**OPINION MINING AND SENTIMENTAL ANALYSIS ON ONLINE PRODUCT REVIEW**

**INTRODUCTION**

Opinion mining and sentiment analysis offer valuable tools for extracting actionable intelligence from this vast trove of online reviews. Opinion mining, also known as sentiment analysis, involves the computational analysis of text to determine the sentiment expressed towards a particular product or service. By leveraging natural language processing (NLP) techniques, sentiment analysis algorithms can categorize reviews as positive, negative, or neutral based on the language used

This paper will explore the methodologies and technologies used in sentiment analysis, including machine learning algorithms, lexicon-based approaches, and hybrid models. Additionally, it will examine the challenges inherent in sentiment analysis, such as dealing with sarcasm, irony, and cultural nuances. Furthermore, the paper will discuss the ethical considerations surrounding opinion mining, including issues related to privacy, bias, and the responsible use of consumer data.

**ABSTRACT**

In today's digital era, online product reviews play a pivotal role in shaping consumer purchasing decisions. However, the sheer volume and diversity of these reviews pose a significant challenge for consumers seeking to distill meaningful insights. Opinion mining and sentiment analysis techniques offer promising solutions to this problem by automatically extracting sentiments expressed in online reviews. This paper presents a comprehensive exploration of opinion mining and sentiment analysis applied to online product review journals.

Through an analysis of sentiment patterns and trends in online product review journals, this research aims to provide valuable insights for businesses and consumers alike. By understanding consumer sentiment, businesses can enhance their products and services to better meet consumer needs and preferences. Meanwhile, consumers can make more informed purchasing decisions based on the sentiment analysis of online reviews

# SENTIMENTANALYSIS TASKS

The work at hand is both broad and demanding, since it encompasses the fields of natural language processing, web mining, and machine learning. Figure 1.1 illustrates the many degrees of sentiment analysis. The job of sentiment analysis is divided into the following tasks. The user's text does not contain any information to rewrite.

Featurebasedsentiment

Document levelsentiment

Sentencelevelsentiment

Word levelsentimentanalysis

SentimentAnalysistasks

* Classification of Subjectivity: In this context, sentences are classified as either expressing subjective opinions or as being objective statements. The current analysis aims to identify the presence of views and other subjective elements inside a text including a series of sentences, which are otherwise intended to objectively convey factual information.
	+ Sentiment Classification: Once the assessment of a sentence's opinionated nature has been conducted, it is necessary to determine the polarity, which indicates whether the expressed view is positive or negative. The sentiment classification encompasses several approaches, including binary classification (distinguishing between positive and negative sentiments), regression analysis, multi-class classification (categorising sentiments as very negative, negative, neutral, positive, or highly positive), and ranking methodologies. The sentiment categorization is often achieved by two primary strategies, namely the Corpus-based approach (CBA) and the Dictionary-based approach (DBA).
* The CBA (Content-based Affect) algorithm is used to assess the emotional affinity of words and to apply a happiness factor to terms based on their frequent occurrence in blog entries that have been labelled as cheerful.
* The DBA system incorporates lexical resources such as WordNet in order to automatically extract words that are associated to emotions.
	+ Tasks that are provided free of charge or as a gesture of goodwill. The aforementioned duties are included by it.
* The process of extracting the opinion holder or source is referred to as object holder extraction.
* Object/Feature Extraction refers to the process of identifying the target entity.

# Machine Learning Techniques

Machine learning algorithms are very valuable for sentiment classification, since they enable the categorization of text into positive, negative, or neutral categories. This methodology use classification algorithms to categorise textual data into distinct classes. Training datasets and testing datasets play a crucial role in the acquisition of knowledge and the evaluation of the effectiveness of the document, respectively. The acquisition of training samples marks the starting phase of Machine Learning, followed by the commencement of the training set to instruct the classifier. Supervised and unsupervised approaches are widely used machine learning methodologies [25]. Supervised machine learning algorithms such as maximum entropy, XG Boost, Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbour (K-NN) are categorised as supervised learning methods. On the other hand, unsupervised machine learning algorithms including Hidden Markov Model (HMM), Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Singular Value Decomposition (SVD) are classified as unsupervised learning techniques [29].

 ***Supervised Methods***

The process of constructing a function based on labelled training data, which consists of a set of training samples, is often referred to as a supervised learning methodology. Supervised learning relies on a dataset that is labelled, meaning that the training procedure of this learning technique involves the provision of labels.

