

Product reviews analysis to extract sentimental insights with class confidence rate using self-organizing map neural network

Sara Ahsain, Yasyn Elyusufi, M'hamed Ait Kbir

Intelligent Automation and BioMedGenomics Laboratory (IABL), Technical Sciences and Medical Sciences (STSM) Doctoral Center, Abdelmalek Essaadi University, Tétouan, Morocco

Article Info

Article history:

Received Apr 22, 2024

Revised Aug 14, 2024

Accepted Aug 20, 2024

Keywords:

Competitive learning

Customer behavior

Personalized marketing

Product recommendations

Self-organizing map

ABSTRACT

Customer data analysis helps companies to understand customer intentions and behaviors better. This study introduces an analysis of product reviews to help managers adopt a more efficient strategy to extract valuable knowledge and help detect segment of customers that need a special attention and products that need improvement or with the most impact. The used dataset is a set of Amazon reviews divided into multiple categories; each review has a target column called 'overall' that takes a value between 1 and 5 (customer's satisfaction). Based on the 'overall' column, multiple labeling methods have been used and compared to get a binary target variable, positive or negative, that affects a class to a review. This dataset contains more than one million reviews and can give companies great insight into products' quality and customers' retention. This work has materialized by using customer segmentation and competitive learning with self-organizing map (SOM) Model and adopting a new approach to explore the generated network/map, it is based on clustering and map nodes labelling using a majority voting process. The results show that the proposed dual approach combining the prior knowledge, related to supervised learning, and the competitive learning abilities enhances the SOM model's capabilities.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Ahsain Sara

Intelligent Automation and BioMedGenomics Laboratory (IABL), Technical Sciences and Medical Sciences

(STSM) Doctoral Center, Abdelmalek Essaadi University

Tétouan, Morocco

Email: sara.ahsain@etu.uae.ac.ma

1. INTRODUCTION

Customer analysis is becoming an increasingly common source of information. Also, strengthening long-term customer relationships through individualized e-commerce offers will depend on the company's ability to use customer data to develop and control customer interactions. In the same context, the revolution of predictive analytics consists of exploiting the mass of data newly available to establish predictive hypotheses from previous customer behavior and interactions. Developing robust sentiment analysis models tailored to product reviews is crucial to enhancing customer loyalty and product quality. These models are invaluable tools for understanding customer sentiments and improving product quality. Various statistical and data mining techniques have been extensively researched in literature to craft predictive models to leverage insights extracted from customer feedback and identify both strengths and areas that require enhancement to ensure deeper customer engagement and to augment overall product satisfaction. The machine learning and artificial intelligence landscape is vast, the same as that of the growing field of customer analysis.

As businesses explore the possibilities of integrating these technologies, leveraging models to strengthen customer relationships and provide personalized e-commerce experiences becomes crucial.

E-commerce landscape keeps growing by using advanced artificial intelligence (AI) techniques such as self-organizing map (SOM) models for sentiment analysis represents a critical frontier in understanding and responding to customer needs. This unsupervised neural network model is a promising solution for a deeper understanding of customer reviews by facilitating the discovery of patterns and correlations that might uncover exciting insights in the data.

Supervised learning models, like classification, regression, or ensemble learning methods, have become established tools for leveraging labeled data to make predictions or uncover relationships. Supervised learning offers a range of benefits due to its versatility, as it includes algorithms designed to tackle different tasks (binary classification/multi-class classification). This type of model, although practical, heavily depends on having labeled training data available. The success of this approach relies on the quality and representativeness of the data used for training.

On the other hand, unsupervised learning techniques, such as clustering and dimensionality reduction, offer a range of approaches that help identify patterns, structures, or relationships from data. This quality makes it exceptionally skilled in revealing concealed insights, recognizing clusters, and simplifying data. This algorithm finds application in domains such as anomaly detection, pattern recognition, and exploratory data analysis. Although unsupervised algorithms have benefits, evaluating their performance can be challenging because they require numerical input and have no clear performance metrics. Furthermore, interpreting the results can be complex since these algorithms identify patterns without guidance.

Models based on artificial neural networks (ANN) have also seen great success; for instance, a multilayer perceptron (MLP) is a supervised learning algorithm that is a feedforward type of neural network, which means that the information travels in one direction from the input layer through several hidden layers to get to the output layer without any cycles or loops. MLPs are mostly known for their capability to learn from complex mappings and non-linear relationships in data. However, this model has certain drawbacks, such as overfitting or underfitting, because it can capture noise in data, which might require regularization and careful consideration of parameters. It also requires a large amount of labeled data, making it a bit hard to use in all domains because the data may not always be readily available. Furthermore, MLP models can be computationally expensive, which would demand substantial resources [1], [2].

A self-organizing map (SOM) is a widely used model in unsupervised learning tasks. The data can be explored to extract regularities owing to the competitive learning algorithm behind the model. In this context, Tsai suggests using two hybrid models that combine two distinct neural network methods to predict churn [3]. These techniques are back-propagation neural networks (ANN) and self-organizing maps (SOM); the first model employs a method to reduce data by filtering out training data that's not representative, and then the result is fed to the predicting model using the second technique. The results show that the combined ANN hybrid model outperforms the other methods. Cuadros *et al.* proposes in [4] a segmentation framework, where the customer lifetime value, customer loyalty calculation, and client segment building are done using a self-organized map. In the same context Asmara *et al.* used in [5] SOM model to analyze interactions among bird diversity, spatial distribution, and land use types in the Kenyir landscape in Malaysia.

In our work, we aim to study possibilities provided by unsupervised models, especially the SOM model, applied to product review data to gain insight into products that the customers endorse. This work is part of a series of research conducted by our team within the context of the development of customer profiling via their account activity by answering specific questions to build an effective e-commerce platform (Ex: whether it is a fake account, detecting their preferences via sentiment analysis; recommending products; detecting churn). In this very context a new approach to customer product appreciation and sentimental insights is proposed in order to extract sentimental insights using SOM neural network. In the first section of this work, we present related works found in the literature. The second section presents our approach to classifying customers into two categories and predicting positive from negative reviews. In the third section, we present and discuss the results obtained. The conclusion comes in the last section.

2. SELF ORGANIZING MAP USE IN LITERATURE

A SOM is an artificial neural network algorithm for clustering and visualizing multi-dimensional data in lower-dimensional space. SOM is a type of unsupervised learning algorithm. It uses a grid of neurons or nodes arranged in two dimensions, where each node represents a prototype or a cluster corresponding to a region of the input space [6]. The SOM is trained by iteratively adjusting the weights of these nodes to match the input data. The adjustment is made using a technique called competitive learning, without bestowing labels, where the node with the closest weight to the input data is chosen as the winner, and its weights are updated to be even closer to the input [7].

SOMs have been widely used in many fields, such as pattern recognition [8], [9], data mining [10], image processing [11], and data visualization [12], [13]. They are beneficial for dimensionality reduction and exploratory data analysis, as they can provide a low-dimensional representation of complex datasets, making

it easier to understand and interpret the underlying patterns and relationships in the data. Neisari *et al.* [14] used a mix of unsupervised learning (SOM) and convolutional neural networks (CNN) to classify reviews in order to detect spam reviews, which resulted in 0.87% accuracy by combining the two models. SOM has also been applied to temperature and precipitation patterns over China in this study [15] to compare between 2021 and 2022 patterns. Dalal *et al.* [16] has used SOM more specifically, as well as an adaptive moving self-organizing map and fuzzy k-mean clustering, for brain tumor segmentation, focusing mainly on extracting the tumor regions.

