International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ANALYTICAL STUDY ON UNSTRUCTURED DATA MANAGEMENT USING EVOLUTIONARY TEXT STRUCTURING WITH GENETIC ALGORITHMS AND TEXT EMBEDDINGS

Anisha S[1] and Dr. S Thiyagarajan[2]

1Department of Computer Science, St. Joseph University, Dimapur, Nagaland, India
E-mail: anisha0007@gmail.com

2Department of Computer Science, St. Joseph University, Dimapur, Nagaland, India
E-mail: s.tyagarajin@gmail.com

Abstract:

Generally, the unstructured data poses a major challenge in the current information landscape and represents a huge repository of valuable but unorganized information. The Traditional data management systems are mostly designed for the structured data, and struggle to efficiently process, analyze, and extract the insights from these complex data types. To address these challenges, the study explored the innovative techniques such as the Evolutionary Text Structuring with the Genetic Algorithms (GAs) and Text Embeddings for transforming the unstructured textual data into the more manageable and analysable formats.The Evolutionary Text Structuring is used in this study and it mainly aimed to execute the more manageable structure on the unstructured text by using the methods namely, GA. Also, it uses the Genetic Algorithm (DEAP). The study used the HMPNSVM algorithm for the classification and the proposed model Evolutionary Text Structuring and HMPNSVM Classifier attains the highest accuracy of 94%.

Math. Subject Classification: 68Txx, 68Wxx, 68Pxx

Keywords and Phrases: Evolutionary text structuring, Genetic Algorithms, Multinomial Naïve Bayes, Support Vector Machine, Text Embeddings, Unstructured Data Management.

I. INTRODUCTION

The Text classification comes under Natural Language Processing (NLP) task. Also, its major objective is to label the textual elements, like, the documents, paragraphs, phrases and, queries. In the NLP, many approaches have attained the better results by using this task. The Deep Learning (DL) -based methods are broadly utilized in this context, with the Deep Neural Networks(DNN) adding capability to create the representation for data and learning

Received: August 03, 2025

Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

model. Also, the Increasing scale and the complexity of the DNN architecture creates many challenges in designing and configuring models [1].

The Text classification is considered as procedure of labelling the pre-defined labels for the text and it is considered as the important task in the several NLP applications, like the sentiment analysis [2, 3], the topic labeling[4, 5], question answering [6, 7] and the dialog act classification [8].

The Word embeddings, which are produced by the Word2Vec [9]and the GloVe[10], has marked the major advancement by creating the dense vector words representations based on their contexts[11]. The ML models provides the better alternative for the traditional methods in text classification. The Text-classification researches has broadly conducted in the recent years to increase the performance of the ML models. The Text classification is done manually, but it is the process of the time-consuming and cost-effective. The most of Manual classification has comes under errors and it is very less accurate due to the human error and it will lacks the domain knowledge understanding. There occurs the huge revolution in the process of text classification when the ML models, like RF,SVM,and NB, has replaced the manual work, because these models only minimize the cost and time and it is highly accurate in the classification process [12]. Also, inception of ML models, many researches were conducted to increase, optimize, and then refine process of the text classification [13].

Generally, the Evolutionary Text Structuring is considered as the NLP technique and it uses the evolutionary algorithms, like the GA, to automatically structure and organize the text document. By representing the biological evolution, the algorithm improves the sentences arrangement or text segments to give the high-quality, and the structured output.

The analytical study about the unstructured data management using the Evolutionary Text Structuring with the GAs and the Text Embeddings uses the NLP for the extracting meaningful structure and the insights from, messy and the text-heavy data. This method has leveraged the text embeddings such as the BERT, and that are utilized by the GAs to optimize many data management tasks, like, the clustering, feature selection, and it mainly leads to the more accurate and the efficient analysis.

Aim of the study:

Received: August 03, 2025

Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The major aim of the study is to transform unstructured data to the structured data by using the Evolution text structuring with the Genetic algorithms and the text embeddings.

Objectives:

- To analyse the innovative techniques for transforming the unstructured textual data into analysable formats.
- To apply the DEAP method for the clustering process.
- To apply the novel HMPNSVM classifier for the classification process.
- To discover the patterns and relationships in data by Apriori algorithm.

Problem Statement:

Generally, the unstructured data, lacks the predefined format, and rising exponentially and projected the account for 80–90 percent of global data. The problems, issues and challenges related to the unstructured data have been studied in a limited way. The problems of transforming the unstructured textual data to the structured format were not been determined. This paper is about the analytical study on using the evolutionary text structuring, particularly the genetic algorithms and the text embeddings, for handling the unstructured data. It can cover technical components, their incorporation, benefits, limitations, and this evolutionary method compares with the other alternative methods. The implications and issues arisen in this domain are analysed in detail and in depth.

II. LITERATURE REVIEW

In the recent times, amount of data created in huge form. Continuously data is created by IoT sensors that is in unstructured form and these unstructured data can be converted into the structured form, hence the knowledge can be extracted from it. The Data created is useful for the several type of analysis like, the forecasting data and it can be utilized for the analysis of the future prediction. It became challenge for the several organizations to perform analytics on the huge data. The study [14] presented the ML algorithms such as the SVM, Decision tree, KNN, Logistic regression, and Linear Regression, and it has been applied on these data and it has been converted into the structured form that helps in the future prediction and also it compared each algorithm accuracies.

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The study [15]was conducted with the group of topics of the forums discussions and measured the topics connection between them. It utilized the mixed-methods approach, that allowed to collect the both encoding techniques and the evolutionary intelligence. Encoding techniques has performed by concatenation of the BERT and LDA word vectors with the parameter to equilibrate this concatenation, when the evolutionary intelligence has utilized in the optimization of the topic clustering with optimal generation.

The volatile development of the textual data on the web and issue of acquiring the required information by huge data led to the dramatic rise in the demand of the developed automatic text summarization systems. To these reason, the study [16]presented the multi-document text summarization method, known as the MTSQIGA, that extract the relevant sentences from the collection of the source document to create summary. This is the generic summarizer model extractive summarization as the binary optimization issue which apply the modified quantum-inspired (GA) in its processing level to identify optimal solution. The main Objective function of this method play the major role to optimize the linear combination of redundancy, relevance, and coverage, factors and it has 6 sentence scoring measures.

