Customer Reviews Sentiment Analysis: A hybrid technique of Lexicon and Machine Learning based Classification model (SVM, NB, Logistic Regression)

MSc Research Project
Data Analytics

Komal Vijay Bhalerao
Student ID: 20135386

School of Computing
National College of Ireland

Supervisor: Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Komal Vijay Bhalerao
Student ID:	20135386
Programme:	MSc Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Christian Horn
Submission Due Date:	16/08/2021
Project Title:	Customer Reviews Sentiment Analysis: A hybrid technique of Lexicon and Machine Learning based Classification model (SVM, NB, Logistic Regression)
Word Count:	7466
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Komal Vijay Bhalerao
Date:	16th August 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Customer Reviews Sentiment Analysis: A hybrid technique of Lexicon and Machine Learning based Classification model (SVM, MNB, LogReg)

Komal Vijay Bhalerao
x20135386

Abstract

The majority of items are available online in our digital age. E-commerce platforms are evolving in order to put products within the reach of online users in order to deliver the highest level of customer pleasure and convenience of use. People nowadays tend to rely on feedback before ordering any goods online, therefore reading hundreds of evaluations takes up a lot of time for customers. Making decisions on enhancing quality of the product and acquiring insights, companies and organizations can obtain lot of data from customer sentiment analysis. A lot of research has previously been implemented on the classification of Sentiment Analysis based on many different aspects and techniques, however, not a lot of research has been done with a combination of Lexicon based and Machine Learning classification model. The process of Sentiment analysis can be tedious since the data available is textual format and it is the most unstructured type of data available. In this research, to enable efficient and outstanding outcome for classification, text pre-processing is carried out and two types of feature extractors are used. In order to fulfil this task, three machine learning models were implemented. The outcome generated by these models were evaluated using different evaluation matrices and the results were compared. SVM provided the best accuracy for classification i.e. 91% using TF-IDF vector.

Keywords- NLP, Sentiment Analysis, Machine Learning, TF-IDF, CountVectorizer, SVM, Naive Bayes, Logistic Regression

1 Introduction

The recognition and classification of Human Sentiment for understanding and monitoring the emotion or sentiment behind a product in order to recommend similar product based on the previous reviews of customers is an imperative task. Recent decade has witnessed the change of manual door to door shopping to the digitization of shopping online and thus the Internet is now the go-to place for customer to buy things online according to their convenience.

1.1 Background and Motivation

The era of customer service has arrived. When any organization meets the needs of it's customers providing the best service, it is said to be growing. Therefore, it is important for an organization to get feedback on each product for improving product manufacturing or correcting any anticipated faults. Reviews assist customer's making better decisions of whether to invest buying a product online or not. The motivation behind this research is to find whether a combination of lexicon and machine learning based classification models like SVM, Naive Bayes, Logistic Regression the addition of word vector based feature extraction techniques like Count Vector and TF-IDF vector achieve better outcome in terms of accuracy for predicting the polarity of sentiments for products on reviews posted by customers.

One of Amazon's most well-known offerings is Amazon Customer Reviews (also known as Product Reviews). Millions of Amazon customers have posted over a hundred million comments to express their thoughts and describe their experiences with products on the Amazon.com website in the nearly two decades since the first review in 1995. This makes Amazon User Reviews a valuable resource for academic scholars working in fields such as Information Retrieval (IR), Natural Language Processing (NLP) and Machine Learning (ML). Sentiment Analysis in other terms, Opinion Mining is a assessable representation of opinions and emotions behind a text. Positive, negative and neutral are the three possible sentiments classification found behind a text. Sentiment Analysis has gained immense popularity in recent years. In Machine Learning, it is termed as one of the types of classification approach of text that is based on Sentimental Orientation (SO) of opinions contained in the form of text (Jha et al.; 2021).

Natural Language Processing (NLP), one of the domains of Machine Learning is the capability of machines of reading, analyzing and extracting the simplification of human language. Due to the exponential advancements in computational power and data access by authorizing in deriving insightful meaning, NLP is advancing immensely in the domains like healthcare, finance, media and human resources. In other words, NLP allows the handling of human emotions carried out in the form of text or speech (May et al.; 2020). There are many techniques used in NLP such as, Text summarization, Aspect Mining, Topic Modeling and Sentiment Analysis. This research focuses on Sentiment Analysis . Handful amount of research has been implemented in the past on classification of sentiments using several methods and techniques. Sentiment analysis divides the subjective effects of the user's feelings into polarities. In general, there are two types of techniques to dealing with this issue.

- **Approaches based on Machine Learning**, One of the most well-known techniques is this one. Researchers are becoming interested in this technology because it is more versatile and precise. There are two types of machine learning methods for sentiment classification: supervised and unsupervised. Classification methods are usually classified as either supervised or unsupervised. Supervised machine learning methods such as Decision trees, Naive Bayes, and others are used for classification, while unsupervised learning techniques such as Neural networks are used for clustering.

- **Lexicon-based approach**, The simplest method for conducting sentiment analysis is to use lexicon-based techniques. This method takes use of a dictionary that contains words that have already been pre-tagged. To perform sentiment analysis, WordNet and SentiwordNet are publicly available dictionaries (Khanvilkar and Vora; 2019).

However, not a lot of work has been done in the past of exploring a combination of Machine-Learning and lexicon based models with different feature extraction techniques suitable for the data available and classifying the sentiment in terms of thorough pre-processing against the ranking/ratings of the products. Classifiers like Support Vector Machine (SVM) and Naive Bayes have proved to be very effective for a classification problem (Shivaprasad and Shetty; 2017). Alongside extraction of features, the outcome provided by several machine learning model is remarkable (Gamal et al.; 2019). However, the scope of the data used is large and the experiments are conducted on variety of datasets. Therefore, classification of Sentiment by extracting feature on a subset from the entire set has not been explored vastly and from the smaller subset, is the scope on which the project is revolved around. This research uses a subset of products for a decent amount of data for accomplishing the task of Sentiment Analysis.

1.2 Project Requirements Specification

This research project aims to classify reviews by conducting Sentiment Analysis of Beauty products available on Amazon.com. This will help customer in a way providing the product with most positive ratings and negative ratings which will help the customer with the decision-making process.

1.2.1 Research Question

The main focus is on comparing and evaluating the production/performance of different Machine Learning Classification models for classifying the customer uploaded feedback and the sentiments or emotions expressed on various beauty products available on Amazon.
RQ: *"How well can Lexicon and Machine Learning based predictive classification models (SVM, Naive Bayes and Logistic Regression) be used on customer reviews for Sentiment Analysis?*

1.2.2 Research Objective and Contributions

The research project includes pre-processing the data, implementation of different feature extraction techniques, implementation of three Machine Learning Classification algorithms such as SVM, Naive Bayes and Logistic Regression. Further, the results produced by these models are evaluated using different evaluation techniques and compared in terms of performance.
The Research Objectives are:

1. **Objective 1:** Critically review the literature on NLP, sentiment analysis using ML and other techniques.

2. **Objective 2:** Download data available online for research and pre-process data in order for it to be ready for feature extraction and classification.

3. **Objective 3:** Implement and Evaluate classification models (SVM, Naive Bayes, Logistic Regression).

4. **Objective 4:** Compare results of the classification models.

Contributions: The project contribution includes implementation of a combination of a hybrid technique (lexicon and machine learning) based classification models with the addition of using two feature extractors. A subset of Amazon product reviews was used.

This will help customer in a way providing the product with most positive ratings and negative ratings which will help the customer with the decision-making process .

The structure of the technical report at hand is as follows. Section 2 highlight and review the most relevant literature review as per the objective of this project. Section 3 explain in details, the methodologies used, the design specification and architecture along with the data pre-processing steps followed. The implementation of the algorithms used, the evaluation and results is described in Section 4. Lastly, in section 5 and 6 the discussion of results and future work is covered, respectively.

