1 = 2 in this case).



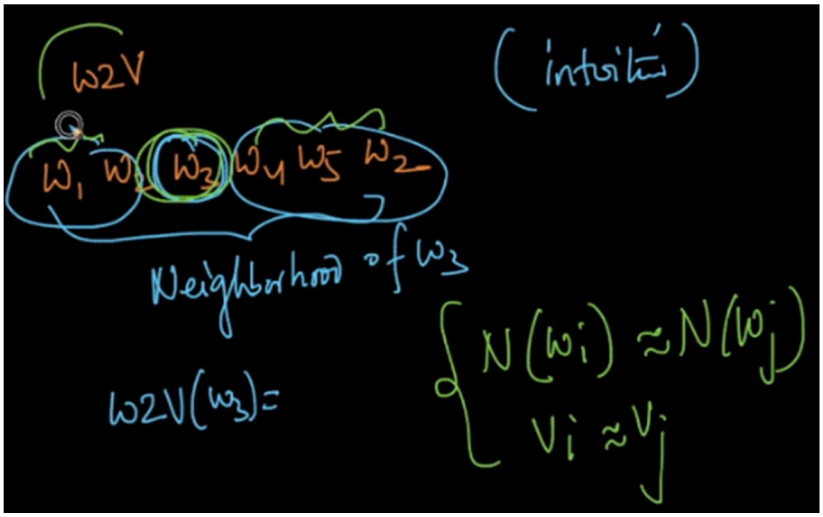
III. TF-IDF

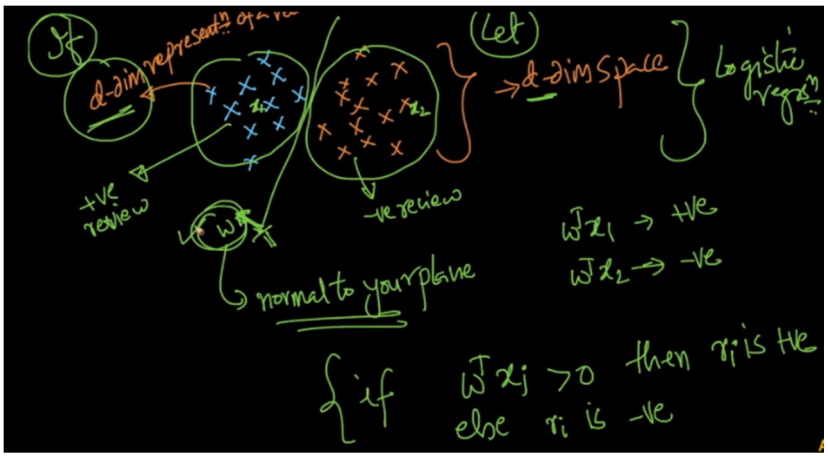
Tf-idf stands for *term frequency-inverse document frequency*, and the tf-idf weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.



IV. WORD2VEC

Word to vector is a text featurization technique which uses the semantics of a sentence and converts a given word into a dense d dimensional vector ,and helps to identify the closeness between two words and also detect relationships amongst them and it helps us to leverage the power of linear algebra for text classification ,it is a deep learning concept developed by google.





Work Done

1. Development Environment

i. Python (Version: 3.6)
ii. Jupyter Notebook

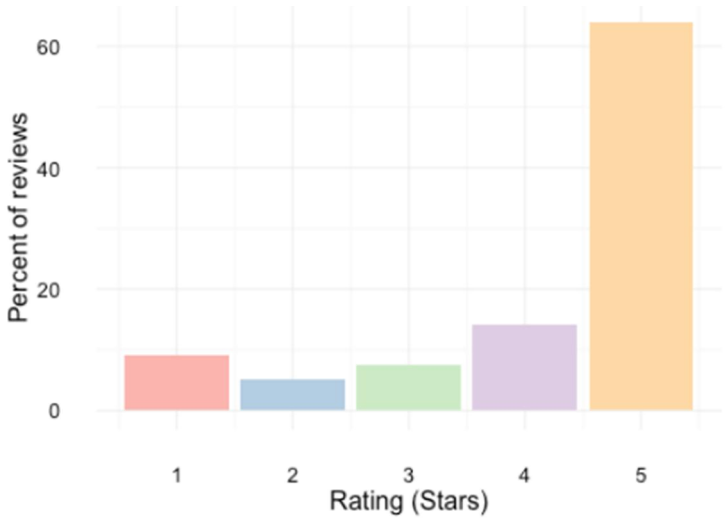
2. Results and Discussion

We used a database of over 500,000 reviews of Amazon fine foods that is available via Kaggle.

