Consumer review Analysis using NLP and Data Mining

Md. Nasimuzzaman
Computer Science and Engineering
School of Data and Sciences
BRAC University
Dhaka, Bangladesh
md.nasimuzzaman@g.bracu.ac.bd

Ahmed Nur Merag
Computer Science and Engineering
School of Data and Sciences
BRAC University
Dhaka, Bangladesh
ahmed.nur.merag@g.bracu.ac.bd

Sumya Afroj
Computer Science and Engineering
School of Data and Sciences
BRAC University
Dhaka, Bangladesh
sumya.afroj@g.bracu.ac.bd

Md. Mustakin Alam
Computer Science and Engineering
School of Data and Sciences
BRAC University
Dhaka, Bangladesh
md.mustakin.alam@g.bracu.ac.bd

Md Humaion Kabir Mehedi
Computer Science and Engineering
School of Data and Sciences
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

Annajiat Alim Rasel
Computer Science and Engineering
School of Data and Sciences
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—The popularity and development of the Internet have greatly increased the amount of information that is easily available. For a sizable portion of the population, it has evolved into the primary forum for opinion expression.Reviews, internet forums, and social networking sites like Twitter and Facebook are all possible venues for sharing these perspectives. Sentiment analysis and information extraction are made possible by the ease of access to all of these viewpoints, which is free. Sentiment analysis is concerned with automatically ascertaining, through the use of computer tools, if a reviewer's purpose is positive, negative, or neutral with regard to a specific good, service, or a problem. Because they have such a positive impact on every particular organization, these strategies are rapidly gaining favor. As a result, we researched multiple methods for classification, data preprocessing, and feature extraction in this study and compared the outcomes using the FAR, FRR, and accuracy measures. The simulation's outcomes show that the SVM classifier produces the greatest accuracy for the specified dataset when used on a dataset of features extracted by the N-gram model.**

***Index Terms*—opinion, possible, accuracy, positive, analysis, data, learning, extraction, feature, methods, techniques.**

I. Introduction

Over the past ten years, internet usage has dramatically expanded, and with it, so has the volume of text data generated. Analysis of these is more important than ever because there are more articles, reviews, and online discussions about a variety of issues. Determining the outlook of customers toward a certain service is one of text data analysis's most crucial applications. Sentiment analysis is the categorization of opinions about specific products, services, or topics in a text (word, sentence, or a document) in order to computationally determine the author's polarity or point of view toward the topic, which may be positive, negative, or neutral. [1]. On specific text data, there are many approaches to perform sentiment analysis, two of which are based on machine learning, and a third on knowledge. [2]. The main goal of sentiment analysis using a knowledge-based technique is achieved with the use of language processing algorithms and lexicon techniques. While the approach to machine learning is focused on the extraction of features, including the word frequency, POS tagging, TF-IDF words of opinion, etc [3]. An identified training-supervised machine learning algorithm receives a dataset and is then capable of classifying and learning the polarity (positive, negative, or neutral) using the learned model, analyzing an unknown text [4]. To build a sentiment analysis model that is more accurate, data preparation and feature extraction are necessary. Because feature extraction accounts for the majority of a model's accuracy, it is essential. Techniques used in data preprocessing include noise reduction, normalization, tokenization, and vectorization [5]. Feature extraction methods like N-gram tagging and term frequency-inverse document frequency have also been investigated in an effort to improve accuracy. When these techniques are combined, the model's accuracy can be significantly improved [4] [3]. The proposed effort examines the impact of various preprocessing methods. On the database of Amazon cell phone reviews and feature extraction methods in order to ascertain how they impact the model's accuracy.Text classification techniques and automatic classification systems have been used to accomplish natural language processing and text analysis and produce the desired results. The following is the essay's organization: The work that has been done in the past utilizing opinion mining is discussed in Section 2. The comparison is examined in Section 3. Between the application of various NLP approaches and the supervised machine learning techniques. Section Four displays the performance and the report on the analysis. The task is concluded in Section VI.

II. Related Works

A succinct synopsis of previous work on opinion mining is provided in this section. The fields of sentiment analysis and opinion mining have seen a lot of research and advancement in recent years. For applications involving sentiment analysis, we carefully analyzed the value of text pre-processing. [6].

By using a chopped method, extraneous characteristics were removed. It was discovered from experimental results proving sentiment analysis with appropriate feature representation and selection that The accuracy of the classifier is greatly improved by adequate text pre-processing. Instead of identifying the text's sentiment before evaluating it, sentiment analysis identifies the text's sentiment first, opinion mining is a method for obtaining and analyzing someone's feelings on something. We examined 54 recent articles and found that there is still opportunity for advancement in the techniques for sentiment categorization and feature selection [7].