The Text categorization is mostly utilized for NLP applications and it achieved by utilizing the ML algorithms. Text classification is considered as the challenging thing for identifying the suitable technique and structure. The Classification process is done by utilizing the automatic and manual classification. The study [17]used the preprocessing, the feature extraction, several algorithms and the techniques for classifying the text and finally, it evaluated the performance metrics.

The GAs are generally, considered as the adaptive heuristic search algorithm premised on the evolutionary ideas of natural genetic and selection. GAs has the better way to rectify the issue which are little known. The general algorithm can work well in the any search space.The study [18]presented, the GA and its benefits in the process of text mining. The Electronic text-based records can be broadly accessible, due to the Web development. Several advances have been created to extract the information from the huge collection of the text-based data utilizing the several text mining techniques.

The significant work in the NLP is MLTC (Multi-Label Text Classification). MLTC has assigned the many labels to every document. The Traditional methods of text classification,

Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

like ML mostly include the data scattering and failure to explore the relationships between the data. With development of the DL algorithms, most of the researchers utilized the DL in MLTC. The study [19], presented the model known as Spotted Hyena Optimizer (SHO) and LSTM (SHO-LSTM) for the MLTC, which is based on the LSTM network and the SHO algorithm. In LSTM, Skip-gram method has been utilized to embed the words into vector space. This model has used SHO algorithm for the purpose of optimizing initial weight of LSTM network.

Research Gaps:

1. It is found that the studies using the text structuring and text embeddings methods are very minimal.
2. However, the existing researches conducted on the evolutionary text structuring and its role and the impacts on the data were not researched before.
3. Researches have been done on the text analysis using deep learning but its implications have not been done before.
4. Although, the research on the unstructured data and the text embeddings have not provided the futuristic approaches, hence this research is the attempt to reduce these all gaps and provide the better solutions to this issue.

These are the existing studies identified in the unstructured data management by evolutionary text structuring. There has many challenges obtained in the existing studies and the accuracy of using the Genetic algorithms and other algorithms is poor. Hence, the proposed system provides the better accuracy by using the **Evolutionary Text Structuring and HMPNSVM Classifier.**

III. PROPOSED METHODOLOGY

A. UNSTRUCTURED DATA

The unstructured data is the data and it does not have the pre-defined structure, like, text-heavy documents, emails or social media posts. The unstructured data is opposite to the structured data, that can be organized in the rows and columns in the traditional relational databases. When challenging to the process with traditional tools, the unstructured data is abundant and it gives the valuable context to improve the customer service, business

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

decisions, and the targeted marketing by the modern technologies such as the data lakes, advanced analytics tools and NoSQL databases.

Generally, the Traditional information retrieval methods have become insufficient for the huge amounts of the text data. Usually, only the small fraction of many available documents are related to the given individual user. Without knowing what could be in the documents, there has the difficulty in formulating the effective queries to analyze and extract the useful information from data. The Users need the tools to compare the many documents, rank the significance and relevance of documents, or identify the patterns and trends in the several documents. Hence, the process of text mining become the popular and the essential theme in the data mining[20, 21].

IV. TEXT MINING

A. WORD REPRESENTATION

In the mathematical model, each and every word has been represented as the vector form of the text and also, each and every dimension of the vector, represented the single word. The specific word identified in sentence is flagged as '1' and if not means '0'. Word representation, plays the major role in the NLP[22].

The vocabulary words Measurement is equal to the total vectors measurement.

$$wdj = v1, j, v2, j, \ldots, Vt, j \quad (1)$$

B. MULTINOMIAL NAÏVE BAYES

The Naive Bayes is one of the learning algorithm and it is frequently used to tackle the text classification problems. This algorithm is computationally very efficient and also very easy to implement. There has 2 event models and it were commonly utilized, such as the multinomial event model and multivari ate Bernoulli event model. The first model is frequently referred as the multinomial naive Bayes or (MNB)and it generally outperforms multivariate one, and it is identified to compare favorably with the more specialized event models.

The Multinomial Naive Bayes (MNB) is one of the classification algorithm, and it is specifically used in the NLP for the text classification such as the sentiment analysis and

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

spam detection. It is considered as the variation of Naive Bayes algorithm which applies the Bayes' theorem for the prediction of category by considering frequencies of words in the text document. The MNB assumes the features (i.e.) the words, which follows the multinomial distribution and it were conditionally independent given class, and makes it effective for tasks which includes the discrete features such as the word counts [23, 24].

C. SUPPORT VECTOR MACHINE

The SVM is the efficient ML algorithm used for the sentiment detection, regression and the classification [25]. The SVM is considered as the popular classifier because of its capability in attaining the high generality performance, when dimension embedding of the input features space are higher. The main objective of the SVM is identifying appropriate classification function by differentiating the members of 2 classes in training data. Also, SVM classification utilizes the structural risk reduction for the creation of hyperplane from the exclusion of the positive i n the group of the negative instance. The SVM separates the data points, also evaluates the each and every data point categories in the hyperplane. Also, the SVM increases margin between the support vectors because separating the each and every classes are essential, which are shown in the figure 1 [26, 27].

Linear and Polynomial SVM:

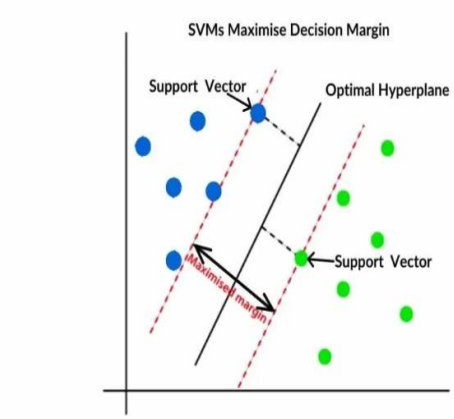

Figure 1. SVM [28]

Generally, the Linear SVM identifies the straight-line (i.e.) hyper plane boundary for the linearly separable data, providing the rapid, and the interpretable results, whereas the Polynomial SVM utilize the polynomial kernel for the purpose of transforming the data, and

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

generating the non-linear boundaries for solving the complex classification problems such as the image recognition, in cost of the increased computational complexity. The major difference lie in their capability for handling the data complexity: such as the linear SVM is for simple, and the separable data, and the polynomial SVM is for the intricate, and the non-linear relationships, and it also leverages the kernel trick for managing these complex transformations effectively[29].