The Naive Bayes method is widely used in document level categorization due to its simplicity and ease of implementation. The primary objective of this classifier is to use the joint probabilities of words and categories in order to calculate the probability of categories in the designated test text. This approach exhibits more efficacy in addressing the complex challenges associated with interdependent aspects. Moreover, these systems exhibit enhanced speed in decision-making processes and do not need a substantial amount of data for training purposes.

The Support Vector Machine (SVM) is based on the notion of structural risk minimization, which aims to minimise the chance of mistake by exploring hypotheses (h). Support Vector Machines (SVM) provide the capability to dynamically change the training patterns, allowing for the acquisition of a vast array of patterns in order to mitigate the challenges associated with categorization complexity [30].

The K-Nearest Neighbours (K-NN) algorithm falls under the category of lazy learning, since it does not make any assumptions about the underlying data distribution. This classifier relies on the category labels of both the training and testing texts to make predictions. The aforementioned approach involves the classification of an item by determining its majority class within a group of k-nearest neighbours.

The Maximum Entropy approach is particularly advantageous in the field of Natural Language Processing (NLP) since it demonstrates superior performance even when the conditions of conditional independence are not met. This approach does not make any assumptions regarding the relationship between the attributes.

Unsupervised learning refers to a machine learning approach where a model is trained on unlabeled data without any explicit guidance or supervision from a human expert. The absence of a categorical structure and the lack of accurate target identification necessitate the reliance on clustering methods.

# Lexicon-BasedApproaches

This approach uses a sentiment dictionary in conjunction with opinion words to compare and assess the polarity of the words throughout the dataset. Lexicon-based methodologies primarily depend on the use of a sentiment lexicon, which encompasses a compilation of established and prearranged sentiment words, phrases, and even idioms. This lexicon is specifically designed for conventional forms of communication, such as the Opinion Finder lexicon.

The category encompasses the Dictionary-based and Corpus-Based approaches.

# RESEARCHOBJECTIVES

* The objective is to assess the attitudes and views of customers within a singular or many domains, pertaining to a certain product or company.
* The objective of this research is to examine the effects of social media sentiment analysis on the target audience, specifically focusing on the aspects of originality and relevancy.
* The aim is to develop a prediction function or objective function that can accurately forecast a desired result based on predetermined input criteria or qualities.
* In order to eliminate noise such as stop words, stemming, and blank spaces from the collected Twitter data, it is necessary to use a suitable pre-processing strategy.
* In order to enhance the accuracy of the classifier, it is important to do a highly pertinent feature extraction for the purpose of key word extraction.
* In order to improve the classification accuracy, it is possible to use efficient optimisation techniques with the aim of minimising the discrepancy between the anticipated output of the classifier and the actual output**.**

**PROPOSED METHODOLOGY**

 The use of artificial intelligence (AI) learning methods enables the automated identification and classification of user behaviour. When training models, these strategies do not need a substantial quantity of data to classify samples. In this part, a decision tree (DT) and a probabilistic graphical model (PRXGB) have been constructed to implement this paradigm. This section includes parameter settings and performance assessment metrics as well. The proposed design is shown in a schematic manner in below picture.

****

**a. Data source**

 Amazon is a prominent multinational corporation based in the United States that operates in the technology sector. The company's many business endeavours include electronic commerce, whereby they engage in the procurement and storage of merchandise, as well as oversee other aspects such as shipping, pricing, customer support, and refunds. The dataset contains information on the ratings and reviews of over 1,000 Amazon products, as sourced from the official Amazon website.

 The characteristics or attributes of a particular entity or phenomenon. Features can refer to physical qualities,

• product\_id - Identification number assigned to a product • product\_name - Official name assigned to a product • category - Classification or grouping of a product based on its characteristics or purpose • discounted\_price - Reduced price offered for a product after applying a discount • actual\_price - Original or full price of a product before any discounts are applied • discount\_percentage - Proportion or ratio representing the amount of discount applied to the product, expressed as a percentage • rating - Evaluation or assessment of a product's quality or performance, typically expressed through a numerical scale or rating system • rating\_count - Total number of individuals who have participated in the voting process to provide a rating for the product on the Amazon platform

 The dataset includes the following variables: "about\_product," which provides a description of the product; "user\_id," which represents the unique identifier of the user who wrote the review for the product; "user\_name," which indicates the name of the user who wrote the review; "review\_id," which represents the unique identifier of the user review; "review\_title," which provides a concise summary of the review; "review\_content," which contains the detailed content of the review; "img\_link," which provides the link to an image of the product; and "product\_link," which provides the official website link of the product.