Zhengtian *et al.* [17] SOM clustering abilities were used to select relevant features from data before applying different models such as (K-NN, SVM, and decision trees) then, it compared the results to other feature selection methods and proved a significant increase in the accuracy of the models. Angulo-Saucedo *et al.* [18] has also used variants of SOM in structural health monitoring to classify damages. Additionally, Yuan *et al.* [19] has used the SOM algorithm to predict the patient outcome and response to therapy by applying cell segmentation, systematic classification, and in silico cell labeling on an image database of breast cancer.

Zhengtian *et al.* [17] used SOM clustering methods on a binary classification problem for feature selection and then applied different models, such as decision trees, resulting in 0.75% accuracy. However, it achieved a higher accuracy when applying SVM 0.85%. On the other hand, the paper [18] uses variants of the SOM model called counter propagation artificial neural network (CPANN), supervised Kohonen (SKN), and X–Y fused Kohonen (XYF), which has resulted in an overall of 0.74% accuracy using SKN and 0.73% using SYF. Additionally, this paper [19] has used an SOM model of 49 nodes with 5000 iterations. Based on only the top five features, it achieved 0.76% precision, 0.79% recall, 0.78% F1, and 0.70% AUC.

In the present paper, we present a SOM-based model to predict customer profiles that are more likely to have a positive feeling towards a product than those who do not. The dataset used in this paper is a collection of Amazon reviews that has been collected since 1996 up to 2018 and is divided into multiple categories. The categories used for this experiment combine three categories: 'Magazines and Subscriptions', 'Software', and 'Beauty' with 12 features. Since the focus is on the natural language processing aspect of the dataset, we only kept the text data that was needed. Before evaluating models, thorough steps were used, such as data pre-processing, robust scaling, and feature selection.

Like many machine learning algorithms, SOM-based models also require parameter tuning (learning rates, neighborhood functions), which can be essential to the final results. Thus, systematic experimentation and cross-validation are crucial for determining optimal parameters. Another challenge is the subjective interpretation of SOM maps, meaning that relevant clusters or patterns can require an expert or risk being wrongly identified or ignored. Quantitative measures and validation techniques should be used to enhance the objectivity of map interpretation.

3. THE PROPOSED APPROACH FOR CUSTOMER RETENTION AND PRODUCT APPRECIATION

This work is a continuation of the research work carried out in [20] this stage we are interested to reviews binary classification. The impact of using the SOM model will be assessed to help the financial organization make decisions about the suitable strategy to discern customer sentiments and refine the overall products quality based on sentimental insights extracted via review analysis. The following sections will be devoted to the preprocessing, modeling and evaluation phases. Some of the key contributions of this research are:

- a. Extensive dataset use: The dataset includes Amazon reviews divided into categories, focusing on specific ones like 'Magazines and Subscriptions,' 'Software,' and 'Beauty.' This comprehensive dataset provides a detailed analysis of customer satisfaction and sentiments;
- b. Combining the prior knowledge and the clustering capabilities: The research employs a SOM model to segment and predict customer profiles based on their sentiments toward products, aiding in identifying customer clusters and their associated sentiments. This can be reached by combining the prior knowledge, reviews are labelled, and the competitive learning power provided by the model. The SOM model is first trained in an unsupervised manner for initial clustering, followed by supervised labeling and classification using a majority voting process. This dual approach enhances the model's predictive capabilities.
- c. Improved customer and product insight: the map generated by the SOM model is explored to detect nodes with some particular profiles. This contributes on a better understanding customers' intentions and behaviors, allows personalized product recommendations and targeted marketing strategies and reduces the need for generic marketing campaigns. This offers also valuable insights for managers to identify which products need improvement and which have the most significant impact, thereby aiding in product quality enhancement.

- d. Objective interpretation of SOM maps: To address the subjectivity in interpreting SOM maps, the study suggests using quantitative measures and validation techniques, ensuring accurate identification of relevant clusters.
- e. Thorough model evaluation and fine-tuning: The research involves extensive preprocessing, data scaling, feature selection, and parameter initialization to ensure a good and appropriate implementation of the SOM model. In fact, detailed fine-tuning is needed, which is accomplished by systematic experimentation and cross-validation to optimize parameters such as learning rates and neighborhood functions.

The findings of this study can help financial organizations make informed decisions about customer sentiment analysis and product quality improvement, leveraging sentimental insights derived from customer reviews. To provide a clear understanding of the research process and its key components, the diagram in Figure 1 illustrates the methodology employed in this study. The model focuses on processing the data and make it ready to be used by the different algorithms while also ensuring a high effectiveness.

Figure 1 outlines the process used to classify sentiments in Amazon product reviews through machine learning. It starts with data collection, explicitly gathering an Amazon review dataset. This data then goes through a preprocessing stage, which includes expanding contractions, converting text to lowercase, removing digits and punctuation, eliminating stop words, and lemmatizing the text to its base forms. After preprocessing, the data is labeled with VADER, a pre-trained machine-learning tool known for sentiment analysis.

Next, term frequency-inverse document frequency (TF-IDF) was used to transform text data into numerical features for feature extraction. These features are selected using the light gradient boosting machine (LGBM) to pinpoint the most relevant ones for classification. Finally, the data is classified using machine learning models, such as a self-organizing map, and compared to the classification and regression trees (CART) decision tree, which categorizes the reviews as positive or negative, thereby identifying customer sentiments towards the products. This research provides a structured method for analyzing customer reviews using the SOM model, enabling better customer segmentation, personalized recommendations, and improvements in product quality.

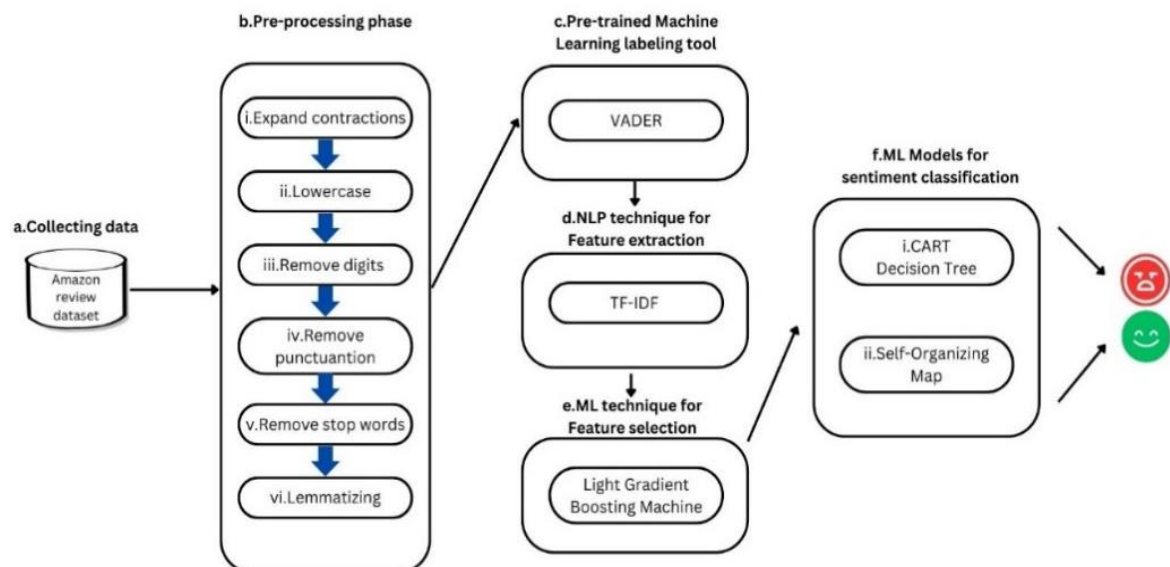


Figure 1. The methodology diagram of this research

3.1. Dataset

Large datasets can help build models that achieve better accuracy and other performances due to their ability to provide various representative samples. This allows the model to collect a broader range of patterns. Also learning from examples helps reduce overfitting, as it minimizes memorizing specific instances or noise, and then improve model robustness against outliers and variations, and facilitate rare event detection. The dataset used to train the model contains product reviews and metadata from Amazon, including 459 436 Software reviews, 89 689 Magazine subscriptions, and 371 345 Beauty reviews spanning from 1996 to 2018 [21].