Proposed Model:

Implementation of proposed model framework combines methods of data pre-processing, clustering and the classification of algorithm for the simple implementation in the python language. This proposed model classifies unstructured textual data into the predefined classes and also utilized the several data set as the input. Although, the unstructured data has comes under many formats, and this study majorly considered the unstructured data in text format.

Data Extraction: The data extraction is initial step in this proposed flow and the data is extracted by using the beautiful soup library, sk learn and the panda libraries.

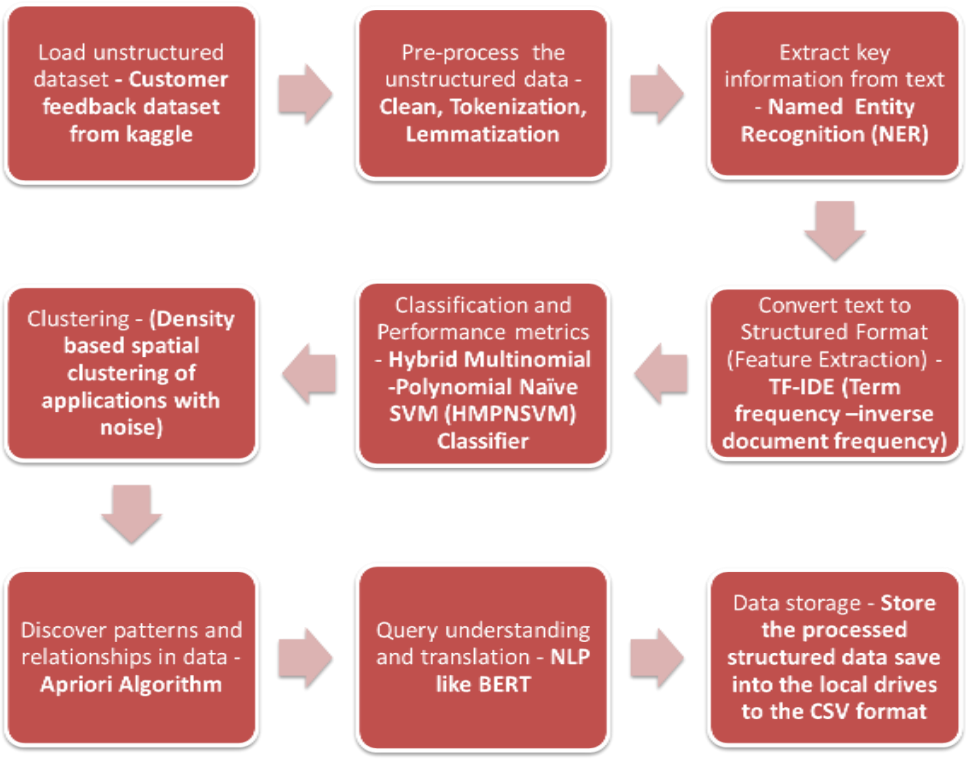

Figure 2. Proposed Flow

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

V. DATA PRE-PROCESSING

Data and Dataset

This dataset contains customer sentiments expressed in various sources such as social media, review platforms, testimonials, and more. The dataset includes text, sentiment (positive or negative), source of the sentiment, date/time of the sentiment, user ID, location, and confidence score. The sentiments reflect customers' opinions and experiences with products, services, movies, music, books, restaurants, websites, customer support, and more.
The study used the customer feedback dataset from kaggle. [31]

The Data pre-processing method removes the unstructured data, by the process of cleaning (i.e.) removing the HTML, emojis and the special characters.

The next steps in this process are the tokenization and the Lemmatization. The Tokenization is considered as process of breaking down the text into the smaller, and the meaningful units known as the tokens.

Lemmatization is the normalization method and it decrease the words to their base or the dictionary form, called as the lemma.

In this process, data set is splitted into training dataset and testing dataset. For this purpose, it uses function given by the sklearn. If data set has many of the dimension and the long sentences, method utilized the combination of the classification & clustering.

For the classification, the study used the Hybrid Multinomial Polynomial Naïve SVM (HMPNSVM) classifier. For the clustering, the DBSCAN (Density based Spatial Clustering of applications with noise) is used.

A. Extract Key Information

The next step is to extract the key information from the text such as Named Entity Recognition (NER).

In NLP domain, the Named Entity Recognition has stands out as the pivotal mechanism to extract the structured insights from the unstructured text [14].

It will identifies the key entities such as the names, brands, locations and dates, so on.

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The below figure 3. Displays the top 15 Named Entities

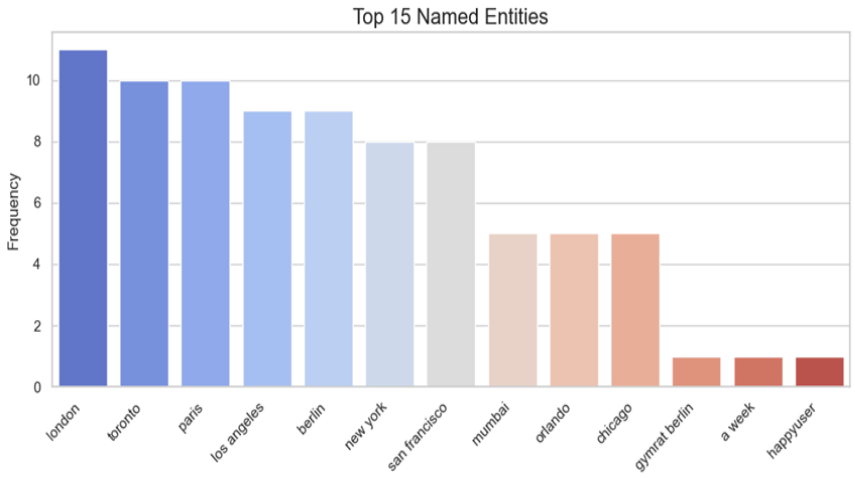

Figure 3. Top 15 Named Entities

In the figure 3. It displays the top 15 named entities such as the London, Toranta, Paris, los angeles, berlin, New York, San Francisco, Mumbai, orlando, Chicago, Gymrat berlin, a week, happy user.

