International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V71I12P209

Volume 71 Issue 12, 78-89, December 2023
© 2023 Seventh Sense Research Group®

Original Article

Optimizing Customer Experience Analysis Across Dataset Size Reduction and Relevant Features Selection

Sara AHSAIN[1*], Yasyn EL YUSUFI[1], M'hamed AIT KBIR[1]

[1]*Intelligent Automation & BioMedGenomics Laboratory (IABL), STSM Doctoral Studies Center, Abdelmalek Essaadi University, Morocco*

[*]*Corresponding Author: sara.ahsain@etu.uae.ac.ma*

Received: 27 July 2023 Revised: 12 September 2023 Accepted: 18 September 2023 Published: 06 December 2023

Abstract *- Today, in an era of data-driven business, customer sentiment analysis is becoming more important. It allows organizations to identify areas in their operations where some services and products can be improved. This can help them to make better decisions and improve their customer experience. The main goal of this study is to classify Amazon customers' reviews. The dataset consists of a collection of product reviews with an overall appreciation. This dataset is a rich source of information for academic researchers in the fields of natural language processing and machine learning that concern customer experience understanding with some products. Despite its diversity in terms of product categories, the huge number of records makes the exploration and the use of this dataset time and resources-consuming. Thus, it is not easy to use computers with standard performances. The proposed approach is centered on selecting a representative subset of the original dataset combined with relevant feature selection, using ensemble learning techniques, on reducing the processed data size while achieving interesting results compared with research interested in the same dataset. In fact, when dealing with the 'Magazine subscriptions' category and using only 12% of the original collection of examples, the proposed approach shows a high level of performance with respect to the following metrics: accuracy (up to 0.94), sensitivity (up to 0.90) and specificity (up to 0.97).*

Keywords *- Classification, Feature extraction, Feature selection, Sentiment analysis.*

1. Introduction

Artificial Intelligence (AI) has greatly impacted the field of digital marketing. Its main objective is to help companies create more effective, personalized marketing campaigns.

In fact, Machine Learning (ML) is a subset of AI that enables machines to learn and improve from data experience. It provided great help with the process of decision-making, customer engagement, and personalization through its capability of processing a vast amount of customers' data like their behavior, preferences and purchasing history. Nevertheless, ML tools should be carefully integrated and must consider data privacy problems, ethics and transparency. It should, without fail, find a balance between personalising content and respecting the customer's privacy. Only the reviewer ID has been used in the dataset used in this research paper, and the rest of the data is related to the review.

One example of ML usage in Morocco for a better understanding of the consumers' needs and customs ways is Marjane Holding [1], a hypermarket chain that debuted in Morocco in 1990 and started this year a digitalization journey. It has, in fact, launched its first marketplace mobile platform at the start of 2023. Additionally, they started using AI to help with the decision-making process, digital marketing and recommendation systems. They rely on the use of data like sales transactions, customer loyalty schemes and the different interactions of the users on their digital platforms [2].

One of the crucial steps to understanding the customer's tendencies is the ability to analyze the sentiment related to the text that a user adds digitally, like product reviews on e-commerce platforms or tweets about a precise trend. Sentiment Analysis (SA) tools can nowadays distinguish the different nuances of the text to get a comprehensive understanding of the customer.

The goal of this research is to optimize the customer experience by detecting their appreciation for products after purchasing. This analysis allows the seller to focus on the winning products, fix the deficiencies of products with potential and completely eliminate weak products. The focus has also been on selecting a representative subset of the dataset while focusing on keeping interesting results.

For an effective analysis of the customers' needs, the research has focused on a dataset that encloses various reviews about a wide range of products, ensuring a wide variety of text data. The dataset was annotated using different text analysis tools based on the customer's general appreciation. Then, the paper focuses on multi-class labeling to capture the text data's nuances.

This research paper is presented as follows: the first section focuses on details of some of the newest related work to this research area. Then, an outline of the goal behind this research paper and more details of the dataset that has been used. Next, a section has been devoted to the understanding of data and the features used for this end, and there is also an explanation of the labeling of the dataset and extracting features from text data. Then, some feature selection methods have been explored to solve the dataset's high dimensionality and reduce resource consumption. An explanation of the choice of the optimal number of features has also been detailed. Finally, before displaying the results, a section details the methods and metrics used to evaluate this work.

This research aims to optimize the customer experience by integrating existing and novel techniques for relevant feature selection. The goal is to enable users to extract meaningful insight from large and diverse datasets. Natural Language Task (NLP) is known for being a very time-consuming task if not dealt with correctly. This paper will explore an innovative approach to reduce the dataset size by avoiding the use of all the available samples and selecting the most relevant features; otherwise, it will be impossible to use machines with usual performances to deal with some problems.

The dataset size reduction and feature selection research has been a critical area of research for businesses to improve the client-enterprise relationship. This paper aims to contribute to addressing this research gap and advance the field of NLP.

2. Related Work

This section discusses the latest innovations related to 'sentiment analysis' and 'Machine Learning methods. The focus is on analysing the feature extraction and selection methods that are most used in this research area and discussing the classification techniques and the metrics chosen to evaluate the model performance.

Table 1 outlays some of the latest research. Researchers have noticed that one of the most used classical methods to extract features and prepare text data to be fed to machine learning models is Term Frequency-Inverse Document Frequency (TF-IDF). It is a very common method used to transform text into a structured representation that captures the word's importance across the document.

Table 1. Related work on techniques used in sentiment analysis

Author	Feature extraction	Feature selection	Classification technique	Evaluation
[3]	TF-IDF	Particle Swarm Optimization	Genetic Algorithm(GA), ant Colony Optimization (AO), Particle Swarm Optimization (PSA)	Precision, recall, accuracy, and F1-score
[4]	Log Term Frequency-based Modified Inverse Class Frequency (LTF-MICF)	Hybrid Mutation-based White Shark Optimizer (HMWSO)	GARN architecture, which combines Recurrent Neural Networks (RNN) and attention mechanisms	Accuracy, precision, recall, f-measure
[5]	Bag of Words, TF-IDF	Not stated in the paper	Support Vector Machines (SVM), Bidirectional Encoder Representations from Transformers (BERT) and Long Short-Term Memory Network (LSTM)	Accuracy, precision, recall, ROC AUC and F1-score
[6]	TF-IDF, bags of words, Word2ec	Not stated in the paper	Naive Bayes NB, Random Forests RF, SVM, Logistic regression and LSTM	Precision, recall, accuracy, and F1-score
[7]	TF-IDF	Information Gain (IG) and Chi-Square	SVM, Decision tree DT and k-Nearest Neighbors (KNN)	Accuracy, precision, recall
[8]	Not stated in the paper	Information Gain	SVM, KNN and NB	Accuracy, ROC AUC
[9]	TF-IDF	Not stated in the paper	NB and SVM	Precision, recall, accuracy, and F1-score

Paper [3] presents a method to perform a perceptive sentiment analysis (FS) by gathering information about the various possible solutions generated by the PSO algorithm. The research aims to identify the most accurate subset of drug reviews for improving sentiment analysis classification. According to the experiment results, PSO performed better than ant colony optimization and a genetic algorithm regarding classification. In terms of precision, recall, F-score, and accuracy, the algorithm had an average of 49.3%, 73.6%, and 57.2%, respectively.

