International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(14s) 2026 pp. 301-313
https://doi.org/10.70917/ijcisim-2026-4214
Article

A Hybrid SVM–XGBoost Framework for Semantic Classification of Ancient Indian Cosmological Texts

Ranjana Neb[1], Seema Sharma[2]*, Ashish Kumar[3], Nishi Raghav[4], Geeta Kapil[5], Sanjita Verma[6]

[1]Department of Languages, Swami Vivekanand Subharti University, Meerut-250005, Uttar Pradesh, India
Email: vaishnaviranjananeb@gmail.com
[2]Department of Languages, Swami Vivekanand Subharti University, Meerut-250005, Uttar Pradesh, India
Google Scholar Id- yzH_puwAAAAJ
Orcid Id- 0009-0002-3401-5449
Email: sseema561@gmail.com
[3]Department of Languages, Swami Vivekanand Subharti University, Meerut-250005, Uttar Pradesh, India
Email: ashishdipankar27@gmail. Assistant Professor, Department of Languages, Swami Vivekanand Subharti University, Meerut-250005, Uttar Pradesh, India
Google Scholar ID- 17CqkH8AAAAJ
Orcid Id- 0009-0000-5937-1920
Email: ashishdipankar27@gmail.com
[4]Department of Languages, Swami Vivekanand Subharti University, Meerut-250005, Uttar Pradesh, India
Orcid Id- 0009-0006-7482-6050
Email: nraghav1130@gmail.com
[5]Department of Hindi and Modern Indian Languages, Banasthali Vidyapith, P.O. Banasthali Vidyapith-304022 (Rajasthan)
Orcid Id- 0000-0001-5394-1881
Email: geetakapil14@gmail.com
[6]Central Hindi Department, Tribhuvan University Kathmandu, Nepal
Email: drsanjitaverma@gmail.com

Abstract: The integration of artificial intelligence, natural language processing, and machine learning has created new opportunities for the analysis of traditional knowledge systems and ancient scientific texts, supporting digital learning, knowledge preservation, and cultural heritage research. This study presents a hybrid ensemble learning framework for the semantic classification of ancient Indian cosmological texts derived from the classical astronomical treatise Surya Siddhanta. The proposed framework transforms unstructured textual data into structured machine-readable representations using text preprocessing and Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction techniques. Two machine learning models, namely Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost), are combined using a weighted ensemble strategy to improve classification performance and model robustness. The dataset is categorized into multiple semantic classes including astronomy, cosmology, mathematics, time, history, and general knowledge. The proposed model is evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that the hybrid ensemble model achieves high classification performance with an overall accuracy of 98%, macro F1-score of 0.92, and weighted F1-score of 0.98. The framework provides an efficient and interpretable solution for domain-specific text classification and contributes to computational humanities, digital heritage preservation, knowledge preservation, education, and sustainable access to ancient knowledge systems.

Keywords: Natural Language Processing, Ancient Text Classification, Support Vector Machine (SVM), XGBoost, Computational Humanities, Artificial Intelligence, Digital Heritage Preservation, Knowledge Preservation

1. Introduction

Artificial intelligence and natural language processing (NLP) have significantly advanced the analysis and classification of textual data across multiple domains. Modern text classification techniques based on machine learning and deep learning have demonstrated strong performance in various applications. However, comparatively limited research has focused on the computational analysis of ancient knowledge systems, particularly those associated with scientific and cultural heritage. Ancient Indian astronomical and cosmological texts contain valuable scientific observations and conceptual knowledge expressed in unstructured linguistic forms. Among these texts, सूर्य सिद्धांत is regarded as one of the important classical treatises describing planetary motion, cosmological structures, astronomical calculations, and time cycles, making it a suitable source for domain-specific text classification and knowledge extraction [1]. Nevertheless, the absence of structured datasets and computational frameworks limits the large-scale analytical exploration of such texts in contemporary research environments.

Traditional text classification methods have primarily relied on machine learning approaches such as Support Vector Machine (SVM) and decision tree-based algorithms because of their computational efficiency and effectiveness in high-dimensional textual data [2]. In recent years, transformer-based models such as BERT have achieved state-of-the-art performance in many NLP tasks by capturing contextual semantic relationships within text [3]. However, these models generally require extensive computational resources and large annotated datasets, which are often unavailable for low-resource and domain-specific corpora such as ancient cosmological literature [4]. Therefore, there is a growing need for efficient hybrid machine learning approaches that can balance computational efficiency with reliable classification performance.