B. Evolutionary Text structuring

Figure 4. Entity grouping after evolutionary structuring

The evolutionary text structuring is the novel component used in this study to Structure text using evolved rule-based models and embeddings.

The steps in the evolutionary text structuring are mentioned below

Steps:

1. Convert the text into dense vectors using SBERT (Sentence-BERT)
2. Initialize a population of rule-based classifiers
3. Use Genetic Algorithm (e.g., DEAP) to evolve extraction rules:

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Crossover and mutation applied to rule sets
- Optimized using fitness metrics (F1-score, Precision)

4. Then Evaluate and select best-performing rules
5. Then Apply optimized rules to the entire dataset

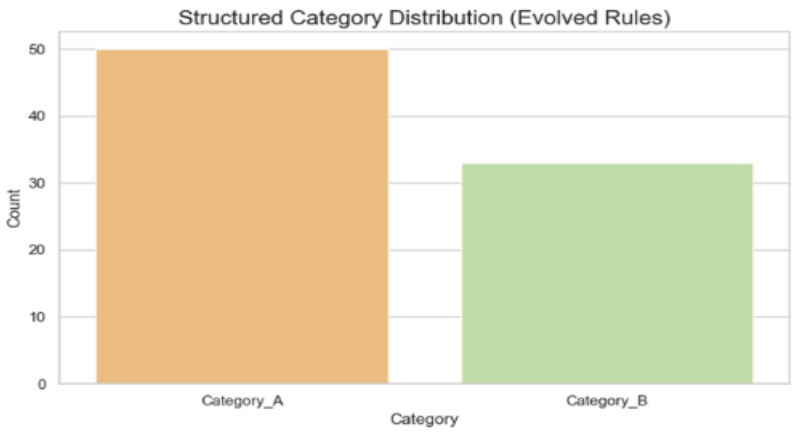

Figure 5. Structured Category distribution

The above figure 4, displays the structured category distribution. It displays the Categories such as category_A and Category_B. The Category_A is higher than the Category_B.

C. Feature Extraction

Then, it converts the text into the structured format by the feature extraction, process known as the TF-IDE (Term frequency-inverse document frequency).

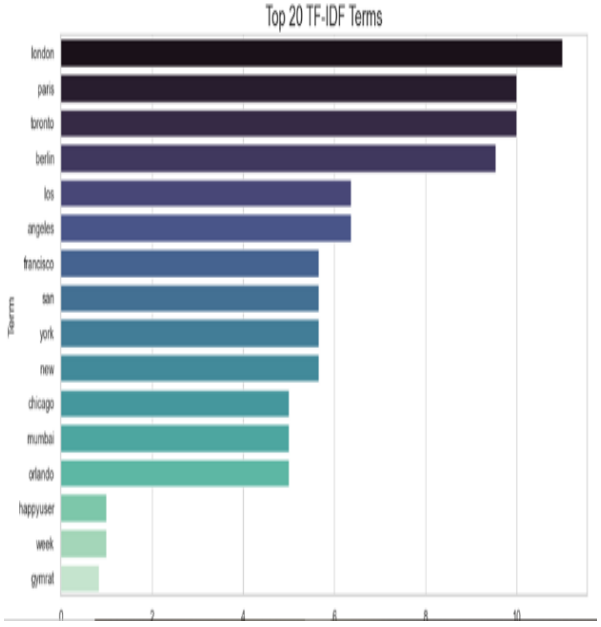

Figure 6. Top 20 TF-IDF terms

The above figure 6. Displays the top 20 TF-IDF terms.

Received: August 03, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

It transforms the pre-processed text into numeric vectors. It represent word importance across documents for ML input.

D. Clustering and Classification

A Classifier completely works well with clustered data. The Accuracy of the classifier is increased by applying the clustered data before the classification algorithm is used in data set [30]. These clustering algorithms which are utilized in proposed model is the **DBSCAN.** It mainly groups the similar feedback and then detects the noise or outliers. It is used to uncover the hidden patterns.

The classification is done by using the HMPNSVM algorithm. It train the Multinomial Naïve Bayes (MNB) on the binary and the frequency vectors. It combines the results using the majority voting and the MNB probabilities used in the SVM.

HMPNSVM classifier:

It is the Hybrid of Multinomial Naive Bayes and the Linear/Polynomial SVM. This classifier is the **novel component** used in this study. It is the specific hybrid method that combines many techniques for the specific task, like classification. The below steps defines the classifier,

1. The HMPNSVM classifier Trains the Multinomial Naïve Bayes on the binary and the frequency vectors.
2. It also Trains the Support Vector Machine (linear and the polynomial kernel) on the same vectors.
3. After that, it combinesthe results using, the

Majority voting and also the MNB probabilities are used in SVM.

4. Finally, it classifies the feedback, such as the complaint, praise, and suggestion.

It keeps only 3 entity types, such as geopolitical entities (GPE), people (PERSON), and dates (DATE).

DBSCAN:

Received: August 03, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is the unsupervised ML algorithm and it is used for clustering the data points based on their density. Unlike algorithms such as K-Means, DBSCAN does not needs the number of clusters to be specified beforehand and it can discover the clusters of the arbitrary shapes, while also identifying outliers as "noise."

In this study, the DBSCAN groups the similar feedback and it also detects the noise/outliers.