Paper [4] proposes a Gated Attention Recurrent Network (GARN) architecture that combines Recurrent Neural Networks (RNN) and attention mechanisms. A novel approach was also proposed to extract features from the Log Term Frequency-based Modified Inverse Class Frequency (LTF-MICF) dataset. The algorithm had an average accuracy of 97.86%, precision of 96.65%, recall of 96.76% and f-measure of 96.70%.

Paper [5] used word-of-bags and TF-IDF to extract features from the dataset. They used Support Vector Machines SVM, Bidirectional Encoder Representations from Transformers (BERT) and Long Short-Term Memory network (LSTM) for the classification task. It reached an accuracy of 88.79%, 90.59% and 90.42% respectively. Paper [6] works on Moroccan text analysis using TF-IDF, bags of words, and Word2ec for feature extraction and proceeded to a comparison between Naive Bayes (NB), Random Forests (RF), SVM, Logistic regression and LSTM. They have reached an accuracy of 68.59%.

Paper [7] worked on Arabic text data as well. This paper aims to solve the dimensional issue by comparing the features selection algorithms Chi-Square and Information Gain. The SVM classifier performed well in the tests and achieved the highest accuracy with the IG algorithm at 85%.

The main observation is that one of the recurrent models used for classification is SVM and LSTM. For the feature extraction, TF-IDF is being used by the majority of research papers cited in this paper.

3. Work Proposed in this Paper

3.1. Goal

This research paper aims to provide an aid to decision-making by applying machine learning to a real-life customer's review dataset in order to guess the consumer's appreciation of the product based on the review. The goal is to work on subjects like sentiment analysis, mining customers' opinions about a given product, etc.

For that, different Machine learning models have been used to get the best results. The paper covers different multi-class classifications based on two target variables.

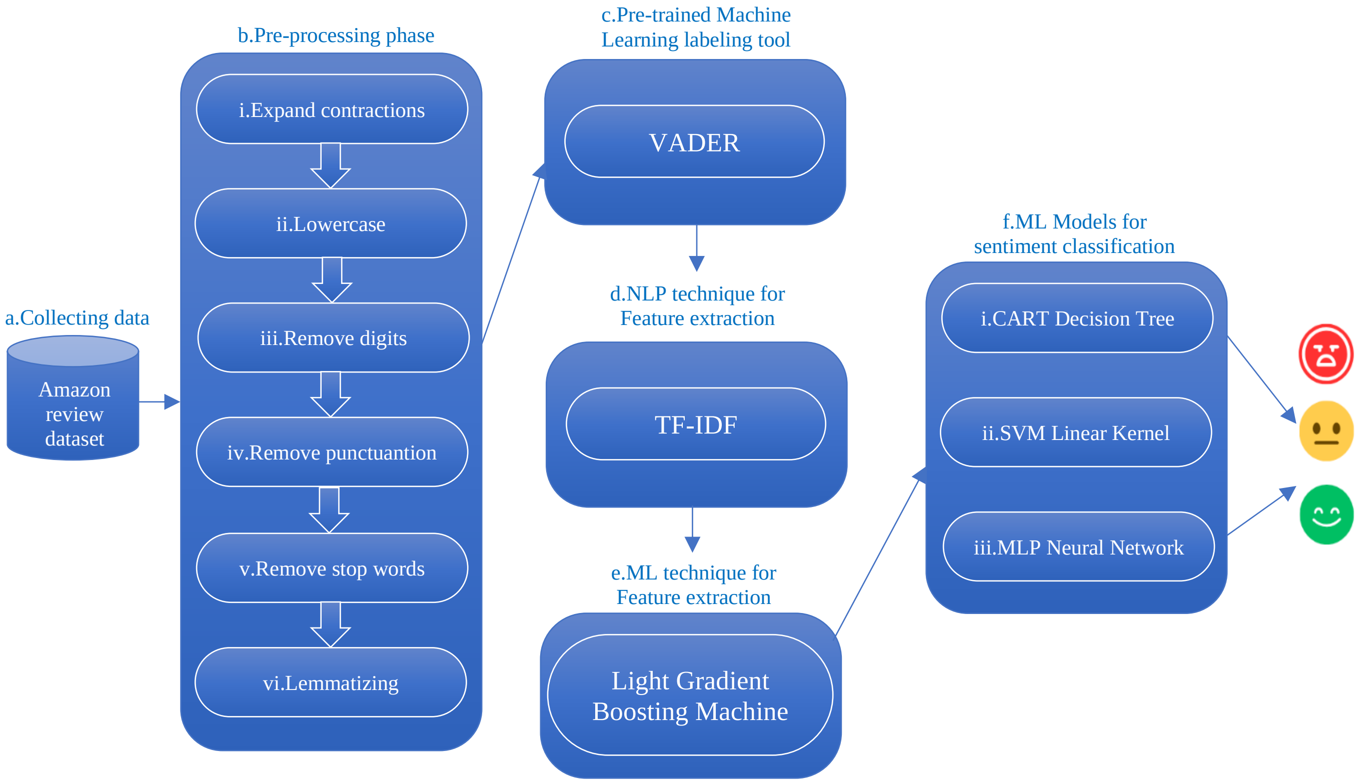

Fig. 1 The global architecture of the proposed solution

This research discusses different text analysis methods aiming to get the best results with Machine Learning algorithms; the goal is to apply these findings to a future dataset that is still being collected and in relation to Moroccan culture and customers. Continuing this research, these findings will generalize on Arabic text data.