2 Related Work

2.1 Introduction

The literature review is an important aspect of the planned research paper since it gives a thorough examination of the work done in the field. It provides an overview of the research topic and the hypothesis to be tested. Numerous Sentiment Analysis studies have been conducted in the past.This review compares several algorithmic models to examine the impact of Machine Learning approaches.The complete review of relevant papers is divided into different categories reviewed below.

2.2 Application of Natural Language Processing in Sentiment Analysis

Objective parameters and subjective parameters are the two types of parameters that are considered for mining the customer reviews while using opinion mining, suggested in the research conducted by (Rajeev and Rekha; 2015). Age and count of reviews, star ratings are contained in the objective parameter. Whereas, sentiment or emotion of the customers regarding a product particularly is contained in the subjective parameter. Implementation of objective parameters is done using mathematical calculations on the other hand, different techniques of Natural Language Processing (NLP) are used for implementing subjective parameters. Based on the adjectives used, these NLP techniques derive the sentiment or opinion conveyed by the customer in terms of reviews to be either positive or negative. The research solely focuses on one category of product i.e Mobile phone from flipkart. Step-wise approach methodology is implemented in this research such as Review extraction, review process, feature extraction of reviews, product score classification and comparing products. Natural Language Toolkit (nltk) and POS tagging is used in order to process the reviews.

According to research conducted by (Shivaprasad and Shetty; 2017) the Sentiment Analysis classification is divided into three levels also suggested by (Solangi et al.; 2018) as a discovery in the research i.e. Document level, sentence level and aspect level. In this research, the framework of Sentiment Analysis is discussed on a basic level and the process is followed including steps such as Data preparation, Review Analysis and classification of sentences, (Rajeev and Rekha; 2015) also followed the same steps in their research conducted. Further, the polarity of the sentiment is classified into three approaches. These three approaches are namely, Binary approach, Multi-level approach and Fuzzy approach or contextual approach, as evident in the research. In Binary approach, as the name suggests the sentiments are classified into two categories i.e. positive and negative. Star-based inference approach, other name for Multi-class approach is where the star-based ratings system is carried out. The star rating system is categorized into lowest(1 star) to highest (5 stars).The three measures of polarity i.e positive, negative and neutral is focused in the third approach i.e. Fuzzy approach or contextual approach. Additional, in continuation the research discussed two different approaches/ methods used for Sentiment Analysis is discussed briefly namely, lexicon-based approach and machine learning approach. In this research, the Binary approach is followed along with the combination of both approaches for sentiment analysis is implemented.

Natural Language Processing (NLP) and tool-kits that are proved to be extremely useful in dealing with tokenization and word segmentation is reviewed in the research conducted by (Solangi et al.; 2018). Text processing techniques such as Segmentation, tokenization, POS Tagging and Conditional Random Fields(CRFs) are briefly described in this research. Many powerful tools in NLP for tokenization and segmentation such as LTP tool, Parser:C++, Gensim python etc are also described explicitly. The research continued with the process of sentiment analysis by dividing the process in three different levels. These three levels included Document-level, sentence-level and fine-grained level. In the Document level i.e the first level, the tone conveyed in a sentence is determined as positive, negative or neutral. Moving forward in the second level i.e. the sentence level, the evaluated outcome in the previous step are transpired to be negative, positive or neutral. In conclusion, the research suggested use of different applicable techniques for retrieving correct information or future modification of NLP.

2.3 Application of Sentiment Analysis on customer reviews using Machine Learning

Three different classification models are employed in this study by (Hermansyah and Sarno; 2020) such as Naive Bayes, TextBolb, and KNN (K-Nearest Neighbors). The study's conclusion includes a comparison of the effectiveness of the three algorithms utilized. The KNN algorithm generated the highest performance, with a total accuracy of 75%, followed by the Naive Bayes algorithm, which scored 70%, and TextBolb, which scored the lowest of all three i.e. 55%. The Naive Bayes algorithm will be used in this study, and the level of accuracy is attempted to increase.

In addition to the NLP approaches mentioned in the preceding section, (Shivaprasad and Shetty; 2017) employed Machine Learning methods like as Naive Bayes, Maximum Entropy, and Support Vector Machine (SVM). When it comes to evaluating sentiment, classifiers are essential. In comparison to SVM, the Naive Bayes method showed to be more efficient and scalable when it came to classifying text. Many scholars, on the other alternative, utilize SVM algorithms as complex classifiers, and the performed experiment shown that SVM achieves higher performance in terms of accuracy. Maximum Entropy, on the other hand, is suitable for probability distribution. To summarize, when compared to other methods, SVM showed to be the most accurate.

According to (Jabbar et al.; 2019), the POST (Parts Of Speech) approach may categorize phrases using adverbs, adjectives, adverbs, and verbs. For evaluations like "not good" and "good," the Negation Phase Identification cation algorithm was used to gain a better understanding of the sentiment behind the feedback. The reviews are gathered further in the study from Amazon.com, namely on products in the Cosmetics and Musical Instruments sections. The SVM (Support Vector Machine) method is used in conjunction with a mobile app to calculate the total percentage of positive and negative evaluations for a product. From the results obtained so far in the field, Support Vector Machine (SVM) is among the successful algorithms employed for Sentiment Analysis. In order to attain better outcomes, SVM is employed in this study.

Three distinct methods, such as Naive Bayes, SentiWordNet & Logistic Regression, were employed in the study by (Jabbar et al.; 2019) to detect the sentiment underlying reviews by categorizing them as positive or negative. The proposed process is broken down into multiple steps, such as Text Extraction, which involves extracting text from the URL of the specified website after providing the credentials. Amazon.com is used to gather reviews in this scenario. In the next step, a list of items is displayed, from which the user can select a product to obtain reviews, as well as the classification of the product as positive or negative. The study concluded that, of the three algorithms tested, the Naive Bayes classification algorithm performed well.

(Gamal et al.; 2019) examined several Machine Learning Models for sentiment classification and review mining utilizing a variety of data-sets including IMDB, Amazon, Twitter, and Cornell. Extraction Of features & Machine Learning Classification are the two steps in the process. The TF-IDF (Term Frequency-Inverse Document Frequency) and N-Gram Algorithms are the primary feature extraction methods utilized. Naive Bayes, Multinomial Naive Bayes(MNB), Bernoulli Naive Bayes(BNB), Passive Aggressive(PA), SGD, AdaBoost, Ridge Regression and Logistic Regression are among the ML algorithms used in this study. K-fold Cross Validation was utilized as the evaluation technique. Finally, PA and Ridge Regression, when combined with various feature extraction techniques, outperformed with accuracy ranging from 87 percent to 99.96 percent.

Machine Learning-based sentiment analysis is a well-studied topic. (Singh and Sarraf; 2020) did a study that used both customer and seller comments and used Random Forest Classification for Sentiment Analysis to classify the reviews. BeautifulSoup, a Python library, is used to get live data from Flipkart.com. The website's live information is extracted using URLs. The BOW (Bag Of Words) method is used to generate output that is either 0 or 1 based on whether the feedback is positive or negative. A User Interface (Application) is designed to compare the overall percentage of negative and positive evaluations for a given product. This implementation can be done on a variety of websites, including EBay, Amazon and many more E-commerce websites.

