# Implementation of Customer Review Sentiment Analysis over Apache Spark

Ansh Mahajan
School of CET
MIT World Peace University
Pune, Maharashtra
1032201496@mitwpu.edu.in

Neha Bhavsar
School of CET
MIT World Peace University
Pune, Maharashtra
1032212195@mitwpu.edu.in

Rijul Bilaiya
School of CET
MIT World Peace University
Pune, Maharashtra
1032201467@mitwpu.edu.in

Aditya Pathak
School of CET
MIT World Peace University
Pune, Maharashtra
1032201493@mitwpu.edu.in

Dipesh Shah
School of CET
MIT World Peace University
Pune, Maharashtra
1032212247@mitwpu.edu.in

**Abstract:**
Sentiment Analysis of large-scale data has become increasingly important and urging researchers to use new platforms and tools that can handle large volumes of data. In this paper, we present new evaluation experiments of sentiment analysis for a large-scale dataset of online customer's reviews under Apache Spark data Processing System. Apache Spark's scalable machine learning library (MLlib) is used and three classification techniques from the library are applied; Naive Bayes, Logistic Regression and Linear Support Vector Classifier (SVC) model. The results are evaluated using the accuracy metric. Experimental results show that the Logistic Regression classifier outperforms Naive Bayes and Linear SVC classifiers.

**Keywords: Sentiment Analysis, Online Customer's Review, Apache Spark, Apache Spark's scalable machine learning library, Naive Bayes, Logistic Regression and Linear SVC model**

---

## 1. Introduction:

Online reviews and recommendations have a big impact on Customers' purchasing opinions, especially now, when people tend to express their opinions and passions further than ever, on virtual communities and social networks. According to a check report, 93 of consumers relate to online reviews before making their purchase opinions.

Purchase is always an commerce between two realities, Customers and business possessors. Customers can use reviews to make better opinions about what products to buy, while businesses, on the other hand, benefit from reviews in terms of gaining useful information about client satisfaction on their products. This information can be also used to estimate their marketing strategies, to ameliorate their products and to enhance their performance.

Classifying large quantities of unshaped data from the Internet is a grueling task. Hence, sentiment analysis, along with Natural Language Processing ways, flourished in recent times, to give a frame for analysis of textual data obtained from reviews. These ways prognosticate the opposition of the opinions( positive or negative), aiding Customers to have a conclusion about a product. On the other hand, companies can understand in this way the position of satisfaction of their Customers.

Sentiment analysis is a natural language processing problem, which implies the discovery and reclamation of knowledge from textual data. The sentiment analysis

follows a sequence of ways similar as the reviews collection, the lowercase conversion, punctuation and fresh spaces junking, stop words junking, tokenization, lemmatization, point birth and eventually bracket.

Amazon is one among the most important e-commerce retailers, used every day for online shopping. The Amazon ranking system ranges from 1 to 5, where" 1" is the worst standing and" 5" is the loftiest standing.

In order to assess the overall semantics of consumer feedback, this paper explores the sentiment bracket into positive or negative passions, for online reviews, using specific styles applied in this sphere. This Paper aims to make a classifier to prognosticate consumers satisfaction, whose performance will be estimated, grounded on the dataset of the Amazon Customers reviews. This has the implicit purpose of helping companies to ameliorate their products and on the other hand, to help implicit Customers to make better coping opinions.

The paper is structured as follows. This paper begins with the preface. Section 2 discusses the affiliated work in the former literature. Section 3 explains both exploration methodology and perpetration independently. Section 4 reports the experimental results in terms of performance criteria for colorful classifiers. Incidentally, Section 5 concludes the findings of the paper and Section 6 exposes the unborn compass or Future Scope.

## 2. Literature Survey:

Various studies focused on the problem of identifying customers' opinions on different products using Amazon reviews. In the following, these papers are reviewed in terms of pre-processing techniques, feature extraction methods, proposed methodologies, and evaluation metrics.