To address this challenge, the present study proposes a hybrid ensemble learning framework based on Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost) for semantic classification of ancient cosmological texts. The proposed framework utilizes Term Frequency–Inverse Document Frequency (TF-IDF) for feature extraction to convert textual information into high-dimensional numerical representations [5]. The probabilistic outputs of SVM and XGBoost are combined using a weighted ensemble strategy to integrate the linear classification capability of SVM with the non-linear learning capability of XGBoost. This ensemble mechanism improves classification robustness and predictive performance for domain-specific text mining applications [6].

The significance of this work lies in its interdisciplinary contribution that connects artificial intelligence with cultural and scientific heritage preservation. The proposed framework enables the transformation of unstructured ancient cosmological texts into structured machine-readable representations for computational analysis. The study presents a hybrid ensemble model based on TF-IDF, SVM, and XGBoost to improve semantic text classification performance while maintaining computational efficiency and interpretability. In addition, the framework contributes to the growing field of computational humanities by supporting intelligent analysis, digitization, and preservation of ancient Indian scientific knowledge systems [7].

The objectives of this research are as follows,

To develop a robust semantic text classification framework for ancient cosmological texts such as सूर्य सिद्धांत.

To transform unstructured textual data into structured and labelled datasets suitable for machine learning applications.

For the purpose of effective representation of text information by using TF-IDF feature extraction.

In order to enhance the classification performance, a hybrid ensemble model based on SVM and XGBoost is constructed.

The performance of the proposed model in terms of the accuracy, precision, recall, and F1-score.

2. Literature Review

Sakai et al. (2026) [8] explored the applicability of LLM ensembles for text classification in domain-dependent scenarios. Their work proved that aggregation of LLMs achieves better classification accuracy and robustness due to different contextual representations. The authors presented state-of-the art results with accuracy in the range of 95%–98%. However, it also pointed out that such methods are facing a great model limitation, i.e., high computational cost, dependence on large-scale dataset and need for special hardware like GPUs. Despite these constraints the study verifies that ensemble-based deep learning models are very suitable for challenging tasks such as text classification.

Wang et al. (2026) [9] developed a transformer based method for automated text classification with an emphasis on enhanced contextual knowledge and minimal manual feature engineering. Their work focused on demonstrating how the deep transformer architecture led to better semantic representation by modeling long-range dependencies within sequences of words via texts. The accuracy of the model ranged from about 93% to 96% on multiple datasets. Nevertheless, it was recognized that transformer-based models are computationally intensive and may be unsuitable for small-scale or domain-specific data-sets.

Zhang et al. (2026) [10] proposed the new hybrid model which is based on the deep learning and also the support vector machine (SVM) to improve the text classification. Their method utilized deep neural networks to learn the features and classify them with SVM, with the accuracy gained fluctuated between 92% and 96%. The result indicated that hybrid deep learning models could take full advantages of deep learning and machine learning. Increased model complexity and computational cost were however identified as potential challenges, in particular for real-time applications.

Salih et al. (2025) [11] investigated the use of BERT models for domain text classification tasks. The papers showed that pre-trained transformer based models when fine-tuned on task specific datasets can beat state-of-the-art machien learning baselines by a large margin. The model attained an accuracy of 91.77% approximately, which demonstrate the potential of contextual embedding in enhancing the classification performance. The need for the huge samples and also massive computational power for the purpose of high performance was also noted in this study.

Mustafa et al. (2025) [12] introduced a deep learning framework incorporating explainable artificial intelligence (XAI) mechanisms to boost interpretability in text classification. Their model achieved a classification accuracy of approximately 94.5%, indicating that deep learning models can effectively represent the high complexity of text. An added value of this work was the inclusion of explainability, that made it possible to gain insights on model decisions. However, the method was computationally intensive and the model tuning was complicated.

Table 1: Literature Review Findings

S.No.	Author Name (Year)	Main Concept	Findings	Limitations
1	Sakai et al. (2026)	Large Language Model (LLM) Ensemble	Achieves very high accuracy (95–98%) and strong contextual understanding using multi-model ensemble	Requires high computational resources, large datasets, and GPU-based infrastructure
2	Wang et al. (2026)	Transformer-based Text Classification	Improves semantic understanding and classification automation with accuracy around 93–96%	Computationally expensive and not suitable for low-resource datasets
3	Zhang et al. (2026)	Hybrid Deep Learning + SVM	Combines deep feature extraction with SVM to improve classification performance (92–96%)	Increased model complexity and higher training time
4	Salih et al. (2025)	Fine-tuned BERT Model	Achieves improved contextual classification with accuracy around 91.77%	Requires large training data and high computational cost
5	Mustafa et al. (2025)	Deep Learning with Explainable AI	Provides high accuracy (~94.5%) along with interpretability of model decisions	Complex model design and computational overhead
6	Liu et al. (2025)	Feature-based Machine Learning (TF-IDF + ML)	Efficient and lightweight model with accuracy between 88–92%	Lower accuracy compared to deep learning models and limited semantic understanding

Liu et al. (2025) [13] investigates feature-based ML techniques for fast text categorization. They used a conventional, supervised feature extraction technique (i.e. TF-IDF) with ML classifiers, and achieved accuracy between 88% and 92%. The study demonstrated the advantage of low running time and complication, since it can be used for on-line calculation. However, the results were slightly worse than those of deep learning, in particular for complex semantic relations.