2.4 Application of Deep Learning in Sentiment Analysis

Natural Language Processing has been frequently used in fields such as text classification, machine translation, speech recognition, and text classification, in addition to Sentiment Analysis. In the domain of sentiment analysis, deep learning techniques are emerging to anticipate sentiment reviews and opinions. In the suggested domain, there is a lot of research going on. (Prabha and Srikanth; 2019) conducted a survey that discusses brie facts about deep learning algorithms at both the target/aspect and sentence level. The disadvantages of doing sentiment analysis utilizing Machine Learning methods are fully explored. The BoW method has a major flaw because it loses the sequence of the words and excludes semantic information, resulting in the sentence's core being lost. The N-gram method, on the other hand, solves the disadvantage of BoW but struggles from data fineness. The survey comes to a close with a review of Deep Learning approaches that may be used with Sentiment Analysis.

(Seetharamulu et al.; 2020) designed and built a Deep Learning for Emotion Recognition method. The effectiveness of deep learning is shown to be largely dependent on the availability of large-scale training data in this study. The goal of this study is to create a Deep Learning-based framework for categorizing customer evaluations as positive or negative. This framework makes use of the many product ratings available to categorize the emotion of the review. A prototype application is created to illustrate the proof of concept. When comparing WDE to baseline approaches such as CNN, it is clear that WDE outperforms and produces beneficial outcomes. The accuracy of the findings is assessed.

Deep learning is a strong approach for sentiment analysis that is employed in machine learning. Multiple levels of data characteristics are used to get the prediction results. Deep learning has gained a lot of traction in the field of sentiment analysis, due to its success in other fields (Zhang et al.; 2018).According to (Gupta et al.; 2020) latest research in this area, maximum scalability may be obtained by adopting Aspect-Based sentiment analysis. The different polarity aspects of a sentence can be found using ABSA (Aspect-Based Sentiment Analysis). The suggested CNN (Convolutional Neural Network) technique was applied in this study, with generic and domain-specific embeddings produced using Word2Vec for training purposes. The polarity of a review is determined using the Textbolb technique. Later, Precision, Recall, Accuracy, and F-measure are used to analyze the Deep Neutral Network architecture. The model has an overall accuracy of 89%.

However, (Yang and Yang; 2020) split Aspect-Based Sentiment Analysis (ABSA) into two sub-categories: Aspect-term sentiment analysis (ATSA) and Aspect-category sentiment analysis (ACSA).Aspect-term sentiment analysis (ATSA) and Aspect-category sentiment analysis (ACSA) are two types of sentiment analysis (ACSA). Models for sentiment analysis such as RNN (Recurrent Neural Network) and LSTM (Long Short-Term Memory) have complicated structures and take more training time, thus in this study, a novel model based on gating mechanism with the combination of Self-attention mechanism and CNN is introduced. Self-attention is utilized in the first stage to extract the structural characteristic of information. Later, CNN takes the original sentence's characteristics and combines them with the result. To generate a final sentiment feature, the initial sentence is merged with an aspect-term or an aspect category from the previous phase. Finally, the experiment proved to be effective. The accuracy of the self-attention gated CNN model for ACSA was 82% and 80 % for 10 experiments. The accuracy of the ATSA model was 72 % and 63 %, respectively.

2.5 Hybrid approach (Lexicon-based and Machine Learning) based Sentiment Analysis

To improve accuracy, (Haque et al.; 2018) used two distinct feature extractors: BoW and TF-IDF & chi-square. The experiment is carried out on an unlabeled supervised sample. Both an active and a manual technique are used to label the dataset. To get positive outcomes, several simulation approaches such as cross validation and train-test ration are used.SVM (Support Vector Machine) produced the best findings of all the classification methods used. In most situations, however, a 10 fold increase in accuracy gave superior outcomes. The total accuracy in this research was 90%, and the assessment procedures Precision, Recall, and F1 were used to obtain 90% results.

Machine Learning methods and Lexicon-Based methodologies are used as a hybrid approach for Sentiment Analysis. Because reviews are often categorized as positive or negative, a hybrid model was used to solve the neutral class (Rajeswari et al.; 2020). Naive Bayes, Logistic Regression, Support Vector Machine and Decision Tree are some of the Machine Learning techniques used. The data-set used in the proposed research work include product, twitter, and movie reviews. SentiWordNet is recommended for classifying neutral reviews, in addition to positive and negative evaluations. The features are retrieved after pre-processing texts utilizing BoW (BagOfWords), n-gram and TF-IDF. Metrics like as AUC, recall, precision, f-measure, and accuracy are used in the assessment. In conclusion, TF-IDF showed to be more accurate than others for feature extraction, and furthermore Logistic Regression provided higher accuracy than the other classifiers employed in the research.

When compared to lexicon-based or Machine Learning approaches, the use of a hybrid method improves sentiment analysis classification results. A hybrid method, which combines Machine Learning and Lexicon-based approaches to produce optimal outcomes, is used to achieve the best results. Different hybrid methods and approaches are described by (Ahmad et al.; 2017). pSenti, SAIL, NILC USP, and Achemy API are among the tools available. It has been observed that Lexicon-based models improve performance when there are obvious borders between emotions, but Machine Learning techniques perform better when no distinct limits are specified. Because of its simplicity of application in opinion mining or sentiment classification, the hybrid model is significant and successful.

2.6 Application of Word Vectors technique in Sentiment Analysis

Word representation is considered a critical component in many Natural Language Processing systems. Vector-based models outperform other models in this assessment. Vector-based models express consistent similitude of words as angle or length between word vectors in a wide expanse (Maas et al.; 2011). When unsupervised vector-based algorithms are combined, it creates useful lexicons, but it lacks the ability to comprehend sentiment information, which is fundamental to many NLP jobs and word meanings. The model provided used a combination of supervised and unsupervised methods. A dictionary of 5000 frequently used phrases is constructed in the form of tokens for movie reviews from IMDB to understand word representation. Later in the evaluation process, the Document Polarity classifier is used to anticipate whether the reviews will be positive or negative.The goal of the study was to limit the representation of words. The word representations might capture both semantics and sentiment if they were included to an unsupervised model.

Glove, an unsupervised technique, has been shown to be effective in interpreting the meaning and emotions behind a sentence or review using vector representation of words.To solve the lack of consistency and effectiveness, (Sharma et al.; 2017) employed a mix of the above skip gram and Continuous Bag of Words and developed the GloVe approach. The dot product of a vector and other vectors is used to measure the correlation between two words. Later, it is determined that the greatest dot product is indistinguishable.

The goal of (Fan et al.; 2016)paper is to test the effects of word vector representations for Sentiment Classification. The tasks were further divided into three sub-tasks in this study: extracting sentiment words, detecting sentiment words polarities, and text sentiment predictions. Using diverse text data on vector representations, the efficiency of domain-dependent vectors is assessed. Accuracy, recall, and F1 were among the measures used to conclude the assessment. On APP reviews for text sentiment analysis, the outcomes are 87 %, 86 %, and 85.5 %, respectively.

2.7 Conclusion

Several research papers were examined in this part with reference to the application of Machine Learning, deep learning, NLP, and other domains in the Sentiment Analysis classification based on rankings. It also emphasizes the small amount of study that has been done on classifying sentiment using various feature extraction methods and techniques. Only a few authors have investigated the lack of implementation of feature extraction techniques in limited cases. Objective 1 has been met in this section. The methodology used in the technical report is depicted in the next section.

3 Sentiment Analysis Methodology and Design

The Knowledge Discovery in Databases (KDD)[1] and CRISP DM[2] techniques are used in most data mining programs. The KDD approach, on the other hand, is used in this research project. The general sequential flow of the project is shown below, which depicts the phases involved in putting the study into action, from data collecting through outcomes evaluation. Figure 1 depicts the KDD flow diagram for Sentiment Analysis.