**Guia, 2019** [8] applied supervised machine learning algorithms such as Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest, to predict the reviews sentiment, based on Amazon Reviews: Unlocked Mobile Phones dataset. In the classification process, the authors used only Rating and Review attributes and removed the instances with neutral reviews. After applying preprocessing steps such as converting the dataset into lowercase, tokenization, removal of punctuation and stop words, the data was splitted into 80% for training and 20% for test. The results for the application of classifiers, show that the Support Vector Machine classifier is the most accurate, with the highest values for all metrics, followed by Random Forest classifier. The authors presented also a statistical study in terms of the impact of brand and price in the reviews polarity. They concluded that ZTE has the most positive reviews rate with 82.9 % positive reviews and in terms of price, more positive reviews were obtained for higher prices, which can be explained by the quality of the phones.

The research conducted by **Aljuhani, 2019** [2], studied the performance of different machine learning algorithms, such as Logistic regression, Naïve Bayes, Stochastic gradient descent and convolutional neural network respectively, using different features extraction techniques such as BOW and TF-IDF, each of them with three variations depending on the number of grams used. The authors divided the data into 70% for training, 15% for testing and 15% for development. The results revealed that convolutional neural networks provided the best results.

In their papers, both **Aljuhani, 2019** [2] and **Bansal, 2018** [4], used unbalanced and balanced datasets. While **Aljuhani, 2019** [2] categorized both balanced and unbalanced data into, five and four stars as positive rating, one and two stars as negative rating, and three stars as neutral rating, **Bansal, 2018** [4] categorized balanced and unbalanced data separately. **Bansal, 2018** [4]used balanced data, meaning that the number of negative reviews (1 and 2 stars) is equal to the number of positive reviews (4 and 5 stars)

and removed neutral reviews. For unbalanced data, they categorized (1 and 2 stars) as negative reviews and (3, 4 and 5 stars) as positive reviews. **Bansal, 2018** [4] applied deep learning methods such as, CBOW and skip-gram, with different machine learning algorithms: SVM, Naïve Bayes, Logistic Regression and Random Forest. The experimental results showed that Random Forest using CBOW achieved the best accuracy. [8] [2] [4]

**Shaheen, 2019** [16] performed a sentiment classification on mobile phone reviews dataset, using seven different classifiers and based on their results, the Random Forest classifier outperformed all other classifiers, with an accuracy of 85% for the given dataset, followed by LSTM and CNN. The authors also exposed the distribution of reviews with respect to their ratings, showing that most reviewers have rated 4 stars and 3 stars. Also, the study concluded that there is a direct correlation between rating and price. [16]

**Ravi, 2019** [13] implemented four algorithms, Naive Bayes, Support Vector Machine, Random Forest and K-Nearest Neighbor, using different sizes of training and test data. They concluded that the Random Forest classifier produced the best accuracy metrics. [13]

**Qaiser, 2021** [12] focused on the comparison of machine learning methods applied in Sentiment Analysis such as Naïve Bayes, Decision Tree, Support Vector Machine and the modern method, Deep Learning. The ML techniques were applied to a dataset of 4289 rows, about technological impact on employment, remaining with 1047 rows, after completing the preprocessing steps **Qaiser, 2021** [12]. They concluded that the deep learning method performed the best, with an accuracy of 96.41%, followed by Naïve Bayes and Support Vector Machine with 87.18% and 82.05% respectively. [12]

**Tan, 2018** [17] conducted a study on a dataset of 34660 instances, from customer reviews of Amazon products. The dataset was divided into a training set of 60%, a validation set of 20% and a test set of 20%. They implemented machine learning algorithms such as Naïve Bayes, Support Vector Machine with Linear Kernel, Support Vector Machine with RBF Kernel, KNN-4, 5, 6 and deep neural networks, such as Recurrent Neural Network. They concluded that the Long ShortTerm Memory generates the most accurate predictions. [17]

## 3. Big Data and Spark Architecture

Big data is defined as the process that is applied when conventional data mining and handling techniques are unable to get the insights and meaning of the underlying data. The relational database management systems engines cannot process unstructured or large data anymore. So dealing with such data types needs a different processing approach called big data, which utilizes enormous Parallelism on readily-available machines. Therefore, Apache Spark developed to make processing and analyzing the data easier. Big data has three main characteristics.