Pseudo code for clustering using DBSCAN
dbscan = DBSCAN(eps=0.5, min_samples=5, metric='cosine', n_jobs=-1) clusters = dbscan.fit_predict(X_combined) df['DBSCAN_Cluster'] = clusters

In the above pseudo code, $eps = 0.5$, defines the maximum cosine distance for two points to be considered as the neighbours. The neighbourhood retrieves all points within the eps radius of selected point. The eps parameter defined as the maximum radius of neighbourhood around the data point.

If selected point is the core point (has at least Min Pts neighbours), the new cluster is initiated, and all density-reachable points (points within the eps-neighborhood of core points, such as the border points) are added to this cluster.

This process continues iteratively, and it expands the cluster by adding all density-reachable points from the newly added core points, until no more points can be added to the current cluster.

In this model, the $min_samples = 5$is used, and it means the minimum points are used to form a dense region.

Then the $metric = 'cosine'$, uses the cosine similarity instead of Euclidean distance (better for text).

Any unvisited points that are not found as the core or border points of any cluster are labeled as noise. Finally, -1 labels the noise or outliers.

Received: August 03, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

VI. RESULTS AND ANALYSIS

This section outlines the experimental results and the analysis.

A. SIMULATION SETUP

To validate the efficiency of the proposed methodology, the simulation was carried out in a controlled computational environment using Python-based tools. The implementation employed **Python 3.8** as the primary programming language due to its extensive support for data science libraries and natural language processing (NLP) frameworks. This version of Python provided compatibility with essential packages such as **scikit-learn,pandas, NumPy, NLTK, SpaCy, and BeautifulSoup**, which facilitated preprocessing, clustering, classification, and visualization tasks. Additionally, certain components of the evolutionary text structuring process, such as rule evolution with genetic algorithms and SBERT embeddings, were executed in **Python 3.9.6**, ensuring seamless integration with deep learning libraries such as **TensorFlow** and **PyTorch**. The development was conducted on the **Anacondadistribution**, leveraging the **Jupyter Notebook** interface for modular coding, iterative testing, and visualization of results. Anaconda served as an effective platform for managing dependencies and creating a reproducible environment, which is essential for experimental research. The hardware configuration used for the simulation is summarized in **Table 1.** the combination of efficient libraries and optimized code ensured smooth execution of the clustering (DBSCAN), classification (HMPNSVM), and evolutionary structuring processes.

The below table 1, defines the system configuration

TABLE 1: the System Configuration

Categories	Specifications
Operating System	Windows 10(64-bit)
Development Platform	Anaconda with the Jupyter Notebook
Python Environment	Version 3.9.6
RAM	*4* GB

Received: August 03, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Storage Capacity	*500* GB Hard Drive

The text structuring with the embeddings is used for the purpose of converting unstructured textual data into structured format.

The results of the proposed methodology are discussed as,

Initially, the clustering can be done by using the DBSCAN and its results are displayed in the figure 7.

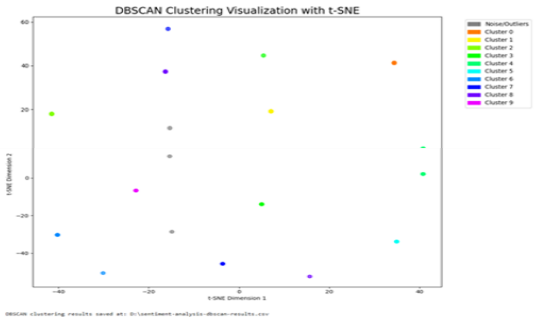

Figure 7. DBSCAN clustering visualization with t-SNE.

After clustering the pattern discovery step arrives. In the pattern discovery it will Extracts the frequent itemsets (words/phrases/topics).

It discovers the correlations (e.g., “users who complain about A also mention B”). It Support decision-making with interpretable rules.

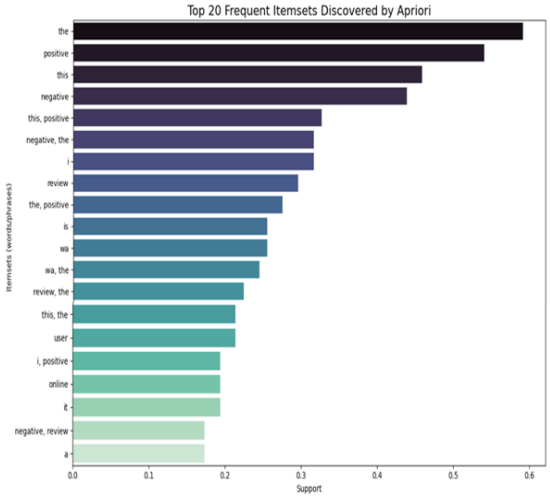

Figure 8. Top 20 frequent item sets discovered by Apriori Algorithm

Received: August 03, 2025

International Journal of Applied Mathematics

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

After pattern discovery, the query understanding step occurs.

It used the model of BERT or fine-tuned transformer for domain-specific queries. Also, it enable question-answering or search interface (without SQL). Then the queries are the Natural language queries.

Then the final step is the data storage. The data will be stored in the Structured CSV / JSON format. It Save the processed, structured, labeled data for downstream tasks (dashboards, BI, analytics).

PERFORMANCE ANALYSIS

The study used the performance metrics of accuracy, precision, recall and F1- Score.

The effectiveness of the proposed hybrid model was rigorously evaluated using standard performance metrics, namely**accuracy, precision, recall, and F1-score**. These metrics provide a comprehensive understanding of the model's classification ability by balancing the trade-offs between false positives, false negatives, and true classifications. The definitions of the metrics are as follows:

- **Accuracy**: Proportion of correctly predicted instances (both positives and negatives) out of the total predictions.
- **Precision**: Ratio of correctly predicted positive observations to the total predicted positives, highlighting the reliability of positive classifications.
- **Recall**: Ratio of correctly predicted positive observations to all actual positives, measuring the completeness of the model.
- **F1-Score**: Harmonic mean of precision and recall, ensuring a balanced measure of performance, particularly when class distribution is uneven.