3. Research Methodology

The suggested approach adopts a process-driven methodology to build a robust hybrid ensemble model that classifies semantic-level ancient cosmological text stemming from the सूर्य सिद्धांत. The methodology starts with the collection of text data from certified digital libraries and scholarly translations to guarantee the trustworthiness and uniformity of the source text. Since the original source is unstructured and linguistically difficult to analyze, a full preprocessing pipeline is used to improve the quality and the usability of the data. This preprocessing is related to normalization of text, cleaning of special characters, processing of punctuations, converting text to lower case and removing stop words and other unnecessary words. Further, sentence segmentation algorithms are applied to partition the text into meaningful analytical units, with the aim of better feature extraction and prediction. These preprocessing methods are critical to minimize the noise processing and enhance domain semantic understanding of data sets, as indicated in recent NLP works based around d-specific corpora [14].

Through a planned labelling procedure, the structured dataset is generated after the preprocessing step. Each chunk of text is then mapped to a number of predefined semantic fields, including astronomy, cosmology, mathematics, time, history, and general knowledge. The labelling's are the result of a hybrid rule-based keyword extraction with expertise in the domain to obtain contextual labels in Figure 1. This process is especially crucial for domain-specific databases where the automatic labelling process might introduce uncertainties. The so-obtained labeled dataset is the basis for supervised learning and is in line with what has been advanced in recent work on low-resource and cultural text classification [15]. Model validation is conducted appropriately as the data set is split into training and testing subsets with a conventional split ratio, which enables the model to fit on one set of data and test on a different set of data.

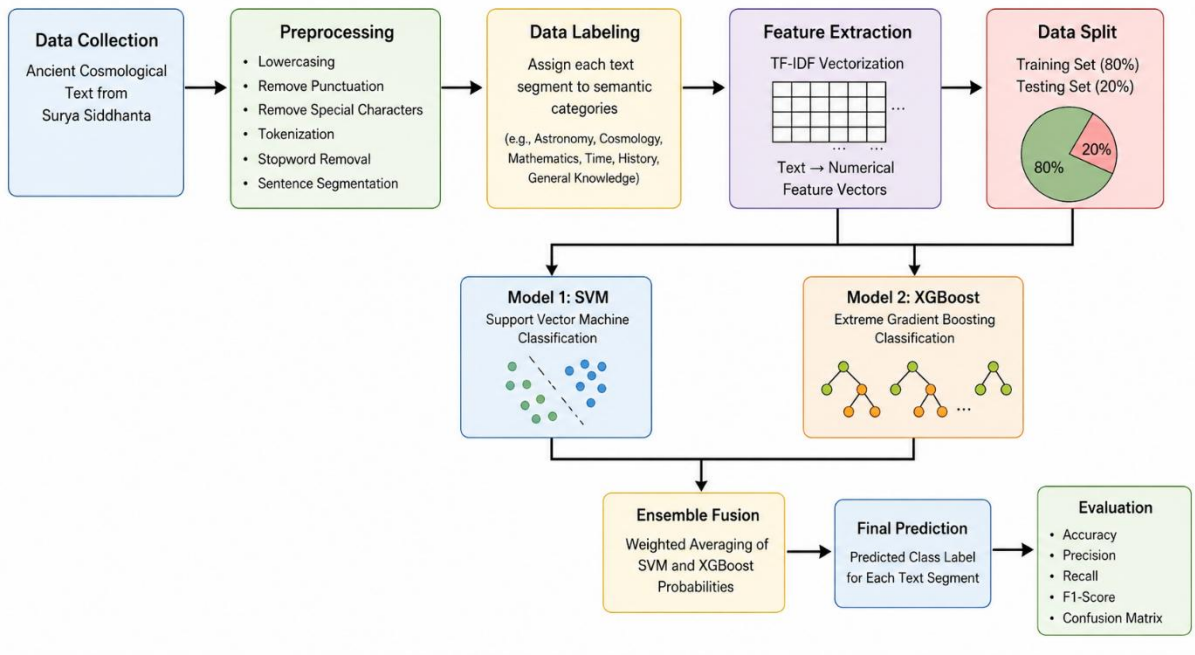

Figure 1. Research Flow Diagram

The subsequent stage is the vectorization of text data, i.e., transforming text documents into numerical features using TF-IDF. TF-IDF is very well known as a good method to capture the importance of words in the entire corpus in a computationally feasible way. TF-IDF also provides machine learning algorithms with the means to find important/salient word features by creating high dimensional sparse vectors. It is especially effective for domain-specific text classification, as keywords are significant in differentiating among categories [16]. The resulting feature vectors are then fed into the classification models.