### A. Big Data Characteristics:

Big data is defined as the process that is applied when conventional data mining and handling techniques are unable to get the insights and meaning of the underlying data. The relational database management systems engines cannot process unstructured or large data anymore .So dealing with such data types needs a different processing approach called big data, which utilizes enormous parallelism on readily-available machines. Therefore, Apache Spark developed to make processing and analyzing the data easier. Big data has three main characteristics.

i. **Volume:**

Big data describes huge volumes of data. Creating and modifying data were one of the main employees' jobs. But Nowadays data is generated with the aid of networks, machines and human interaction on systems such as ecommerce websites, so the quantity of data to be analyzed becomes vast.New research statistics show that Amazon dataset includes reviews over 233.1 million (142.8 million in 2014).

ii. **Velocity:**

Today Data is generated too fast and also needs to be processed fast. Ecommerce is one of the most important means of providing data. Ecommerce sites are continuously producing a complex unstructured and semi-structured shape of data and presently created 90% of data in the last two years. Increasing the velocity of big data is based on using modern devices such as smart phones and more advanced technology. The Internet is the primary factor for gathering enormous data. Amazon Web Services Kinesis is an example for data velocity handler streaming application.

iii. **Variety:**

Sources and kinds of structured and unstructured data are multiple and various.Sources such as spread-sheets and databases have been used in the previous years to store and keep data. Now data comes in many different forms like photos, audio, videos, emails, PDFs, monitoring equipment, etc. Due to the great diversity of unstructured data, many problems arise during storage, mining and analyzing data. All organizations, companies and manufacturers currently have nearly 85% of the data, structured and semi-structured and unstructured shape of data .

## B. Spark Framework:

Apache Spark is one of the newer open-source, lightning fast big data distributed processing platforms based on the same principles as Hadoop, which is developed in a manner to evolve the computational speed. Hadoop has some problems relating to performance in specific cases, like graph based algorithms tasks or repetitive tasks. Also, 172 Hadoop can't cache intermediate data for getting high performance however rather; it flows the data to the disk between each step read and write. Spark minimizes the number of read and write cycles making it 10 times faster than Hadoop on disk in applications running; also, keeping intermediate data in-memory makes Spark 100 times faster when dealing with memory . Resilient distributed data set (RDD) may be the key theoretical idea in spark, which represents a read-only collection of objects divided across a set of nodes, with the ability to rebuild them if an occurrence of losing any partition occurs. Spark is best known for its capacity for accomplishing batch, interactive, and machine learning in addition to streaming all in the same cluster. One of the main features of Spark is its scalability, so the cluster can be provided with a number of nodes. Another main feature in Spark is language flexibility so Spark developers can use (API) in Scala, Python, Java and R programming languages. The architecture of the Spark framework is portrayed in Figure. 1

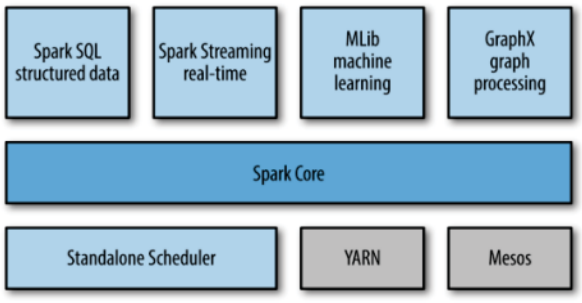

Figure 1. Apache Spark Framework

## 4. Proposed System:

The three classifiers in our proposed solution were implemented by using PySpark language as a favored language. The overall architecture of the suggested solution is portrayed in Figure 2, taking into consideration the previously discussed aspect.

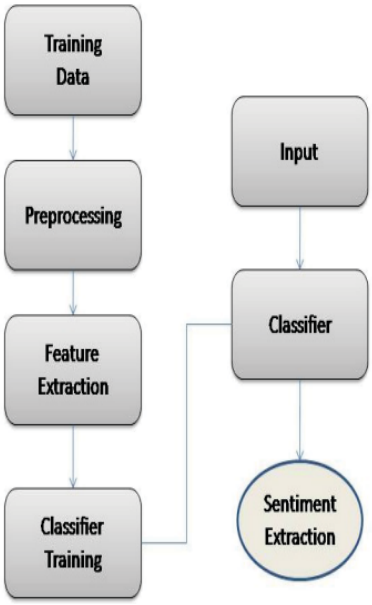


Figure 2. System Architecture

## 5. Research Methodology:

This section presents the methodology and techniques used for the classification of product reviews.

### A. Dataset:

Our dataset (Amazon Kindle Dataset , 2019), consists of 568,450 instances, fetched from 2004 to 2019. The data was retrieved from Amazon.com and focuses on reviews for both unlocked and locked carriers.

