@SE-190018

UGC & Govt. Approved
Sonargaon University [SU]
সোনারগাঁও ইউনিভার্সিটি (এসইউ)
WE WILL RISE UP WE WILL SHINE
147/I, Green Road, Tejgaon, Dhaka

Sentiment Analysis on Choice Home Warranty Product & Service Reviews

Submitted by

Kazi Saiduzzaman
ID: CSE1601007030

Md. Riajur Rahman
ID: CSE1601007031

Md. Aftabuzzaman
ID: CSE1601007062

Shafirul Islam
ID: CSE1601007028

Supervised By
Emam Hossen
Lecturer
Department of Computer Science and Engineering
Sonargaon University

This project has been submitted to the Department of the Computer Science & Engineering at Sonargaon University in the partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering

September, 2019

Department of Computer Science & Engineering
Sonargaon University, Bangladesh

BONA FIDE CERTIFICATE

Certified that this project report **"Sentiment Analysis on Choice Home Warranty Product & Service Reviews"** is the bona fide work of "Kazi Saiduzzaman (CSE1601007030), Md. Riajur Rahman (CSE1601007031), Md. Aftabuzzaman (CSE1601007062), Shafirul Islam (CSE1601007028)" who carried out the project work under my supervision. This is to further certify to the best of my knowledge, that this project has not been carried out earlier in this institute and the university.

Signature of the Supervisor

Emam
5.9.19

Emam Hossen

Lecturer

Department of Computer Science & Engineering

Certified that the above-mentioned project has been duly carried out as per the norms of the college and statutes of the university.

Bulbul Ahamed

Associate Professor & Head

Department of Computer Science & Engineering

ACKNOWLEDGEMENTS

First and for most, we would like to all praises and thanks to the almighty Allah for giving us courage to complete this project.

We would also like to express our sincere thanks to our supervisor **Emam Hossen**, Lecturer, Department of Computer Science & Engineering, Sonargaon University for his continual guidance, support and encouragement throughout the duration of this project, we consider it a privilege to have had the opportunity to work with him and share his valuable knowledge and expertise. We are deeply grateful to him for detailed and constructive comments and for his important support throughout this work.

During this study, we have collaborated with many friends for whom we have greater regard. We wish to extend our warmest thanks to all those who have helped and supported ours. We express our sincerest thanks to the Department of Computer Science & Engineering, who gave us the opportunity to work with team and gave us untiring help during difficult moments.

We warmly thank to **Bulbul Ahamed,** Associate Professor & Head, Department of Computer Science & Engineering, Sonargaon University for his valuable advice and moral support. His expensive discussions around work and interesting explorations in operations have been very helpful for this study. We also express our thanks to Vice Chancellor **Professor Dr. Md. Abul Bashar,** who has given us the opportunity to study in such a decent environment.

Finally, we would like to dedicate this project to our parents for their love, encouragement and help throughout the project.

ABSTRACT

"Choice Home Warranty Product & Service" has emerged as one of the online platforms to provide online product selling and warranty services to the customers. This is a project to implement a "Sentiment Analysis on Choice Home Warranty Product & Service Reviews". To classify the reviews of every customer either positive or negative is difficult and time consuming and this situation can make obstruct against a corporate progress. This project is basically analyzing of the customer reviews to determine the customer reactions on the basis of polarity and subjectivity. This polarity is then converted to a scale of -1.000 to 1.000 (where 0.000 is average) and thus the rating of individual product or service is obtained. The main purpose of "Sentiment Analysis on Choice Home Warranty Product & Service Reviews" is to improve the relationship with the customers and provide the better services to customers as per customers demand by analyzing the reviews either positive or negative.

"Sentiment Analysis on Choice Home Warranty Product & Service Reviews" in this project is included few modules. The modules included are Data Scrapping, Data Analyzing, Natural Language Processing, Data Cleaning, Lemmatization, Review Classification, Graph Generating. Data Scrapping to import the web data from "Choice Home Warranty Product & Service Reviews" web page and sort out the reviews. Data Analyzing module note out the minimum unit of reviews. Natural Language Process basically defines the numerical value from each word of reviews. Data Cleaning module reduces the anomaly and special characters from the reviews. Lemmatization refers to doing thing properly with the use of vocabulary and morphological analysis of words to remove the inflectional endings and to return the dictionary form of a word. Review Classification is the target module to classify the reviews on basis of polarity and subjectivity. Finally, a bar chart is produced from the polarity to make visualize the analysis result by the Graph Generating module.

The tools used to develop this "Sentiment Analysis on Choice Home Warranty Product & Service Reviews" included Jupyter Notebook, Python packages: bs4, pandas, numpy, nltk.corpus, textblob, matplotlib.pyplot, and random.

Table of Contents

Chapter 01 Introduction 1

1.1 Motivation and Problem Statement **2**

1.2 Project Scope **2**

1.3 Project Objective **3**

1.4 Impact, Significant and Contribution **3**

1.5 Background Analysis **4**

Chapter 02 Literature Review 6

2.1 Introduction **7**

2.1.1 Natural Language Processing 7

2.1.2 Natural Language Toolkit (NLTK) 7

2.1.3 Stop Words 7

2.1.4 Lemmatization 8

2.1.5 TextBlob Package 9

Chapter 03 System Implement Process 10

3.1 Process **11**

3.2 Tools **12**

3.3 Implementation Issues and Challenges **14**

3.4 Timeline **15**

3.5 Requirement Specification **16**

Chapter 04 System Implementation 17

4.1 System Implementation **18**

4.1.1 Data Scrapping 18

4.1.2 Data Analyzing 20

4.1.3 Natural Language Processing (Count Stop Words) 21

4.1.4 Data Cleaning 22

4.1.5 Lemmatization 26

4.1.6 Review Classification by Polarity & Subjectivity 28

4.1.7 Graphical Representation of Polarity 30

Chapter 05 Result Analysis 31

5.1 Result Analysis **32**

Chapter 06 Conclusion 33

6.1 Conclusion **34**

6.2 Future Enhancement **34**

References 35

List of Figure

Figure – 3.1: Process of System Implementation ---------- 11
Figure – 3.2: Relation among the implementation tools ---------- 14
Figure – 4.1.1.1_1: Process of Python requests package ---------- 18
Figure – 4.1.1.1_2: Process to input https link for scrapping web data ---------- 18
Figure – 4.1.1.1_3: Status code for scrapping operation ---------- 19
Figure – 4.1.1.1_4: Gathered Source code as continuous text ---------- 19
Figure – 4.1.1.2: Order List of review data from source code ---------- 20
Figure – 4.1.2: Summarized parameters by analysis ---------- 21
Figure – 4.1.3_1: List of English stop words filtered by nltk.corpus library ---------- 21
Figure – 4.1.3_2: Output of stop word ratings ---------- 22
Figure – 4.1.4_1: Tokenization of a sentence ---------- 24
Figure – 4.1.4_2: Lowercase of review data ---------- 24
Figure – 4.1.4_3: Output after reducing of punctuations ---------- 25
Figure – 4.1.4_4: Output after data cleaning process ---------- 26
Figure – 4.1.5_1: Inflected form of word ---------- 27
Figure – 4.1.5_2: Output of lemmatized data ---------- 28
Figure – 4.1.6_1: Output of polarity and subjectivity parameters ---------- 29
Figure – 4.1.6_2: Output of finalized summary ---------- 30
Figure – 4.1.6_3: Graphical representation based on polarity ---------- 30

List of Table

Table-3.4_1: Gantt chart of System Implementation ---------- 15

Chapter 01

Introduction

1.1 Motivation and Problem Statement

Start in Today's era the e-commerce is developing rapidly these years, buying products online has become more and more fashionable owing to its variety of options, low cost value (high discounts) and quick supply systems, so abundant folks intend to do online shopping. In the meantime, the standard and delivery of merchandise is uneven, fake branded products are delivered. Before buying any product from online customers are more likely to check the reviews of that product. There are a large number of review websites. These sites or social networking sites like Facebook page provide reviews of various products and services. Every type of product has some reviews from the consumers.