The analysis conducted in this paper can give more insights to the sellers, retailers, etc., about which products are popular and liked by the users quickly and effectively. It can also help the sellers improve the quality of their products by directly targeting unsatisfied clients.

By exploring existing techniques' effectiveness, the main goal of this study is to create a new methodology that will advance sentiment analysis and the NLP field in general by understanding and creating a more accurate sentiment analysis model.

The importance of this research paper revolves around optimizing the customer experience analysis by working on the problem of dataset size reduction and relevant feature selection. In the upcoming sections, multiple methods have been compared before choosing the best one. In fact, this paper gives a detailed description of relevant feature selection methods and a comparison study that justifies the use of 1,000 best features only when machine learning models are staged.

Figure 1 outlays the global architecture of the solution proposed in this research paper and the different steps the dataset has passed through.

3.2. Dataset

The dataset used for this study represents Amazon reviews of multiple products spanning from 1996-2018. The initial dataset contains more than 233.1 million reviews [10], also available in subset datasets divided by categories. The table below shows how the data is divided into multiple categories.

In this research, many sub-categories have been examined and tested before settling on the 'Magazine subscriptions' category with 89,689 reviews.

Table 2. Amazon review dataset distribution

Dataset	Number of reviews
All Reviews	233.1 million
Movies and TV	8,765,568
Books	51,311,621
Cell Phones and Accessories	10,063,255
Clothing, Shoes and Jewelry	32,292,099
Digital Music	1,584,082
Electronics	20,994,353
Magazine Subscriptions	89,689
Software	459,436
Home and Kitchen	21,928,568

Table 3. Dataset feature explanation

Feature	Significance
Overall	Rating of a product
Verified	Shows if the purchase was verified
Review time	Raw time of the review
Unix review time	Unix time of the review
Reviewer ID	Identification of the reviewer
Asin	Identification of the product
Style	Dictionary of the product metadata
Reviewer name	Name of the reviewer
Review text	Text written by the reviewer as a review of a given product.
Summary	Title or summary written by the reviewer
Vote	Helpful votes of reviews by other reviewers
Image	Attached is an image of the review

Because of the limited resources to work on such a voluminous dataset and with the aim of having acceptable results, the research has focused on randomly selecting an even number of 9,000 reviews for each class in order to obtain a balanced dataset.

Table 2 presents 10 of the most interesting categories with the number of reviews they enclose.

The dataset holds 12 features containing a mix of text and numeric data. This research has focused on the natural language processing task and used only the text data and a column 'overall' that represents the rating of the product that a user gave.

The main features this research has focused on for text analysis are the 'overall' column to precise the appreciation of the client and the 'review text' column containing the review text that the customer had added.

A suitable dataset for sentiment analysis should have high-quality annotations. Each data point should be correctly labeled with the corresponding sentiment (positive, negative, neutral, etc.) to ensure the dataset's quality and integrity. Also, having an adequate dataset reduces overfitting, and a balanced label distribution is important as it avoids biases.

Real-world data reflects sentiments expressed in authentic user-generated content like social media posts or product reviews. This ensures that the model can handle the nuances, irregularities, and challenges found in actual user sentiments.

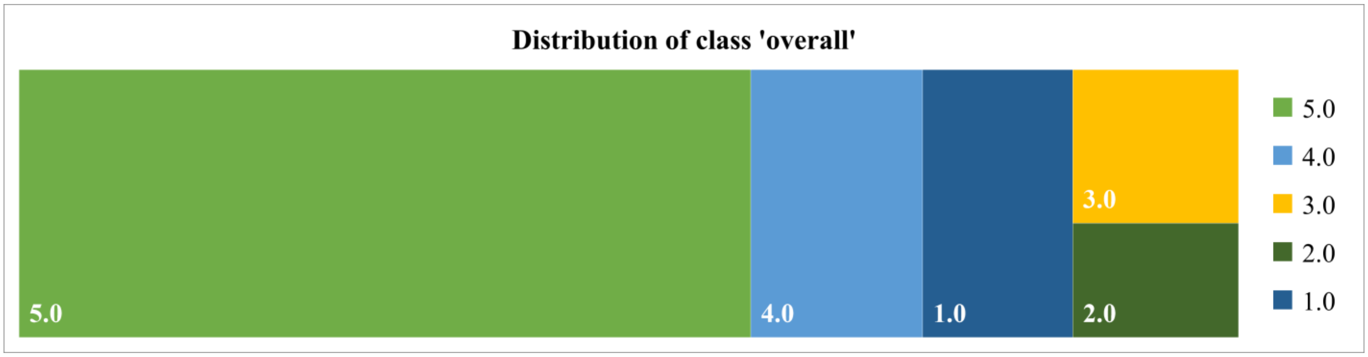

Fig. 2 Distribution of the target column 'overall'

4. Materials and Methods

4.1. Understanding the Data

The dataset used is a subset of the Amazon reviews dataset [10] that originally extended from 1996 to 2018 and has more than 233 million reviews [11].

The subset database is a Magazine subscriptions category with approximately 89,000 reviews. This section covers the exploration of the data and the cleaning process.

In the preliminary exploratory data analysis step, it has been noticed that a total of 11 features with some columns having missing values like 'vote', 'style' and 'image'. The target column is 'overall' with 5 classes (1.0, 2.0, 3.0, 4.0, 5.0).

Figure 2 outlays the 5 classes with an uneven distribution, and it is also obvious that the majority of the reviews are positive.

Also, after verifying the features' unique values, it shows that the 89,689 reviews were written just about 2,428 products by 58,399 reviewers.

Although the dataset contains 89,689 rows and 12 columns, only text data is needed for the Sentiment Analysis Task covered in this paper; thus, only keeping the review text and overall values is important.

Before starting the data cleaning process, the dataset had to be verified for null values to be dropped (only 49 empty reviews were found).

The cleaning process will focus on the column 'reviews' since it contains the main text on which the rest of the analysis will be based. To determine the data cleaning steps that will be adapted to the dataset, the best approach is to look at some random reviews from the dataset.

Based on that, the text has some contractions like (it is, would have, etc.), numbers, punctuations (',' ';' '!'), and uppercase text. In order to work on that, these instructions have been followed:

- Expand contractions:
  Contractions are a reduced version of some words like would have for would have. These expressions are used to shorten text but do not help in text analysis. A dictionary of some contractions has been created and mapped to their expanded version using regular expressions for better results.