The dataset contains the following attributes:

i. Id
ii. ProductId
iii. UserId
iv. ProfileName
v. HelpfulnessNumerator
vi. HelpfulnessDenominator
vii. Score
viii. Time
ix. Date
x. Summary
xi. Text
xii. sentiments

In our analysis, we will focus only on the 'Score' i.e rating and 'Text' features, as these are the most useful and relevant for model building. In order to have an overview of the dataset, the 'Score' distribution is shown in Figure 3 The classes are imbalanced, as classes 2, 3 and 4 have very small amounts of reviews, compared to class 5.

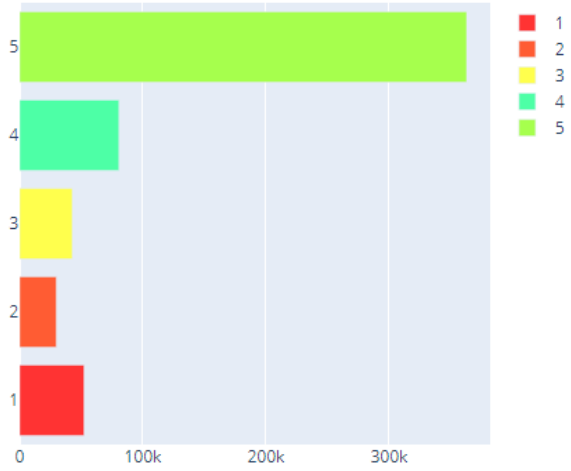


Figure 3. Distribution of Reviews Rating

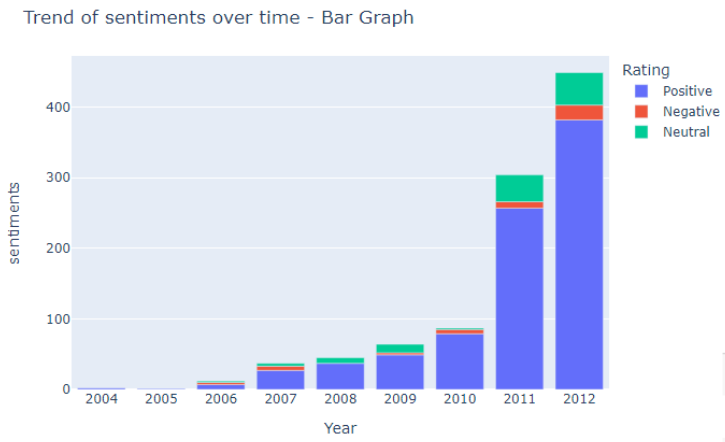


Figure 4. Trend of Sentiments over time

Before applying the preprocessing tasks, we are going to categorize the dataset, as follows:

i. Rating with a value less than or equal to 3, is labeled as "Negative".
ii. Rating with a value greater than 3, is labeled as "Positive".

## B. Data Pre-Processing:

The performance of a classifier can be highly increased by preprocessing the data. Considering this, the preprocessing phase, applied to our dataset, included the following steps:

i. Convert the uppercase letters into lowercase.
ii. Remove all the URLs starting with HTTP.
iii. Remove all the special characters.
iv. Remove all single characters.
v. Remove single characters from the start.
vi. Substitute multiple spaces with single space.
vii. Remove all the punctuation.
viii. Remove the stop words, such as “the”, “a”, “in”, using Stopwords Corpus for English words, from the NLTK library.
ix. Tokenization, the process of splitting the original text in the form of sentences into words.
x. Lemmatization, the process of transforming the word into its significant base structure, using Wordnet Lemmatizer from NLTK library.

| | id | pos_keywords | neg_keywords |
|---|---|---|---|
| 0 | B007JFMH8M | [tasty, christmas, walmart, baking, vegan] | [corn, tasty, sawdust, cholesterol, pancakes] |
| 1 | B002QWP89S | [walmart, steak, puppies, lovesssssssssssssssssss, tunafish] | [corn, cancer, diarrhea, puppies, petsmart] |
| 2 | B002QWP8H0 | [lovesssssssssssssssssss, steak, puppies, tunafish, petsmart] | [cancer, butter, puppies, dinner, petsmart] |
| 3 | B002QWHJOU | [cheapest, cats, puppies, steak, tunafish] | [killing, puppies, diarrhea, google, petsmart] |
| 4 | B0026RQTGE | [walmart, steak, puppies, lovesssssssssssssssssss, tunafish] | [corn, cancer, diarrhea, puppies, petsmart] |

Figure 5. Positive and Negative keyword Extraction

After Data Cleaning on ‘Text’ attribute, the dataset is splitted into 70% for the training set, used to learn the models and 30% for the testing set, used to calculate the model's performance in terms of classification report which includes Area under Receiver Operating Characteristic (ROC) curve, Accuracy, Precision and Recall.