Performed some basic exploratory analysis to better understand reviews:

I. Ratings distribution

I first looked at the distribution of ratings among all of the reviews. We see that 5-star reviews constitute a large proportion (64%) of all reviews. The next most prevalent rating is 4-stars(14%), followed by 1-star (9%), 3-star (8%), and finally 2-star reviews (5%)

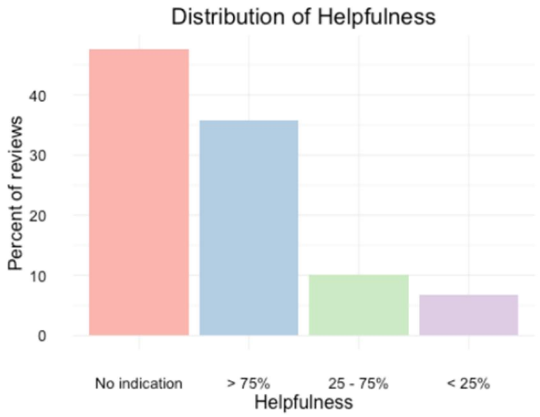


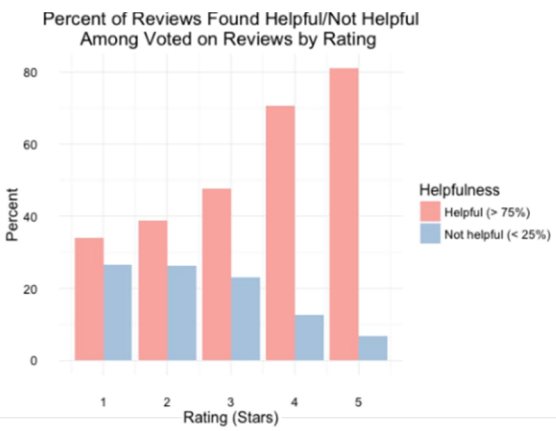
Positive words (Most Commonly distributed)

Negative words (Most commonly used)

II. Helpfulness

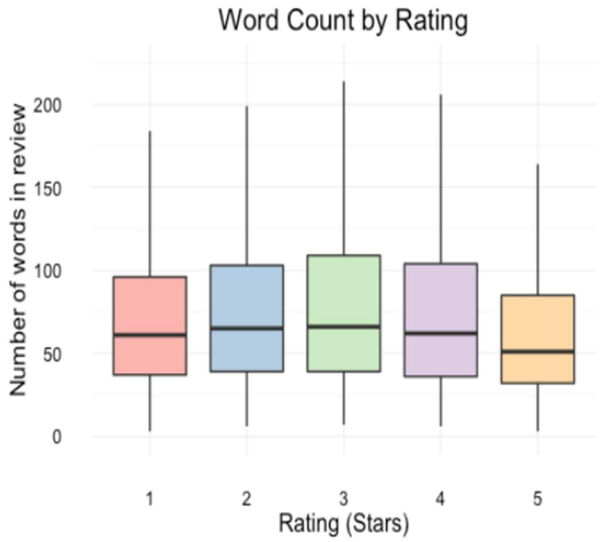
Reviews are voted upon based on how helpful other reviewers find them. The most helpful reviews appear near the top of the list of reviews and are hence more visible. As such, I was interested in exploring the properties of helpful reviews.

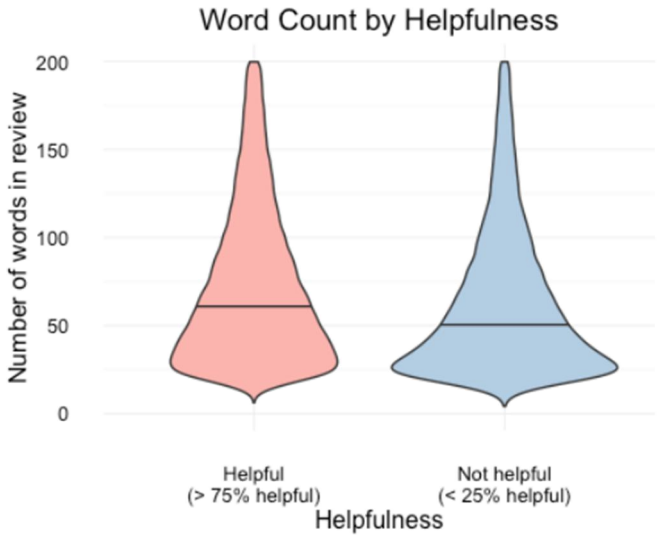




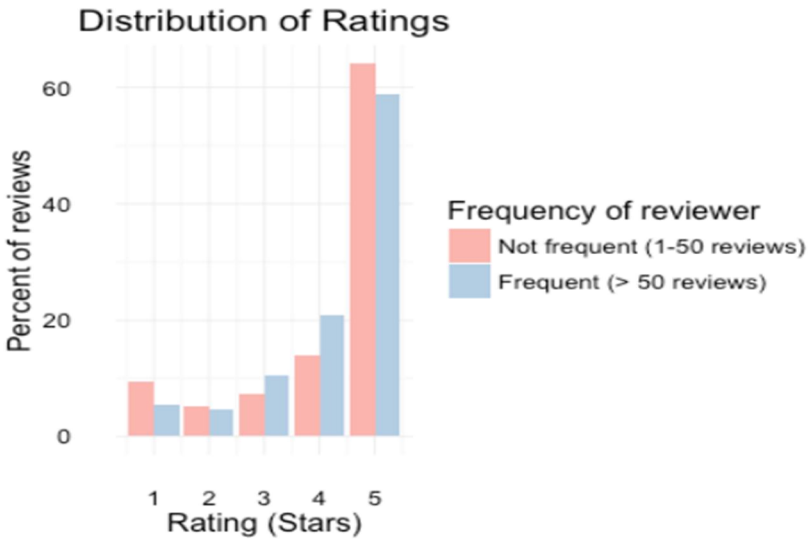
III. Word count

I wanted to see how word count related to the other properties of reviews already discussed, including rating and helpfulness.