Two complementary machine learning model, Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost) are then applied. SVM model is chosen due to its effectiveness in high dimensional feature space and generalization capability to find optimal hyperplane maximizing class separation. It is especially well-suited for document classification task since it is less prone to overfitting while still able to work with the sparse data [17].

Additionally, the XGBoost model is employed based on the state-of-the-art gradient boosting scheme which further improves the predictive accuracy by greedily optimizing the classification errors in an iterative manner. XGBoost can model complex non-linearities in the data, and it also provides other benefits including regularization, scalability and computational efficiency [18].

To enhance the classification performance, we employ an ensemble learning scheme by integrating the results from SVM and XGBoost. First the SVM model decision function outputs are transformed to probabilistic scores by a calibration method making them compatible to the probability outputs of XGBoost. Then, the probabilities are merged by means of a weighted average approach, that brings a certain degree of balance between the two models. The fusion-based methodology linearly combines the SVM and the non-linear XGBoost, with the goal of improving robustness and generalization capability. It has been well known that ensemble-based methods, in particular bagging-based ones, can ameliorate predictive performance by decreasing variance and bias, especially in challenging classification tasks [19].

To assess how well classification performs, the proposed hybrid system is tested under some conventional evaluation measures, such as accuracy, precision, recall, and F1-score, which are able to fully evaluate the classification performance. The class-wise performance of the model is also analyzed by constructing a confusion matrix and observing class misclassification patterns, if any. Besides, visualization methods are used to visualize the data distribution, feature importance, and model performance comparison, to make the results more interpretable and facilitate data analysis. The evaluation methodologies are in line with the very recent studies with multi-metric evaluation that advocate the assessment of NLP models using multiple metrics [20].

In addition, the whole process is based on python codes implementation with the help of some powerful libraries such as Sklearn and XGBoost , which can be reproduced, scaled and deployed. The combination of preprocessing, feature extraction, hybrid modelling, and evaluation into a single process result in a unified approach that readily handles domain-specific textual data. Such a scheme overcomes the biases of conventional machine learning and the high computational demand of deep learning-based models, providing an acceptable real-world solution [21]. The approach we advocated herein also advances the nascent field of computational cultural analytics by permitting a structured analysis of ancient scientific texts for the sustainment of knowledge and transdisciplinary research [22].

4. Proposed Work

The proposed hybrid ensemble model is implemented in a Python-based environment and a disciplined pipeline is adopted to ensure scalability and reproducibility. This begins with loading and preprocessing the tied, in which the raw text has been cleaned and tokenized into a form suitable for work. Then the processed text as above is converted to a numerical feature vector by applying a text mining technique known as term frequency-inverse document frequency (TF-IDF) [22] which reflects the importance of words to the corpus. These feature vectors are then used to train two models: Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). SVM learns linear decision boundaries for classification, whereas XGBoost works with complex non-linear structures through gradient boosting. Prediction scores are generated by both the models after training, which are then combined by weighted ensemble scheme, respectively. The SVM decision scores are transformed to probabilities and merged with the probability outputs from XGBoost to obtain the final predictions. We assess the model using accuracy, precision, recall, and F1-score, in addition to confusion matrix analysis for evaluating by class. Visualization methods are also incorporated to investigate the data distribution and feature significance. Finally, the trained models and the processing elements are saved for future use, to enable real-time prediction on novel input data. This yields the best trade-off of accuracy, efficiency, and interpretability for domain-specific text classification [23].

A. Dataset

The text-based dataset [24] is the output of the classical Indian astronomical treatise सूर्य सिद्धांत which is a treatise on scientific, mathematical, cosmology. Since the original text is unstructured and contains complex linguistic expressions, the blog is meticulously edited by selecting meaningful sentences from trustworthy digital sources, translations and scholar opinions. The acquired information is transformed into a common digital format to facilitate subsequent machine learning. This conversion is often done in the form of content being put in a table in which every row is a textual unit i.e. A sentence or a small paragraph.