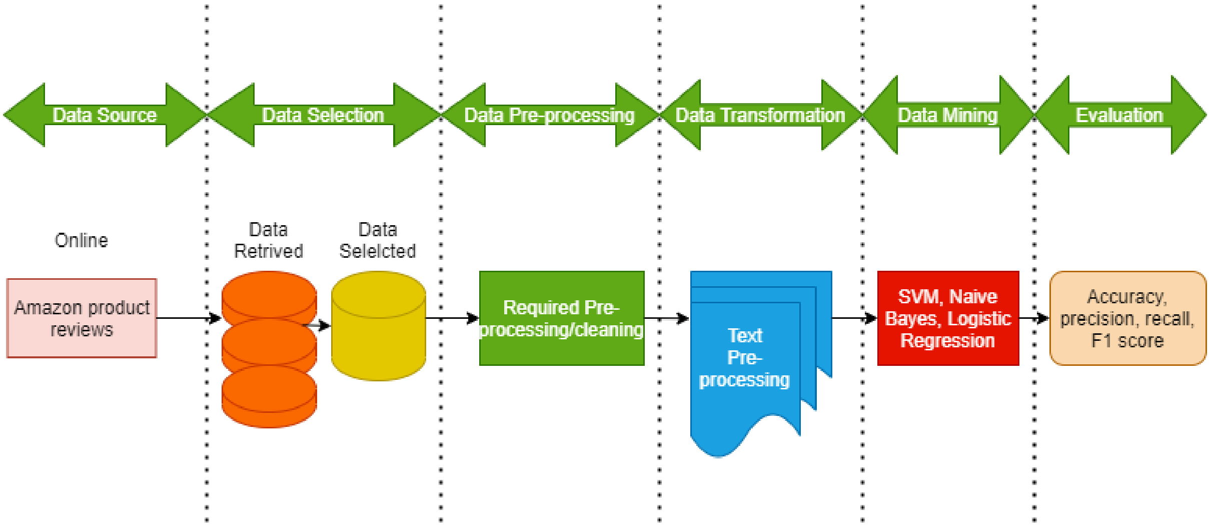


Fig 1. KDD Process flow for Sentiment Analysis

3.1 Data Selection

THe dataset chosen for the research was available online, published by a renowed author for research purposes [3].From 1995 through 2015, a collection of comments and associated metadata were written on the Amazon.com marketplace. This is to make it easier to research the qualities (and evolution) of customer reviews, which could include how individuals evaluate and share their experiences with large-scale items. (Over 130 million

[1] https://www.javatpoint.com/kdd-process-in-data-mining

[2] https://www.datascience-pm.com/crisp-dm-2/

[3] http://deepyeti.ucsd.edu/jianmo/amazon/index.html

customer reviews). The dataset is distributed in different categories of products available on amazon and is accessible on request basis. The downloaded data however is a sub-category. Sub-category of ”All Beauty” products is used for this research project among other categories. Both original data (3271,345 reviews) and meta data (32,992 products) has been used in this project respectively.

3.2 Data Pre-processing and Transformation

Data Pre-Processing:
The data that was available for the project was in the json format. As a result, the initial step in data pre-processing was to convert the json format to csv in order to use pandas to construct dataframes. To improve the findings, the original data and meta data were blended to create a combination of columns. The data was in raw format, it was necessary to clean it before applying Machine Learning models to it. Data pre-processing has been implemented, including the removal of null values, renaming columns, removing irrelevant or unnecessary columns, and converting to datetime format etc.

Data Transformation:
The data was ready for transformation after pre-processing. Text pre-processing is carried out in this step. A subset of ”perfumes” as a product is gathered for this project from the numerous beauty items available. Using the ratings column, a rating class is created to categorize the ratings as good or bad. As shown in Figure 2, the reviews were rated on a scale of 1 to 5, with 1 being the lowest and 5 being the highest, based on the number of reviews available online.

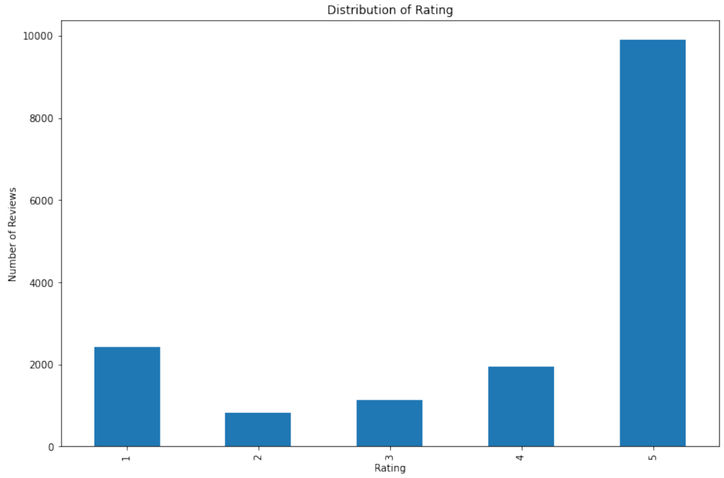


Fig 2. Distribution of Ratings against Number of reviews

Following that, several text pre-processing procedures such as the removal of stop words, the removal of special characters, the conversion to lower case, the removal of punctuation, and the removal of special characters has been implemented using libraries such as Spacy, gensim, textBolb, nltk and more [4]. Detailed description is provided in the respective implentation sections. To obtain clean text from the summary text, the normalize and lemmaize methods are utilized. Further the data is split into train and test with the ratio of 75(train):25(test).

[4] https://www.infoworld.com/article/3519413/8-great-python-libraries-for-natural-language-processing.html

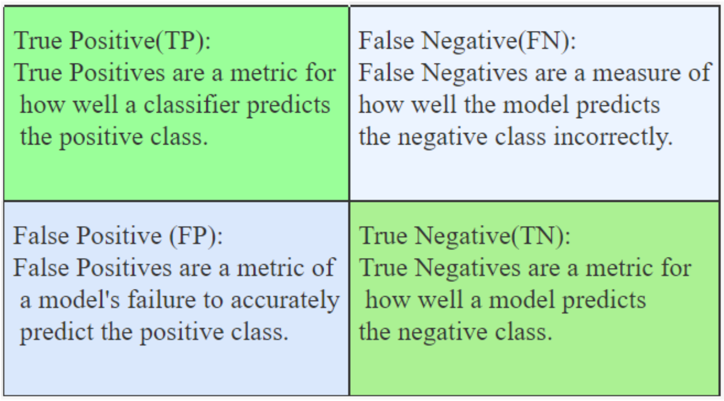
3.3 Data Mining

Feature extraction is conducted externally in traditional machine learning techniques. Feature extractions were performed for the SVM, Logistic Regression and Naive Bayes in this study, and various features were extracted. The Feature extraction technqiues used are CountVector and TF-IDF.Term Frequency-Inverse Document Frequency (TF-IDF) is an Information Retrieval method that considers both Term Frequency (TF) and Inverse Document Frequency (IDF). It reflects the prominence of a particular word in a group of documents. The tf-idf value rises proportionally as the number of times a word appears in a document. This extraction method gave good results in a research conducted by (Haque et al.; 2018). Where as CountVector is used to convert a text into a vector based on the frequency (count) of each word that appears throughout the text. Algorithms like SVM, Naive Bayes and Logistic Regression provide better results with the implementation of Feature extraction for Sentiment Analysis. Therefore, TF-IDF and CountVectorizer feature extractors has been implemented for three Machine Learning models used for classification of reviews.

3.4 Data Interpretation and Evaluation

The results derived from the models are presented and compared in section 6. A confusion matrix and a classification report are used to depict the algorithm's outcomes. For measuring model performance, the relevant F1 scores, accuracy, and precision are computed and given in the implementation part of each of the algorithms. The major elements of these evaluation matrices, as well as the formulas for calculating them, are presented below.

Primary Components



True Positive(TP): True Positives are a metric for how well a classifier predicts the positive class.	False Negative(FN): False Negatives are a measure of how well the model predicts the negative class incorrectly.
False Positive (FP): False Positives are a metric of a model's failure to accurately predict the positive class.	True Negative(TN): True Negatives are a metric for how well a model predicts the negative class.