To categorize product evaluations using data from Twitter, a number of machine learning techniques that incorporate semantic analysis are used. The Naive Bayes method was found to perform better when paired with an unigram model than when used alone. Furthermore, accuracy increased following the use of WordNet's semantic analysis tool, which was subsequently followed by the earlier method [8]. In addition to neural networks, SVMs, and lexicon-based approaches, many novel techniques, such as random forest, the rule miner, radial basis function neural network, etc., have not been fully employed. [9].

The issue of a computer being able to predict and comprehend a person's sentiment or a contextual opinion on something is known as sentiment analysis. According to his narration, The data and the language employed in it determine the modifications required to enhance the classifier's performance. When transformations are performed, as well as when the least important data is filtered away, the machine learning approach learns more successfully and generalizes better [10].

Using data from Twitter, we conducted a poll on sentiment analysis. Using existing accessible approaches, such as machine learning techniques, unstructured, heterogeneous opinions that are occasionally good, occasionally negative, or neutral. He got to the conclusion that accuracy is impacted by data cleanliness. Furthermore, he asserted that using the Bigram model produced better results than using other techniques like SVM and Naive Bayes [11].

Another idea that has been presented in detail is sentiment analysis on Twitter data, which is carried out using machine learning algorithms in a step-by-step manner. [12].

In order to analyze Twitter data, this research recommended using a scalable, rapid, and flexible text framework called Apache Spark. Out of all the algorithms used, a decision tree's accuracy, precision, recall, and F1-Score were all 100 percent. [2].

We have discussed the challenges he faced in his work performing sentiment analysis relevant to the techniques and tactics used. The investigation was concluded by claiming that the review structure and issue nature are the elements determining the proper challenges for the sentiment analysis evaluation. The corpus was painstakingly elucidated on three levels: (1) performance (standard vs. dialect), (2) discussion (as opposed to "non chat"), and (3) valence (which indicated whether the writer was coming from a positive, negative, both, neutral, or n/a perspective). According to this study, fewer classifications increase accuracy (up to 87 for a two-fold categorization). Accuracy rose to 94.6 percent when the feature sets retrieved from "only non-chat" and "just standard" were used. The skewness of the corpus caused them to only receive fair results [13]. We also employed a machine learning strategy to classify tweets as favorable or negative using Twitter sentiment analysis about electrical equipment like laptops, mobile phones, etc. An updated feature vector's classification accuracy was assessed employing ensemble classifiers, maximum entropy, Naive Bayes, and SVM. They found that the modified feature vector is effective for electrical goods and that all of the classifiers' accuracy was essentially the same. [14].

III. Architecture and Modeling

The suggested work offers numerous stages for assessing the opinion using a database of product reviews. The graphic displays the architecture design for the suggested method below. The subsequent paragraphs outline the numerous steps.

1.

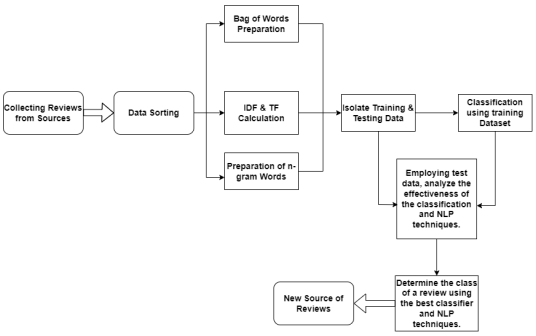


Fig. 1: Architectural Schematics

Regarding the suggested approach, each review can be thought of as a record, which can be represented mathematically as a grouping of words as given in Equation 1. Equation 2 now brings together all of these sets in a disjoint union. To illustrate, consider the remarks " I adore hockey," "I love football," and "I love cricket." Then, "cricket, football, hockey, I, love" would make up the group of different words. By utilizing the integers 0 or 1 to denote the presence or absence of words, binary vectors are produced as shown. "I love cricket" = [1, 0, 0, 1, 1] "I love football" = [0, 1, 0, 1, 1] "I love hockey" = [0, 0, 1, 1, 1] The bag of words is made up of a collection of such vectors. We can observe that there are a lot of zeros, which indicates that the matrix containing the words is sparse. The terms with the lowest frequency are removed from the bag since people rarely express their ideas using those words.