## C. Word clouds of reviews for Most Reviews Brand:

Word cloud is a widely used data visualization technique used for representing text data, in which the size of words indicates their frequency or importance. Both types of reviews contain some common words like “buy”, “battery” or “one”. Figure 6 shows that the most frequent words encountered in positive reviews are: “great”, “good”, “loved”, “baked”, “soft”, etc. On the other hand, the most frequent negative review words are “moist”, “bad”, “calorie”, “little”, as seen in Figure 7.


Figure 6. Word Cloud visualization for positive reviews


Figure 7. Word Cloud visualization for negative reviews

### D. Feature Extraction:

When dealing with text features, the original text needs to be converted into a document-term, since the machine learning algorithms do not support text features. Thus, after the preprocessing stage, data will be vectorized, using the following methods: CV (CountVectorizer) and IDF (Inverse Document Frequency). The result from each method will be a matrix that represents the text as vectors, which can be fed to the machine learning algorithms to build classification models.

### E. Spark's MLlib Classification Models:

Spark MLlib is defined as a distributed machine learning framework on top of Spark Core. The Spark excels at iterative algorithms which makes it perfect for executing Machine Learning algorithms since the majority of Machine Learning techniques depend on repeated jobs. The main aim behind MLlib design is to make machine learning scalable and easier processing for data. MLlib has a high performance as a result of its design to run in parallel on clusters.

In language processing, most part of classifications are performed using supervised machine learning, and this will be the subject of this paper. Text classification can be done using different algorithms. In this context, the algorithms implemented for classification are named classifiers.

This section describes three of the most used supervised classifiers in text classification.

i. **Naive Baiye:** Naive Bayes technique will select the best class for a document, based on the probability that the terms in the document belong to that class. From a mathematical point of view, the probability of classifying the document in a class c is:

$$C_{NB} = \text{argmax } P(c) \prod_{1 \leqslant k \leqslant n_d} P(t_k \mid c) \quad \text{.....(1)}$$

Where,

- P (c) is the probability that a document belongs to class c (based on training data), also called previous class probability.
- $P(t_k \mid c)$ is the probability that a term t from position k in document d, will be found in documents of class c.
- $n_d$ is the number of terms in document d.

ii. **Logistic Regression:** Logistic Regression is a classification that serves to solve the binary classification problem. The result is usually defined as 0 or 1 in the models with a double situation. The equation used in this algorithm is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta(\text{num}) \quad \text{.....(2)}$$

iii. **Linear Support Vector Classifier:** Linear Support Vector Classifier tries to find the optimal hyperplane which could separate the data into two classes in the case of binary classification. From a geometric point of view, given two types of points in a space, it tries to minimize the distance from one of the points to the other. This minimization problem is equivalent to the following problem:

$$\min(\frac{1}{2}|w|^2) \quad \ldots..(3)$$

Where w is the direction of a vector.

## F. Evaluation Metrics or Classification Report:

Evaluation metrics play an important role to measure the classification performance. In order to evaluate the results of the three algorithms two of the most popular measures are used: Accuracy and Area under ROC curve. These two metrics are explained in the following:

i. Accuracy predicts how often the classifier makes the correct prediction. Accuracy is the ratio between the number of correct predictions and the total number of predictions.
ii. The Area Under the ROC Curve is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the ROC, the better the performance of the model at distinguishing between the positive and negative classes.
iii. Precision is how good the model is at predicting a specific category.
iv. Recall tells you how many times the model was able to detect a specific category.

## 6. Experimental Results:

The reviews have been classified as positive and negative, based on the star rating.

There were several machine learning algorithms employed in this paper such as Naive Bayesian, Logistic Regression and Linear Support Vector Classifier. Different feature selection techniques were applied on classifiers, such as IDF and CV.