Table 1 shows the different distributions of the Amazon dataset, including the number of reviews it with holds. To find the best-reviewed products and the most willing customers to re-purchase and help the decision-making process be aware of the precautions the financial organization should undertake, we decided to apply the SOM model and evaluate the performance. Table 1 shows the features used.

This dataset has 12 features, with different data types described in detail in Table 2. This research paper works on a natural language processing task, which leads to a focus only on text data. The final dataset mixes three sub-datasets belonging to three categories in order to work on the classification of reviews by rating.

Table 1. Amazon reviews distribution per category

Dataset	Number of reviews
All reviews	233.1 million
Movies and TV	8 765 568
Software	459 436
Books	51 311 621
Cell phones and accessories	10 063 255
Digital Music	1 584 082
Magazine subscriptions	89 689
Beauty	371 345
Grocery and gourmet food	5 074 160
Home and Kitchen	21 928 568

Table 2. Dataset features

Feature name	Signification	Type
Overall	Rating of the product	Float
verified	True if the purchase was verified	Boolean
Review time	Raw date time of the review	Date
Reviewer ID	The ID of the reviewer	String
Asin	ID of the product	String
Style	Dictionary of the product metadata, e.g., "Format" is "Hardcover"	Array
Reviewer name	Name of the reviewer	String
Review text	Text of the review	String
Summary	Summary of the review	String
Unix review time	Unix time of the review	Unix time
Vote	Helpful votes of reviews by other reviewers	Number
Image	Attached image to the review	Array

Data scientists need an appropriate dataset to perform well to gain insight into customers' behavior. This dataset should contain sufficient samples from different document categories. It ensures the usage of real-world reviews categorized into different product categories after purchase.

3.2. Data pre-processing

Data preprocessing and cleaning are essential steps for a dataset based on text analysis. Irrelevant information like HTML tags, punctuation, or special characters should be removed. Addressing standard text preprocessing tasks such as removing stop words, normalizing text (lowercasing, stemming, lemmatization), and handling spelling mistakes or abbreviations can improve the dataset's quality.

To train machine learning models accurately, the data must be cleaned and preprocessed. Irrelevant data, such as null and poorly formatted data, special characters, punctuation, should be discarded. Additionally, other steps were used, such as lowercasing, lemmatization, and stemming. Changing the effectiveness of one or more of these steps can significantly increase the model's accuracy.

First, the flow was initialized by formatting the dataset's attributes to fit the paper's needs. It combines the review text and the title into one column, and then we added the product category column based on which category the review belongs to. Before performing exploratory data analysis (EDA) and converting the dataset to a format that is adequate for models, the following transformation pipeline was adopted: i) cleaning and feature engineering, ii) cleaning stop words, iii) removing nulls, iv) removing punctuation; label encoding, and v) lemmatizing.

An EDA has been performed to comprehensively understand the dataset and assess data quality. It aids in feature selection, validates assumptions, informs model selection and design, and helps detect outliers. It allows more profound insights into the dataset's characteristics and relationships. Figure 2 shows that the original dataset was divided into 5 rating values; each target value can be considered as a class; it is also

unevenly distributed between the classes and reviews with a positive meaning that is categorized as the class ‘positive’ represents the majority class with over than 75% of the dataset.

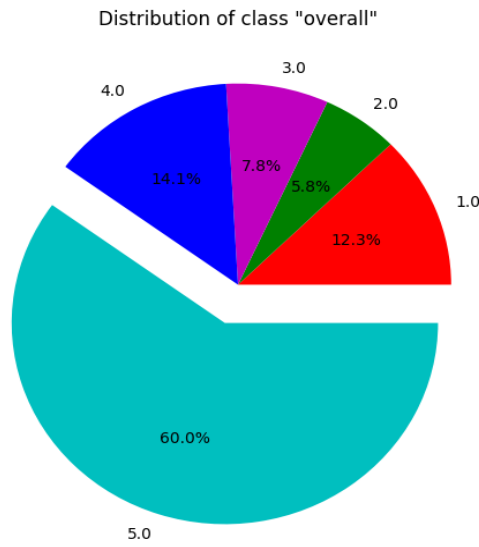


Figure 2. Distribution of class ‘overall’ in the dataset

This paper covers diverse natural language processing techniques, focusing on classifying reviews by rating. To do that, two classes (positive/ negative) were combined (9,500 random reviews were chosen from each class). Here is an example of a review before and after applying the cleaning pipeline:

- My husband wanted to reading about the Negro Baseball and this a great addition to his library\n Our library does not have information so this book is his start. Thank you
- Husband wants to read about negro baseball this good addition to library does not have information books start thank you

Some additional examples were added in Table 3 by category after the pre-processing step.

Table 3. Example of reviews after data pre-processing

Text review	Category
Can software company please come out alternative quickbooks so small business owner be release from constant force upgrade and removal of basic functionality	Software
The product be mediocre at good and quickbooks remove basic functionality from the base product on each upgrade and force its customer to pay extra for its horrendous business practice, mediocre product, in business because of lack of competition	
It is a good product but the only thing be it just smell really bad if your look for a smooth removal then this be it good product	Beauty
Of course, I will have to wait a year to see what happen but when I try to cancel auto renewal on another magazine what a hassle will wait and see how this one go auto renewal	Magazine subscription

Figure 3 shows the results of mixing different categories of reviews to get more accurate outcomes. Three dataset categories were combined to diversify the type of clients by choosing reviews that belong to diverse centers of interest and thus capture the nuances of language use through different types of clients.

Word distribution frequency per category has also been visualized to explore this dataset further. The five most frequent words by each category are displayed in Figures 4, 5, and 6. The results show that most top words are related to the category subject. Examples of Software include software, hardware, and computer.

The class ‘overall’ labeled the data and was set as the target column. Data labeling enables the model to differentiate patterns and connections within the data during the last comparison phase. The Amazon review dataset used in this study lacks a default label column. However, a human annotator can infer the general sentiment through the ‘overall’ column. Nevertheless, this research does not solely rely on manual labeling and evaluates well-established machine learning and text mining techniques pre-trained on English vocabulary. Ahsain *et al.* [22], a comparison between manual labeling, Vader tool [23] and TextBlob

tool [24] used on the same dataset, concluded that the results of these three methods were close. For this paper's continuation, the dataset was labeled using the Vader tool and was classified into two classes (negative and positive reviews).