A. Data Pre-Processing

The dataset under investigation is formatted in tabular values (TSV). When utilizing this design, each time a tab is encountered, new data is added to a new column. Our algorithm was trained using reviews left on the Amazon by cell phone users. Since all of the implemented methods are supervised in nature, the dataset must first be categorized before our algorithm can be trained on it. To import the dataset, we use Python 3.6's Pandas library. Preparing the text for use by computers in activities like analysis, prediction, etc. is a process known as text preprocessing. Text preprocessing involves a variety of steps, but in this article we will focus on removing reviews with short word counts, translating all reviews into English, changing all reviews to lowercase, removing stop words and special characters, and implementing stemming of words using a variety of libraries. These are the words that are employed as articles or connectors in English grammar but do not convey any emotion. As a result, the second stage entails getting rid of these stop words. To eliminate these stop-words from the review, we employed the NLTK corpus. Stop-words such as 'not' are typically removed during this step. But in opinion mining, the existence or absence of something matters a lot. For instance, the review claims that the cell phone companies offered here aren't terrible. The transformation of a favorable evaluation into an unfavorable one is evident. In order to prevent issues like these from occurring, related terms are left in the NLTK corpus and are not removed from it after this stage. Making root words from the original words is the last stage in pre-processing. Without a prefix or suffix, a word is said to be a root word. While opinion mining, we are more interested in the sentiment found in the text than in grammar. For example, the root of the word loved. Consequently, the task of sentiment analysis has been made easier by this conversion. For the purpose of reducing every term in the dataset to its root words, the Porter-Sweeney algorithm is used.

2.

B. Term Frequency-Inverse Document Frequency (TF-IDF)

The TF-IDF is produced by combining the inverse document frequency and term frequency. The number of times a term (t) appears in a document (d) is referred to as its term frequency. Equation 1 can be used to express this.

$$TF(t,d) = Ft,d \tag{1}$$

$$IDF(t,d) = log\frac{N}{|t\epsilon d, d\epsilon D|} \tag{2}$$

The information density of a word, as well as its frequency or rarity across all texts, are measured by inverse document frequency. It can be determined by multiplying the ratio of the number of documents N in the corpus D by the number of documents d that include the phrase t to obtain the logarithm. The answer is given in equation 4. The final result is the TF-IDF, which is given by equation 3.

$$TFIDF(t,d) = TF(t,d) * IDF(t,d) \tag{3}$$

C. N-gram model

An "n-gram" is defined as a continuous string of n terms from a given sample of text. In text processing, the n-gram is also known as "shingles." A development of the bag of words model is the n-gram. The n-gram technique produces a sparse matrix that resembles a bag of the words. The distinction is that the newly created columns list both singular words and a set of singular words based on the value of n. A unigram model, which is nothing more than a general collection of words, is what is known when the value of n is 1. The model is known as a bi-gram model when n is 2, and so forth. Take the statement "I read a book about the history of America" as an example. A, a book, about, about the, book, book, history, history of, I, I read, of, of America, read, read a, the, the history would thus be the collection of distinct phrases in the bi-gram model. The existence or absence of words is indicated by a 0 or 1, much like a bag of words. The most rarely occurring words are eliminated from the model, which results in another sparse matrix.

D. Bi-Gram Model

A statistical language model called a "bi-gram model" accounts for the likelihood that two words would appear one after the other. To determine the likelihood of a word given the words that come before and after it, a bi-gram model employs a method known as Markov modeling. For example, if we were to look at the bi-grams of the sentence "The cat ran away" we would get the following bi-grams: "The cat", "cat ran", "ran away". From this, we can calculate the probability of each bi-gram, which in this case would be 1/3 for each bi-gram. This probability is then used to calculate the probability of the whole sentence. The bi-gram model is commonly used in applications such as automatic speech recognition and text summarization.

E. Theme Creation

Bi-gram frequencies allow us to identify motifs that provide context for the product. This can be accomplished manually by choosing the themes that will contribute the most insight to the investigation. Using bi-gram models involves extracting two-word combinations from text data in order to identify topics of discussion. A bi-gram model can be used to identify the most frequently occurring combinations of words in a review, which can then be used to create themes or topics that are discussed in the review. This can help to provide more insight into the sentiment of a review and the topics that the reviewer has discussed.

F. Isolation of Training and Test Data

The procedure of training and testing comes after pre-processing, which is the hardest phase. In this step, a small amount of data is provided to the algorithm, which is then

trained to recognize the data pattern. After the learning process, a certain amount of data must be provided to gauge the algorithm's level of understanding. It is usual to divide the data into two 4:1 subgroups for training and testing. Additionally, this ratio may change depending on the classification algorithm.