Regarding this scenario, sentiment analysis is a brilliant term of artificial intelligence for understanding the customer or audience's sentiment in the basis of polarity and subjectivity. Choice Home Warranty Product and Service has a lot of customers in online market and the reviews of a particular product or service from the customers are increasing day to day. "Sentiment Analysis on Choice Home Warranty Product & Service Reviews" scraps the review data as data set and completes the classification of review texts by subjectivity/objectivity and negative/positive attitude of customers. This project presents an empirical study of efficacy of classifying product review by tagging the keyword. Thus, hereby this is a completely different approach by removing the unstructured data and then classifying reviews employing Natural Language Processing (NLP) toolkit.

1.2 Project Scope

Sentiment analysis is a method of studying the customers using linguistic tones, content, their likes, hashtags, shares, etc. to analyze the emotion conveyed. Sentiment Analysis is not only a social analytic tool but also an interesting field of study. It is a field that is still being studied, although not at great lengths due to the intricacy of this analysis. This project work can help the brands to accommodate the sentiment analysis tools into areas of customer feedback, marketing, CRM and ecommerce.

Sentiment analysis methods till now have been used to detect the polarity in the thoughts and opinions of all the users that access social media. Researchers and Businesses are very interested to understand the thoughts of people and how they respond to everything happening around them. Companies can use this project to evaluate their products and services and to improve their products and services. There is so much potential in machine learning, overtaking some of the manual labor-

based tasks that are labor intensive. Collecting opinions on the web will still requires processing that can filter out un-opinionated user-generated content and also to test the trustworthiness of the opinion and its source. Sentiment analysis by this project has more possibility and accuracy rather than manual analysis of sentiment data virtually.

There is a lot of scope in analyzing the customers review on the web. Now-a-days, there are lots of website whose post the reviews of different product or services provided by different reputed companies. Sentiment analysis will have to pace up with this change. "Sentiment Analysis on Choice Home Warranty Product & Service Reviews" can help the companies to change strategies based on customer feedback.

1.3 Project Objective

"Sentiment Analysis on Choice Home Warranty Product & Service Reviews" is a strategy for study the customer sentiment either positive or negative. A sound relationship with customer gives the service marketers a competitive edge. To get more profit by providing better product and service should be the basic target of a company. For this reason, there is need for a long-term relationship between the companies and the customers. "Sentiment Analysis on Choice Home Warranty Product & Service Reviews" is the process of customer review data mining, analyzing of customers feedback and display a graph of the result which presents the polarity values basis of positivity and negativity.

1.4 Impact, Significant and Contribution

The ability to understand sarcasm, hyperbole, positive feelings, or negative feelings has been difficult, for machines that lack feelings. Companies want to accommodate the sentiment analysis tools into areas of customer feedback, marketing, CRM, and ecommerce. Collecting opinions on the web will still requires processing that can filter out un-opinionated user-generated content and also to test the trustworthiness of the opinion and its source. This project has been able to predict with more than 80% accuracy the feelings portrayed by customers. Yet with so many limitations this is one field which is growing at great pace within many industries. Al though, by having this project, the companies will able to track the customers and fulfill their parameter. The companies can easily detect the sentiment of the customers by analyzing their review data.

1.5 Background Analysis

Sentiment analysis is a process to find opinions, emotions from text, reviews, tweets and other sources of natural languages. All the opinion/emotion is captured using natural language processing. Now-a-days 'Natural Language Processing' is becoming more and more popular because of the amount of data is getting bigger and bigger. Using natural language processing we can find out trends, popularity etc of any product, brand, or corporate. This is big field for research thus we have done our research on generating product and service reviews of "Choice Home Warrantee Product & Service" from customer/client sentiment collected from a web site whose posts the customer reviews of "Choice Home Warrantee Products & Services". In our research we have selected all the latest reviews of customers regarding product and service as our experimental datasets.

In most of the review sites available over the Internet, the reviews of products are mostly given by technical wizards. People who have immense knowledge about technologies are bound to know much about any given product as they know every little detail that is needed to be known in order to understand the efficiency and performance of a product. These types of people do not even need to use a product to judge, just a specification list works fine for them to give positive or negative reaction on the product. In the review sites, most of the reviews are given by the sort of people, which means a much more technical review. This is not actually a bad thing as their judgment is far more accurate than any average person but the problem is most of the people are not well aware of a lot of attributes mentioned on the reviews. Most of the people do not understand the reason behind a review being positive or negative. Then again it is not like general reviews are not posted on these review sites but as we know the best reviews are prioritized by the service. That is why general reviews are found by facing a lot of difficulty but people tend to look over 2-3 reviews and make their decision. Mostly non-technical people who have less knowledge of technologies tend to share their reviews and opinion on review sites. The reviews are actually their personal opinions about a product or service which are more informal and less technical in most cases. Their posts reflect the reasons of their liking or disliking a product. There is a problem though. These posts are really hard to keep track of. Even though using hashtags and other methods posts regarding any specific topic or things can be seen, keeping track of these posts is difficult. Also, most of these posts are not even taken into account due to less followers, audience type. Due to all these reasons mentioned above we thought of collecting all these posts of reviews and analyzing it to create a rating system. This system will take all the reviews of customers from "Choice Home Warrantee Product & Service"

and then process them using Natural Language Toolkit (NLTK), do a sentiment analysis on these reviews to get the polarity and subjectivity which further leads us to understand of a review.

Chapter 02

Literature Review

2.1 Introduction

In this literature review, this project going to review on the article, book and internal resources to study about the "Sentiment Analysis on Choice Home Warranty Product & Service Reviews".

2.1.1 Natural Language Processing

Natural language processing is focuses on machines and human language. It basically enables machines to understand human languages. Its main challenges are recognizing speech, understanding human language etc. The main syntaxes of natural language processing are parts of speech tagging, parsing, word segmentation, sentence breaking and terminology extraction. The main processes that we used will be described in the NLTK section.

2.1.2 Natural Language Toolkit (NLTK)

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 [3].

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language." Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more.

2.1.3 Stop Words

In computing, stop words [1] are words which are filtered out before processing of natural language data (text). Stop words are generally the most common words in a language; there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list. Some tools avoid removing stop words to support phrase search.

Any set of words can be chosen as the stop words for a given purpose. For some search engines, these are some of the most common, short function words, such as "the", "is", "at", "which", and "on". In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as "The Who", "The The", or "Take That". Other search engines remove some of the most common words—including lexical words, such as "want"—from a query in order to improve performance.

2.1.3.1 NLTK Stop Words

Natural Language Processing with Python Natural language processing (nlp) is a research field that presents many challenges such as natural language understanding [2]. Text may contain stop words like 'the', 'is', 'are'. Stop words can be filtered from the text to be processed. There is no universal list of stop words in NLP research, however the NLTK module contains a list of stop words.

2.1.4 Lemmatization

Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form [4].

In computational linguistics, lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatization depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. As a result, developing efficient lemmatization algorithms is an open area of research.

Lemmatization mapping as follows:

Am, are, is => be

Car, cars, car's, cars' => car

So, the result of this mapping of text will be something like:

the boy's cars are different colors => the boy car be different color

2.1.5 TextBlob Package

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more [5].