Tables no. 1 show the results for all mentioned classifiers, using different features extraction techniques.

| S. No | Classifier | ROC (in %) | Precision (in %) | Recall (in %) | Accuracy (in %) |
|---|---|---|---|---|---|
| 1. | Naive Baiye | 88.18 | 86.72 | 91.71 | 88.34 |
| 2. | Logistic Regression | 90.38 | 89.55 | 92.54 | 90.47 |
| 3. | Linear Support Vector Classifier | 90.39 | 90.15 | 91.72 | 90.45 |

Table no 1: Result and Analysis of all classifiers on training set

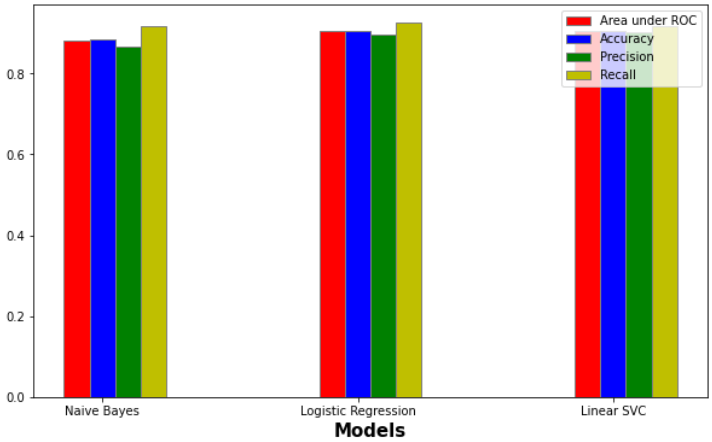

Figure 8: Bar Graph Representing all evaluation matrix of all the applied models

For Naive Bayes, we can see in Table no. 1 that Accuracy achieved with a value of 88.34%. As seen in Table no.1, Logistic Regression Classifier obtains the best performance when using the pipeline approach, achieving 90.47% for accuracy. As shown in Table no. 1, the Linear Kernel was used to evaluate Support Vector Classifiers achieved 90.45% for accuracy. So, the performance of these two classifiers is very high and comparable to each other.

From all the experiments the best results of all three algorithms is obtained by Logistic Regression followed by Linear Support Vector Classifier followed by Naive Baiyes.

## 7. Conclusion and Future Work:

Reviews are essential for both customers and companies. From consumers' point of view, reviews help them to make better decisions when buying products. On the other hand, companies benefit from reviews, by knowing the level of consumer satisfaction about their products and acting accordingly.

In this paper, we proposed a machine learning approach for text sentiment analysis. We performed sentiment analysis on Amazon reviews dataset, using different types of machine learning algorithms, such as Naive Bayes, Logistic Regression and Support Vector Classifier. We used different feature extraction approaches such as CV and IDF analyzed the classifiers results, based on two performance metrics: Accuracy, Precision, Recall and Area under ROC curve. We described the basic theory behind the models, approaches used and the performance metrics for the conducted experiments. We went through different research papers on sentiment analysis over text-based datasets.

Overall, we were able to achieve promising results for classifiers, based on the performance metrics obtained. We found that the pipeline approach, which combines idf Transformer and CountVectorizer, achieves the highest metrics results, for almost all the classifiers. We can also observe that, the models applied to CV and IDF, give better results for all classifiers. The highest accuracy is 90.47%, obtained by the Logistic Regression Model. Our results show that Naïve Bayes is also a classifier to consider, being lower than the Support Vector Classifier with Linear Kernel, with its highest accuracy of 90.45%.

As future work, we plan to continue to study other algorithms applied in the Sentiment Analysis field and to evaluate them. Also, another future direction would be to collect more data, in order to test the performance of classifiers on a massive dataset and see if there are improvements in results. Another point to consider in the future, would be to adjust and explore more parameters for the classifiers, which could contribute to even better results. And not lastly, in the future we intend to explore more methods of linguistic analysis, such as semantic analysis.

## 8. Future Scope:

A. **Identifying and Predicting Market Trends:** It enables you to analyze large amounts of market research data in order to spot emerging trends and better understand consumer buying habits. This type of practice can help you navigate the complicated world of stock market trading and make decisions based on market sentiment.