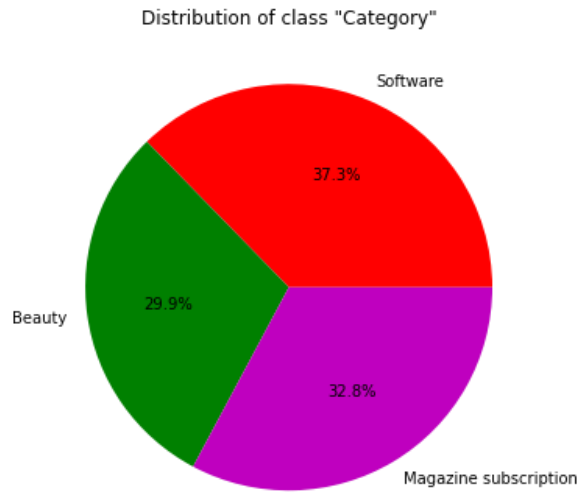


Figure 3. Distribution of class 'category' after mixing datasets

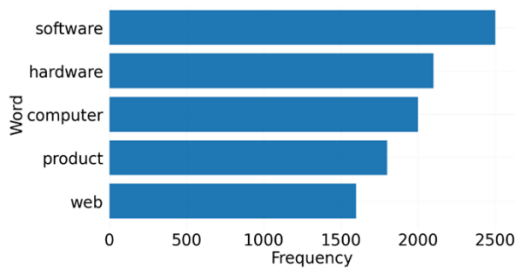


Figure 4. The five most frequent words in the 'software' category

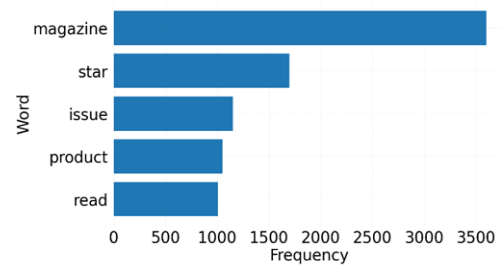


Figure 5. The five frequent words of the category 'magazine subscription'

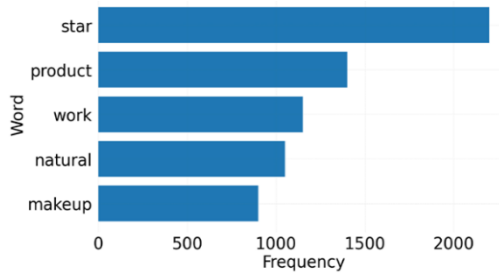


Figure 6. The five top frequent words of the category 'beauty'

3.2.1. Feature extraction

Within natural language processing (NLP), the feature extraction step is a fundamental process that aids machine learning systems in deciphering textual data by converting it into numerical representations. This not only facilitates machine comprehension of textual information but also serves to have a minimalist data representation. This step is important because of the role it plays in mitigating the challenges of high dimensionality, which results in the enhancement of computational efficiency. It also contributes to the generalization capabilities of the framework.

This study used the term frequency-inverse document frequency (TF-IDF) algorithm. TF-IDF is a widely adopted technique for transforming textual content into a structured representation while ensuring the significance of words across the entire document. The paper aims to encapsulate the importance of words within the text and contribute to the overall efficacy of the feature extraction process.

The TF-IDF algorithm, as indicated by the papers [25]–[27] captures a word's uniqueness by comparing its occurrence in a document and its prevalence across various documents. The term “TF” stands for Term Frequency, the number of times a word appears in a document; “IDF” stands for Inverse Document Frequency, the number of documents in which a specific word has appeared [28].

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

TF-IDF selectively captures terms that are frequent within a singular review but not across the entirety of reviews [22].

3.2.2. Feature selection

The feature selection step is crucial to getting 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 will probably have fewer results because a large part of the features is generic and does not influence the prediction of the customer's overall appreciation.

Based on the results of the feature selection step in the paper [20], light gradient boosted machine LightGBM represents one of the lightest and most effective methods for feature selection and classification. The results in the paper [22] a demonstration of LightGBM was applied to this data. The initial corpus consisted of 25,374 features, with a target class consisting of positive or negative reviews. LightGBM was used to select the best number of features. It selects an increasing number of the most relevant features to feed the reviews' corresponding vectors to the models: CART, SVM, and MLP. Figure 7 shows that the results stagnate at a corpus of 100 features.

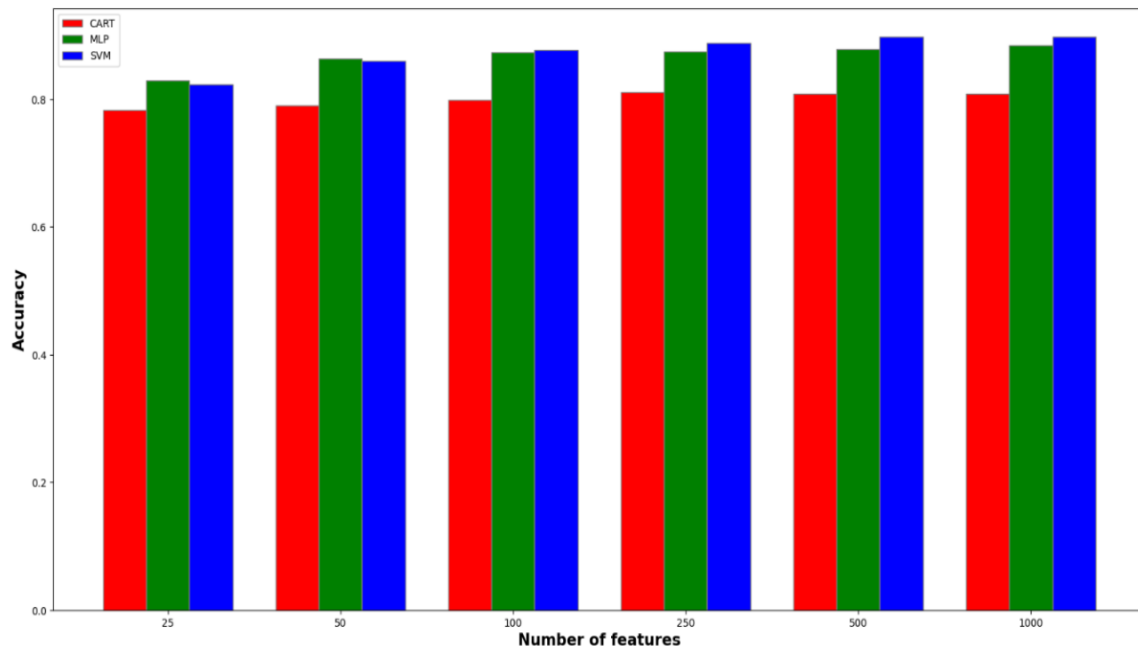


Figure 7. Best features using LightGBM

3.2.3. How self-organizing map model works

SOM model uses an unsupervised learning type called competitive learning, a self-organizing procedure used to match each input vector with a neuron in a 2D grid/map of neurons [29]. The key idea of the SOM model is that nodes located close to each other in the map have weight vectors corresponding to data samples situated close to each other in the data space. Training samples are introduced one at a time to the iterative training algorithm to update weights, initialized randomly, and move the corresponding vectors toward dense regions of the data space.

During the training process, the SOM progressively maps the higher-dimensional input space to a 2D map, preserving the topological properties of the input data. This means that similar or neighboring inputs will be represented by nearby nodes on the map. As a result, SOM can be used to cluster and visualize the relationships among the input data.

Taking an input data of size (m, n) where m is the number of training samples and n is the number of features. We start with by the initialization of the weight's matrix, of size (n, L, C) , where $L \times C$ is the number of nodes/neurons of the map, L and C are the number of lines and columns of the map. Then, iterating over the input data, each training sample updates the winning node/neuron, weight vector with the shortest distance from the training sample, and its neighbors [30].