TextBlob provides the following features:

1. Noun phrase extraction
2. Part-of-speech tagging
3. Sentiment analysis
4. Classification (Naive Bayes, Decision Tree)
5. Language translation and detection powered by Google Translate
6. Tokenization (splitting text into words and sentences)
7. Word and phrase frequencies
8. Parsing
9. n-grams
10. Word inflection (pluralization and singularization) and lemmatization
11. Spelling correction
12. Add new models or languages through extension
13. WordNet integration

In this project we use the 'Sentiment analysis' and 'Word and phrase frequencies' features to calculate the polarity and subjectivity values.

Chapter 03

System Implement Process

3.1 Process

There are 7 steps in the system model. Those are:

1. The first step is data scrapping from "https://www.consumeraffairs.com/homeowners/choice_home_warranty.html" using a script to scrape reviews as dataset.
2. The second step is analyzing data to sort out the number of words, characters and other attributes to be needed for next steps.
3. The third step is stop words count for natural language processing.
4. The fourth step is data cleaning for reducing of unwanted data/symbols or special characters from the dataset.
5. The fifth step is lemmatization of the words for grouping together the inflected forms of a word.
6. The sixth step is to classify the reviews on the basis of polarity and subjectivity.
7. The final step is to generate a rating based on graph of the polarity values.

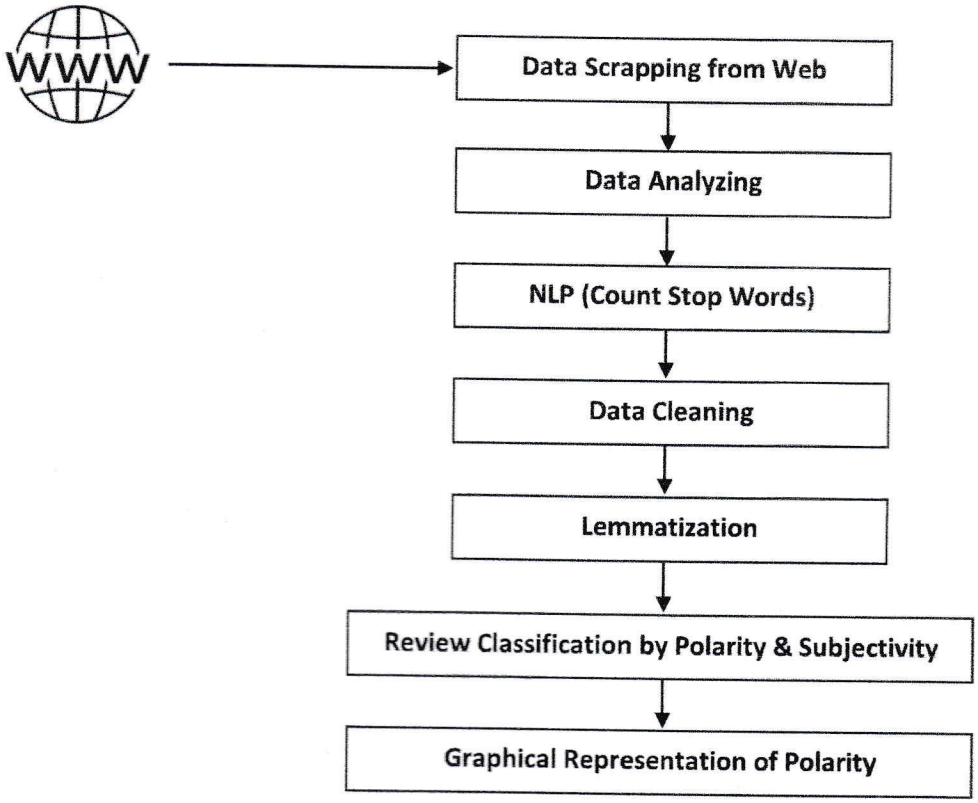


Figure – 3.1: Process of System Implementation

3.2 Tools

The main tools or skill used to develop this project is Python (OOP Language), Python library, package and module: BeautifulSoup, bs4, numpy, pandas, nltk.corpus, stopwords, textblob, TextBlob, Word, matplotlib.piplot, random and Jupyter Notebook.

Python: Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects [6].

BeautifulSoup: BeautifulSoup is a Python library for pulling data out of HTML and XML files. It works with the parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work [7].

numpy: numpy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of numpy, numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created numpy by incorporating features of the competing numarray into Numeric, with extensive modifications [8].

pandas: pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language [9].

nltk.corpus: The nltk.corpus package defines a collection of corpus reader classes, which can be used to access the contents of a diverse set of corpora [10].

matplotlib.pyplot: matplotlib.pyplot is a state-based interface to matplotlib. It provides a MATLAB-like way of plotting. pyplot is mainly intended for interactive plots and simple cases of programmatic plot generation [11].

random: This module implements pseudo-random number generators for various distributions. For integers, there is uniform selection from a range. For sequences, there is uniform selection of a

random element, a function to generate a random permutation of a list in-place, and a function for random sampling without replacement [13].

On the real line, there are functions to compute uniform, normal (Gaussian), lognormal, negative exponential, gamma, and beta distributions. For generating distributions of angles, the von Mises distribution is available [13].

Almost all module functions depend on the basic function random(), which generates a random float uniformly in the semi-open range [0.0, 1.0). Python uses the Mersenne Twister as the core generator. It produces 53-bit precision floats and has a period of 2**19937-1. The underlying implementation in C is both fast and threadsafe. The Mersenne Twister is one of the most extensively tested random number generators in existence. However, being completely deterministic, it is not suitable for all purposes, and is completely unsuitable for cryptographic purposes [13].

The functions supplied by this module are actually bound methods of a hidden instance of the random.Random class. Class Random can also be subclassed if anyone want to use a different basic generator of his/her own devising: in that case, override the random(), seed(), getstate(), and setstate() methods. Optionally, a new generator can supply a getrandbits() method — this allows randrange() to produce selections over an arbitrarily large range.

Jupyter Notebook: Jupyter Notebook (formerly IPython Notebooks) is a web-based interactive computational environment for creating Jupyter notebook documents [12]. The "notebook" term can colloquially make reference to many different entities, mainly the Jupyter web application, Jupyter Python web server, or Jupyter document format depending on context. A Jupyter Notebook document is a JSON document, following a versioned schema, and containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media, usually ending with the ".ipynb" extension.

A Jupyter Notebook can be converted to a number of open standard output formats (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python) through "Download As" in the web interface, via the nbconvert library or "jupyter nbconvert" command line

interface in a shell. Jupyter Notebook provides a browser-based REPL built upon a number of popular open-source libraries.