- Lowercase:
  When working with 'Natural language processing', models can interpret words like (good and Good) as two different words.

- Digits:
  Digits will only confuse the model, and its accuracy will be reduced. For that, digits have been removed and words with digits as well.

- Punctuations:
  Punctuation is important for English grammar but does not make a difference in text analysis. As a result, all punctuation and special characters from the text have been removed.

- Stop words:
  Stop words are considered common words in English grammar, so they were removed for better search performance.

- Lemmatizing:
  It is the process of stemming the words to their original state (Ex, best for good, etc.). The goal is to return to the base of a given word. After multiple tests, the spaCy [12] library has been chosen as it gives the most accurate results.

This section explores the dataset further after the pre-processing phase. The word cloud [5, 13, 14] method has been used to represent a visually appealing visualization method for text. Its main goal is to provide an overview of the words that appear in the text with a high frequency. The word cloud has a method that eliminates stop words in the visualization process, but this step has already been performed in the data pre-processing step.

Figure 3 shows the 'word cloud' Python library, used to visualize the frequency of words present in reviews.

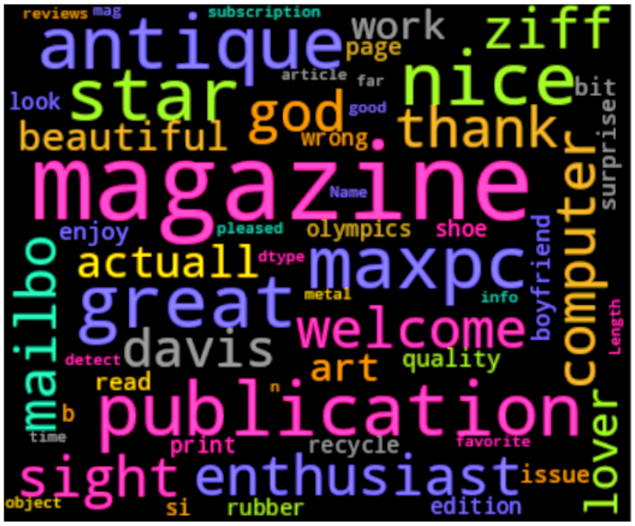

Fig. 3 Word cloud representation of all reviews

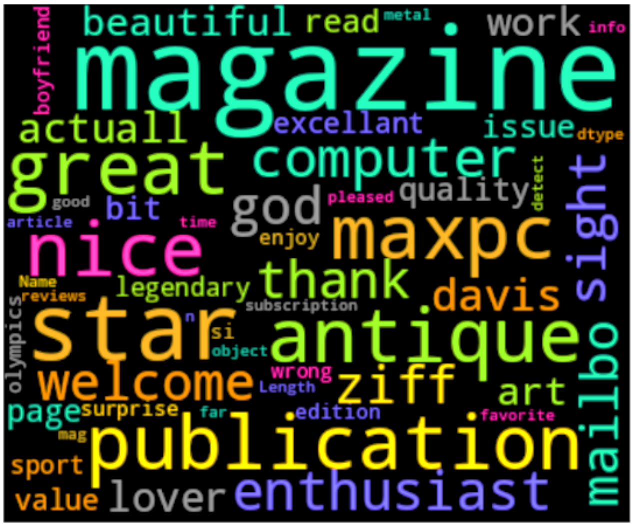

Fig. 4 Word cloud representation of positive reviews

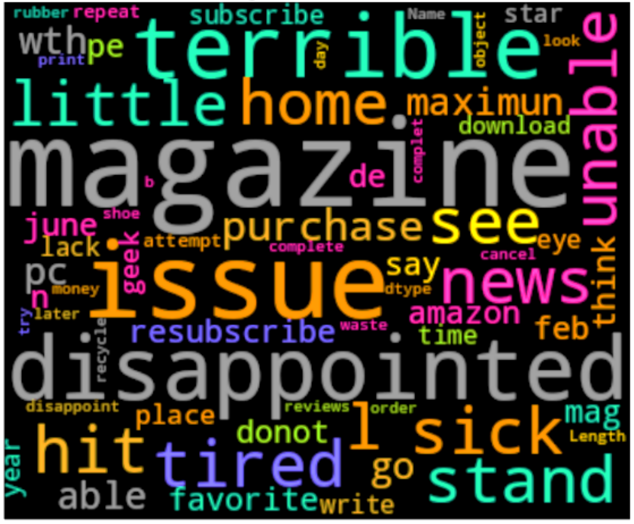

Fig. 5 Word cloud representation of negative reviews

The conclusion that can be drawn from this representation is that a small number of positive words such as ‘nice’, ‘great’, ‘beautiful’, ‘enthusiast’, ‘enjoy’, etc. But it also shows that many neutral words in the reviews would not be useful with the appreciation analysis of the consumer reviews.

Nevertheless, it gives an overview of the general satisfaction of the consumers, and it shows that this dataset has a majority of positive reviews.

To get more insight from the dataset, this research has started with a word cloud visualization of positive and negative reviews. Figures 4 and 5 show the emergence of some positive and negative words, respectively. It also shows that some neutral words (that will most likely not alter the consumer's appreciation) are repeated through reviews like ‘unfailingly’, ‘subscribe’ and ‘inconsistency’.

To gain more insight into the data and for it to be used in this paper, the dataset has been labeled for a classification and a regression based on the ‘overall’ column.

4.2. Labeling Data

This paper will use the dataset to train machine learning models by considering this problem as a classification task. The process involves determining the various emotional tones and subjective opinions expressed in a given text, whether positive, negative or neutral. Then, it has been compared to the original ‘overall’ column. It represents a customer's score given to a product and can be 1 to 5 stars. Having done the right data labeling, it can be expected that good efficiency will result from the model.

Data labeling is necessary for the model to learn patterns and relationships between data in the training phase.

The Amazon review dataset used in this paper does not have a label column by default. But a human annotator can deduct the general impression by the column ‘overall’.

Nevertheless, this research relied not only on manual labeling but also tested renowned Machine Learning and text mining methods that have been trained preliminary on English vocabulary.

This section will discuss and compare labeling methods to choose an adequate method for the rest of this research.

- VADER - Valence Aware Dictionary for Sentiment Reasoning tool [15]:
  A Machine Learning algorithm is a rule-based and lexicon-based approach to analyze the sentiment enclosed in the text by assigning scores to specific words or sentences. It considers the intensity and valence of the words in expressing sentiments. This is a pre-trained model on English data that allows text sentiment classification.