G. Classification

We first decompose the dataset into its component components, and then, for the algorithm to learn how to classify the data, we provide training data. The Naive Bayes classifier, the SVM classifier, the Decision Tree classifier, the Random Forest classifier, and other classification techniques have all been utilized. The decision tree creates a tree-like structure with necessary classes at the leaf nodes using a set of questions and criteria. Equation 4 is used to determine entropy and choose between the tree's roots.

$$H = -\sum p(X) log_p(X) \tag{4}$$

A conditional probability model is the foundation of the Naive Bayes method. Equation 5 gives the probability if the data set to be classified is represented as a vector x=(X1,..., Xn) of n independent features. The Naive Bayes classifier presupposes feature independence.

$$P\left(X_1, ...X_n\right) = P\left(C_k\right) \prod_{i}^{n} = P\left(X_i | C_k\right) \tag{5}$$

In this case, Ck stands for the K'th class name. The SVM classifier classifies the input data by establishing one hyperplane in between the set of points x specified by Equation 6.

$$\vec{w}.\vec{x} - b = 0 \tag{6}$$

Here, is the hyper-plane's normal vector. Each classification algorithm has advantages and disadvantages, and depending on the data-set's characteristics, they perform differently. Each NLP technique is applied to various algorithms for checking the accuracy performance.

H. Performance Evaluation

The test data-set is fed to a trained model to determine the performance of any classification technique. By following this process, We can evaluate how well an algorithm has internalized the information and is able to forecast classes of new data. The confusion matrix that we develop contains the total number of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values generated by the supplied data-set. Equation 7 gives the false accept rate (FAR) and the false reject rate (FRR).

$$FAR = \frac{FP}{FP + TN} and FRR = \frac{FN}{FN + TP} \tag{7}$$

Equation 8 can be used to calculate the classification algorithm's precision. How well the algorithm agrees with human judgment is measured by accuracy. Based on accuracy, a suitable algorithm is utilized for subsequent prediction.

$$Accuracy = \frac{TN + TP}{TP + FP + TN + FN} \tag{8}$$

I. Class Prediction

In order to determine which class a new dataset belongs to, it can be passed into the selected algorithm. Since it already understands the nature of the dataset, the machine can construct the most suitable class. An algorithm is updated whenever a new consumer writes a review because we used customer reviews of Amazon smartphones as our dataset, which determines whether it is a favorable or unfavorable review of the cell-phone.

IV. Implementation and Performance Analysis

Python 3.6, Pandas, Google Translate, NLTK, RE, Matplotlib, Seaborn, and Itertools libraries and packages are used to implement the proposed approach. The Kaggle Amazon cellphone review data-set To train the algorithm, to determine if a review is favorable or unfavorable, additional preprocessing is applied to the Kaggle data. We took into consideration at least 1500 reviews with a possible rating for the experiment. Using the modeling outcomes from these 1500 reviews, various confusion matrices are constructed using various algorithms. Table 1 displays the results of the simulation. Table 1 shows that the N-gram model produced superior outcomes. Additionally, when compared to other classification algorithms, the SVM classifier's results are superior. Table 2 displays the confusion matrix produced by the SVM classifier.

Firstly, in this figure, finding the count of words in each review and storing them under the column named "title"

	Index	asin	name	rating	date	verified	title	body	helpfulVotes
0	0	1		3	11-Oct-05	False	Def not best, but not worst	I had the samsung a600 for awhile which is abs...	1.0
1	1	2		1	7-Jan-04	False	Text Messaging Doesn't Work	due to a software issue between nokia and sprt...	17.0
2	2	1		5	30-Dec-03	False	Love This Phone	this is a great, reliable phone. I also purcha...	5.0
3	3	3	ami	5	16-Mar-04	False	Love the Phone, BUT...!	I love the phone and all, because I really did...	1.0
4	4	1		4	28-Aug-05	False	Great phone service and options, lousy case!	the phone has been great for every purpose it ...	1.0
...	...	...	...	...	...	...	...	...	...
1494	1494	2	white	1	25-Aug-14	True	One Star	was reported as lost as soon as it reached ups	NaN
1495	1495	1		1	8-Apr-15	True	One Star	malo	NaN
1496	1496	2	adam	5	24-Sep-14	True	Five Stars	very good	NaN
1497	1497	1		5	11-Apr-13	True	bueno	me llego a tiempo muy bueno, excelente product...	NaN
1498	1498	1		4	26-Apr-11	True	Good product	for sure is not an iphone, but is good to be d...	2.0

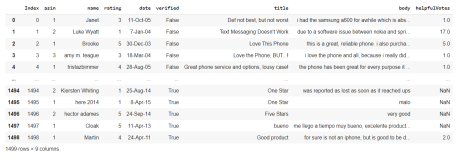
1499 rows × 9 columns

Fig. 2: Count of words in each review and storing

In this figure, all rows with a count of words in each review greater than 15 are added to the new data-frame.