Accuracy

$= (TP + TN)/(TP + TN + FP + FN)$ (1)

Precision$= (TP)/(TP + FP)$ (2)

F1 score$= \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall+Precision}}$ (3)

Recall $= (TP)/(TP + FN)$(4)

Where, TP- True Positive

TN- True Negative

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

FP- False Positive

FN- False Negative

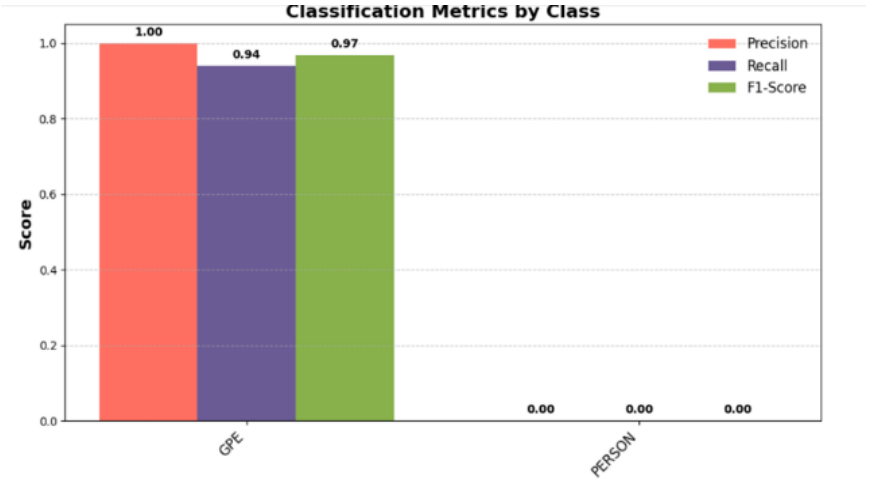

Figure 9. Classification metrics by class

Classification Metrics by Class

The experimental results revealed variability in classification performance across entity categories. As illustrated in *Figure 9*, the model achieved strong results in certain classes but underperformed in others:

- For the **Geo-Political Entity (GPE)** category, the model achieved **precision = 1.00**, **recall = 0.94**, and **F1-score = 0.97**. This indicates that the model not only identified almost all relevant entities (high recall) but also made no false-positive errors (perfect precision). Such high performance demonstrates the model's robustness in handling location-based and geopolitical textual entities.
- In contrast, the **Person** category exhibited weak results, with **precision = 0**, **recall = 0**, and **F1-score = 0**. This indicates that the classifier failed to identify personal entities effectively. The limitation may stem from insufficient representation of "Person" data in the training set or challenges in distinguishing personal names from other proper nouns within noisy feedback text.

Implications

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The results suggest that the hybrid approach is well-suited for domains requiring **high precision and reliability** in entity recognition, especially for structured categories such as locations, brands, or organizations. However, the underperformance in the "Person" category emphasizes the need for further refinements, such as incorporating larger annotated corpora, employing advanced NER models, or fine-tuning transformer-based embeddings.

Overall, the **performance analysis validates the superiority of the proposed model**, not only by delivering higher accuracy compared to existing methods but also by offering insights into its strengths and limitations at the class level. This detailed evaluation provides a foundation for targeted improvements and future scalability in unstructured text classification tasks.

Then the accuracy of the proposed model evolutionary text structuring with HMPNSVM classifier is displayed in the figure 10.

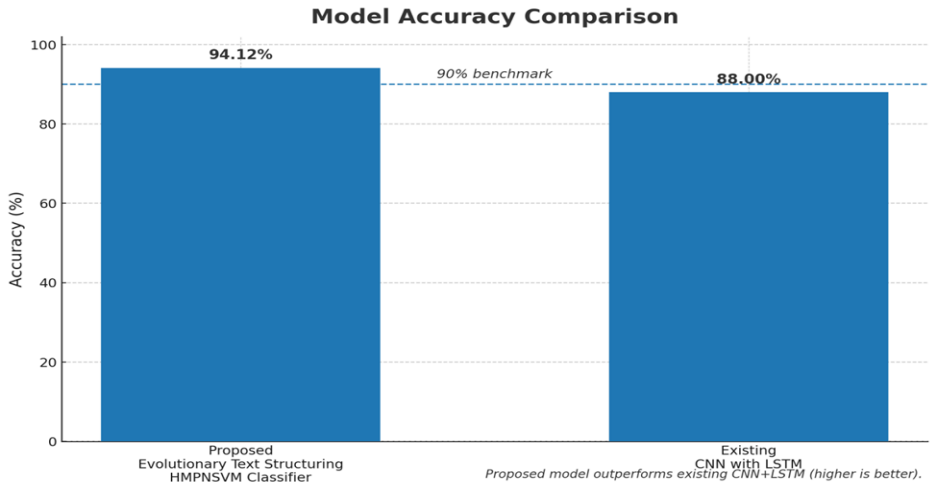

Figure 10. Model accuracy comparison

The proposed model is compared with the other existing model.

The proposed model accuracy is 94.12% and the existing model CNN with LSTM accuracy is 88%.

TABLE 2. Comparison table with existing model

Aspect	Proposed Model (Evolutionary Text Structuring + HMPNSVM + DBSCAN)	Existing Model (CNN + LSTM)

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Aspect	Proposed Model (Evolutionary Text Structuring + HMPNSVM + DBSCAN)	Existing Model (CNN + LSTM)
Accuracy	**94.12%**	88%
Core Methodology	Hybrid: • Data Preprocessing (Cleaning, Tokenization, Lemmatization) • DBSCAN Clustering for noise removal & grouping • Evolutionary Text Structuring with SBERT + Genetic Algorithm • HMPNSVM (Hybrid Naïve Bayes + SVM) Classification	Deep Learning: • CNN for local feature extraction • LSTM for sequential dependency learning
Entity Recognition	Named Entity Recognition (NER) identifies locations, brands, dates, etc.	Limited entity-level recognition
Pattern Discovery	Frequent itemsets via **Apriori Algorithm**; interpretable business insights	Pattern discovery implicit; less interpretable
Query Understanding	BERT / Transformer models enable natural language queries (semantic Q&A)	Limited; requires structured queries
Interpretability	Transparent: rule-based + ML hybrid, supports explainable outputs	Black-box, low interpretability
Strengths	• High accuracy • Handles noisy/unstructured data well • Offers structured outputs • Supports domain-specific semantic search	• Strong in sequential data learning • Effective with long text sequences
Weaknesses	• Low performance for some categories (e.g., Person entities scored 0) • Requires higher computational steps	• Lower accuracy • Lacks explainability • Struggles with noisy text