The intensity and polarity of the words or phrases used in a text are assessed by assigning polarity labels to each word or phrase and then categorizing them into various sentiment categories. It can be used to analyze the overall sentiment score of a text, or it can be applied to the specific words or phrases in a document.

- TextBlob for Sentiment Analysis [16]:
  It offers a pre-trained built-in sentiment analysis tool (also based on Machine Learning methods) that can be used to analyze the various aspects of a text. It can assign a polarity score ranging from -1 to 1 and measure the text's degree of objectivity or subjectivity. It can automatically classify English text as neutral, positive, or negative.

  TextBlob simplifies sentiment analysis tasks, making it suitable for social media monitoring and opinion-mining applications.

The figures below show the results of data labeling on an original dataset of 89,689 reviews. First, the different automatic labeling methods and comparison through a bar plot in Figure 6.

This figure shows that the text analysis tools divided the dataset into three separate categories: 'Positive', 'Negative' and 'Neutral'. It also shows that the majority of the reviews were classified as positive, which endorses the conclusion made in the previous section, where the frequent words across all the reviews and the occurrence of some positive vocabulary have been visualized.

Based on these results, it was mandatory also to try manual labeling of the reviews using the target column. As a reminder, 'overall' is the target column in this dataset. It has 5 distinct values that are unevenly distributed, and it represents the product's rating.

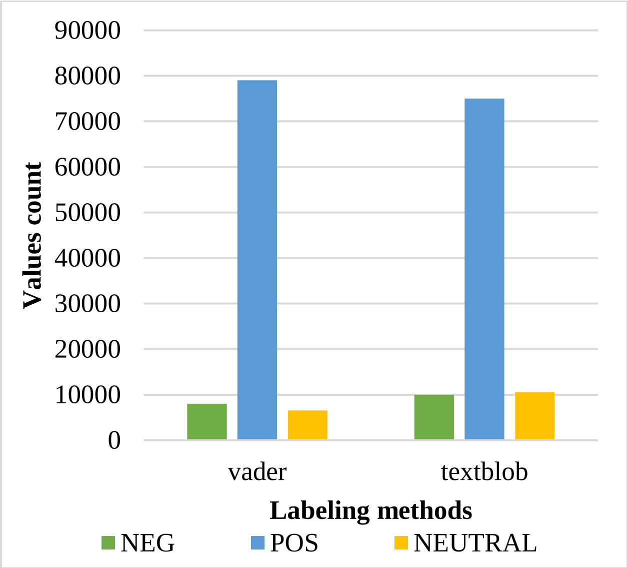

Fig. 6 Visualization of labeling results of text analysis tools Vader and Textblob in the bar plot

All the reviews with an overall rating equal to or higher than 4.0 have been labeled 'positive'. Those with an overall rating equal to or less than 2.0 have been labeled as 'negative' and 'neutral' in all the reviews, with an overall rating equal to 3.0.

Figure 7 shows the representations of the methods used in this paper for the data labeling. It reveals that the results are generally similar; most of the data was labeled positive in the three methods. The smallest class across all the tools is neutral, and the middle class is negative.

In detail, the results were 66,466 positives, 16,252 negatives and 6,971 neutrals using the manual labeling method. Conversely, a result of 74,469 positives, 9,817 negatives and 5,403 neutrals with the VADER labeling tool. Moreover, finally, a total of 75,498 positives, 7,712 negatives and 6,430 neutrals using the TextBlob Sentiment Analysis method.

This paper used the labeling results of the VADER tool, mainly for its specialized focus on sentiment analysis in social media and informal text, ensuring accurate results for this dataset. In addition, the VADER tool has a well-established reputation due to its community support and extensive research.

4.3. Feature Extraction

In Natural Language Processing, feature extraction is essential as it helps machine learning systems process text data by extracting its numerical representations. It also condenses data, which helps reduce its high dimensionality and improve computational efficiency. Picking relevant features can enhance the frameworks' generalization capabilities.

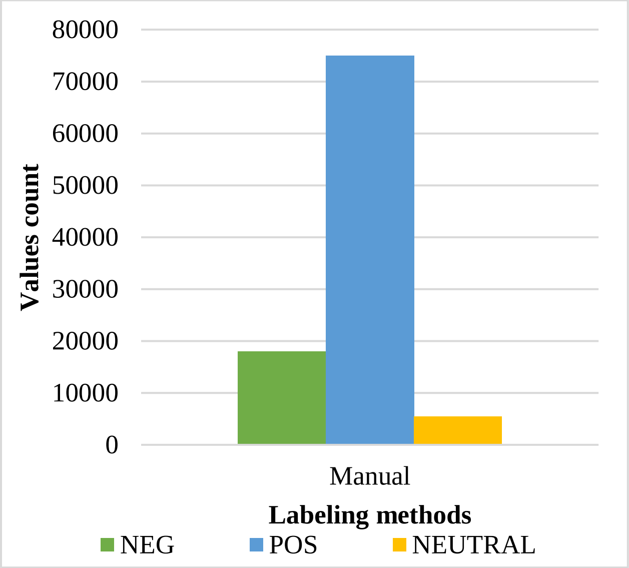

Fig. 7 Comparison of labeling tools results against manual labeling

This paper has used Term Frequency Inverse Document Frequency TF-IDF algorithm to achieve this. It is a very common method used to transform text into a structured representation that captures the words' importance across the document.

The TF-IDF algorithm [3, 17, 18] measures the originality of a word by applying a comparison between the number of occurrences of a word in a document (or, in this case, a review that is represented as a sentence and sometimes can be a paragraph), and the number of documents in which it has occurred.

TF represents the term frequency, which is the occurrence of the word in the review. IDF stands for inverse document frequency, the number of documents a word has appeared in.

$$tf - idf(t) = tf(t, d) * idf(t) \qquad (1)$$

In summary, TF-IDF captures only the terms that are frequent in a review but not across all the reviews.

Upon using these methods, a result of 15,000 reviews with a corpus of 16,000 words has been achieved. In this text analysis case, the words are considered features that grade each sentence.

The large number of features has led us to consider feature selection methods to choose an optimal number of features without impacting the model.