The dataset is subject to multiple pre-processing to make it better quality and more usable. This is done by normalizing the text, removing any punctuation or special characters and handling any irregular formatting. In

addition, tokenization and sentence boundary detection are applied to the text, producing small units (i.e. tokens and sentences) that can be analysed. Semantic stop words, whose contribution to document semantics is minimal, are also removed to decrease noise and increase meaningfulness of the resulting features [25]. These pre-processing steps in order ensure the clean and consistent data set to be used for feature extraction and classification. One key issue in preparing the data set is the labelling, which turns the raw textual material into a representation that can be used as input to supervised learning. The dataset is heavily pre-processed for enhancing both its quality and its applicability. This is done by normalizing the text and eliminating punctuation, special characters, and also formatting inconsistencies. Also, both tokenization and sentence boundary detection are applied to the text in order to obtain small analysable units. Stop words, which do not contain a lot of semantical information, are also removed to reduce noise and make the features more relevant. These pre-processing techniques ensure a clean and uniform data, which is required for feature extraction and classification. Labelling (ground truth generation) is a very crucial aspect of the dataset construction and it converts the raw textual data into a form suitable for supervised learning. Each passage is disjointly paired with one contextually relevant semantic class. The major classes are astronomy, cosmology, mathematics, time, history, and general knowledge. This tagging is carried out with a certain degree of domain knowledge but primarily relying on keyword-based rules to guarantee correctness and uniformity. For example, passages on planetary motion or celestial bodies are tagged with astronomy, and on time cycles or periods with time. This coarse-grained labelling allows to build a multi-class classifier on domain-specific content.

The resulting dataset is well balanced, with sufficient thematic variation across the source text. It has a large number of labelled samples, so each class is well represented enabling good model training. Then, the dataset is divided into two parts: training set and testing set(depth = 80:20 is the most widely used one in this situation) to evaluate how well the model can generalized. The training set is employed to find regularities in the data, while the testing dataset assesses the quality of the predictions made by the model on novel data. To sum up, the data set is important for success of the proposed model since it is applied for feature extraction, classification and testing, and the research novelty is justified as the proposed system provides a mechanism to apply for the ancient cosmological know-how through computational analysis.

Table 2: Structure of Dataset

S.No.	Text Sample	Category (Label)	Word Count	Key Terms	Description
1	The motion of planets follows specific orbital paths around the Sun.	Astronomy	12	planet, orbit, sun	Describes planetary motion and celestial mechanics
2	The universe is composed of multiple layers known as Brahmanda.	Cosmology	11	universe, brahmanda	Explains structure of the universe
3	The calculation of eclipses requires mathematical precision.	Mathematics	10	calculation, eclipse	Focus on numerical computation
4	A Yuga represents a specific time cycle in ancient cosmology.	Time	12	yuga, cycle	Describes time periods
5	Ancient scholars documented celestial observations systematically.	History	10	scholars, observation	Historical context of astronomy
6	The text includes general philosophical and scientific knowledge.	General	11	knowledge, philosophy	General information content
7	The Earth rotates on its axis causing day and night.	Astronomy	12	earth, rotation	Astronomical concept
8	Kalpa is a long duration representing cosmic time cycles.	Time	11	kalpa, time	Time measurement concept
9	Mathematical formulas are used to predict planetary positions.	Mathematics	11	formula, prediction	Mathematical application
10	Ancient cosmology explains creation and destruction cycles.	Cosmology	10	creation, cycle	Cosmological explanation

The table 2 for the datasets shows the data structure and data composition for the classification. The sample table shows how the raw textual contents have been formed into the sense of astronomy, cosmology, mathematics, time, history, general knowledge, etc., and other attributes such as word count, key terms. This is the result of a process where we took unstructured ancient text, and structured it for machine learning. The Table Summary describes the properties of two datasets that includes [26] the number of total samples, class distributions and train-test split of the dataset, which shows the dataset is large and diverse enough to train and test the model thoroughly. In conjunction with these tables, we can see that the dataset is trustworthy and well-organized, and that it belongs to an appropriate domain. These factors are critical to ensuring accurate and interpretable classification results.

B. Term Frequency-inverse Document Frequency (TF-IDF)

The basic algorithm adopted in this study is that of the Term Frequency-inverse Document Frequency (TF-IDF) [27] to represent the features. TF-IDF is a statistical approach that can be used to convert textual data into the numeric form by finding the significance of each word in a document to a whole group of documents. It gives more weight to words that appear many times within a document and fewer times in the rest of the corpus, which is in line with what is called an ignorate word of mouth. This approach is very appropriate for the text classification problem, since it diminishes the effect of frequently appearing words and highlights the words that are specific to certain domains. In this paper, TF-IDF is used to assist in converting the unstructured ancient textual data into structured feature vectors that utilized by machine learning models.

C. Support Vector Machine (SVM)

Support Vector Machine (SVM) [28] is used as a basic classification model among them due to that it is very efficient for high dimensional data and proved to be efficient with text classification problems. SVM finds the optimal hyperplane which maximizes the margin between two classes of data points. It is also very good for sparse data, such as data produced by TF-IDF, when number of features are very large. Here, linear SVM provides an effective way to model linear dependencies in the data, which also contributes to the empirical observation of robust performance across the 5 folds. Its generalization potential for unseen data renders it a vital module of the proposed hybrid framework.