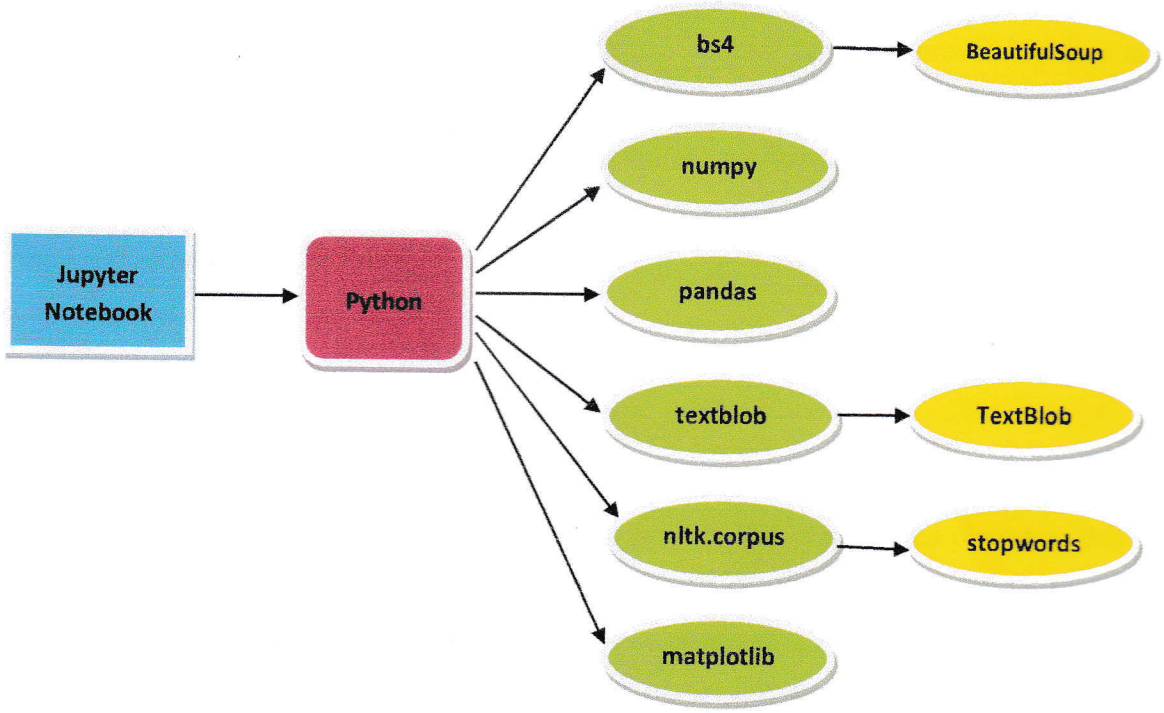


Figure – 3.2: Relation among the implementation tools

3.3 Implementation Issues and Challenges

There are several defined elements in a piece of text that factor into sentiment analysis: the object, the attributes, the opinion orientation, and the opinion strength.

- **Object**: The product, service, individual, organization, event or topic being analyzed.
  - Example: iPhone
- **Attributes**: The specific components and properties of the object
  - Component examples: battery, touch screen, headphone jack
  - Property examples: size, weight, processing speed
- **Opinion holder**: The person or organization who's expressing the sentiment
  - Example: the person who purchased the iPhone
- **Opinion orientation (polarity)**: The general position of the opinion
  - Examples: positive, negative or neutral
- **Opinion strength**: The level, scale or intensity of the opinion
  - Examples: ecstatic > joyous > happy > contented

To obtain complete, accurate and actionable information from a piece of text, it's important to not only identify each of these five elements individually but to also understand how they work together to provide the full context and sentiment. Because keyword processing only identifies the sentiment reflected in a particular word, it typically fails at providing all of the elements necessary to understand the complete context of the entire piece. Natural language processing uses machine learning and data mining to provide a more complete picture, but the inherent complexity of language makes it difficult to ensure algorithms accurately analyze tone and context. While humans are generally better equipped to identify all five of the elements needed to accurately interpret the opinions expressed in a piece of text, manual processing presents its own challenges, primarily in regards to speed and scale.

3.4 Timeline

This project was estimated to be complete in a period of around 8 months. The following Gantt chart will show the timeline for each phases of the project. There are seven phases in this project and take different duration to complete. The initial planning will take around 30 days to complete, planning will take 10 days. After planning, the prototyping and design will carry out concurrently which prototyping is estimated to complete in 20 days and testing prototype is estimated to complete in 40 days. After testing prototype phase will be the implementation phase which take the longest days estimated as 80days. When the implementation was done, the testing and debugging will carry out, each of it will use around 30days.

Sl.	Task Name	Duration	Start Date	End Date
1	Initial Planning	30 Days	Jan 1, 2019	Jan 30, 2019
2	Planning	10 Days	Jan 31, 2019	Feb 9, 2019
3	Prototyping	20 Days	Feb 10, 2019	Mar 1, 2019
4	Testing Prototype	40 Days	Mar 2, 2019	Apr 10, 2019
5	Implementation	80 Days	Apr 11, 2019	Jun 29, 2019
6	Testing	30 Days	Jun 30, 2019	Jul 29, 2019
7	Debug & Enhancement	30 Days	Jul 30, 2019	Aug 28, 2019

Table-3.4: Gantt chart of System Implementation

3.5 Requirement Specification

Data Collection: This system should collect the web data as real time operation. Only web link has to be provided to the IDE and automatically the review data will be scrapped from the target web page. After scrapping, all the review of that particular source page will be grouped together as dataset.

Data Analysis: It is needed to analyze the most probably attributes of each review which are essential to calculate the rating parameters of each review.

Polarity & Subjectivity: Project output should be denoted by polarity and subjectivity values which will be the main objective of this study.

Polarity & Subjectivity Scale: Polarity should be in three categories: positive, negative and neutral. If the polarity value seems more than 0 (zero), it will be treated as positive polarity and if the polarity value seems less than 0 (zero), it will be treated as negative polarity, and if the polarity value is near about 0 (zero) (approximately three decimal after point), it will be treated as neutral polarity. The benchmark value for subjectivity should be from 1 (one) to 0 (zero). 1 (one) will define the highest subjectivity and 0 (zero) will define the lowest subjectivity.

4.1 System Implementation

4.1.1 Data Scrapping

4.1.1.1 The requests library

The first thing we needed to do to scrape a web page is to download the page. We can download pages using the Python requests library. The requests library will make a GET request to a web server, which will download the HTML contents of a given web page for us. There are several different types of requests we can make using requests, of which GET is just one. For gathering data, we used a script to scrape reviews from Choice Home Warranty Product & Service review posting website using the BS4 Python package which requests to retrieve the contents and BeautifulSoup4 to parse the retrieved content by using requests.get( ) Python library function to scrap all web data as source data.

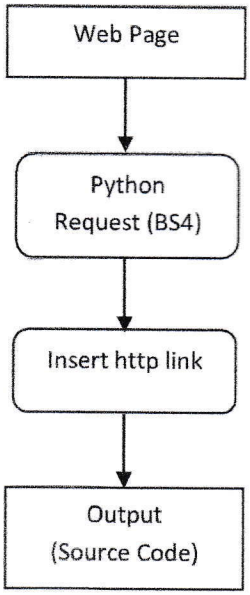


Figure – 4.1.1.1_1: Process of Python requests package

```
r=requests.get('https://www.consumeraffairs.com/homeowners/choice_home_warranty.html?page=4#sort=top_reviews&filter=none')
```

Figure – 4.1.1.1_2: Process to input https link for scrapping web data

After running our request, we get a Response object. This object has a status_codeproperty, which indicates if the page was downloaded successfully:

```
r.status_code
```

```
200
```

Figure – 4.1.1.1_3: Status code for scrapping operation

A status_code of 200 means that page downloaded successfully. We won't fully dive into status codes here, but a status code starting with a 2 generally indicates success, and a code starting with a 4 or a 5 indicates an error.

After retrieve source data from the particular web page, it is stored as continuous text form.