4.4. Feature Selection

Feature selection is a crucial step to get interesting results. As detailed in the previous section, the number of features outgrows the number of reviews. This implies that the model training will require a large computational capacity and probably have fewer results because a big part of the features is generic and does not influence the prediction of the overall appreciation of the customer.

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \qquad (2)$$

ni_j : node importance j;
w_j : node: j Segment of statements;
C_j : J node impurity;
Left and right (j): separating a node j results in a child node.

Each feature's value is computed by taking into account the equation (3), where is the relevance of fi_i feature is:

$$fi_i = \frac{\sum_{j: node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k} \qquad (3)$$

To normalize this result, it can be split on the totality of the important values of the features:

$$normfi_j = \frac{fi_j}{\sum_{j \in all\ features} fi_j} \qquad (4)$$

The Random Forest grade's final feature relevance is determined by all the trees' mean. The significance worth of each tree is then divided by the total number T:

$$RFfi_i = \frac{\sum_{j \in all\ trees} norm fi_{ij}}{T} \qquad (5)$$

Norm (RFfii): each feature in the tree j is normalized;

Based on the feature selection results in the paper [19], Light Gradient Boosted Machine LightGBM represents one of the lightest and most effective methods for best features.

Figure 8 outlays the manipulation of selecting an increasing number of word frequency datasets and fed the reviews to the models selected in this research paper, CART, as a decision three-based model, MLP, and both are neural network models.

Figure 8 shows that the results begin to stagnate, starting at a corpus of 1,000 features.

4.5. Evaluation Metrics

Evaluation metrics are important to measure a model's performance because they provide objective and quantitative measures of how well the model is performing. These metrics help assess accuracy, precision, recall, F1 score and other relevant measures, allowing researchers to compare models, identify weaknesses, and make informed decisions for model improvement. In addition to these metrics, other expressive metrics have been used, like Sensitivity, Specificity and Cohen's Kappa, which are essentially based on the confusion matrix [19].

The majority of these metrics rely on True positives (TP), positive values that were correctly predicted; False positives (FP), values that were falsely predicted as positive; True Negatives (TN), negative values that were correctly predicted; and False Negative (FN), values that were wrongly predicted negative.

- Accuracy: Identifies the percentage of correct predictions that the model could perform.[20]

$$\frac{TP+TN}{Total} \qquad (6)$$

- Sensitivity: Identifies the rate of correct positive results that were predicted.

$$\frac{TP}{TP+FN} \qquad (7)$$

- Specificity: Unlike the Sensitivity metric, specificity identifies the percentage only of correct negative predictions.

$$\frac{TN}{TN+FP} \qquad (8)$$

- Compute Cohen's kappa [21]: a statistic that measures inter-annotator agreement.

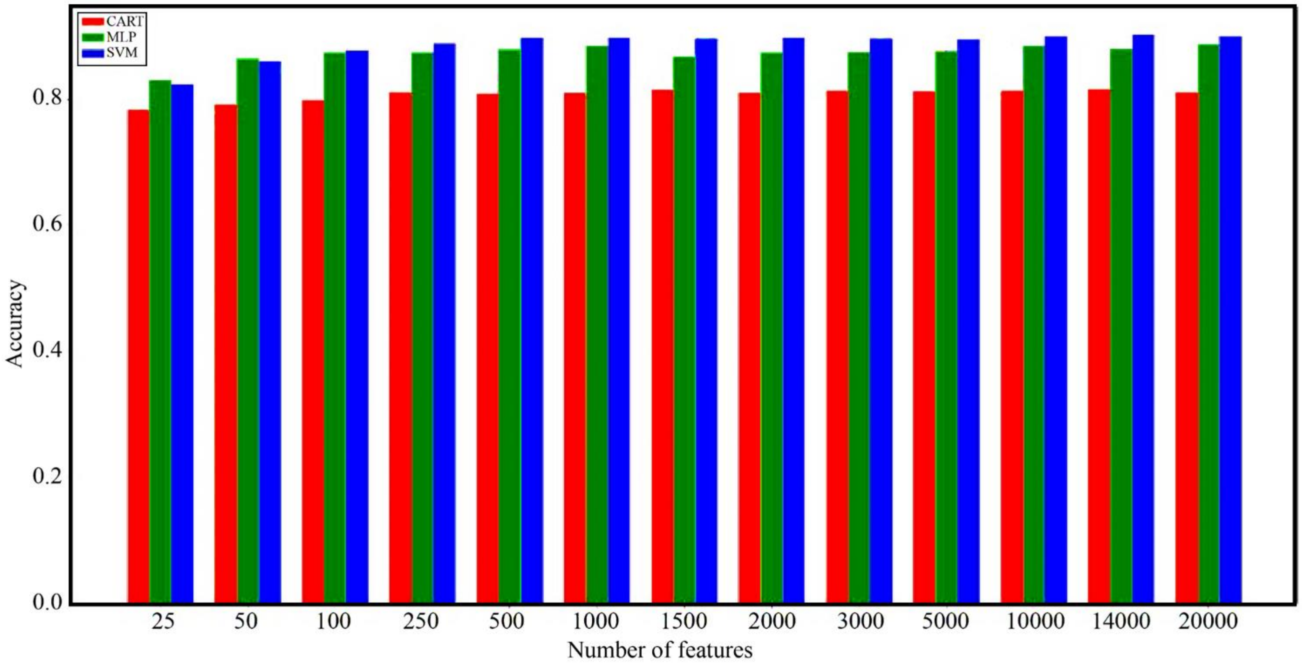

Fig. 8 Training models based on different numbers of features selected using LightGBM

This function computes a score expressing the agreement level between two annotators on a classification problem. It is annotated using the symbol $\boldsymbol{\kappa}$, it has arranged -1 to 1, and it depends on the probability of agreement minus the probability of disagreement.

$$\frac{(p_o - p_e)}{(1 - p_e)} \quad (9)$$

4.6. Modeling

After gaining more insight into this data, pre-processing it, and labeling it, the next step focuses on preparing the text data as features to be fed to the Machine Learning Algorithms used in this research paper.

To do so, the first step was to split the dataset into a 90% training subset and 10% for the test phase. In the continuation of this work, it will assess the impact of three Machine Learning algorithms: Decision Tree [22], Support Vector Machine (SVM) [23] and Multi-Layer Perceptron (MLP) [24].