**b. Preprocessing**

 Data is kept into log files in a plain text format. The task involves eliminating extraneous data and analysing the occurrence rates of several violation categories, including data frequency, network frequency, and error frequency, based on the provided logs. The severity of errors may vary, and their occurrence frequency can impact this variability. Furthermore, the significance of the erroneous term as a contributing factor. In a mathematical context,

 $E\_{web}= \sum\_{n=0}^{M-1}ω\_{n}d\_{n}$

 The weight factor of a tweet category ranges from 0 to 1, denoted as ω\_n, and represents the frequency of visits or use of the category. The variable M represents the aggregate quantity of tweet categories under observation, including several domains such as electronic mail and e-commerce, secure and insecure websites, social media platforms, as well as entertainment-related tweets, among others. The organisation assigns a weight to each category, which may vary across different organisations. In order to measure the frequency of user logs, a metric is used.

 $E\_{net}= \sum\_{n=0}^{N-1}ω\_{b}d\_{b}$

 The variable B represents the cumulative number of network actions being monitored, including but not limited to FTP, shared files, and user areas. The parameters ωb and db denote the weight and frequency of access for each respective network type. Machine log frequencies may be expressed in a comparable fashion to:

 $E\_{Mac}= \sum\_{o=0}^{o-1}ω\_{o}d\_{o}$

 In this context, O represents the overall quantity of user review actions that are now under observation. The variables ω\_o and d\_o denote the respective weight and frequency of accessing each individual log category. It is important to standardise the weighted frequency score for each category due to the potentially substantial number of frequencies. Normalisation is achieved by dividing the weighted frequency components by both the total number of categories and the maximum frequency within the relevant category. Subsequently, the frequency of each kind will be normalised by the implementation of a normalisation procedure. The normalisation method takes into account both the total number of categories within a type and the maximum frequency of the type. Therefore, equations 3.4-3.6 represent$BE\_{Web}$, $BE\_{Net}$ and $BE\_{Mac}$, namely, normalized user tweet data frequencies, respectively. These are given by:

 $BE\_{Web}= \frac{D\_{web}}{N\*max⁡(d\_{1},d\_{2}, …,d\_{N})}$

 $BE\_{Net}= \frac{D\_{Net}}{B\*max⁡(d\_{1},d\_{2}, …,d\_{B})}$

 $BE\_{Mac}= \frac{D\_{Mac}}{O\*max⁡(d\_{1},d\_{2}, …,d\_{O})}$

 The variables N, B, and O denote the aggregate count of tweet categories inside each respective category. To maintain consistency with the input format of the constraint mean system, it is necessary to confine the frequency factor within the range of 0 to 1.

**c. Feature extraction**

 The procedure of decomposing connected variables into their constituent metrics is a fundamental component of the Principal Component Analysis (PCA) cycle. Principal component analysis (PCA) is a technique used to reduce the dimensionality of fused feature (FF) vectors, hence enhancing their classification accuracy. The construction of the model include the consideration of characteristics that demonstrate appropriate classification of the data. The game FF consists of n distinct dimensions.In the initial F=$z\_{1},z,…,z\_{n}$the concept under consideration has several dimensions, necessitating its reduction to the variable "kn." The Principal Component Analysis (PCA) algorithm proceeds through a series of phases in order to obtain a collection of reduced fused features. (RFFs):

1. **Data scaling:**

$O\_{j}^{i}=\frac{O\_{j}^{i}-\overbar{O\_{j}}}{σ\_{j}}$

1. **Co-variance matrix computation:**

$$\sum\_{}^{}=\frac{1}{m}\sum\_{i}^{m} \left(O\_{i}\right)\left(O\_{i}\right)^{T},\sum\_{}^{}\in R^{n×n}$$

1. **Third, we calculate the eigenvector and eigen value.:**

$V^{T}\sum\_{}^{}=λμU=\left[\begin{matrix}∣&∣&∣\\V\_{1}&V\_{2}…&V\_{n}\\∣&∣&∣\end{matrix}\right],V\_{i}\in R^{n}$

1. **Eigen value selection.The formula used in the selection process of the top 100 Eigen values in K-space for implementation into our classification system was as follows.**

$z\_{i}^{new }=\left[\begin{matrix}V\_{1}^{T}z^{i}\\V\_{2}^{T}z^{i}\\\cdots …..\\…..\\V\_{k}^{T}z^{i}\end{matrix}\right]\in R^{k}$