Fig 3. Primary components measures used for the research

Evaluation metrices

- **Accuracy:** The ratio of total accurate/correct predictions to the total number of instances/observations is termed as Accuracy.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

- **Precision:** The number of accurately predicted positive classes divided by the total number of positive predictions generated by the model(correct and incorrect) is known as Precision.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- **Recall:** The ratio of predicted positive instances to all the instances that belongs to the positive class is Recall.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

- **F1-score:** The harmonic average of precision and recall is calculated as F1-score.

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

F1 score can be used as a good evaluation matrix because the dataset is not well balanced and it is adequate to use as a evaluation metric in this case. All of the matrices listed above, however, are computed for each model implemented.

The sentiment analysis classification methodology was developed in line with the requirements of the project. A subset is extracted from the entire dataset which contained millions of customer reviews and the subset was considered for the final implementation. The train and test data was split in the ratio of 75:25, 75% for training and 25% for testing. In this research, the clean text retrieved and rating class is split into train and test cases.

4 Design Specification

Figure 4 depicts the three-tier architecture utilized to implement this project, briefly describing the procedures taken and the technologies and tools employed.

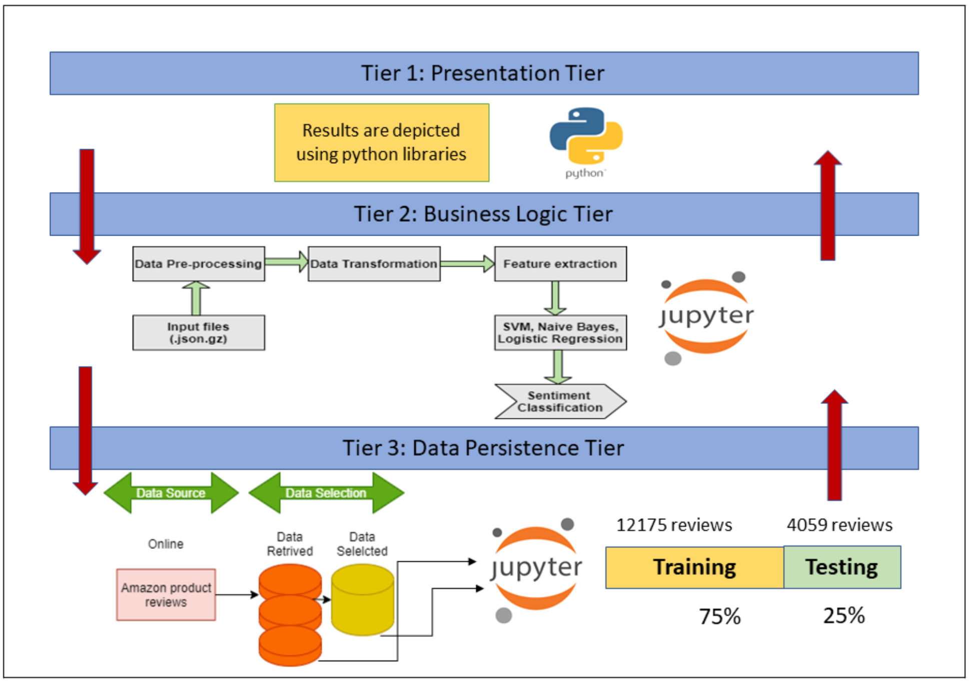


Fig.4 Sentiment Analysis Classification Design

TIER 3: Data Persistence Tier

The data is selected and gathered in this stage. The dataset used for this project was available online. Among the data retrieved, a subset of data is selected to use in this research. Later, the data gathered is fecthed to a jupyter notebook to create dataframes in order to perform operations.

TIER 2: Business Logic Tier

In this tier, the procedure is followed sequentially. Firstly, the input files fetched in jupyter notebook was in json.gz format. The original data and the meta data file is merged to carry out pre-processing. Pre-processing procedures are implemented such as discarding irrelevant columns, dropping null vales, changing datetime object etc. The clean data with NAs removed was then transformed. As the project's aim is to analyse sentiments, a clean text in terms of text pre-processing to apply classification algorithms was mandatory therefore, the transformation of data included text pre-processing such as removing special characters, excluding stop words and more. All the text pre-processing has been carried out using python libraries including nltk, spacy, gensim and re. A clean data was generated and the application of feature extractors such as TF-IDF vectorizer and Count Vectorizer is used for each ML classification algorithms. All the steps are conducted in Jupyter Notebook.

TIER 1: Presentation Tier

The outcome generated in the previous stage is visualized in this stage. The visualizations are in the form of Classification Report, confusion matrix, bar plot, comparison table etc and has been generated using python libraries such as matplotlib and seaborn.

5 Implementation of Sentiment Analysis Classification Models

5.1 Introduction

The implementation, evaluation and results of the models used for Sentiment Analysis from the Amazon product reviews is discussed in this section. The primary focus of the project is the text pre-processing since among all the data available, text is one of the most unstructured form of data. The data was formed in a way the labels were created in order to classify the sentiment as Good or Bad. In terms of modeling, the review ratings were not distributed normally, the ratings were classified. Rating 1 and 2 have been classified as 'BAD' and rating class from 3, 4 and 5 have been classified as 'GOOD'.The process of feature extraction and the methods used is also discussed in the section. For evaluating the performance of the models implemented, Confusion matrix and Classification reports are plotted [5]. F1 score, precision, recall and accuracy are used to determine the performance of machine learning models for each models. A comparison of developed models with feature extractors is carried out additionally, a comparison of both developed models and the existing models is also presented. Out of all the models, the model with best performance is selected.

5.2 Machine Learning Models

5.2.1 Support Vector Machine

Sentiment Analysis is a natural language processing technique that analyzes text to identify if the author's intents toward a given topic, product, or other entity are positive or negative. SVM is a supervised(feed-me) machine learning technique that may be applied to classification and regression problems. SVM conducts classification by locating the hyper-plane that separates the n-dimensional classes we generated.The scikit learn library's **SVC** function was used to implement this approach. Traditional approaches, such as SVM, do not necessitate a separate validation set, hence all of the data was saved in the same dataframe. The binary classification for reviews in terms of ratings were encoded as follows: [0: "bad", 1: "good"]

5.2.2 Naive Bayes

One of two famous naive Bayes variations used in sentiment analysis is Naive Bayes, which executes the naive Bayes algorithm for multinomial distributed data. This technique is a variant of the well-known naive Bayes algorithm, which is designed for predicting and classifying tasks with more than two classes. The same process as SVM is followed for Naive Bayes model. The scikit learn library' **MultinomialNB** function was used to implement this approach.

5.2.3 Logistic Regression

Logistic regression is a linear model used for classification. Other name for logistic regression are maximum entropy classification, logit regression or sometimes log-linear classifier. A logistic function is used to model the probability of the probable outputs of a single

[5] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

trial. Again as SVM and Naive Bayes is followed for logistic regression as the functions of train and test are globally defined. The scikit learn library' **LogisticRegression** function was used to implement this approach.

5.3 Feature Extraction

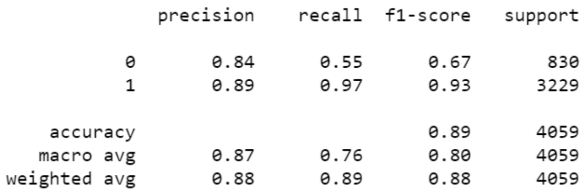
For conventional Machine Learning models like SVM, Naive Bayes and Logistic Regression feature extraction is a mandatory task to achieve better results, unlike deep learning algorithms. Therefore, in this research two feature extractors are used namely, CountVectorizer and TF-IDF are used for each models to determine the performance separately. The goal of the CountVectorizer approach is to transform textual data into vectors, with each document resulting in a vector that indicates the frequency of all unique words found in the documentation vector space for that particular document. The TF-IDF score (Term Frequency-Inverse Document Frequency) was added to our CountVectorizer (BoW) model to help us focus on more important words. TF-IDF weights words based on how uncommon they are in our dataset, ignoring terms that are overused and contribute to the noise. A comparison of performance of both the feature extractors along with ML models is presented below.