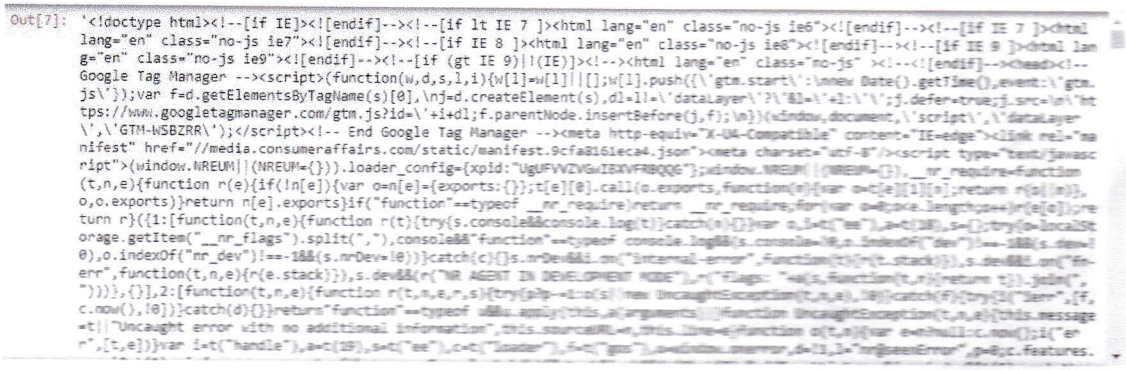


Figure – 4.1.1.1_4: Gathered source code as continuous text

We did not use the Review search API. In within few seconds plenty of review data are scraped for use.

4.1.1.2 Parsing Page with BeautifulSoup

We have downloaded an HTML document. We use the BeautifulSoup library to parse this document, and extract the text from the <p> tag. We first import the library, and create an instance of the BeautifulSoup class to parse our document.

After that, we get the review data from the continuous source data as an order list of reviews.

Out[14]:

	reviews
0	Fixing my clothes dryer. I was very pleased wi...
1	1st Complaint: On 7-31-2019 filed claim with C...
2	I purchased a 5 year contract with Choice Home...
3	I was pleased with the quick service that Kell...
4	A repair person came to look at our cooktop. H...
5	I bought my house 2 years ago and it fills lik...
6	We are very pleased with Lopez A/C. They came ...
7	Your service are awesome. When I called reques...
8	Thank you so much for coming to my aide when I...
9	Contacted customer care to place claim. Operat...
10	****WARNING**** Contract expired today. Sales ...
11	I called to renew my home warranty. The first ...
12	DO NOT TRUST CHOICE HOME WARRANTY FOR ONE SECO...
13	I can say that so far I have been very impress...
14	Aaron was very Knowledgeable and Courteous. He...
15	Quick response, quick analysis. Can't beat it ...
16	Anytime I communicated with Choice the staff w...
17	I purchased the full plan from Choice Home War...
18	The technician assigned was great, let my drye...

Figure – 4.1.1.2: Order list of review data from source code

4.1.2 Data Analyzing

In short, data analysis (text mining and textual analysis) is the automated process that allows machines to extract and classify information from data or text. Here, we try to count several parameters from the textual data as a summary extraction of the reviews. Number of words, number of characters and average word length has been summarized for each review data. The summarized parameters are to be used for the next operation.

In this project analyzing the data is an automated process. Choice Home Warranty Product & Service review data is increasing day by day from their regular customers. So, it is time consuming and expensive to analyze the parameters by manually or human analysis. For this reason, it was the first challenge towards the successive implementation of the project work.

Out[20]:

	reviews	word_count	char_count	average_word_length
0	Fixing my clothes dryer. I was very pleased wi...	52	260	4.019231
1	1st Complaint: On 7-31-2019 filed claim with C...	91	530	4.824176
2	I purchased a 5 year contract with Choice Home...	80	413	4.175000
3	I was pleased with the quick service that Kell...	122	615	4.049180
4	A repair person came to look at our cooktop. H...	42	222	4.309524
5	I bought my house 2 years ago and it fills lik...	71	368	4.197183
6	We are very pleased with Lopez A/C. They came ...	44	265	5.045455
7	Your service are awesome. When I called reques...	41	232	4.682927
8	Thank you so much for coming to my aide when I...	78	379	3.871795
9	Contacted customer care to place claim. Operat...	46	297	5.478261
10	****WARNING**** Contract expired today. Sales ...	79	464	4.886076
11	I called to renew my home warranty. The first ...	58	381	4.602941
12	DO NOT TRUST CHOICE HOME WARRANTY FOR ONE SECO...	58	340	4.879310
13	I can say that so far I have been very impress...	71	386	4.450704
14	Aaron was very Knowledgeable and Courteous. He...	55	366	5.672727
15	Quick response, quick analysis. Can't beat it ...	50	325	5.520000
16	Anytime I communicated with Choice the staff w...	47	261	4.574468
17	I purchased the full plan from Choice Home War...	151	766	4.978470
18	The technician assigned was great, let my drye...	48	280	5.525000

Figure – 4.1.2: Summarized parameters by analysis

4.1.3 Natural Language Processing (Count Stop Words)

Stopwords are words those are generally considered useless. Most search engines ignore these words because they are so common that including them would greatly increase the size of the index without improving precision or recall. NLTK comes with a stopwords corpus that includes a list of 179 English stop words. Let's see what happens when we filter out these words.

```
print(stop_words)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'y
ourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those',
'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'a
n', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'b
etween', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'of
f', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both',
'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'ar
en', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "have
n't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "should
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

Figure – 4.1.3_1: List of English stop words filtered by nltk.corpus library

The target work of this step is to determine the stop_word_rate. For this process we used the output parameters from the data analysis step.

The following calculation is done for the determination of stop word rate:
stop_word_rate = (number_of_stop_word)/ (number_of_word).

After that, we get the following output for stop_word ratings:

Out[27]:

	reviews	word_count	char_count	average_word_length	stopword_count	stopword_rate
0	Fixing my clothes dryer. I was very pleased wi...	52	260	4.019231	25	0.480769
1	1st Complaint: On 7-31-2019 filed claim with C...	91	530	4.824176	36	0.395604
2	I purchased a 5 year contract with Choice Home...	80	413	4.175000	45	0.562500
3	I was pleased with the quick service that Kell...	122	615	4.049180	65	0.532787
4	A repair person came to look at our cooktop. H...	42	222	4.309524	23	0.547619
5	I bought my house 2 years ago and it fills lik...	71	368	4.197183	29	0.408451
6	We are very pleased with Lopez A/C. They came ...	44	265	5.045455	21	0.477273
7	Your service are awesome. When I called reques...	41	232	4.682927	20	0.487805
8	Thank you so much for coming to my aide when I...	78	379	3.871795	43	0.551282
9	Contacted customer care to place claim. Operat...	46	297	5.478261	16	0.347826
10	****WARNING**** Contract expired today. Sales ...	79	464	4.886076	32	0.405063
11	I called to renew my home warranty. The first ...	68	381	4.602941	34	0.500000
12	DO NOT TRUST CHOICE HOME WARRANTY FOR ONE SECO...	58	340	4.879310	21	0.362069
13	I can say that so far I have been very impress...	71	386	4.450704	39	0.549296
14	Aaron was very Knowledgeable and Courteous. He...	55	366	5.672727	24	0.436364
15	Quick response, quick analysis. Can't beat it ...	50	325	5.520000	20	0.400000
16	Anytime I communicated with Choice the staff w...	47	261	4.574468	22	0.468085
17	I purchased the full plan from Choice Home War...	151	766	4.079470	70	0.463576
18	The technician assigned was great, let my drye...	40	260	5.525000	15	0.375000

Figure – 4.1.3_2: Output of stop word ratings

4.1.4 Data Cleaning

Text is just a sequence of words, or more precisely, a sequence of characters. But when we usually deal with language modelling, or natural language processing, we are more concerned about the words as a whole, instead of just worrying about character-level depth of our text data. One reason behind that is that in the language models, individual characters don't have a lot of "content". Characters like 'd', 'r', 'a', 'e' don't hold any context individually, but when rearranged in the form of a word, they might generate the word "read", which might explain some activity what we are probably doing right now.

For data cleaning operation we consider the following issues:

- The dataset is comprised of only English reviews.
- All text has been converted to lowercase.
- There is white space around punctuation like periods, commas, and brackets.
- Text has been split into one sentence per line.

First of all, we get the reviews as dataframe (df) which is an ordered list of review data.