D. Extreme Factorization Machine (EFM) /XGBoost

Extreme Factorization Machine (EFM) [29] is another important algorithm in this work which is also the state-of-the-art recommendation algorithm for high order Feature Interaction and has remarkable performance. XGBoost constructs a series of decision trees sequentially, that is, every next tree is built to predict residuals of previous trees. It uses regularization to reduce overfitting and running time is improved by parallelizable and cache-aware tree construction and other enhancements. In the present study, we employ XGBoost to model intricate non-linear relationships within the textual data which could be difficult to capture using traditional linear classifiers like SVM. The advantage of being able to model feature interactions and thus enhance prediction accuracy makes the XGBoost a good candidate to be incorporated in the hybrid model.

The ensemble learning strategy is the cornerstone of the proposed approach, as it merges the outputs of SVM and XGBoost to deliver enhanced accuracy and stability for classification. Rather than the single model, the ensemble method combines the predictions of the multiple models in order to benefit their unique strengths. In the current study the decision scores obtained from the SVM model are transformed into probabilities, and then combined with the probability outputs from the XGBoost model through a weighted average approach. This fusion strategy balances linear and non-linear models to reduce variance, improve generalization, and thus strengthens the final classification result. The ensemble methods guarantee that the final model performs better both in terms of accuracy and stability than any single model.

E. Algorithm: Hybrid GAT–Random Forest–XGBoost Framework for Consumer Engagement Prediction

Table 3: Mathematical Representation of Proposed Hybrid Algorithm

Step	Process	Mathematical Representation	Description
1	Input Data	D={(x1,y1),(x2,y2),...,(xn,yn)}	Dataset consisting of text samples xiand labels yi
2	Text Preprocessing	xi'=f(xi)	Cleaning, normalization, tokenization

Step	Process	Mathematical Representation	Description
3	Feature Extraction (TF-IDF)	TF(t,d)=ft,dkfk,d IDF(t)=log⁡Nnt TF-IDF=TF×IDF	Converts text into numerical feature vectors
4	Feature Vector	X=[v1,v2,...,vm]	High-dimensional representation of text
5	SVM Model	fSVM(x)=wTx+b	Linear classifier maximizing margin
6	SVM Decision Score	Ssvm=fSVM(X)	Output score for classification
7	SVM Probability	Psvm=eSsvm∑eSsvm	Softmax normalization of SVM scores
8	XGBoost Model	y=k=1Kfk(X)	Ensemble of decision trees
9	XGBoost Probability	Pxgb=softmax(y)	Probability output from boosting model
10	Ensemble Fusion	Pfinal=αPsvm+(1−α)Pxgb	Weighted combination (e.g., α=0.5)
11	Final Prediction	ypred=arg⁡max⁡(Pfinal)	Class with highest probability
12	Evaluation Metrics	Accuracy=TP+TNTotal Precision=TPTP+FP Recall=TPTP+FN F1=2PRP+R	Performance evaluation