6 Evaluation and Results

As discussed in previous section, the evaluation of the model is carried out using different matrices such as Precision, recall, f1-score and accuracy. A total of over 16k reviews were considered for this implementation with a train test split of 75:25. 75% train set and 25% test set making it 12175 and 4059 for train and test sets, respectively. For all the experiments the same training set and test set was used.

6.1 EXPERIMENT 1: Classification models with CountVector.

6.1.1 SVM



```
              precision    recall  f1-score   support

           0       0.84      0.55      0.67       830
           1       0.89      0.97      0.93      3229

    accuracy                           0.89      4059
   macro avg       0.87      0.76      0.80      4059
weighted avg       0.88      0.89      0.88      4059
```

Fig. 5 Classification report of SVM using CountVectorizer

With the help of a classification report and confusion matrix, the performance of the model was evaluated. As depicted in figure 6, the classification report shows the overall accuracy of 0.89 provided by SVM using CountVectorizer. The precision, recall and f1 score are 0.84, 0.55 and 0.67 respectively. The accuracy of the model depicts how well our model performed. In this case, the SVM performed well.

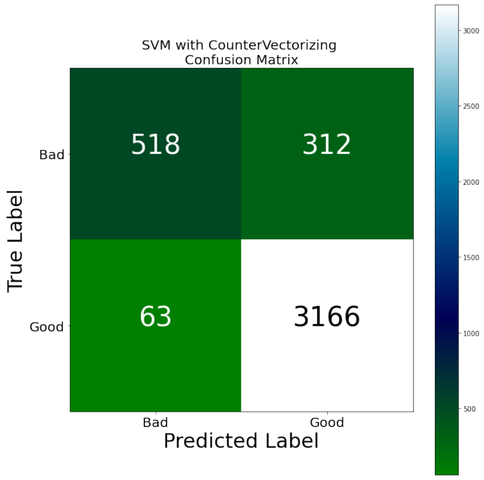
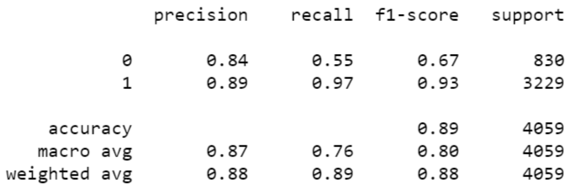


Fig.6. SVM Confusion Matrix with CountVectorizer

The confusion matrix in figure 6 depicts the True label against the predicted label. The SVM model correctly predicted 518 as bad reviews (TP) and 3166 as good reviews(TN). The reviews wrongly predicted as good reviews but are actually bad is 63.

6.1.2 Naive Bayes



```
              precision    recall  f1-score   support

           0       0.84      0.55      0.67       830
           1       0.89      0.97      0.93      3229

    accuracy                           0.89      4059
   macro avg       0.87      0.76      0.80      4059
weighted avg       0.88      0.89      0.88      4059
```

Fig. 9. Classification report of Naive Bayes using Count Vector

As depicted in figure 6, the classification report shows the overall accuracy of 0.89 provided by Naive Bayes using CountVector. The precision, recall and f1 score are 0.84, 0.55 and 0.67 respectively. 89% of the overall is predicted by the model.

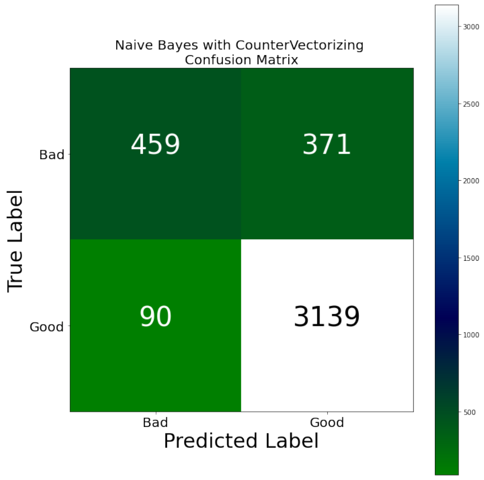
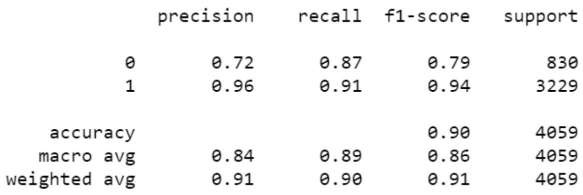


Fig. 10. Confusion Matrix of Naive Bayes using Count Vector

The confusion matrix in figure 8 depicts the True label against the predicted label. The Naive Bayes model correctly predicted 459 as bad reviews (TP) and 3139 as good reviews(TN). The reviews wrongly predicted as good reviews but are actually bad is 90. In comparison with Naive Bayes using TF-IDF, Count Vector predicted nearly better.

6.1.3 Logistic Regression



```
              precision    recall  f1-score   support

           0       0.72      0.87      0.79       830
           1       0.96      0.91      0.94      3229

    accuracy                           0.90      4059
   macro avg       0.84      0.89      0.86      4059
weighted avg       0.91      0.90      0.91      4059
```

Fig. 11. Classification report of Logistic Regression using Count Vector

As depicted in figure 6, the classification report shows the overall accuracy of 0.90 provided by Logistic Regression using CountVector. The precision, recall and f1 score are 0.72, 0.87 and 0.79 respectively. 90% of the overall is predicted by the model.

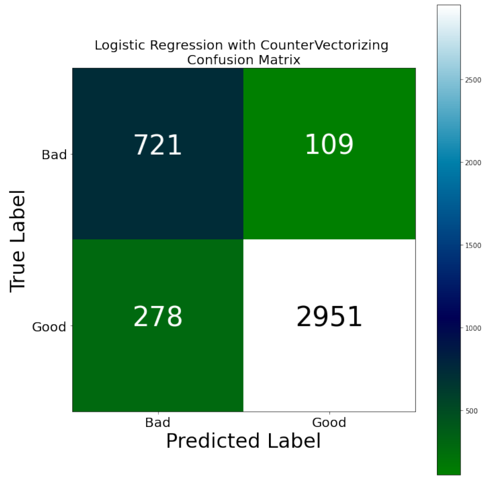
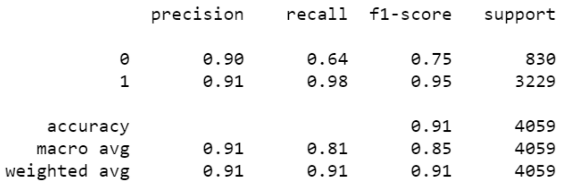


Fig. 12. Confusion Matrix of Logistic Regression using Count Vector

The confusion matrix in figure 8 depicts the True label against the predicted label. The Naive Bayes model correctly predicted 721 as bad reviews (TP) and 2951 as good reviews(TN). The reviews wrongly predicted as good reviews but are actually bad is 109. In comparison with Logistic Regression using TF-IDF, Count Vector predicted gave better results in terms of accuracy, precision, recall and f1 score.