As the training progresses, the SOM organizes itself so that nearby neurons on the grid respond to similar input vectors. This means that neighboring neurons reflect relationships between input vectors. The neuron that most closely matches the presented input pattern is referred to as the winner neuron or best matching unit. This neuron, along with its neighbors as defined by the algorithm [5], updates its weight vectors based on the SOM learning rules as (2):

$$w_{ij}(t + 1) = w_{ij}(t) + \alpha(t)h_{jc}(t)[x_i(t) - w_{ij}(t)] \quad (2)$$

Here, $w_{ij}(t)$ represents the weight between node i , in the input layer, and node j , in the output layer, at the iteration time t . $\alpha(t)$ is the learning rate, which decreases over time. $h_{jc}(t)$ is the neighborhood function, which defines the size of the neighborhood around the winning node to be updated during the learning process. In the final stage, the weight vectors of all activated neurons are updated accordingly [5].

The map provides a visual representation of the relationships between input vectors. The neighboring neurons on the map correspond to similar inputs, which allows for visualization of clusters in the data. Each class was given a symbol 'x', 'o', or 'rectangle' with different color intensities, grey scale levels. The symbols 'o' and the 'rectangle' were attributed to the class 0 or 1, respectively; the symbol 'x' showed equal voting for both classes (can be called neutral).

4. RESULTS AND DISCUSSION

4.1. Using self-organizing map model

The SOM map size is fixed, specifying the number of rows and columns equal to 15 in the present experiment. Figure 8 shows the neuron locations in the map and the cluster's topology, resulting from the voting process, when all the training samples are ventilated over the map according to the winner neuron computing rule. SOM is first trained in an unsupervised manner to map input data into clusters, grouping them by feature vector similarity. Then, in a supervised manner, labels are allocated to SOM nodes, each cluster with an independent label. The training samples are ventilated over the map according to the winner neuron computing rule used in the training phase. After this, the majority voting is applied to determine the most common label for data samples associated with each node. New data is classified by assigning the nearest node label to it.

The SOM neural network model is implemented using the MiniSOM library of Scikit-learn. The library allows users to explore how often neurons have won the competition. When passing through the dataset samples, fine-tuning is made to give all neurons the chance to participate in contests.

The initialization method sets empirical parameters such as network dimensions: (5×5) , (10×10) , (15×15) , and (20×20) , as well as the number of training cycles (100). The clusters will be mapped to the problem class using the training samples labels. The result is a mesh grid that illustrates the neighborhood relationships between neurons. The algorithm employs a winner-takes-all approach, where the neuron-related weight vector, the closest to the input pattern (the fittest neuron), is updated along with its neighbors. After training, the network can be used for labeling and classification tasks. The learning rate initial value is $\alpha(0)=0.9$, a decay function is used to reduce the learning rate over iterations [31], as illustrated in Figure 8.

$$\alpha(t) = \alpha(0)\left(1 - \frac{t}{\max_iterations+1}\right) \quad (3)$$

The same processes have been run on different dimensions before choosing the optimal generated map to ensure that the map's configuration effectively meets the analytical goals and provides robust and interpretable results. In Figure 9. different dimensions were processed in Figure 9(a) an example of 5×5 dimensions, in Figure 9(b) the dimensions were slightly increased to 10×10 , then 15×15 in Figure 9(c) and finally, the results were also processed on a 20×20 map.

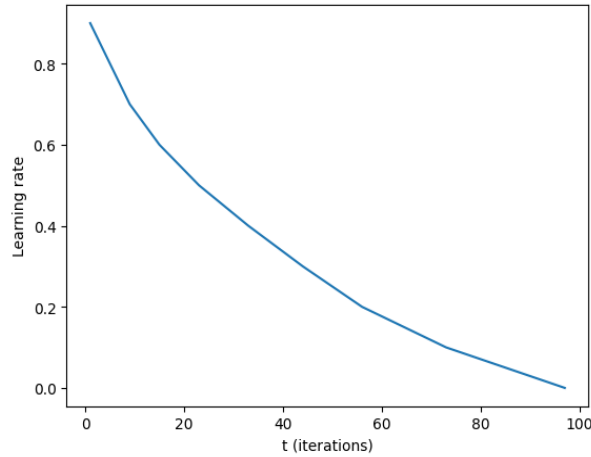


Figure 8. Decay function of the learning rate

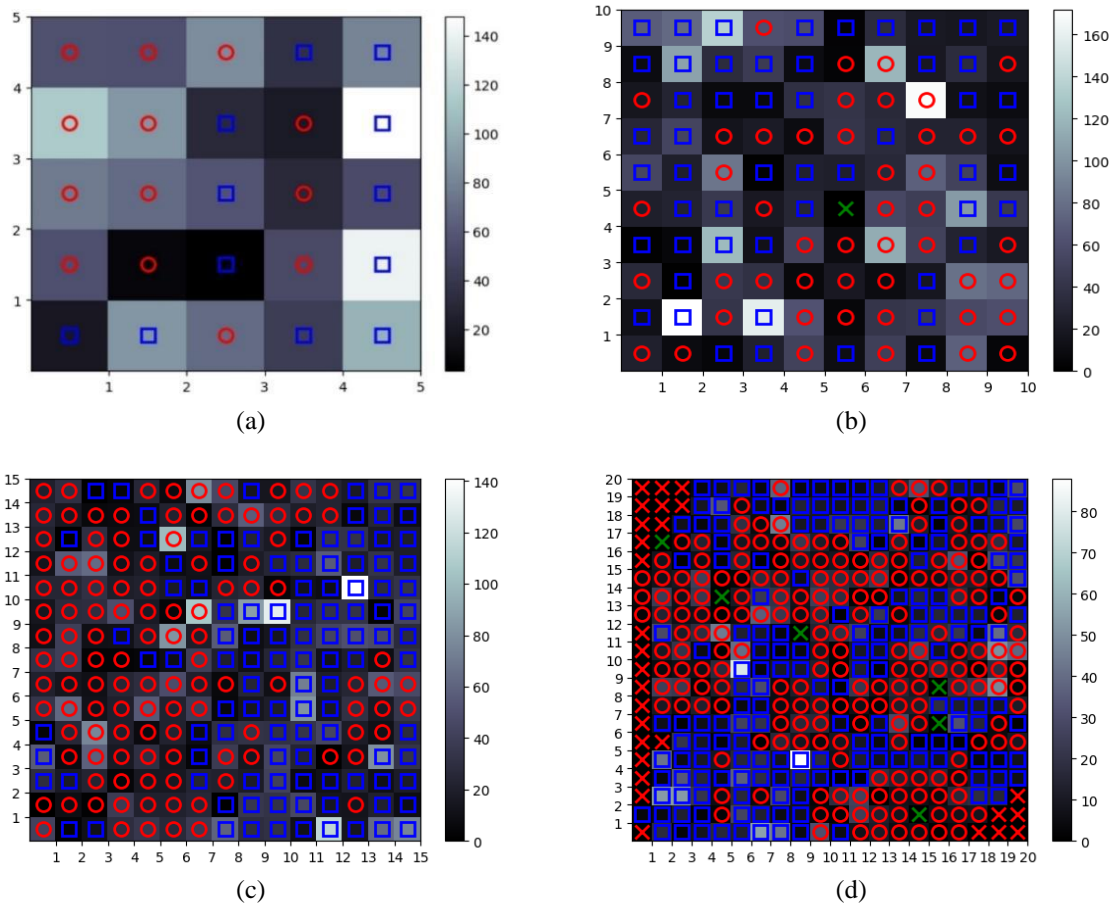


Figure 9. SOM map output with multiple sizes: (a) 5×5, (b) 10×10, (c) 15×15, and (d) 20×20