1. **At last the feature matrix was constructed.**

**d. Classification**

**1. PRXGB**

 The data obtained in the previous iterations was used for the purpose of categorising the reviews provided by users. The categorization procedure started by conducting an examination of the essential characteristics of the product, in conjunction with an assessment of the user's individual preferences. One such approach to analysing an input involves the use of source characteristics that have been adjusted in terms of scaling and translation. Finite impulse response (FIR) filter banks may be used to categorise features into portions of varying importance, namely low and high. In the prioritisation of data characteristics, more emphasis is placed on attributes that exhibit quick changes, whereas attributes that exhibit sluggish variations are assigned a lower priority. The use of row and column low priority classification may be employed to get coefficients that represent the overall count of features present in the dataset. When assigning low priority classification to the values in the rows and high priority classification to the values in the columns, it is possible to see various vertical nuances in these coefficients. By assigning priority to both rows and columns, one may determine the coefficients that effectively preserve the horizontal properties of the given data. The categorization of rows and columns with high priority results in the determination of fine-scale coefficients, which reveal the diagonal features of the tweet data. In addition to Twitter traits, these details are used for classifying people' viewpoints.

 In the suggested technique, the training data samples are first assigned equal weights. The weight is adjusted for each instance based on the output of the classifiers. Instances that are classified correctly have a decrease in value, whereas instances that are misclassified experience an increase in value due to the redistribution of weights. The provided data is then inputted into another classifier, which aims to accurately categorise the instances that carry more weight in the algorithm during the subsequent iteration. The recalibration of weights is performed once again based on the output of the newly implemented classifier. The act of recalculating weights in order to maintain a constant total is often referred to as "normalisation." Ultimately, the determination of the hypothesis value occurs at completion of all iterations. The hypothesis proposed by the suggested Classifier in Equation 3.11 argues that it may take on values of either 0 or 1.

$user review class=\frac{K}{I}=\frac{K}{\sum\_{i=1}^{i\_{0}}I\_{i}}$

Where,

$I\_{i}=K×(\sqrt{\frac{E}{Z}})^{2}$

 The linear model, denoted as N, is widely recognized and accepted in academic circles due to its conceptual and quantitative merits. In the field of classification, many predictive traits have been associated with N.

$$ϕ\left(K\right)=ϕ(H\left(p\right),B\left(p\right))$$

$z\left(K\right)=f(ϕ\left(p\right);(H\left(p\right),B(p))$

Here, $π$=conditional probability of {$P(K=1|M\_{1},…,M\_{k})$}.

Inverse transformation is:

$P\left(M\_{1},…,M\_{k}\right)=\frac{e^{α+\sum\_{j=1}^{k}β\_{j}M\_{j}}}{1+e^{α+\sum\_{j=1}^{k}β\_{j}M\_{j}}}$

**Step 1:** In this study, we will choose a weighted feature index, determine the cluster number of categories, and consider the number of data samples.  and cluster center.

**Step 2:** Determine the classification function for selecting the user type.

$\hat{c}\_{i}=tant(℧\_{c}GC\_{θ\_{q}}\left(q\_{i}\right)+W\_{c}t\_{i-1}+b\_{c})$

 The equation provided is used for the purpose of determining the characteristic

function. The below equation may be used to ascertain the user type of the database.,

****

 The maximum rate of incitement is denoted as , whereas the ideal weight of the movement initiated is within the range of [0, 1].,

**Step 3:** Compute the user review updation

 The resolution of this issue necessitates the use of algorithms capable of navigating subjective multidimensional spaces. For instance, the Lagrangian pertaining to a decision space of n dimensions may be succinctly expressed.

$E\_{b}+E\_{w}=1$

Where,
$$E\_{b} :=\{Q\in E|Q\left(T\right)=b\}$$

$$D\_{w} :=\{Q\in E|Q\left(T\right)=w\}$$

The XG boost classifier is used to develop a robust classifier for efficiently categorising messages from Twitter users.

**2. Decision tree**

 A decision tree (DT) is a simple classification algorithm that may be used to categorise data into distinct categories. Within the field of data technology (DT), the process of segmenting data occurs in a continuous manner, according to a pre-established criteria. The decision trees (DTs) are well recognised and acknowledged within the domain of supervised categorization. Artificial intelligence systems provide high efficacy in performing categorization tasks, exhibiting streamlined decision-making processes, and may be rapidly and effortlessly developed via the use of an efficient algorithm. Due to its early introduction as a prominent method in the study of predictive modelling, this methodology has gained significant recognition and popularity within the area”.