```
df['reviews']
```

```
0     Fixing my clothes dryer. I was very pleased wi...
1     1st Complaint: On 7-31-2019 filed claim with C...
2     I purchased a 5 year contract with Choice Home...
3     I was pleased with the quick service that Kell...
4     A repair person came to look at our cooktop. H...
5     I bought my house 2 years ago and it fills lik...
6     We are very pleased with Lopez A/C. They came ...
7     Your service are awesome. When I called reques...
8     Thank you so much for coming to my aide when I...
9     Contacted customer care to place claim. Operat...
10    ****WARNING**** Contract expired today. Sales ...
11    I called to renew my home warranty. The first ...
12    DO NOT TRUST CHOICE HOME WARRANTY FOR ONE SECO...
13    I can say that so far I have been very impress...
14    Aaron was very Knowledgeable and Courteous. He...
15    Quick response, quick analysis. Can't beat it ...
16    Anytime I communicated with Choice the staff w...
17    I purchased the full plan from Choice Home War...
18    The technician assigned was great, let my drye...
19    I was please with service performed and the am...
20    Living in a remote area proves to be a challen...
21    I signed up with Choice for my summer home. Th...
22    AC Technician showed up next morning at 8:30 a...
23    After much comparison shopping online I decide...
24    Great initial experience. Kayleigh ** was exce...
25    AC repair. Technician was great. He was friend...
26    Horrible experience just trying to buy a plan....
27    Going through the site is easy enough. However...
28    I bought their warranty and waited for weeks a...
```

Figure – 4.1.4: Ordered list of review data

4.1.4.1 Removing Unwanted Characters

This is a primary step in the process of text cleaning. After getting the dataframe we needed to get rid of all the punctuation, non-alphabets, and any other kind of characters which might

not be a part of the language. The general methods of such cleaning involve regular expressions, which can be used to filter out most of the unwanted texts.

There are some systems where important English characters like the full-stops, question-marks, exclamation symbols, etc are retained. Consider an example where you want to perform some sentiment analysis on human generated reviews, and we want to classify the reviews are negative, neutral and positive.

4.1.4.2 Tokenization and Capitalization/De-capitalization

Tokenization is just the process of splitting a sentence into words.

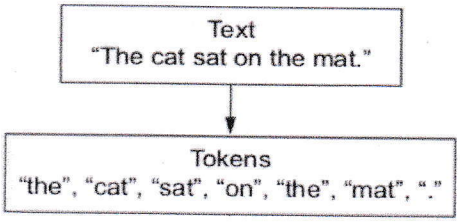


Figure- 4.1.4_1: Tokenization of a sentence

We realized that the above example doesn't only tokenize the sentence, but also makes sure that all the words are lowercase.

So, we convert the dataframe into lowercase of each word and get the following result:

```
df.head(10)
```

	reviews	word_count	char_count	average_word_length	stopword_count	stopword_rate	lowercase
0	Fixing my clothes dryer. I was very pleased wi...	52	260	4.019231	25	0.480769	fixing my clothes dryer. i was very pleased wi...
1	1st Complaint: On 7-31-2019 filed claim with C...	91	530	4.824176	36	0.395604	1st complaint: on 7-31-2019 filed claim with c...
2	I purchased a 5 year contract with Choice Home...	80	413	4.175000	45	0.562500	i purchased a 5 year contract with choice home...
3	I was pleased with the quick service that Kell...	122	615	4.049180	65	0.532787	i was pleased with the quick service that kell...
4	A repair person came to look at our cooktop. H...	42	222	4.309524	23	0.547619	a repair person came to look at our cooktop. h...
5	I bought my house 2 years ago and it fills lik...	71	368	4.197183	29	0.408451	i bought my house 2 years ago and it fills lik...
6	We are very pleased with Lopez A/C. They came ...	44	265	5.045455	21	0.477273	we are very pleased with lopez a/c. they came ...
7	Your service are awesome. When I called reques...	41	232	4.682927	20	0.487805	your service are awesome. when i called reques...
8	Thank you so much for coming to my aide when I...	78	379	3.871795	43	0.551282	thank you so much for coming to my aide when i...
9	Contacted customer care to place claim. Operat...	46	297	5.478261	16	0.347826	contacted customer care to place claim. operat...

Figure – 4.1.4_2: Lowercase of review data

4.1.4.3 Removing Punctuation

Considering the two sentences with the same semantic meaning:

"This food is good.", and "This. Food. Is. Good!!!!!!!!".

Same words, but totally different sentiments, and the only information which can help us to see the different is the overused punctuation, which shows some sort of an "extra" feeling. Emoticons, which are made up of non-alphabets also play a role in sentiment analysis. ":), :(, -_-, :D, xD", all these, when processed correctly, can help with a better sentiment analysis. Here we implement a system that may classify whether some phrase is sarcasm, or not sarcasm, such little details can be helpful.

Apostrophe is one important punctuation character, which needs to be dealt carefully, because a lot of text can be based on apostrophes. Terms like "aren't, shouldn't, didn't, would've, mightn't've, y'all'd've" are infiltrating the online documents just like a disease, and luckily, we do have a cure for that as well.

After that, we reduce the white space around punctuation like periods, commas, and brackets from the output data.

Out[33]:

	reviews	word_count	char_count	average_word_length	stopword_count	stopword_rate	lowercase	punctuation
0	Fixing my clothes dryer. I was very pleased wi...	52	260	4.019231	25	0.480769	fixing my clothes dryer. i was very pleased wi...	fixing my clothes dryer i was very pleased wit...
1	1st Complaint: On 7-31-2019 filed claim with C...	91	530	4.824176	36	0.395604	1st complaint: on 7-31-2019 filed claim with c...	1st complaint on 7312019 filed claim with choi...
2	I purchased a 5 year contract with Choice Home...	80	413	4.175000	45	0.562500	i purchased a 5 year contract with choice home...	i purchased a 5 year contract with choice home...
3	I was pleased with the quick service that Kell...	122	615	4.049180	65	0.532787	i was pleased with the quick service that kell...	i was pleased with the quick service that kell...
4	A repair person came to look at our cooktop. H...	42	222	4.309524	23	0.547619	a repair person came to look at our cooktop. h...	a repair person came to look at our cooktop he...
5	I bought my house 2 years ago and it fills lik...	71	368	4.197183	29	0.408451	i bought my house 2 years ago and it fills lik...	i bought my house 2 years ago and it fills lik...
6	We are very pleased with Lopez A/C. They came ...	44	265	5.045455	21	0.477273	we are very pleased with lopez a/c. they came ...	we are very pleased with lopez ac they came ou...
7	Your service are awesome. When I called reques...	41	232	4.682927	20	0.487805	your service are awesome. when i called reques...	your service are awesome when i called request...
8	Thank you so much for coming to my aide when I...	78	379	3.871795	43	0.551282	thank you so much for coming to my aide when i...	thank you so much for coming to my aide when i...

Figure – 4.1.4_3: Output after reducing of punctuations

4.1.4.4 Removing/Retaining Stopwords

This cleaning step also depends on what we'll eventually be doing with the data after preprocessing. Stopwords are the words which are used very frequently, and they're so frequent, that they somewhat lose their semantic meaning. Words like "of, are, the, it, is" are some examples of stopwords. In this sentiment classification, where keywords are more important than general terms, removing stopwords can be a good idea.

There are two common approaches of removing the stopwords, and both are fairly straight forward. One way is to count all the word occurrences, and providing a threshold value on the count, and getting rid of all the terms/words occurring more than the specified threshold value. The other way is to have a predetermined list of stopwords, which can be removed from the list of tokens/tokenized sentences.

Therefore, we are ready to get the cleaned data.