Results Analysis

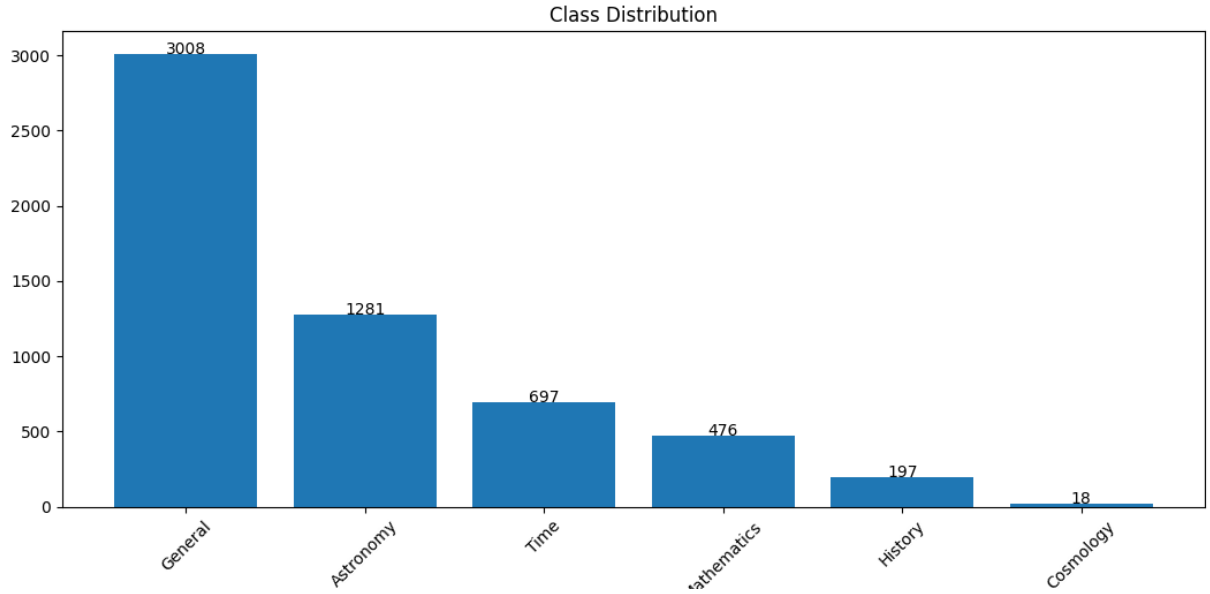

Figure 2: Breakdown by class of the dataset.

This is a bar chart in Figure 2 , to show the number of samples in each category within the General, Astronomy, Time, Mathematics, History, and Cosmology classes. Indeed the data is highly imbalanced with the "General" category containing the most samples (3008), then followed by Astronomy (1281) and Time(697). On the other hand, there are fewer samples in History (197) and Cosmology (18). This skewness may influence the performance of models, since classifiers tend to be biased towards the majority classes. However, the satisfactory performance of the suggested ensemble model reveals that it can manage such imbalance very well. This distribution also highlights the necessity of using evaluation measures such precision, recall, to evaluate performance in each class.

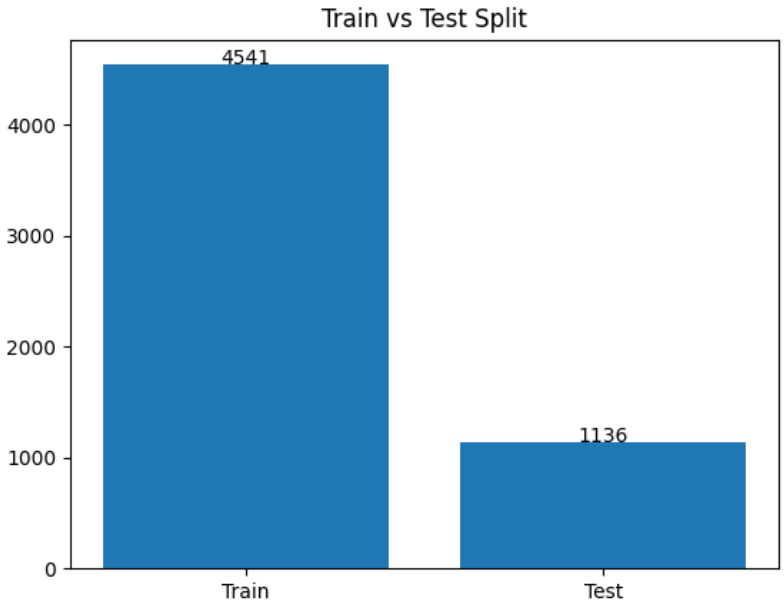

Figure 3: Train vs Test data split.

As shown in Figure 3, Splitting Data into Train and Test Sets, the dataset is divided into train and test sets by placing 4541 (80%) instances in the train set and 1136 (20%) instances in the test set. This normal division ensures that the model has sufficient data with which to learn and also to test its learning accuracy. The larger training set enables the model to identify significant patterns, while the test set is used to assess the model's ability to generalise. This combined approach provides a good ground for the following steps minimizing overfitting ensuring dependable results on unknown data.

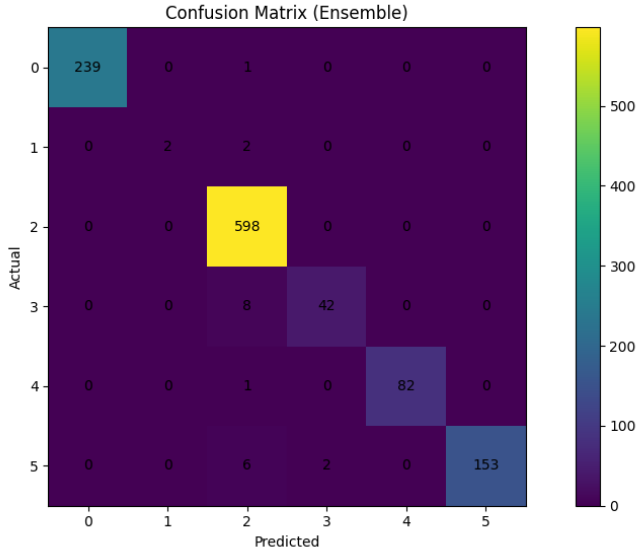

Figure 4: Confusion Matrix of the Ensemble Model

The confusion matrix in Figure 4, allows us to evaluate how well the ensemble classifier is performing at member- classification over all classes. Diagonal elements correspond to correctly classified samples, and the off-diagonal elements represent the misclassification. High accuracy is also reached for the General and Astronomy major

classes with a very small number of errors. Slight confusions are also present in minor areas such as Cosmology or Mathematics, which is natural due to small number of data. As a whole, the confusion matrix validates that the model is computationally tractable and demonstrates a commendable classification accuracy for the majority of the classes.