Choosing the proper dimensionality for a SOM holds great importance as it directly influences the map’s capacity to accurately represent and display complex data relationships. The decision to go with 15-dimension SOM instead of other alternatives, such as 5, 10, or 20, can be influenced by various factors. The dataset size and level of complexity are not adequately captured by smaller maps, such as 5 or 10 dimensions, yet it does not require the additional granularity and interpretation complexity that a 20-dimensional map would provide. The 15-dimension map is also considered to be a more manageable and visually accessible map compared to larger maps. The goal is to spot patterns that could be challenging to

discern if there are too many nodes. In the 15-dimensional map represented in Figure 9(c), it is also noticed that all the clusters have chosen a class successfully, and in no case were the votes equal or null. It also shows a balance in the attribution of classes with high-intensity concentration for both cases, demonstrating the model's ability to classify both intents.

To choose a class for each cluster, votes were compared for both classes. If the two sums are equal, meaning no class was designated, the representation on the map would be '2'. On the other hand, if the first-class votes are higher than the second, it would be represented with '0', and if the second class won, it would be represented with '1'. Nodes that have never won the competition are also defined by '-1'. One of the biggest challenges when using the SOM model is ensuring that all neurons/nodes of the map participate to the completion and reach a map without neurons labeled with '-1' or '2'.

Votes related to each node can be analyzed to compute the nodes' confidence when affected by the available classes. In our experimentation, we used the MiniSOM library which is a NumPy library-based implementation related to self-organizing maps algorithms and their applications. After many simulations that consisted of minimizing the equal or null votes in the map. The map below demonstrates that the model can categorize product reviews into two categories (positive/negative) with an accuracy of 0.83.

Two visual indications were considered to represent the data in Figure 10. The first one is the intensity of the vote, which means that a class is associated with a neuron with a big difference (high confidence). It will have a high grayscale level, and the intensity will diminish gradually to black when a class is associated with a neuron with a low difference (low confidence).

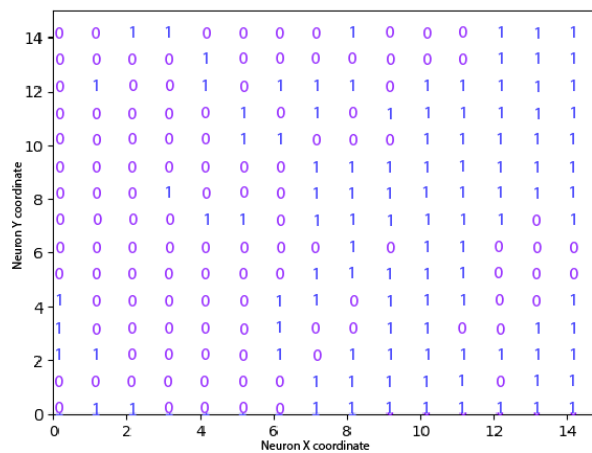


Figure 10. SOM map grid for output neuron reflecting results in Figure 9(c)

The second visual indication represents the associated class label ('X' if both classes' corresponding votes are equal, 'O' if the first class corresponds to the negative class, and 'square' in the other case). The approach adopted here allows the use of clustering methods for labeled data. This is of great importance because it results in an output map where the significant and most representative reviews of a class are highlighted, which allows them to be taken as prototypes to serve as a basis for further analysis. This is not the case when a supervised classification is considered because the output is flat information (target label), and no priority is established between examples that are assigned the same label.

Some neurons stand out in this representation, such as the neurons (10, 12), (0, 11), (5, 10), (3, 13), (12, 5), and (8, 5). The intensity of the white color is higher in these neurons, meaning that the data holders should focus on improving the remarks invoked by the highly highlighted circles, keeping the same quality, and retaining the customers highlighted in the square representations. Table 4 shows a breakdown of the high-intensity neurons by number of votes (negative or positive) and class of confidence.

4.2. Discussions

To add another step of verification for the obtained results using self-organizing maps, a decision tree model has also been used to cluster the data in a tree format, thus enabling the possibility of verifying classification results in both models. The GridSearch method was used for hyperparameter tuning to get the best results possible based on the data used in this paper. In Table 5 results of the performance of SOM and CART models.

Table 4. Breakdown of the high-intensity neurons (the case of map with (15,15) dimensions)

Neuron	Classification from Figure 7	Number votes	Number of negative votes	Number of positive votes	Class confidence
(10, 12)	Positive (second class)	147	3	144	0.97
(0, 11)	Positive (second class)	125	10	115	0.92
(5, 10)	Positive (second class)	100	9	91	0.91
(3, 13)	Positive (second class)	90	5	85	0.94
(12, 5)	Negative (first class)	116	108	8	0.93
(8, 5)	Negative (first class)	100	91	9	0.91

Comparing SOM to CART models rather than another model can be insightful due to their complementary nature. SOM is an unsupervised learning model that visualizes and groups data by finding patterns without predefined labels. In contrast, CART is a supervised learning model that provides transparent and interpretable decision trees that categorize data based on known outcomes. Since the dataset used in this research is already labeled, we compared the results of the SOM model that defined patterns without labels against CART, a model with labeled training data to use as a basis. This comparison serves as a validation of the unsupervised clusters formed by SOM through CART's structured decision-making.

The results presented from the training of SOM and classification models on Amazon reviews demonstrate a sophisticated approach to sentiment analysis. However, the SOM model gives a more selective view, which makes it possible not to be satisfied with binary information as output and subsequently to treat all the data in the same way. Indeed, class confidence is an additional measure that provides further information, allowing the examples to be processed according to their importance and possibly providing additional processing for examples with low confidence. Table 6 represents a breakdown of these highlighted reviews along with the products they are linked to.

Table 5. The performance of SOM and CART models

Model	Precision	Recall	F1-score	Accuracy
Self-organizing map	0.83	0.81	0.80	0.81
CART	0.84	0.81	0.81	0.82

Table 6. Breakdown of the highlighted reviews and their products

Neuron	Impression	Product name	Reference	Total reviews	Product impression
1	Positive	National Geographic Partners LLC	B01F2MKW0I	92	4.5/5
2	Positive	Hearst Magazines	B00005N7PN	607	4.1/5
3	Positive	Norton Security (For 5 Devices)	B00MHZ6Z64	1677	4/5
4	Positive	Hallmark Card Studio 2016 Deluxe	B013X957AM	191	3.9/5
5	Positive	Bitdefender Total Security 2009 1Yr/1Pc	B001C31P4E	5	3.6/5
6	Negative	TurboTax Home & Business 2014 Fed + State + Fed Efile Tax Software - Win [Download] OLD VERSION	B00NG7JYYM	772	2/5
7	Negative	Kidswatch Internet Security Parental Control	B000IZBL9G	25	2.6/5
8	Negative	Meredith	B000FI91TS	5	2.7/5
9	Negative	Bonnier Corporation	B000S5NWAC	30	2/5
10	Negative	Meredith	B00XB73LW1	129	2.8/5

Some of the reviews that were highlighted in the map for the highly intense positive class:

- Timeless. I remember reading at my grandpa's house. I'm excited to be an adult now and have my own subscription! I love the jokes and the stories. I hope it is around for my kid's kids!
- My favorite magazine filled with inspiration.
- Very good program at a low price.
- Love this! Very user friendly.
- This is terrific security software. I replaced Norton, & AVG with Bitdefender and what an improvement in computer performance this provided. Highly recommend.