- Decision Tree DT is a supervised learning algorithm that uses features or attributes of data to create a tree-like structure that predicts outcomes or classifies new data based on previous patterns. Decision trees are similar to human-level thinking, allowing users to interpret the data easily. The goal of this method is to construct a tree that can process all the outcomes at every leaf. DT algorithm is used as an Ensemble Learning method, proving its efficiency [7, 25, 26], which results in an improved sentiment analysis performance.
- Support Vector Machine SVM [23] results are represented by a line that separates the different classes from each other by maximizing the margin and space between the data points. This representation helps classify new data based on its position within the line. The choice of this model is based on its numerous uses in different research studies [7, 8, 9] for improving sentiment analysis accuracy.
- The Multi-Layer Perceptron (MLP) model, explained in a simple manner, is a network of interconnected nodes that work together to process and transform data. It can learn data patterns based on information passed between multiple layers of patterns. The choice of the model is based on solid results obtained in [19, 27] for churn prediction and sentiment analysis.

5. Experimental Results

After the feature selection process, the best results were recorded using the best 1,000 LightGBM features.

Models that were used are:

- SVM with Linear kernel with a decision function shape ‘one-vs-one’ to train multi-class model and default regularization method.
- MLP with stochastic gradient descent solver for weight optimization with an adaptive learning rate of 0.3, max iteration of 250 and 2 hidden layers with 15 hidden units.
- Finally, CART, a Decision Tree Classifier with default values.

This research has achieved respectable results in comparison to other studies in the field, resulting from the deep analysis of all the methods cited before. Every tool used in this research was compared with multiple tools used for the same purpose in multiple related works. Only the best suitable methods were picked for the final results.

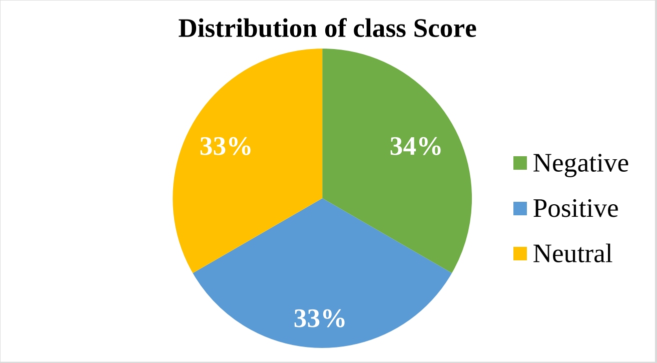

Fig. 9 Visualization of the distribution of the class 'score' after pre-processing and random sampling

5.1. Positive, Negative and Neutral Classification

In Figure 9, the dataset was sampled randomly to 9,000 reviews from each class, with the goal of getting acceptable and interesting results when generalized to the rest of the dataset.

The 3 sentiment classes used in this section resulted from the ML tool VADER for classification. Neutral data usually influences the accuracy of the models.

Table 4. Classification results using 1,000 best features

Name	Accuracy	Sensitivity	Specificity	Cohen'sKappa
CART	0.84	0.84	0.84	0.68
SVM	0.94	0.90	0.97	0.85
MLP	0.92	0.90	0.94	0.85

In Table 4, the SVM model scored the best results with 94% accuracy and all other metrics. It also shows that SVM and MLP results were approximately the same, while CART scored the lowest results.

5.2. 'Overall' Rating Multi-Classification

An overall column represents the rating given to a product by a client. This column has only 5 distinct classes, which has allowed us to treat it as a classification problem, as seen in Figure 10. For this experiment, 5,000 reviews per class were selected, with a sum of 25,000 reviews.

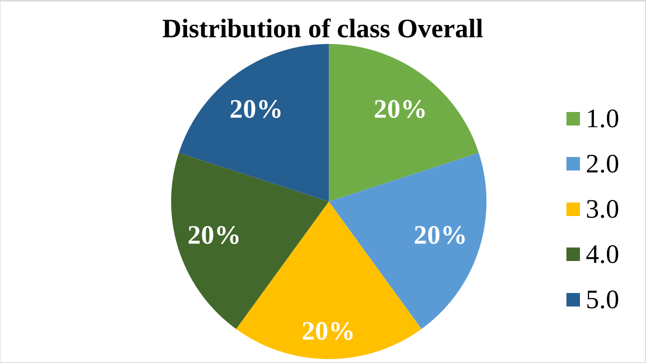

Fig. 10 Visualization of the distribution of the class 'overall' after pre-processing and random sampling

Table 5. Overall classification results using 1,000 best features

Name	Accuracy	Sensitivity	Specificity	Cohen'sKappa
CART	0.72	0.63	0.77	0.41
SVM	0.82	0.71	0.90	0.63
MLP	0.80	0.71	0.86	0.58

Table 5 shows that the accuracy has dropped significantly for the three algorithms. In this case, SVM scored the best results with 82% accuracy, with approximately the same results as MLP.

The CART algorithm scored the least results in both classification scenarios, with 72% for the "5 class" classification task.

With respect to this dataset, [28] this paper has worked on a model that can extract insights from customer reviews and has achieved 91% accuracy in binary classification using the KeyPhrase embedding model against 48% accuracy for multi-class. [29] used the LSTM model and had achieved 90% of accuracy. The paper [30] worked on an enhanced feature selection based on ANOVA and an extended genetic algorithm for online customer review analysis with an accuracy of 78%.

The results above are obtained with a standard performance computer: a system with Core i7, 32 GB of RAM and an SSD-based drive, while [28] worked with a virtual environment with 16-core AMD EPYC 7452 processor, 128 GB RAM and 400 GB of storage and for operations with transformer-based neural networks a GPU-optimized virtual environment with 12-core Intel Xeon E5-2690 processor, 224 GB RAM, 1474 GB (SSD) cache, and two NVIDIA Tesla P100 (32GB) GPUs.

This research analyzed different models for the classification task with various data, which provides a comprehensive understanding of their features and weaknesses for sentiment analysis. By going through different comparison levels across the paper to determine the most effective flow for the dataset at hand, this research has pushed the limits of existing methods and provided valuable insight into their effectiveness.

This research paper intends to help in developing the field's knowledge and help in the process of developing accurate and reliable tools for the sentiment analysis task.

6. Conclusion

This paper summarizes the work performed on a subset of reviews related to the Amazon dataset. It started by pre-processing data, finding the adequate method to label the data, and then the feature extraction and selection.

In each step, several methods have been tried to verify the pertinence of the results before proceeding to the next step.