B. **Keeping an eye on the brand's image:** Sentiment analysis is frequently used to investigate user perceptions of a product or topic. You can also use it to conduct a product analysis and provide all relevant data to the development teams.

C. **Examining public opinion polls and political polls:** To predict the outcome of an election, anyone can use sentiment analysis to compile and analyze large amounts of text data, such as news, social media, opinions, and suggestions. It takes into account

how the general public feels about both candidates.

D. **Data from customer feedback is being analyzed.:** Data from customer feedback can be used to identify areas for improvement. Sentiment analysis can help you extract value and insights from customer feedback data, as well as develop effective customer satisfaction strategies.

E. **Observing and analyzing conversations on social media:** Conversations on social media are a gold mine of information. Look at conversations about your brand on social media to see what your customers are saying with sentiment analysis; this can help any company plan its future strategies much more effectively.

F. **Employee Turnover Reduction:** Analyze large amounts of employee feedback data to determine employee satisfaction levels. The insights are used by the sentiment analysis tool to boost morale and productivity while also informing you of how your employees are feeling

---

## References:


1. Alharbi, N. M., Alghamdi, N. S., Alkhammash, E. H. and Al Amri, J. F., 2021. Evaluation of Sentiment Analysis via Word Embedding and RNN Variants for Amazon Online Reviews. Mathematical Problems in Engineering, 2021, pp.1-10.
2. Aljuhani, S. A. and Alghamdi, N. S., 2019. A comparison of sentiment analysis methods on Amazon reviews of Mobile Phones. International Journal of Advanced Computer Science and Applications, 10(6), pp. 608-617.
3. Al-Sheikh, E. S. and Hasanat, M. H. A., 2018. Social media mining for assessing brand popularity. International Journal of Data Warehousing and Mining, 14(1), pp. 40-59.
4. Bansal, B. and Srivastava, S., 2018. Sentiment classification of online consumer reviews using word vector representations. Procedia computer science, 132, 1147-1153.
5. Bengfort, B., Bilbro, R. and Ojeda, T., 2018. Applied text analysis with python: Enabling language aware data products with machine learning. Sebastopol: O'Reilly Media, Inc.
6. Dadhich, A. and Thankachan, B., 2021. Sentiment Analysis of Amazon Product Reviews Using Hybrid Rule-based Approach. International Journal Engineering and Manufacturing, 2021, pp. 40- 52.
7. Fan, S., 2018. Understanding the mathematics behind Support Vector Machines, Shuzhan Fan, [online]. Available at: [Accessed 20 May 2021].
8. Guia, M. S., Silva, R. and Bernardino, J., 2019. Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis. Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019), pp. 525-531.
9. Guillot, D., 2017. Mathematical Techniques in Data Science Random forest. University of Delaware [online]. Available at: [Accessed 15 May 2021].
10. Kaemingk, D., 2021. Online reviews statistics to know in 2021, Qualtrics XM // The Leading Experience Management Software, [online]. Available at: [Accessed 20 May 2021].
11. Martin, M., 2017. Predicting ratings of amazon reviews-techniques for imbalanced datasets. Dissertation (Master). HEC University of Liège.
12. Qaiser, S., Yusoff, N., Ali, R., Remli, M. A. and Adli, H. K., 2021. A Comparison of Machine Learning Techniques for Sentiment Analysis. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(3), pp. 1738-1744.
13. Ravi, A. K., Khettry, A. R. and Sethumadhavachar, S. Y., 2019. Amazon Reviews as Corpus for Sentiment Analysis Using Machine Learning. International Conference on Advances in Computing and Data Sciences, 1045, pp. 403-411.

14. Ronaghan, S., 2018. The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark, Towards Data Science, [online]. Available at [Accessed 12 May 2021].
15. Roshan, R. P. J., 2020. Amazon Reviews Sentiment Analysis: A Reinforcement Learning Approach. Dissertation (Master). Griffith College Dublin.
16. Shaheen, M. A., Awan, S. M., Hussain, N. and Gondal, Z. A., 2019. Sentiment Analysis on Mobile Phone Reviews Using Supervised Learning Techniques. International Journal of Modern Education and Computer Science, 2019, 7, pp. 32-43.
17. Tan, W., Wang, X. and Xu, X., 2018. Sentiment analysis for Amazon reviews. International Conference, 2018, pp. 1-5.