6.2 EXPERIMENT 2: Classification models with TF-IDF.

6.2.1 SVM



```
              precision    recall  f1-score   support

           0       0.90      0.64      0.75       830
           1       0.91      0.98      0.95      3229

    accuracy                           0.91      4059
   macro avg       0.91      0.81      0.85      4059
weighted avg       0.91      0.91      0.91      4059
```

Fig. 7. Classification report of SVM using TF-IDF

As depicted in figure 6, the classification report shows the overall accuracy of 0.91 provided by SVM using TF-IDF. The precision, recall and f1 score are 0.90, 0.64 and 0.75 respectively. The accuracy of the model depicts how well our model performed. In this case, the SVM performed better with TF-IDF vector as compared to Count Vector.

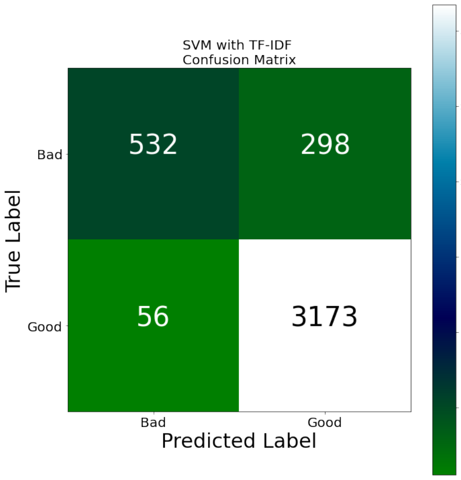
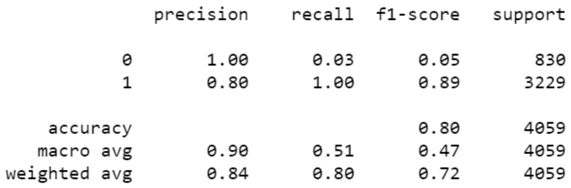


Fig.8. SVM Confusion Matrix with TF-IDF

The confusion matrix in figure 8 depicts the True label against the predicted label. The SVM model correctly predicted 532 as bad reviews (TP) and 3173 as good reviews(TN). The reviews wrongly predicted as good reviews but are actually bad is 56. In comparison with SVM using Count Vector, TF-IDF predicted nearly better.

6.2.2 Naive Bayes



```
              precision    recall  f1-score   support

           0       1.00      0.03      0.05       830
           1       0.80      1.00      0.89      3229

    accuracy                           0.80      4059
   macro avg       0.90      0.51      0.47      4059
weighted avg       0.84      0.80      0.72      4059
```

Fig. 7. Classification report of Naive Bayes using TF-IDF

As depicted in figure 6, the classification report shows the overall accuracy of 0.80 provided by Naive Bayes using TF-IDF.The accuracy of the model depicts how well our model performed. In this case, the Naive Bayes performed better with Count Vector as compared to TF-IDF.

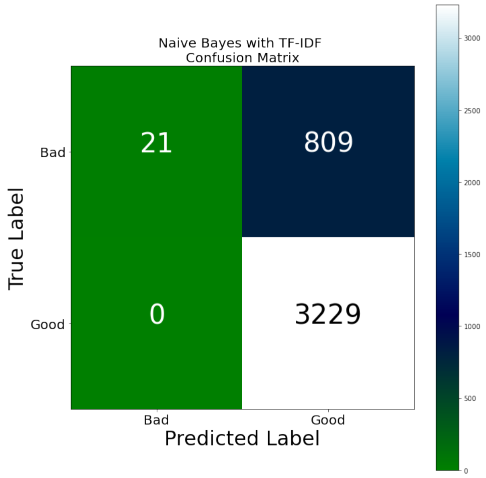
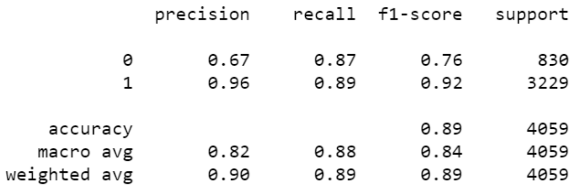


Fig.8. Naive Bayes Confusion Matrix with TF-IDF

The confusion matrix in figure 8 depicts the True label against the predicted label. The Naive Bayes model correctly predicted 21 as bad reviews (TP) and 3229 as good reviews(TN). The reviews wrongly predicted as good reviews but are actually bad is 0. In comparison with Naive Bayes using TF-IDF, Count Vector predicted nearly better.

6.2.3 Logistic Regression



```
              precision    recall  f1-score   support

           0       0.67      0.87      0.76       830
           1       0.96      0.89      0.92      3229

    accuracy                           0.89      4059
   macro avg       0.82      0.88      0.84      4059
weighted avg       0.90      0.89      0.89      4059
```

Fig. 7. Classification report of Logistic Regression using TF-IDF

As depicted in figure 6, the classification report shows the overall accuracy of 0.89 provided by Logistic Regression using TF-IDF. The precision, recall and f1 score are 0.67, 0.87 and 0.76 respectively. The accuracy of the model depicts how well our model performed. In this case, the Logistic Regression performed better with Count Vector as compared to TF-IDF .

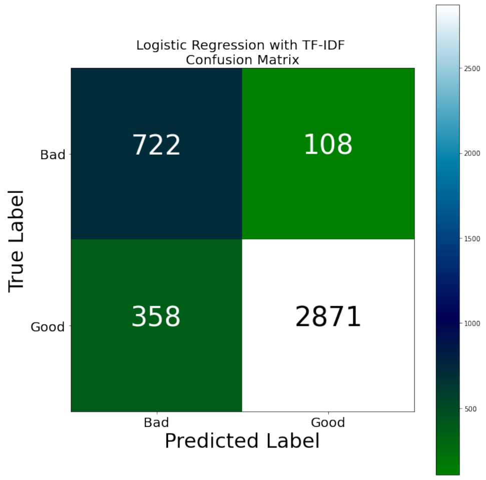


Fig.8. Logistic Regression Confusion Matrix with TF-IDF

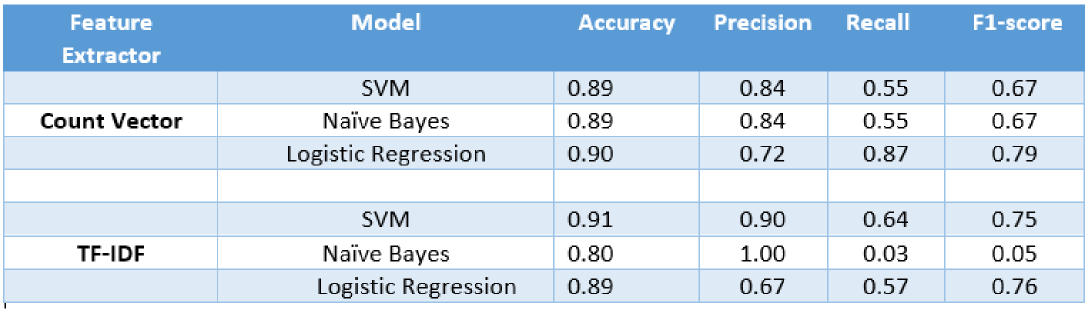
The confusion matrix in figure 8 depicts the True label against the predicted label. The Logistic model correctly predicted 722 as bad reviews (TP) and 2871 as good reviews(TN). The reviews wrongly predicted as good reviews but are actually bad is 358. In comparison with Logistic Regression using TF-IDF, Count Vector predicted nearly better with a slight difference.

6.3 Conclusion

Section 5 and Section 6 fully answers the research question mentioned in section 1.2.1. The performance of the models were good as these models have proved to be working efficient by many researchers for these problems of classification. In this research project, SVM model with TF-IDF feature extractor performed well and produced best results among all the other models. Therefore, for classification task SVM is the model selected.

6.4 Comparison of Implemented models

In table below, the performance comparison of the models implemented for the project namely, SVM, Naive Bayes and Logistic Regression is depicted with training set (75%) and testing set(25%) i.e. 12175 and 4059 for train and test sets.