As well as some reviews for the high intense negative class:

- I cannot be honest about my problems as Amazon will not post it. purchase disk only
- The product was impossible to set up and ridiculously cumbersome to use. I returned it.
- Never have received even one issue!
- Even though I got this magazine at a discount, I wish I had not have wasted my money. I now flip through in less than a few minutes and then trash it.

The first article that my daughter read to me had information about teenage pregnancy, abortions, and the pill. And it was inferred that this is acceptable behavior. It is not acceptable behavior in my opinion. I cancelled the magazine subscription and I plan to contact the editor of this magazine.

The qualitative analysis of reviews identified as belonging to the “positive class” and “negative class” further illustrates the models' capabilities. Positive reviews are characterized by solid praise and satisfaction, using words like “timeless”, “favorite”, and “inspiration” reflecting accurately classified sentiments. On the contrary, negative reviews contain explicit expressions of disappointment and disinterest, with phrases such as “problems”, “impossible”, and “cancelled” showcasing the models' ability to discern nuanced expressions of dissatisfaction.

The reviews in Table 5. belong to high-intensity (positive or negative) classes with class confidence higher than 0.9. Most reviews classified as positive match the overall product impression; nevertheless, some of the negative reviews in the table have a general impression that can be considered high and can categorize the product as an average-rated product. The reason is that a customer experience may vary widely due to individual expectations, usage context, or personal preferences. A product can generally meet or exceed the expectations of most customers. However, it might still not align with everyone's specific needs or expectations, which would lead to some negative feedback. Other reasons could also be linked to isolated product defects, delivery issuers, or even customer service interactions. This means that negative reviews can coexist with positive overall product impressions. However, it also pinpoints some of the product's deficiencies that the seller might not be aware of. The high accuracy and precision of the model, as confirmed by the verification step, indicate its suitability for extracting actionable insights from customer feedback. Thus, businesses can better understand and respond to consumer needs.

This approach stands out from recent methods due to its thorough preprocessing, which includes steps like expanding contractions and lemmatizing to ensure clean data. It utilizes VADER for reliable initial labeling and TF-IDF for effective feature extraction. LightGBM is employed for efficient feature selection, boosting the model's accuracy. Combining SOM and CART decision tree, the approach leverages unsupervised and supervised learning for robust sentiment classification. Rigorous parameter tuning and cross-validation ensure optimal model performance, while SOM enhances interpretability. Additionally, this method is tailored to specific product categories, providing more targeted and relevant insights. Table 7 shows a comparative analysis with papers cited in section 3. The papers below used the SOM models with different datasets.

Table 7. Comparative study with papers from the related work section

Paper	Publication date	Accuracy	Dataset
Paper [17]	2023	0.75	Pinna Indian Diabetes, Wisconsin Breast Cancer, MUSK “CLEAN I”, LSVT Voice Rehabilitation, Olivetti Faces [32]
Paper [18]	2022	0.74	Manually collected dataset from experiments
Paper [19]	2021	0.76	Breast progression dataset and IDC-negative and IDC concurrent DCIS dataset. [33]
This paper	2024	0.85	Amazon Reviews [23]

In fact, it is possible to detect among data samples the ones that deserve to have more attention or need to have a particular treatment, thus avoiding treating equally all the samples belonging to a particular class. Nodes corresponding to those particular data samples are highlighted in the map and can be easily detected as shown in Table 6. Those points of interests are highlighted in the map and can also easily be detected manually.

5. CONCLUSION

This paper aims to study possibilities provided by unsupervised machine learning models, especially SOM neural networks, applied to customer product review data to gain more insight into customer-endorsed products. An Amazon review dataset was collected from 1996 until late 2018 and divided into multiple product categories with 12 features. This experiment focused only on the text data and a mix of three product categories.

This paper introduces a novel approach to customer product appreciation and sentimental insights, leveraging the SOM model, also known as the Kohonen neural network, for unsupervised learning. A competitive learning algorithm generates a grid of neurons/nodes organized in a two-dimensional space. Each node is associated with a weight vector representing a point in the input data space. During the training process, the SOM adjusts the weights of its nodes based on the input data to create a topological map of the input data. The unique aspect of this approach is its ability to use a clustering model for labeled data. Instead

of providing flat output information indicating the class associated with each example, the most representative reviews of a class are highlighted in the map. This additional information allows the extraction of the most relevant prototypes from the package for further processing or searches for the most pertinent information about a given class.

This research emphasizes the strength of SOM models in managing large datasets commonly found in Amazon reviews, providing an efficient sentimental analysis method. Moreover, this study adds value to AI and Machine Learning by showcasing the applications of SOM models in real-world scenarios beyond theoretical boundaries. It paves the way for an investigation to enhance the model's precision and assess its usefulness across various sectors and languages.

Finally, this research has achieved a significant accuracy of 0.83, marking a promising milestone. Future development of this work holds great potential to further enhance the quality of the results, particularly through the use of the ontologies approach during the feature engineering phase. This future development prospect holds promise for even more accurate and insightful results.