```
In [42]: df.head(35)
```

Out[42]:

	reviews	word_count	char_count	average_word_length	stopword_count	stopword_rate	lowercase	punctuation	stopwords	cleanreview
0	Fixing my clothes dryer. I was very pleased wi...	52	260	4.019231	25	0.480769	fixing my clothes dryer i was very pleased wi...	fixing my clothes dryer i was very pleased wit...	fixing clothes dryer pleased repair man asked...	fixing clothes dryer pleased repair man asked...
1	1st Complaint On 7-31-2019 filed claim with C...	91	530	4.824176	36	0.395604	1st complaint on 7-31-2019 filed claim with c...	1st complaint on 7312019 filed claim with choi...	1st complaint 7312019 filed claim choice home...	1st complaint 7312019 filed claim machine work...
2	I purchased a 5 year contract with Choice Home...	80	413	4.175000	45	0.562500	i purchased a 5 year contract with choice home...	i purchased a 5 year contract with choice home...	purchased 5 year contract choice home warranty...	purchased 5 year contract nothing issues 4 cha...
3	I was pleased with the quick service that Kell...	122	615	4.049180	65	0.532787	i was pleased with the quick service that kell...	i was pleased with the quick service that kell...	pleased quick service kelly still website tryi...	pleased quick service kelly still website tryi...

Figure – 4.1.4_4: Output after data cleaning process

4.1.5 Lemmatization

Languages we speak and write are made up of several words often derived from one another. When a language contains words that are derived from another word as their use in the speech changes is called Inflected Language. By this step we reduce the Inflected form of word from the dataframe.

Lemmatization is the process of converting a word to its base form and also lemmatization is text normalization (or sometimes called word normalization) techniques in the field of Natural Language Processing that is used to prepare text, words, and documents for further processing. Lemmatization has been studied, and algorithms have been developed in Computer Science since the 1960's [14]. Python nltk package is the Natural Language Toolkit package provided by Python for Natural Language Processing tasks.

Lemmatization can be performed by using the following python packages:

1. Wordnet Lemmatizer
2. Spacy Lemmatizer
3. TextBlob
4. CLiPS Pattern
5. Stanford CoreNLP
6. Gensim Lemmatizer
7. TreeTagger

Here, we use the Word module from Python TextBlob library package for lemmatizing the dataframe. TexxtBlob is a powerful, fast and convenient NLP package as well. Using the Word and TextBlobobjects, its quite straight-forward to parse and lemmatize words and sentences respectively.

The degree of inflection may be higher or lower in a language. As we read the definition of inflection with respect to grammar, we can understand that an inflected word(s) will have a common root form. Let's look at a few examples:

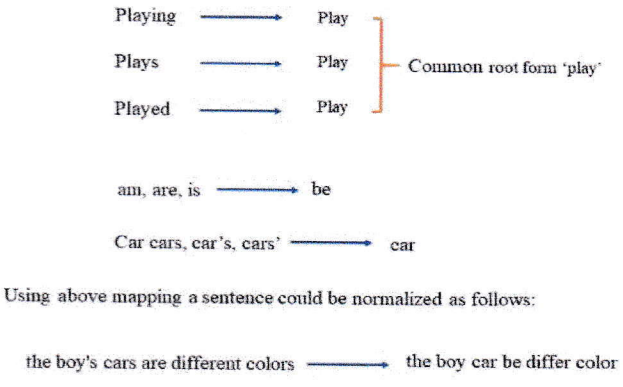


Figure- 4.1.5_1: Inflected form of word

Given examples must have helped to understand the concept of normalization of text, although normalization of text is not restricted to only written document but to speech as well. Lemmatization helps us to achieve the root forms (sometimes called synonyms in search context) of inflected (derived) words.

Out[45]:

	reviews	word_count	char_count	average_word_length	stopword_count	stopword_rate	lowercase	punctuation	stopwords	cleanreview	lemmatized
0	Fixing my clothes dryer. I was very pleased wi...	52	260	4.019231	25	0.480769	fixing my clothes dryer. i was very pleased wi...	fixing my clothes dryer i was very pleased wit...	fixing clothes dryer pleased repair man asked ...	fixing clothes dryer pleased repair man asked ...	fixing clothes dryer pleased repair man asked ...
1	1st Complaint: On 7-31-2019 filed claim with C...	91	530	4.824176	36	0.395604	1st complaint: on 7-31-2019 filed claim with c...	1st complaint on 7312019 filed claim with choi...	1st complaint 7312019 filed claim choice home ...	1st complaint 7312019 filed claim machine work...	1st complaint 7312019 filed claim machine work...
2	I purchased a 5 year contract with	80	413	4.175000	45	0.562500	i purchased a 5 year contract with	i purchased a 5 year contract with choice	purchased 5 year contract choice home	purchased 5 year contract nothing issues 4	purchased 5 year contract nothing issue 4

Figure- 4.1.5_2: Output of lemmatized data

4.1.6 Review Classification by Polarity & Subjectivity

The TextBlob package for Python is a convenient way to do a lot of Natural Language Processing (NLP) tasks. This tells us that the English phrase "not a very great calculation" has a polarity of about -0.3, meaning it is slightly negative, and a subjectivity of about 0.6, meaning it is fairly subjective.

In addition to the polarity and subjectivity mentioned before. When calculating sentiment for a single word, TextBlob uses a sophisticated technique known to mathematicians as "averaging". TextBlob handles negation. Negation multiplies the polarity by -0.5, and does not affect subjectivity. TextBlob also handles modifier words. For an example, here is the summarized record for "very":

word	polarity	subjectivity	intensity
very	0.2	0.3	1.3

Recognizing "very" as a modifier word, TextBlob will ignore polarity and subjectivity and just uses intensity to modify the following word:

```
TextBlob("very great").sentiment
## Sentiment(polarity = 1.0, subjectivity = 0.9750000000000001)
```

The polarity gets maxed out at 1.0, but the subjectivity is also modified by "very" to become 0.75·1.3 = 0.9750.75·1.3 = 0.975.

Negation combines with modifiers in an interesting way: in addition to multiplying by -0.5 for the polarity, the inverse intensity of the modifier enters for both polarity and subjectivity.

TextBlob("not very great").sentiment

Sentiment(polarity = -0.3076923076923077, subjectivity = 0.5769230769230769)

So, polarity = $-0.5 \cdot 1.1.3 \cdot 0.8 \approx -0.31$ polarity = $-0.5 \cdot 1.1.3 \cdot 0.8 \approx -0.31$ and subjectivity = $1.1.3 \cdot 0.75 \approx 0.58$ subjectivity = $1.1.3 \cdot 0.75 \approx 0.58$.

Out[49]:

	reviews	word_count	char_count	lowercase	punctuation	stopwords	cleanreview	lemmatized	polarity	subjectivity
0	Fixing my clothes dryer. I was very pleased wi...	52	260	fixing my clothes dryer. i was very pleased wi...	fixing my clothes dryer i was very pleased wit...	fixing clothes dryer pleased repair man asked ...	fixing clothes dryer pleased repair man asked ...	fixing clothes dryer pleased repair man asked ...	0.261905	0.511905
1	1st Complaint: On 7-31-2019 filed claim with C...	91	530	1st complaint: on 7-31-2019 filed claim with c...	1st complaint on 7312019 filed claim with choi...	1st complaint 7312019 filed claim choice home ...	1st complaint 7312019 filed claim machine work...	1st complaint 7312019 filed claim machine work...	0.100000	0.366667
2	I purchased a 5 year contract with Choice Home...	80	413	i purchased a 5 year contract with choice home...	i purchased a 5 year contract with choice home...	purchased 5 year contract choice home warranty...	purchased 5 year contract nothing issues 4 cha...	purchased 5 year contract nothing issue 4 char...	0.175000	0.325000
3	I was pleased with the quick service that Kiel...	122	615	i was pleased with the quick service that kiel...	i was pleased with the quick service that kiel...	pleased quick service kelly still website tryi...	pleased quick service kelly still website tryi...	pleased quick service kelly still website tryi...	0.231336	0.489050

Figure- 4.1.6_1: Output of polarity and subjectivity parameters

4.1.6.1 Observation

We can see the output is categorized between two — Polarity and Subjectivity.

- **Polarity** is a float value within the range [-1.0 to 1.0] where 0 indicates neutral, +1 indicates a very positive sentiment and -1 represents a very negative sentiment.
- **Subjectivity** is a float value within the range [0.0 to 1.0] where 0.0 is very objective and 1.0 is very subjective. Subjective sentence expresses some personal feelings, views, beliefs, opinions, allegations, desires, beliefs, suspicions, and speculations where as Objective sentences are factual.

```
In [51]: df.describe()
```

Out[51]:

	word_count	char_count	average_word_length	stopword_count	stopword_rate	polarity	subjectivity
count	30.000000	30.000000	30.000000	30.000000	30.000000	30.000000	30.000000
mean	70.033333	379.233333	4.550677	33.666667	0.473058	0.190975	0.448485
std	28.187162	132.084791	0.556469	15.521583	0.065612	0.232942	0.204626
min	40.000000	222.000000	3.811881	15.000000	0.347826	-0.533333	0.000000
25%	46.250000	265.000000	4.056753	21.000000	0.422799	0.101389	0.338194
50%	63.500000	353.000000	4.380114	30.500000	0.479021	0.175000	0.463572
75%	87.500000	469.250000	4.884385	44.500000	0.524120	0.380161	0.558631
max	151.000000	766.000000	5.672727	70.000000	0.580357	0.540000	0.833333

Figure- 4.1.6_2: Output of finalized summary

4.1.7 Graphical Representation of Polarity

This system will provide the facility for automatic generating of graphical representation of the polarity. In this project we show the graphical representation process based on the polarity output of the reviews.

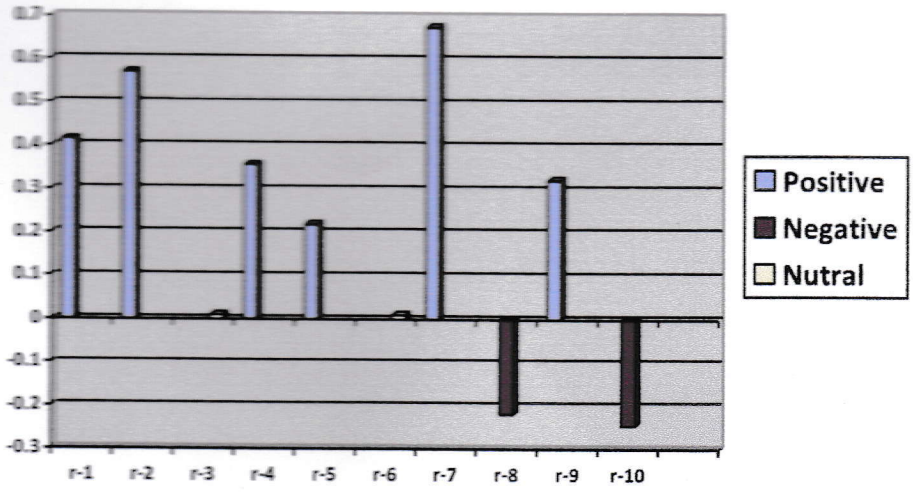


Figure- 4.1.6_3: Graphical representation based on polarity

Chapter 05

Result Analysis

5.1 Result Analysis

In the Review Classification step, we have calculated the polarity and subjectivity. The polarity range is (-1.0 to 1.0) and the subjectivity range is (0.0 to 1.0) and if the polarity is less than 0 then the sentence is negative. If the polarity of the sentence is 0.000 then the sentence is neutral. Thus, if the polarity greater than 0.000, the sentence is positive.

- Positive (>0.000)
- Neutral (0.000)
- Negative (<0)

Relatively, if the subjectivity is 0.000, the sentence is not subjective. Thus, the subjectivity is greater than 0.000, the sentence is subjective and the accuracy of subjectivity depends on the ratio of subjectivity value.

- Subjective (>0.000)
- Not Subjective (=0.000)

Figure-4.1.6_1 shows the summarized table for the review of each customers of "Choice Home Warranty Product & Service". The summarized table shows the positive, neutral and negative polarity and the subjectivity of the sentiment.

Chapter 06

Conclusion

6.1 Conclusion

In this paper, we have proposed a customer review rating system based on the opinion of regular customers of "Choice Home Warranty Product & Service" online service provider. This can be widely used in the future to get proper reviews of any product or service to get the best review possible for a product or service. This project will allow user to briefly analyze bunch of products or services based on their rating and choose the best one depending on the requirement of the user. This project is reliable as the rating is basically generated based on customer opinion and this rating will reflect the actual state of the product and service provided by "Choice Home Warranty Product & Service". Also, this project will be a helpful study for the company or organization to provide their best product and service to the customer.

6.2 Future Enhancement

We have implemented this project to work on "Choice Home Warranty Product & Service" customer review classification. Our basic plan was to implement the project for classifying the customer reviews of any brand or organization. In future we wish to work to integrate the system so that it can generate classification for any brand or organization. That would not require much effort as we can always collect data from different online service provider and we can store the classified result to our own database. We wish to make this project as a complete API and user more friendly interfacing.

We look forward to ease the process of decision making about buying any product or service for the online customers. This project would be just a few steps towards digitalization. We belief it will make people's lives easier as this project will allow users to see an accurate rating of review of any product and service from any reputed online product and service provider and save them from the misery of finding a good review from various review sites.

References

1. Stop Words, https://en.wikipedia.org/wiki/Stop_words
2. NLTK stop words, https://pythonspot.com/nltk-stop-words/
3. NLTK, https://www.nltk.org/
4. Lemmatization, https://en.wikipedia.org/wiki/Lemmatisation
5. TextBlob, https://textblob.readthedocs.io/en/dev/
6. Python (programming language), https://en.wikipedia.org/wiki/Python
7. BeautifulSoup Documentation, https://www.crummy.com/software/BeautifulSoup/bs4/doc/
8. NumPy, https://en.wikipedia.org/wiki/NumPy
9. pandas 0.25.1 documentation, https://pandas.pydata.org/pandas-docs/stable/
10. Corpus Readers, http://www.nltk.org/howto/corpus.html
11. matplotlib, https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.html
12. Project Jupyter, https://en.wikipedia.org/wiki/Project_Jupyter
13. random- Generate pseudo-random numbers, https://docs.python.org/3/library/random.html
14. Stemming and Lemmatization in Python, https://www.datacamp.com/community/tutorials/stemming-lemmatization-python