After the pre-processing pipeline, the dataset was labeled using the VADER text analysis tool. Then, it passed to the feature extraction task using TF-IDF methods and afterwards addressed the dimensionality problem using the LightGBM algorithm. This step has allowed a clarity of choice to select the most relevant features to reduce feature vector dimension and computational cost while ensuring interesting results against related work that deals with the same dataset—at the same time, using standard computers for solution implementation and working with reduced datasets to cope with the heavy calculations.

The results of this paper help the process of better decision-making for different actors in the customer satisfaction journey. The approach followed here can deal with the analysis of text data as part of the task of Natural Language Processing.

For future research, the main objective is to work further with deep learning methods and compare their results to these methods. Using methods like LSTM, BERT, or the two combined might give more interesting insights into the dataset in use. Additionally, the goal is to generalize these results on a dataset that is still being collected based on Arabic text data to classify and extract insight from text messages which can be of great support to decision-makers when it comes to understanding customers' profiles and behaviors in Morocco or the northwestern population of Africa in general.

References

[1] Marjane Holding-Leader in Mass Distribution, Marjane. [Online]. Available: https://www.marjane.ma/corporate/corporate

[2] Haleem, Abid, et al. "Artificial Intelligence (AI) Applications for Marketing: A Literature-based Study." *International Journal of Intelligent Networks*, vol. 3, pp. 119-132, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Afifah Mohd Asri, Siti Rohaidah Ahmad, and Nurhafizah Moziyana Mohd Yusop, “Feature Selection using Particle Swarm Optimization for Sentiment Analysis of Drug Reviews,” *International Journal of Advanced Computer Science and Applications,* vol. 14, no. 5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Nikhat Parveen et al., “Twitter Sentiment Analysis using Hybrid Gated Attention Recurrent Network,” *Journal of Big Data,* vol. 10, no. 1, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Tarun Jain et al., “Sentiment Analysis on COVID-19 Vaccine Tweets using Machine Learning and Deep Learning Algorithms,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Mouaad Errami et al., “Sentiment Analysis on Moroccan Dialect based on ML and Social Media Content Detection,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Maria Yousef, and Abdulla ALali, “Analysis and Evaluation of Two Feature Selection Algorithms in Improving the Performance of the Sentiment Analysis Model of Arabic Tweets,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 705-711, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Reza Maulana et al., “Improved Accuracy of Sentiment Analysis Movie Review Using Support Vector Machine Based Information Gain,” *Journal of Physics: Conference Series*, vol. 1641, no. 1, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] KaiSiang Chong, and Nathar Shah, “Comparison of Naive Bayes and SVM Classification in Grid-Search Hyperparameter Tuned and Non-Hyperparameter Tuned Healthcare Stock Market Sentiment Analysis,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Jianmo Ni, Jiacheng Li, and Julian McAuley, “Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects,” *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, pp. 188–197, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Amazon Review Data, 2018. [Online]. Available: https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/

[12] English, English Spacy Models Documentation, Spacy. [Online]. Available: https://spacy.io/models/en#en_core_web_sm

[13] Mueller, Wordcloud: A Little Word Cloud Generator, Github. [Online]. Available: https://github.com/amueller/word_cloud

[14] Vijaylakshmi Sajwan et al., “Sentiment Analysis of Twitter Data Regarding the Agnipath Scheme of the Defense Forces,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 3, pp. 1643–1650, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] C.J. Hutto, and Eric Gilbert, “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216-225, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[16] Ditiman Hazarika et al., “Sentiment Analysis on Twitter by Using TextBlob for Natural Language Processing,” *Annals of Computer Science and Information Systems*, vol. 24, pp. 63-67, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Aigerim Toktarova et al., “Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Stephen Robertson, “Understanding Inverse Document Frequency: On Theoretical Arguments for IDF,” *Journal of Documentation*, vol. 60, no. 5, pp. 503-520, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[19] Khattak, Asad, et al. "Customer Churn Prediction Using Composite Deep Learning Technique." *Scientific Reports*, vol. 13, no. 1, pp. 1-17, 2023. [CrossRef] [Google Scholars] [Publisher Links]

[20] Sklearn.Metrics.Accuracy_Score, Scikit-Learn. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

[21] Jacob Cohen, “A Coefficient of Agreement for Nominal Scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37-46, 1960. [CrossRef] [Google Scholar] [Publisher Link]

[22] Arno De Caigny, Kristof Coussement, and Koen W. De Bock, “A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees,” *European Journal of Operational Research*, vol. 269, no. 2, pp. 760-772, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[23] Theodoros Evgeniou, and Massimiliano Pontil, “Support Vector Machines: Theory and Applications,” *Advanced Course on Artificial Intelligence*, vol. 2049, pp. 249-257, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[24] Marius-Constantin Popescu et al., “Multilayer Perceptron and Neural Networks,” *WSEAS Transactions on Circuits and Systems*, vol. 8, no. 7, pp. 579-588, 2009. [Google Scholar] [Publisher Link]

[25] Harsh H. Patel, and Purvi Prajapati, “Study and Analysis of Decision Tree Based Classification Algorithms,” *International Journal of Computer Sciences and Engineering*, vol. 6, no. 10, pp. 74-78, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[26] Masoud Amini Motlagh, Hadi Shahriar Shahhoseini, and Nina Fatehi, “A Reliable Sentiment Analysis for Classification of Tweets in Social Networks,” *Social Network Analysis and Mining*, vol. 13, no. 1, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] D. Elangovan, and V. Subedha, “Firefly with Levy Based Feature Selection with Multilayer Perceptron for Sentiment Analysis,” *Journal of Advances in Information Technology*, vol. 14, no. 2, pp. 342-349, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[28] Robert Lakatos et al., “A Cloud-Based Machine Learning Pipeline for the Efficient Extraction of Insights from Customer Reviews,” *arXiv*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[29] Naveen Kumar Gondhi et al., “Efficient Long Short-Term Memory-Based Sentiment Analysis of E-Commerce Reviews,” *Computational Intelligence and Neuroscience,* vol. 2022, pp. 1-19, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[30] Gyananjaya Tripathy, and Aakanksha Sharaff, “AEGA: Enhanced Feature Selection Based on ANOVA and Extended Genetic Algorithm for Online Customer Review Analysis,” *Journal of Supercomputing,* vol. 79, pp. 13180-13209, 2023. [CrossRef] [Google Scholar] [Publisher Link]