Fig. 3: Table containing words greater than 15

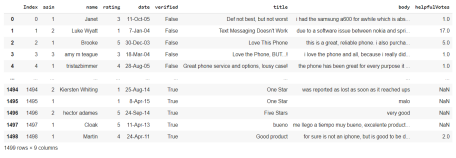
We used the Regular Expressions (RE) library to remove special characters and numbers from these reviews.



	Index	asin	name	rating	date	verified	title	body	helpfulVotes
0	0	1	Janet	3	11-Oct-05	False	Def not best, but not worst	i had the samsung a600 for awhile which is abs...	1.0
1	1	2	Luke Wyatt	1	7-Jan-04	False	Text Messaging Doesn't Work	due to a software issue between nokia and spri...	17.0
2	2	1	Brooke	5	30-Dec-03	False	Love This Phone	this is a great, reliable phone. i also purcha...	5.0
3	3	3	amy m teague	3	18-Mar-04	False	Love the Phone, BUT..!	i love the phone and all, because i really did ...	1.0
4	4	1	trixtadimmer	4	28-Aug-05	False	Great phone service and options, lousy case!	the phone has been great for every purpose it ...	1.0
...	...	...	...	...	...	...	...	...	...
1494	1494	2	Kiersten Whiting	1	25-Aug-14	True	One Star	was reported as lost as soon as it reached ups	NaN
1495	1495	1	here 2014	1	8-Apr-15	True	One Star	malo	NaN
1496	1496	2	hector adames	5	24-Sep-14	True	Five Stars	very good	NaN
1497	1497	1	Cloak	5	11-Apr-13	True	bueno	me llego a tiempo muy bueno, excelente product...	NaN
1498	1498	1	Martin	4	24-Apr-11	True	Good product	for sure is not an iphone, but is good to be d...	2.0

1499 rows × 9 columns

Fig. 4: Table after using RE



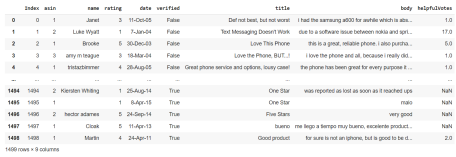
	Index	asin	name	rating	date	verified	title	body	helpfulVotes
0	0	1	Janet	3	11-Oct-05	False	Def not best, but not worst	i had the samsung a600 for awhile which is abs...	1.0
1	1	2	Luke Wyatt	1	7-Jan-04	False	Text Messaging Doesn't Work	due to a software issue between nokia and spri...	17.0
2	2	1	Brooke	5	30-Dec-03	False	Love This Phone	this is a great, reliable phone. i also purcha...	5.0
3	3	3	amy m teague	3	18-Mar-04	False	Love the Phone, BUT..!	i love the phone and all, because i really did ...	1.0
4	4	1	trixtadimmer	4	28-Aug-05	False	Great phone service and options, lousy case!	the phone has been great for every purpose it ...	1.0
...	...	...	...	...	...	...	...	...	...
1494	1494	2	Kiersten Whiting	1	25-Aug-14	True	One Star	was reported as lost as soon as it reached ups	NaN
1495	1495	1		1	8-Apr-15	True	One Star	malo	NaN
1496	1496	2	hector adames	5	24-Sep-14	True	Five Stars	very good	NaN
1497	1497	1	Cloak	5	11-Apr-13	True	bueno	me llego a tiempo muy bueno, excelente product...	NaN
1498	1498	1	Martin	4	24-Apr-11	True	Good product	for sure is not an iphone, but is good to be d...	2.0

1499 rows × 9 columns

Fig. 5: Convert to lowercase

After that we convert all the amazon cell-phone reviews to lowercase letters.

We use the list of stop-words that we downloaded from NLTK and manually add additional words to it. Following that, we go through each and every review to exclude stop words, special characters, and digits.