Feature Extractor	Model	Accuracy	Precision	Recall	F1-score
	SVM	0.89	0.84	0.55	0.67
Count Vector	Naïve Bayes	0.89	0.84	0.55	0.67
	Logistic Regression	0.90	0.72	0.87	0.79
	SVM	0.91	0.90	0.64	0.75
TF-IDF	Naïve Bayes	0.80	1.00	0.03	0.05
	Logistic Regression	0.89	0.67	0.57	0.76

The comparison between the models implemented for Sentiment classification

7 Discussion

The primary objective of the project was to perform sentiment analysis by building models that provide accurate and efficient outcome. After a thorough literature review, a basic understanding of current limitations and gaps were discovered. Most of the researchers developed either a lexicon-based model for classification or used labelled data to apply Machine Learning algorithms. A hybrid of Lexicon and ML based models was the area less explored. In this research project, two feature extraction techniques namely CountVector and TF-IDF were used to perform classification using SVM, Naive Bayes and Logistic Regression. The best results among all were obtained by Support Vector Machine(SVM) with the overall accuracy of 91%. The precision, recall and f1-score for the same was 0.90, 0.64 and 0.75 respectively. The scope of the project included building the best suited classification model for Sentiment Analysis using Amazon product reviews by extracting features. In terms of performance, all the classification models performed reasonably well. From large dataset, a subset is used and pre-processing is carried out. Additionally, text pre-processing is implemented as the data available was in textual format and in order to apply machine learning models it was converted to numeric form for classification. A comparison of various approaches is provided in this research. It also highlights the poor and in-effective performance of conventional classification approaches.

8 Conclusion and Future Work

Given the significance of customer happiness these days, it's critical to concentrate on upgrading existing systems to make them more user-friendly and time-saving. Sentiment Analysis has been shown to be quite effective for customers and businesses in several studies and techniques. From this experiment conducted, the best classifier was able to classify the ratings with 91 percent accuracy. The model performed reasonably well as compared to the existed models.

Although the objectives achieved by the research was a success, in course of development it was not limited and faced some problems. The major problem included the unstructured textual format unlabelled data. It was a tedious task to transform the data. Also the classification models as the name suggests were trained on binary class of ratings. A subset was used in order to implement models but the scope could have been expanded to explore more subsets. One of reasons for using a small subset was to reduce the computational time required for training. Two types of feature extraction techniques were used, although there are plenty of other techniques that could be used to generate better outcomes.

Sentiment Analysis from product reviews is a difficult and tedious task as there are numerous parameters to account for and it may vary from small subset to large set of data. The classification machine learning techniques provided notable results and improvement was observed with the implemented hybrid technique. In conclusion, with a small subset of data with limited text pre-processing and feature extraction techniques the models performed well.The dataset used for the implementation of this project was not labelled and was prone to lot of noise. It is a good idea to expand the scope by exploring more techniques and methods on large datasets by using multiple subset or categories.

In future, a larger dataset should be used and more text pre-processing techniques can be implemented to achieve better results. It would be interesting to see the models performance on labelled dataset in huge expansion. A large dataset may lead to superior outcomes. Deep learning techniques can also be implemented with the scope same as this project, the results derived from this would also be interesting. Additionally, application of a Recommender System using either user-based or item-based collaborative filtering is also a potential idea.

9 Acknowledgement

My sincere gratitude to our college, National College of Ireland, and the MSc in Data Analytics department for enabling me to complete this research project effectively over the last three months. My supervisor and mentor, Prof. Christian Horn, for his assistance throughout the project. His constant support made it possible for me to finish and submit this thesis. I value his time and, in particular, his thoughtful observations. I consider him as a great mentor. Finally, I'd want to thank my friends and family back home for their constant support and encouragement during this study process.

References

Ahmad, M., Aftab, S., Ali, I. and Hameed, N. (2017). Hybrid tools and techniques for sentiment analysis: a review, *Int. J. Multidiscip. Sci. Eng* **8**(3): 29–33.

Fan, X., Li, X., Du, F., Li, X. and Wei, M. (2016). Apply word vectors for sentiment analysis of app reviews, *2016 3rd International Conference on Systems and Informatics (ICSAI)*, IEEE, pp. 1062–1066.

Gamal, D., Alfonse, M., M El-Horbaty, E.-S. and M Salem, A.-B. (2019). Analysis of machine learning algorithms for opinion mining in different domains, *Machine Learning and Knowledge Extraction* **1**(1): 224–234.

Gupta, M., Mishra, A., Manral, G. and Ansari, G. (2020). Aspect-category based sentiment analysis on dynamic reviews, *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, IEEE, pp. 492–496.

Haque, T. U., Saber, N. N. and Shah, F. M. (2018). Sentiment analysis on large scale amazon product reviews, *2018 IEEE international conference on innovative research and development (ICIRD)*, IEEE, pp. 1–6.

Hermansyah, R. and Sarno, R. (2020). Sentiment analysis about product and service evaluation of pt telekomunikasi indonesia tbk from tweets using textblob, naive bayes & k-nn method, *2020 International Seminar on Application for Technology of Information and Communication (iSemantic)*, IEEE, pp. 511–516.

Jabbar, J., Urooj, I., JunSheng, W. and Azeem, N. (2019). Real-time sentiment analysis on e-commerce application, *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)*, IEEE, pp. 391–396.

Jha, B. K., Sivasankari, G. and Venugopal, K. (2021). Sentiment analysis for e-commerce products using natural language processing, *Annals of the Romanian Society for Cell Biology* pp. 166–175.

Khanvilkar, G. and Vora, D. (2019). Smart recommendation system based on product reviews using random forest, *2019 International Conference on Nascent Technologies in Engineering (ICNTE)*, IEEE, pp. 1–9.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. and Potts, C. (2011). Learning word vectors for sentiment analysis, *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150.

May, J., Mokgalaka, A. L., Murali, Kumar, M., Haas, E. and Simran (2020). Machine learning (ml) for natural language processing (nlp).
URL: *www.lexalytics.com/lexablog/machine-learning-natural-language-processing*

Prabha, M. I. and Srikanth, G. U. (2019). Survey of sentiment analysis using deep learning techniques, *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, IEEE, pp. 1–9.

Rajeev, P. V. and Rekha, V. S. (2015). Recommending products to customers using opinion mining of online product reviews and features, *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*, IEEE, pp. 1–5.

Rajeswari, A., Mahalakshmi, M., Nithyashree, R. and Nalini, G. (2020). Sentiment analysis for predicting customer reviews using a hybrid approach, *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*, IEEE, pp. 200–205.

Seetharamulu, B., Reddy, B. N. K. and Naidu, K. B. (2020). Deep learning for sentiment analysis based on customer reviews, *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, pp. 1–5.

Sharma, Y., Agrawal, G., Jain, P. and Kumar, T. (2017). Vector representation of words for sentiment analysis using glove, *2017 international conference on intelligent communication and computational techniques (icct)*, IEEE, pp. 279–284.

Shivaprasad, T. and Shetty, J. (2017). Sentiment analysis of product reviews: a review, *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, IEEE, pp. 298–301.

Singh, S. N. and Sarraf, T. (2020). Sentiment analysis of a product based on user reviews using random forests algorithm, *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, pp. 112–116.

Solangi, Y. A., Solangi, Z. A., Aarain, S., Abro, A., Mallah, G. A. and Shah, A. (2018). Review on natural language processing (nlp) and its toolkits for opinion mining and sentiment analysis, *2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, IEEE, pp. 1–4.

Yang, J. and Yang, J. (2020). Aspect based sentiment analysis with self-attention and gated convolutional networks, *2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)*, IEEE, pp. 146–149.

Zhang, L., Wang, S. and Liu, B. (2018). Deep learning for sentiment analysis: A survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **8**(4): e1253.