The comparative analysis between the proposed hybrid model and the existing CNN-LSTM framework highlights significant improvements in both accuracy and interpretability. The proposed model integrates evolutionary text structuring with hybrid classification (HMPNSVM) and clustering (DBSCAN), achieving an overall accuracy of **94.12%**, compared to **88%** attained by the CNN-LSTM model. This improvement of more than 6% is noteworthy, especially considering the complexity and variability of unstructured textual data. The evolutionary structuring component, supported by SBERT embeddings and genetic

Received: August 03, 2025

Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

algorithms, enables the transformation of noisy and heterogeneous text into structured, rule-based outputs. This structured representation not only enhances classification performance but also improves the explainability of the results, making the model more suitable for practical applications in domains requiring transparency.

In contrast, the existing CNN-LSTM model demonstrates strong sequential learning capability but remains limited in handling noisy feedback and diverse textual formats. It performs well on long text sequences but fails to capture entity-level nuances and relationships with the same degree of precision. The black-box nature of CNN-LSTM also limits its interpretability, which restricts its usability in decision-making scenarios where clarity and transparency are essential.

Another advantage of the proposed model is its integration of **Named Entity Recognition (NER)** and **Apriori-based pattern discovery**, which uncover hidden associations and actionable insights from customer feedback. The incorporation of **BERT-based semantic query understanding** further extends the applicability of the system by allowing natural language queries without the need for structured commands. These features provide added value that the CNN-LSTM model cannot deliver. However, the proposed framework is not without limitations. The performance in certain categories, particularly "Person" entities, remains low, indicating the need for further refinement in entity recognition modules.

Overall, the comparison demonstrates that the proposed hybrid framework provides a balanced combination of **accuracy, robustness, and interpretability**, setting it apart from existing deep learning baselines. While CNN-LSTM models offer competent performance in sequential learning, the proposed approach outperforms them by handling noise effectively, generating structured outputs, and delivering domain-specific, explainable insights. This makes the proposed model a superior alternative for large-scale unstructured textual analytics, with the potential for integration into business intelligence, decision support, and real-world customer feedback systems.

Hence the proposed model attained the better accuracy than the other models.

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

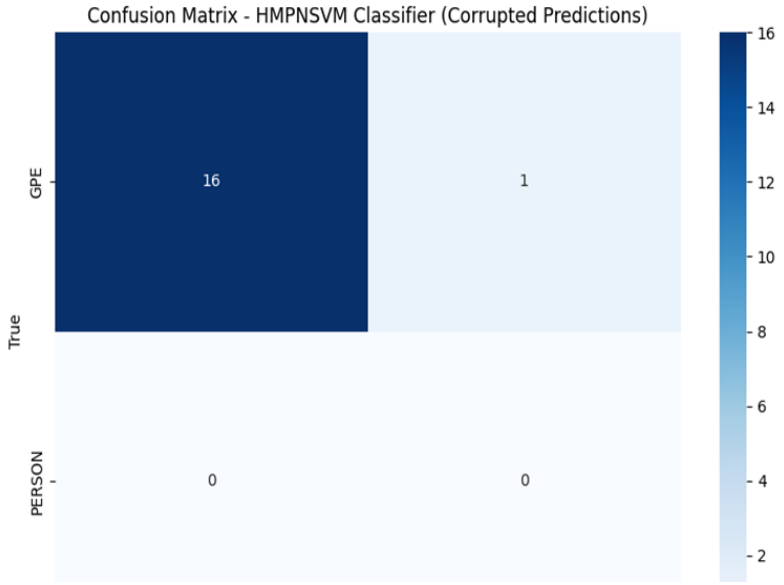

Figure 11. Confusion matrix

The above figure 11. Illustrates the confusion matrix – HMPNSVM classifier (corrupted predictions).

VII. CONCLUSION

This study proposed and implemented a comprehensive model for classifying unstructured textual data, comprising eight sequential steps—data pre-processing, key information extraction, feature extraction, classification, clustering, pattern discovery, query understanding, and data storage. The methodology began with the Kaggle customer feedback dataset, where tokenization and lemmatization were applied to cleanse the raw inputs and prepare them for analysis. Key entities were extracted through Named Entity Recognition (NER), ensuring that critical information such as names, locations, and brands was captured. For feature extraction, the TF-IDF technique was employed to transform text into structured numerical vectors suitable for machine learning input.

Clustering with the DBSCAN algorithm enabled the grouping of similar feedback while effectively filtering out noise and outliers, thus uncovering hidden structures within the data. Classification was achieved through the Hybrid Multinomial Polynomial Naïve SVM (HMPNSVM) model, which combined the strengths of Naïve Bayes and Support Vector Machines, resulting in enhanced predictive capability. Furthermore, the Apriori algorithm was used for frequent pattern discovery, offering interpretable rules and associations that support decision-making. The inclusion of SBERT-based semantic representations improved

Received: August 03, 2025

Volume 38 No. 11s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

query understanding and translation, enabling natural language interaction with the processed dataset. Finally, the data was stored in structured formats such as CSV and JSON, ensuring usability for downstream analytics and visualization.

The experimental results demonstrated that the proposed model achieved an accuracy of **94.12%**, significantly outperforming the existing CNN-LSTM model, which attained 88%. This improvement highlights the robustness and adaptability of the proposed hybrid approach, particularly in handling noisy, unstructured data. The model not only improves predictive performance but also contributes to interpretability, scalability, and business relevance, making it a suitable choice for real-world applications such as customer feedback analysis, business intelligence, and decision-support systems.