REFERENCES

- [1] J. Tang, C. Deng, and G.-B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 809–821, Apr. 2016, doi: 10.1109/tnnls.2015.2424995.
- [2] A. A. Heidari, H. Faris, I. Aljarah, and S. Mirjalili, "An efficient hybrid multilayer perceptron neural network with grasshopper optimization," *Soft Computing*, vol. 23, no. 17, pp. 7941–7958, Jul. 2018, doi: 10.1007/s00500-018-3424-2.
- [3] C.-F. Tsai and Y.-H. Lu, "Customer churn prediction by hybrid neural networks," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12547–12553, Dec. 2009, doi: 10.1016/j.eswa.2009.05.032.
- [4] A. J. Cuadros and V. E. DomÍnguez, "Customer segmentation model based on value generation for marketing strategies formulation," *Estudios Gerenciales*, pp. 25–30, Mar. 2014, doi: 10.1016/j.estger.2014.02.005.
- [5] S. M. Asmara, G. David, M. T. Abdullah, W. I. S. Wan Din, D. N. a/l Eh Phon, and A. F. Z. Abidin, "Self-organizing map (SOM) for species distribution modelling of birds species at Kenyir landscape," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 6, pp. 5235–5243, Dec. 2019, doi: 10.11591/ijece.v9i6.pp5235-5243.
- [6] S. Misra, H. Li, and J. He, "Robust geomechanical characterization by analyzing the performance of shallow-learning regression methods using unsupervised clustering methods," in *Machine Learning for Subsurface Characterization*, 2020, pp. 129–155.
- [7] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982, doi: 10.1007/bf00337288.
- [8] R. Jha *et al.*, "Self-organizing maps for pattern recognition in design of alloys," *Materials and Manufacturing Processes*, vol. 32, no. 10, pp. 1067–1074, Feb. 2017, doi: 10.1080/10426914.2017.1279319.
- [9] U. G. Inyang, O. O. Obot, M. E. Ekpenyong, and A. M. Bolanle, "Unsupervised learning framework for customer requisition and behavioral pattern classification," *Modern Applied Science*, vol. 11, no. 9, Aug. 2017, doi: 10.5539/mas.v11n9p151.
- [10] J. Malone, K. McGarry, S. Wermter, and C. Bowerman, "Data mining using rule extraction from Kohonen self-organising maps," *Neural Computing and Applications*, vol. 15, no. 1, pp. 9–17, Aug. 2005, doi: 10.1007/s00521-005-0002-1.
- [11] R. Ponnmalai and C. Kamath, *Self-organizing maps and their applications to data analysis*. 2019.
- [12] J. Qian *et al.*, "Introducing self-organized maps (SOM) as a visualization tool for materials research and education," *Results in Materials*, vol. 4, Dec. 2019, doi: 10.1016/j.rinma.2019.100020.
- [13] S. Licen, S. Cozzutto, M. Angelucci, and P. Barbieri, "Self-organizing map algorithm as a tool for analysis, visualization and interpretation of electronic nose high dimensional raw data," *Chemical Engineering Transactions*, vol. 68, pp. 313–318, 2018, doi: 10.3303/CET1868053.
- [14] A. Neisari, L. Rueda, and S. Saad, "Spam review detection using self-organizing maps and convolutional neural networks," *Computers and Security*, vol. 106, Jul. 2021, doi: 10.1016/j.cose.2021.102274.
- [15] Z. Zhang *et al.*, "Application of the self-organizing map method in February temperature and precipitation pattern over China: comparison between 2021 and 2022," *Atmosphere*, vol. 14, no. 7, Jul. 2023, doi: 10.3390/atmos14071182.
- [16] S. Dalal *et al.*, "An efficient brain tumor segmentation method based on adaptive moving self-organizing map and fuzzy k-mean clustering," *Sensors*, vol. 23, no. 18, Sep. 2023, doi: 10.3390/s23187816.
- [17] Z. Zhengtian, R. Zhiyuan, and D. Xiaoyan, "Feature selection for binary classification based on class labeling, SOM, and hierarchical clustering," *Measurement and Control*, vol. 56, no. 9–10, pp. 1649–1669, 2023, doi: 10.1177/00202940231173748.
- [18] G. A. Angulo-Saucedo, J. X. Leon-Medina, W. A. Pineda-Muñoz, M. A. Torres-Arredondo, and D. A. Tibaduiza, "Damage classification using supervised self-organizing maps in structural health monitoring," *Sensors*, vol. 22, no. 4, Feb. 2022, doi: 10.3390/s22041484.
- [19] E. Yuan, M. Matusiak, K. Sirinukunwattana, S. Varma, Ł. Kidziński, and R. West, "Self-organizing maps for cellular in silico staining and cell substate classification," *Frontiers in Immunology*, vol. 12, Oct. 2021, doi: 10.3389/fimmu.2021.765923.
- [20] Y. ELYUSUFI and M. A. I. T. KBIR, "Churn prediction analysis by combining machine learning algorithms and best features exploration," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022, doi: 10.14569/ijacsa.2022.0130773.
- [21] J. Ni, J. Li, and J. McAuley, "Justifying recommendations using distantly-labeled reviews and fine-grained aspects," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 188–197, doi: 10.18653/v1/D19-1018.
- [22] S. Ahsain, Y. E. L. Yusufi, and M. A. I. T. Kbir, "Optimizing customer experience analysis across dataset size reduction and relevant features selection," *International Journal of Engineering Trends and Technology*, vol. 71, no. 12, pp. 78–89, Nov. 2023, doi: 10.14445/22315381/ijett-v71i12p209.
- [23] M. Suyal and P. Goyal, "A new classifier model on drug reviews dataset by VADER sentiment analyzer to analyze reviews of the dataset are real or fake based on machine learning," *International Journal of Engineering Trends and Technology*, vol. 70, no. 7, pp. 68–78, Jul. 2022, doi: 10.14445/22315381/ijett-v70i7p208.
- [24] D. Hazarika, G. Konwar, S. Deb, and D. J. Bora, "Sentiment analysis on Twitter by using TextBlob for natural language processing," in *Proceedings of the International Conference on Research in Management & Technovation 2020*, Jan. 2020, vol. 24, pp. 63–67, doi: 10.15439/2020km20.




- [25] A. 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, doi: 10.14569/ijacsa.2023.0140542.
- [26] A. M. Asri, S. R. Ahmad, and N. M. M. 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, doi: 10.14569/ijacsa.2023.0140530.
- [27] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, Oct. 2004, doi: 10.1108/00220410410560582.
- [28] scikit-learn developers "TfidfTransformer," *Scikit-learn*. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html (accessed Jul. 13, 2024).
- [29] A. M. Kalteh, P. Hjorth, and R. Berndtsson, "Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application," *Environmental Modelling and Software*, vol. 23, no. 7, pp. 835–845, Jul. 2008, doi: 10.1016/j.envsoft.2007.10.001.
- [30] J. C. Burguillo, "Using self-organizing maps with complex network topologies and coalitions for time series prediction," *Soft Computing*, vol. 18, no. 4, pp. 695–705, Nov. 2013, doi: 10.1007/s00500-013-1171-y.
- [31] A. K. Guèye *et al.*, "Weather regimes over senegal during the summer monsoon season using self-organizing maps and hierarchical ascendant classification. Part II: interannual time scale," *Climate Dynamics*, vol. 39, no. 9–10, pp. 2251–2272, Jun. 2012, doi: 10.1007/s00382-012-1346-8.
- [32] Wi. Wolberg, Jul. 14, 2024, "Breast cancer Wisconsin (original) [dataset]," *UCI Machine Learning Repository*, 1990. <https://doi.org/10.24432/C5HP4Z>.
- [33] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1550–1560, Jul. 2017, doi: 10.1109/tmi.2017.2677499.

BIOGRAPHIES OF AUTHORS






Sara Ahsain    is a Ph.D. student and Software Engineer at the Faculty of Science and Technologies (FST) of Tangier, University Abdelmaled Essaâdi, Morocco. She received a Eng. degree in software engineering in the same University. As a member of LIST Laboratory since 2019, her research focuses on - artificial intelligence (machine learning, deep learning, planning, and search strategies) and digital marketing. Sara has actively contributed to international conferences and authored ongoing work in international journals. She also has over 5 years of experience in software engineering across different companies. She can be contacted by email: sara.ahsain@etu.uae.ac.ma.



Yasyn Elyusufi    is a Computer Science Professor at the Faculty of Sciences and Technologies, Abdelmalek Essaadi University, Tangier, Morocco since 2018. His primary research interests include artificial intelligence, machine learning, and big data, with a particular focus on social media and marketing analysis. He earned his Ph.D. in computer science from the same university in 2016 and holds an engineering degree from the National School of Applied Sciences in Tangier, Morocco in 2008. Yasyn has authored numerous publications in international journals and actively contributed to prestigious international conferences. He also worked as a lecturer, senior developer, and system administrator. He can be contacted at email: yelyusufi@uae.ac.ma.



M'hamed Ait Kbir    is a full professor at the Computer Science Department of the Faculty of Sciences and Technologies (FST) of Tangier, since 2001, University Abdelmalek Essaâdi, Morocco. As a member of LIST Laboratory, since 2007, his research works focus on three main areas: Computer vision (multimedia flow optimization, multimedia document content watermarking, object recognition, 3D contents indexing and retrieval, 3D reconstruction) - artificial intelligence (machine learning, deep learning, planning and search strategies) - bioinformatics (micro-array data decision making, biological data integration, biological networks analysis). He is a member of scientific committees of many international conferences and journals. As an expert, he participates in the evaluation of public and private education programs for the ANEAQ and the ministry of higher education and scientific research. He can be contacted by email: maitkbir@uae.ac.ma.