$X=\left[D\_{x},D\_{y}\right]$

where $D\_{x}$ and $D\_{y}$are the factors that go into the equation,

$D\_{x}=\frac{1}{3}\frac{\sum\_{i=0}^{n-1}  \left(X\_{i}+X\_{i+1}\right)\left(Y\_{i}X\_{i+1}-Y\_{i+1}X\_{i}\right)}{\sum\_{i=0}^{n-1}  \left(Y\_{i}X\_{i+1}-X\_{i+1}X\_{i}\right)}$

And

$C\_{y}=\frac{1}{3}\frac{\sum\_{i=0}^{n-1}  \left(X\_{i}+X\_{i+1}\right)\left(XY\_{i+1}-X\_{i+1}Y\_{i}\right)}{\sum\_{i=0}^{n-1}  \left(X\_{i}x\_{i+1}-X\_{i+1}Y\_{i}\right)}$

**RESULT AND DISCUSSIONS**

 It is our goal in this part to analyze the model's performance to other techniques provided in the literature. The whole experimentation was carried out under python environment.

****

**Feature attributes**

****

**Rating score analysis of the classifiers**

Depend upon the user review the rating score was given by the suggested classifier and the classification was done precisely as depicted in above picture

** Epoch Vs Loss**

**Accuracy:**

It determines the number of users that are correctly classified. It decides how closely the outcomes match the actual ground truth results.

$ACCU =\frac{TP+TN}{TP+TN+FP+FN}$

**Precision**

It determines how accurate the suggested technique's behavior is by separating required user reviews from the dataset .

$Pre=\frac{TP}{TP+FP}$

**Recall**

The ratio of correctly predicted instances and all instances.

Recall = $\frac{RO}{RO+DB}$

**F1 score**

With the help of two measures (precision and recall), we caneasily calculate the F-measure.

**Classification report for the PRXGB classifier**

**Classification report for the DT classifier**

 The overall obtained classification output of the PRXGB and the DT classifier are depicted in the above picture. As of that image The PRXGB shows satisfied performance than the DT classifier.

**** **ROC calculation of the DT classifier**

****

**AUC analysis of the PRXGB classifier**

**CONCLUSION AND FUTURE WORK**

 The transition from traditional offline marketplaces to digital markets has led to a significant rise in client reliance on online reviews. The use of online reviews has emerged as a significant medium for cultivating trust and exerting influence on consumer purchasing behaviours. Given the extent of reliance on reviews, it is essential to effectively manage the substantial quantity of evaluations and provide trustworthy reviews to consumers. The objective of our study is to accomplish this goal by the use of sentiment analysis on evaluations of electronic devices, with the purpose of categorising the reviews into either positive or negative sentiment. The data was balanced by ensuring a nearly equal distribution of positive and negative evaluations. Subsequently, two classification models were used to categorise the reviews. Among the two classifiers, namely Decision Tree and PRXGB, it has been determined that PRXGB exhibits superior performance. The accuracy results have undergone cross-validation, revealing that the PRXGB model earned the greatest accuracy value of 93% compared to the other models. Several potential future enhancements may be included to enhance the model's efficacy and applicability in actual scenarios. Our next endeavours include the use of Independent Component Analysis (ICA), specifically in the active learning process, to achieve a more comprehensive automation of the data labelling process, hence reducing the need on help from the oracle. The model has the capability to be integrated with software applications that facilitate interactions between customers and their pursuit of evaluating the quality of a certain product. By using a dataset of significant magnitude, it becomes feasible to implement the model on local market platforms, therefore enhancing both accuracy and usability. Lastly, we will endeavour to extend this study until we can apply this approach to other types of text-based reviews and comments.

**Reference Websites:-**

1. <https://data.world/crowdflower/brands-and-product-emotions>
2. https:/[/www.kaggle.com/thoug](http://www.kaggle.com/thoughtvector/customer-support-on-twitter#twcs.csv)h[tvector/customer-support-on-twitter#twcs.csv](http://www.kaggle.com/thoughtvector/customer-support-on-twitter#twcs.csv)
3. https:/[/www.kaggle.com/ka](http://www.kaggle.com/kazanova/sentiment140)z[anova/sentiment140](http://www.kaggle.com/kazanova/sentiment140)