In conclusion, the study emphasizes that hybrid approaches, which combine clustering, classification, evolutionary structuring, and semantic representation, are more effective than traditional deep learning pipelines alone. While the proposed model shows strong promise, it also reveals areas for further exploration, such as improving recognition of specific entity types (e.g., “Person” entities) and testing the framework with varied datasets across multiple domains. Future research should also focus on comparing a wider range of machine learning and deep learning algorithms to validate generalizability and enhance the model’s robustness. Ultimately, this work contributes to advancing text analytics by demonstrating that carefully designed hybrid frameworks can deliver both accuracy and interpretability, bridging the gap between theoretical research and practical applications in unstructured data classification.

Conflict of interest

The author declares that there is no conflict of interest regarding the publication of this paper. This research is part of the author’s study and was conducted independently without any financial or personal relationships that could influence the work reported.

REFERENCES

[1] D. Magalhães, R. H. Lima, and A. Pozo, "Creating deep neural networks for text classification tasks using grammar genetic programming," *Applied Soft Computing,* vol. 135, p. 110009, 2023.

[2] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075,* 2015.

Received: August 03, 2025

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[3] X. Zhu, P. Sobihani, and H. Guo, "Long short-term memory over recursive structures," in *International conference on machine learning*, 2015: PMLR, pp. 1604-1612.

[4] J. Chen, Z. Gong, and W. Liu, "A Dirichlet process biterm-based mixture model for short text stream clustering," *Applied Intelligence,* vol. 50, no. 5, pp. 1609-1619, 2020.

[5] J. Chen, Z. Gong, and W. Liu, "A nonparametric model for online topic discovery with word embeddings," *Information Sciences,* vol. 504, pp. 32-47, 2019.

[6] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188,* 2014.

[7] P. Liu, X. Qiu, X. Chen, S. Wu, and X.-J. Huang, "Multi-timescale long short-term memory neural network for modelling sentences and documents," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2326-2335.

[8] J. Y. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," *arXiv preprint arXiv:1603.03827,* 2016.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781,* 2013.

[10] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.

[11] A. Petukhova, J. P. Matos-Carvalho, and N. Fachada, "Text clustering with large language model embeddings," *International Journal of Cognitive Computing in Engineering,* vol. 6, pp. 100-108, 2025.

[12] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information,* vol. 10, no. 4, p. 150, 2019.

[13] A. Palanivinayagam, C. Z. El-Bayeh, and R. Damaševičius, "Twenty years of machine-learning-based text classification: A systematic review," *Algorithms,* vol. 16, no. 5, p. 236, 2023.

[14] S. Verma, K. Jain, and C. Prakash, "An unstructured to structured data conversion using machine learning algorithm in Internet of Things (IoT)," in *Proceedings of the International Conference on Innovative Computing & Communications (ICICC)*, 2020.

[15] I. Bouabdallaoui, F. Guerouate, and M. Sbihi, "Hybrid text embedding and evolutionary algorithm approach for topic clustering in online discussion forums," *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal,* vol. 13, pp. e31448-e31448, 2024.

[16] M. Mojrian and S. A. Mirroshandel, "A novel extractive multi-document text summarization system using quantum-inspired genetic algorithm: MTSQIGA," *Expert systems with applications,* vol. 171, p. 114555, 2021.

[17] A. Bhavani and B. S. Kumar, "A review of state art of text classification algorithms," in *2021 5th international conference on computing methodologies and communication (ICCMC)*, 2021: IEEE, pp. 1484-1490.

[18] R. Dahiya and A. Singh, "A survey on text mining using genetic algorithm," *International Journal Of Innovative Research And Development,* vol. 3, no. 5, 2014.

Received: August 03, 2025

Volume 38 No. 11s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[19] H. Khataei Maragheh, F. S. Gharehchopogh, K. Majidzadeh, and A. B. Sangar, "A new hybrid based on long short-term memory network with spotted hyena optimization algorithm for multi-label text classification," *Mathematics,* vol. 10, no. 3, p. 488, 2022.

[20] F. S. Gharehchopogh and Z. A. Khalifelu, "Analysis and evaluation of unstructured data: text mining versus natural language processing," in *2011 5th International Conference on Application of Information and Communication Technologies (AICT)*, 2011: IEEE, pp. 1-4.

[21] A. C. Eberendu, "Unstructured Data: an overview of the data of Big Data," *International Journal of Computer Trends and Technology,* vol. 38, no. 1, pp. 46-50, 2016.

[22] Z. Liu, Y. Lin, and M. Sun, "Word Representation," 2020, pp. 13-41.

[23] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial Naive Bayes for Text Categorization Revisited."

[24] M. Abbas, K. A. Memon, A. A. Jamali, S. Memon, and A. Ahmed, "Multinomial Naive Bayes classification model for sentiment analysis," *IJCSNS Int. J. Comput. Sci. Netw. Secur,* vol. 19, no. 3, p. 62, 2019.

[25] S. Suthaharan, "Support vector machine," in *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*: Springer, 2016, pp. 207-235.

[26] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine learning*: Elsevier, 2020, pp. 101-121.

[27] B. AlBadani, R. Shi, and J. Dong, "A novel machine learning approach for sentiment analysis on Twitter incorporating the universal language model fine-tuning and SVM," *Applied System Innovation,* vol. 5, no. 1, p. 13, 2022.

[28] N. V. Otten, "Support Vector Machines (SVM) In Machine Learning Made Simple & How To Tutorial," in *Spot Intelligence*, ed, 2024.

[29] A. Fadlil, I. Riadi, and F. Andrianto, "Improving Sentiment Analysis in Digital Marketplaces through SVM Kernel Fine-Tuning," *International Journal of Computing and Digital Systems,* vol. 16, no. 1, pp. 159-171, 2024.

[30] Y. K. Alapati and K. Sindhu, "Combining clustering with classification: a technique to improve classification accuracy," *Lung Cancer,* vol. 32, no. 57, p. 3, 2016.

[31] https://www.kaggle.com/datasets/vishweshsalodkar/customer-feedback-dataset

Received: August 03, 2025