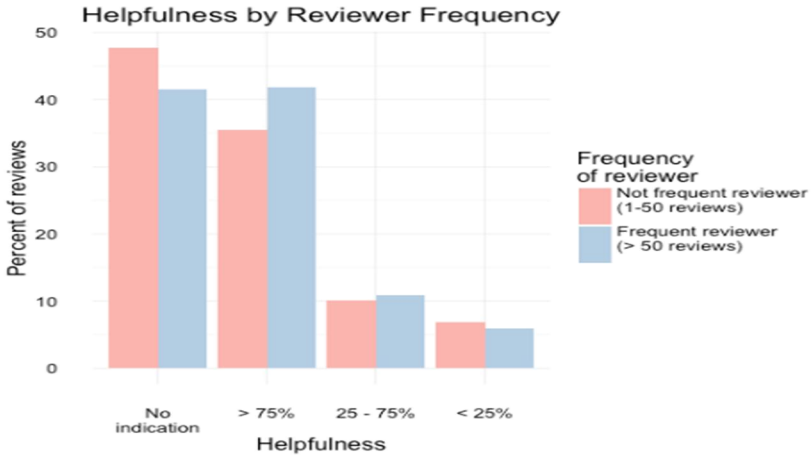




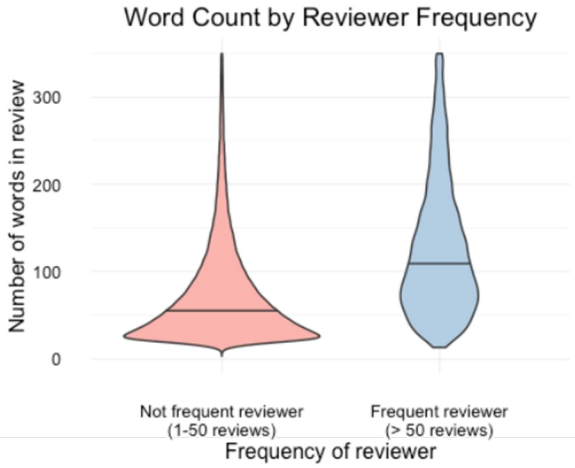
IV. Distribution of ratings



V. Helpfulness by Reviewer Frequency



VI. Word count by Reviewer Frequency



VII. Model Evaluation

```
In [21]: print(model.predict(vect.transform(['Food is good','Food is bad'])))
         [1 0]

In [20]: print(model.predict(vect.transform(['Food is not good','Food is not bad'])))
         [0 1]
```

```
1  lst = model.predict(vect.transform(['Cholle Bhature is not tasty','Cholle Bhature is not bad']))
2
3  print(lst)
4
5  for i in lst:
6      if i == 1:
7          print('This review is postive.')
8      else:
9          print('This review is negative.')
10
```

```
[0 1]
This review is negative.
This review is postive.
```

3. Individual Contribution of project members

I. **Saksham Agarwal** - Applying techniques such as lemmitization and snowball stemming on the data set and preprocessing the text using techniques such as N-Gramming and TF-IDF that help in improving the efficiency and accuracy of the model and hence help the model to classify the text accurately as positive and negative and evaluating the model with the help of test data and AUC technique.

II. **Shardul Negi** - Performed data cleaning techniques such as deduplication and getting rid of duplicate entries and removing the entries that had suspected illegal entries and removing html tags and dividing the data into tokens using tokenisation .Applying the bag of words text featurisation technique and calculating the frequency of words and developing a word to vector model using gensim and finally applying logistic regression analysis .

FUTURE WORK

1. Detecting unfair reviews given by malicious algorithms.

2. Eliminating reviews which are positive but given a less rating.

3. Eliminating reviews which are negative but given a more rating.

4. Detecting and Eliminating artificial reviews given for data training purposes.

5. Improving the efficiency of the current implemented systems.

CONCLUSION

1. The application provides the retailer the rating of his/her product irrespective of what users have mentioned in the amazon website.

2. This application also provides an oversight of how many users were interested in the products, which could be biased using the number of Male and Female reviewers in the database.

3. The interestingness with respect to the product can also be measured with the number of positive, neutral and negative reviews obtained.

4. This analysis could help the retailers make subtle changes either to the products service or with how the product looks according to the number of users.