	Index	asin	name	rating	date	verified	title	body	helpfulVotes
0	0	1	Janet	3	11-Oct-05	False	Def not best, but not worst	i had the samsung a600 for awhile which is abs...	1.0
1	1	2	Luke Wyatt	1	7-Jan-04	False	Text Messaging Doesn't Work	due to a software issue between nokia and spri...	17.0
2	2	1	Brooke	5	30-Dec-03	False	Love This Phone	this is a great, reliable phone. i also purcha...	5.0
3	3	3	amy m teague	3	18-Mar-04	False	Love the Phone, BUT..!	i love the phone and all, because i really did ...	1.0
4	4	1	trixtadimmer	4	28-Aug-05	False	Great phone service and options, lousy case!	the phone has been great for every purpose it ...	1.0
...	...	...	...	...	...	...	...	...	...
1494	1494	2	Kiersten Whiting	1	25-Aug-14	True	One Star	was reported as lost as soon as it reached ups	NaN
1495	1495	1		1	8-Apr-15	True	One Star	malo	NaN
1496	1496	2	hector adames	5	24-Sep-14	True	Five Stars	very good	NaN
1497	1497	1	Cloak	5	11-Apr-13	True	bueno	me llego a tiempo muy bueno, excelente product...	NaN
1498	1498	1	Martin	4	24-Apr-11	True	Good product	for sure is not an iphone, but is good to be d...	2.0

1499 rows × 9 columns

Fig. 6: NLTK and Parsing

V. Results and Discussion

We used Porter's stemmer to implement the stemming of words in each review. Then, we create an empty list by appending all the words in all of the reviews to the list. After applying for loop in Python 3.6, the list now contains 2545 words, some of which may have more than one occurrence. We generated the following graph using Matplotlib and Seaborn to draw the graphs of word frequencies.

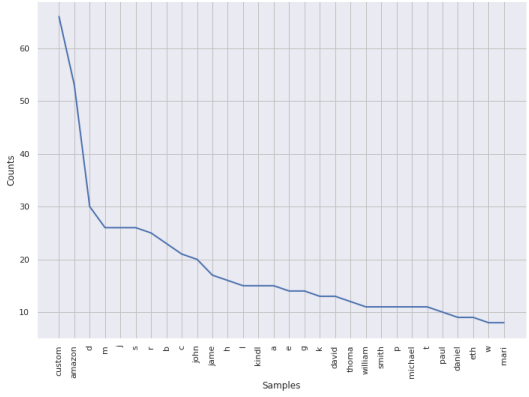


Fig. 7: Graph of word frequency

The function called hapaxes returns, a list of words having a frequency of 1. We find that the length of the frequency is

TABLE I: Converting of the dictionary to a list of tuples

	0	1
0	**janet**	3.0
1	**ami**	6.0
2	**m**	26.0
3	**white**	26.0
4	**owner**	2.0
5	**matt**	3.0
6	**charl**	2.0
7	**cook**	2.0
8	**amazon**	53.0
9	**custom**	66.0
10	**o**	5.0
11	**thoma**	12.0
12	**william**	11.0

TABLE II: Frequency table in descending order

	0	1
226	**jack**	3.0
153	**walker**	3.0
168	**steve**	3.0
175	**rick**	3.0
238	**jess**	3.0
160	**willi**	2.0
161	**schulman**	2.0
163	**helen**	2.0
164	**hawkin**	2.0
281	**jean**	2.0
238	**jess**	3.0
160	**willi**	2.0

1279. From this, we can remark that a huge number of words occur only once, and hence they will not provide any valuable insight to the analysis; hence, we can add this list to the list of stop-words and remove it from the reviews, reducing the computation time of codes. The length of b is 1266.

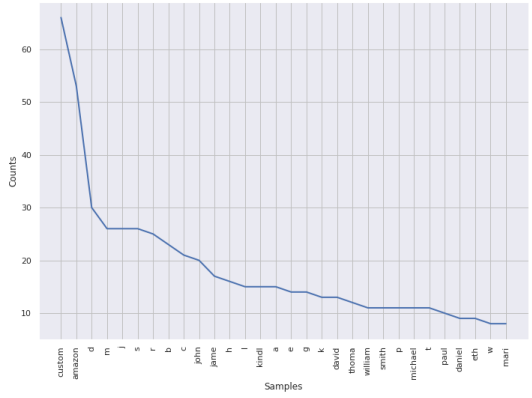


Fig. 8: Hapaxes function graph

NLP words contains all the words in the list, stored in a dictionary format with the word as the key and its frequency as the value. The next step is converting the dictionary to a list of tuples. After that, we create an empty data frame and store the values from list D into it.

Now, we sort this in descending order of frequency, and the following table shows that.

At this point 9, we tried making a list of all the keywords.

The length of the keyword is 282. Also, we found the number of reviews that contained a particular word from the new list of keywords.

	Index	asin	name	rating	date	verified	title	body	helpfulVotes
0	0	1	janet	3	11-Oct-05	False	Def not best, but not worst	i had the samsung a600 for awhile which is abs...	1.0
1	1	2		1	7-Jan-04	False	Text Messaging Doesn't Work	due to a software issue between nokia and spri...	17.0
2	2	1		5	30-Dec-03	False	Love This Phone	this is a great, reliable phone. i also purcha...	5.0
3	3	3	ami	3	18-Mar-04	False	Love the Phone, BUT...!	I love the phone and all, because i really did...	1.0
4	4	1		4	28-Aug-05	False	Great phone service and options, lousy case!	the phone has been great for every purpose it ...	1.0
...	...	...	...	...	...	...	...	...	...
1494	1494	2	white	1	25-Aug-14	True	One Star	was reported as lost as soon as it reached ups	NaN
1495	1495	1		1	8-Apr-15	True	One Star	malo	NaN
1496	1496	2	hector adam	5	24-Sep-14	True	Five Stars	very good	NaN
1497	1497	1		5	11-Apr-13	True	bueno	me llego a tiempo muy bueno, excelente product...	NaN
1498	1498	1	martin	4	24-Apr-11	True	Good product	for sure is not an iphone, but is good to be d...	2.0

1499 rows × 9 columns

Fig. 9: List of all keywords

At this point10, we remove all words from the reviews except the top 100 frequency words by adding them to the list of stopwords. The length of the keyword is 100 in this case.

	Words	Frequency	Indices
0	custom	66.0	[9, 27, 45, 64, 109, 116, 136, 174, 193, 199, ...
1	amazon	53.0	[9, 27, 45, 109, 116, 136, 174, 193, 199, 218,...
4	j	26.0	[5, 46, 89, 122, 176, 190, 192, 194, 296, 319,...
6	r	25.0	[26, 154, 157, 202, 210, 305, 350, 356, 429, 4...
7	b	23.0	[15, 97, 99, 140, 270, 315, 320, 323, 346, 428...
...	...	...	...
142	jordan	3.0	[141, 793, 1419]
140	robin	3.0	[132, 847, 1240]
139	sue	3.0	[179, 523, 818]
137	mike	3.0	[128, 694, 1232]
136	ann	3.0	[190, 957, 1022]

100 rows × 3 columns

Fig. 10: Frequency of top 100 words

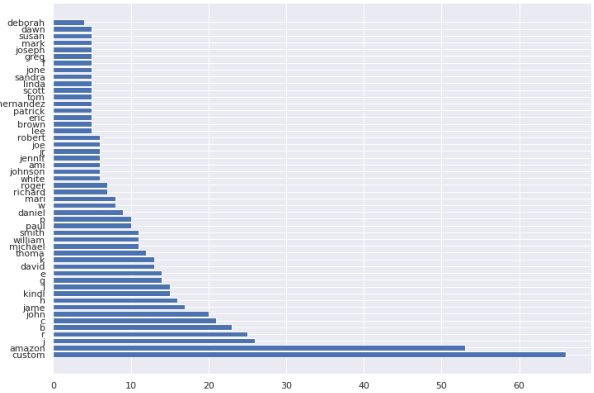


Fig. 11: Words vs Frequency plot

To make word combinations, we now import the Itertools library12.The length of the pair is 4950.

From the graphs, we can observe the pair frequencies and hence decide the themes to look for in the review.

	0	1	2	3
0	custom	amazon	50.0	['45', '695', '922', '646', '230', '966', '354...
1	custom	j	0.0	[]
2	custom	r	0.0	[]
3	custom	b	0.0	[]
4	custom	c	0.0	[]
...	...	...	...	...
4945	robin	mike	0.0	[]
4946	robin	ann	0.0	[]
4947	sue	mike	0.0	[]
4948	sue	ann	0.0	[]
4949	mike	ann	0.0	[]

4950 rows × 4 columns

Fig. 12: Word combinations - 1

	Word1	Word2	Frequency	Indices
0	custom	amazon	50.0	['45', '695', '922', '646', '230', '966', '354...
8	custom	kindl	15.0	['291', '907', '791', '491', '274', '426', '21...
1554	william	roger	2.0	['331', '263']
296	r	john	2.0	['464', '26']
305	r	thoma	2.0	['489', '742']
...	...	...	...	...
1705	smith	jordan	0.0	[]
1704	smith	adam	0.0	[]
1703	smith	pablo	0.0	[]
1702	smith	jim	0.0	[]
4949	mike	ann	0.0	[]

4950 rows × 4 columns

Fig. 13: Word combinations - 2

	Word1	Word2	Frequency	Indices	Words
0	custom	amazon	50.0	['45', '695', '922', '646', '230', '966', '354...	custom amazon
8	custom	kindl	15.0	['291', '907', '791', '491', '274', '426', '21...	custom kindl
1554	william	roger	2.0	['331', '263']	william roger
296	r	john	2.0	['464', '26']	r john
305	r	thoma	2.0	['489', '742']	r thoma
672	jame	h	2.0	['448', '471']	jame h
207	j	david	2.0	['375', '1061']	j david
1213	david	smith	2.0	['560', '1301']	david smith
1629	smith	paul	2.0	['394', '520']	smith paul
227	j	lee	1.0	['190']	j lee
1097	g	ali	1.0	['1160']	g ali
217	j	mari	1.0	['690']	j mari
220	j	white	1.0	['5']	j white
221	j	johnson	1.0	['383']	j johnson
496	c	william	1.0	['983']	c william
224	j	jr	1.0	['46']	j jr
3630	dawn	michel	1.0	['1062']	dawn michel
226	j	robert	1.0	['355']	j robert
1194	e	nanci	1.0	['754']	e nanci
770	h	thoma	1.0	['183']	h thoma
493	c	k	1.0	['122']	c k
230	j	patrick	1.0	['573']	j patrick
3525	mark	lewi	1.0	['607']	mark lewi
1088	g	man	1.0	['1160']	g man
1087	g	timothi	1.0	['1466']	g timothi
490	c	g	1.0	['1399']	c g
489	c	l	1.0	['1195']	c l
1082	g	rosemari	1.0	['365']	g rosemari
239	j	joseph	1.0	['296']	j joseph
228	j	brown	1.0	['1092']	j brown
211	i	william	1.0	['46']	i william

Fig. 14: 50 selected word combinations

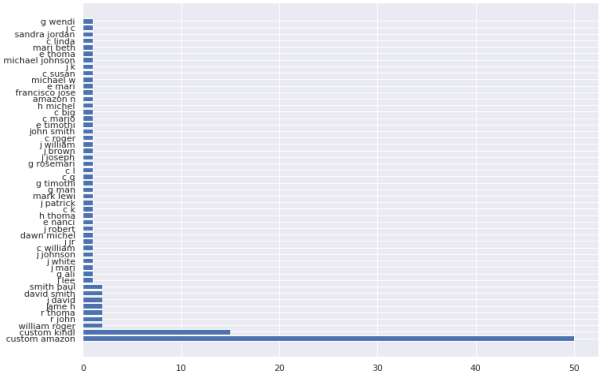


Fig. 15: Words vs Frequency plot-2

VI. Conclusion

The performance of several algorithms and NLP techniques on a dataset of Amazon cell-phone reviews is compared in this study, and the algorithm with the best performance is examined. The SVM classifiers have the maximum accuracy in this case, as we've seen. However, there is always a cost associated with FAR and FRR when determining the algorithm's accuracy. As a result, SVM achieves the greatest accuracy, it is therefore the best classification algorithm for the dataset in question. Different kernel functions and settings can be tested in order to further enhance SVM performance. In this, research we are looking for information from product reviews to determine what the main drawbacks and benefits are. To accomplish the same, we have used data mining and natural language processing concepts. Python 3.7 was used in this instance.

References

[1] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artificial intelligence review*, vol. 52, no. 3, pp. 1495–1545, 2019.

[2] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 2016, pp. 628–632.

[3] Y. Yang, "Research and realization of internet public opinion analysis based on improved tf-idf algorithm," in *2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)*. IEEE, 2017, pp. 80–83.

[4] ——, "Research and realization of internet public opinion analysis based on improved tf-idf algorithm," in *2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)*. IEEE, 2017, pp. 80–83.

[5] V. Singh and B. Saini, "An effective tokenization algorithm for information retrieval systems," 2014.

[6] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia computer science*, vol. 17, pp. 26–32, 2013.

[7] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: a survey. ain shams eng. j. 5 (4), 1093–1113 (2014)."

[8] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in *2014 Seventh international conference on contemporary computing (IC3)*. IEEE, 2014, pp. 437–442.

[9] R. Ravi, "Ravi k., ravi v," *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.

[10] Ò. R. Llombart, "Using machine learning techniques for sentiment analysis," 2017.

[11] V. Kharde, P. Sonawane, *et al.*, "Sentiment analysis of twitter data: a survey of techniques," *arXiv preprint arXiv:1601.06971*, 2016.

[12] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 2016, pp. 628–632.

[13] D. M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," *Journal of King Saud University - Engineering Sciences*, vol. 30, no. 4, pp. 330–338, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1018363916300071

[14] A. Addiga and S. Bagui, "Sentiment analysis on twitter data using term frequency-inverse document frequency," *Journal of Computer and Communications*, vol. 10, no. 8, pp. 117–128, 2022.