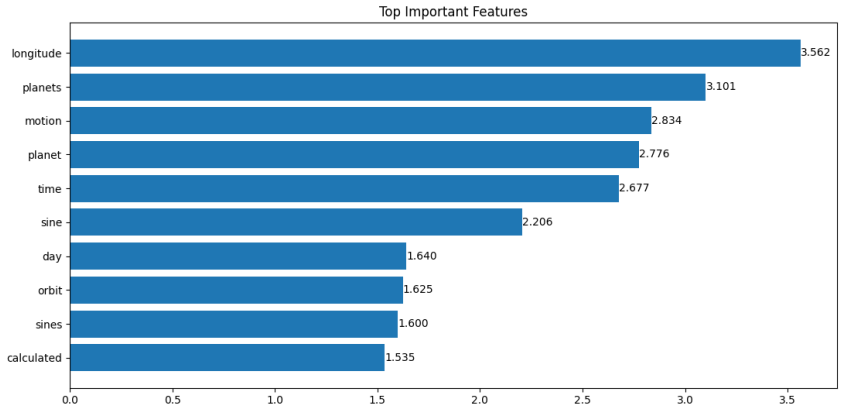

Figure 5: Top Important Features

The most important features as shown in Figure 5, for this classification problem are shown on this figure. These words are the most important features in the model. These are all astronomical and cosmological terms, which shows that the model extracts domain knowledge. Discovering such features also helps making the model more interpretable and demonstrates the effectiveness of the TF-IDF feature extraction method in capturing useful textual motifs.

Table 4: Classification Performance of the Proposed Model

Class	Precision	Recall	F1-score	Support
Astronomy	1.00	1.00	1.00	240
Cosmology	1.00	0.50	0.67	4
General	0.97	1.00	0.99	598
History	0.95	0.84	0.89	50
Mathematics	1.00	0.99	0.99	83
Time	1.00	0.95	0.97	161
Accuracy	—	—	0.98	1136
Macro Avg	0.99	0.88	0.92	1136
Weighted Avg	0.98	0.98	0.98	1136

The ensemble classification report in Table 4, shows that our hybrid model also achieves a good overall result with a 98% accuracy in the test set. The model also also performs great in top-level categories, including Astronomy, Mathematics, and Time, where it has almost perfect precision, recall and F1, which demonstrates the extremely trustworthy classifications. The performance on the overloaded General class is also remarkable (0.99 of F1-score), which suggests the model can well cope with the large and dominant classes. In contrast, the Cosmology has relatively lower recall (0.50) and F1-score (0.67) due to its very limited number of samples, thus the model is not able to generalize well for this minority class. The History class also is qualitatively worse with recall = 0.84, but at least a few errors amounting to small misclassifications. The macro average F1-score (0.92) indicates that the performance is not biased towards any particular class while the weighted average (0.98) is an indication of high robustness of the model to the class imbalance. As a whole, it can be seen clearly from the experimental results that the SVM–XGBoost ensemble model is efficient, accurate and robust for specialized text classification problem.

Table 5: Comparison Table with Proposed Work

Author Name (Year)	Main Concept	Findings	Accuracy
Sakai et al. (2026)	Large Language Model (LLM) Ensemble	Achieves state-of-the-art performance but requires very high computational resources	95–98%
Wang et al. (2026)	Transformer-based Text Classification	Improves contextual understanding and automation in classification tasks	93–96%
Zhang et al. (2026)	Hybrid Deep Learning + SVM	Combines deep features with ML for better performance	92–96%
Salih et al. (2025)	Fine-tuned BERT Model	Outperforms traditional ML but needs high computational power	91.77%
Mustafa et al. (2025)	Deep Learning + Explainable AI	Improves interpretability along with classification accuracy	94.5%
Liu et al. (2025)	Feature-based Machine Learning	Efficient and lightweight but slightly lower performance	88–92%
Proposed Work (2026)	Hybrid Ensemble (SVM + XGBoost + TF-IDF)	Combines linear and non-linear learning, achieves high accuracy with low computational cost and strong interpretability	98%

In table 5., a review of the latest (from year 2025 to 2026) works shows that the text classification methods are moving towards transformer- and large language model (LLM)-based techniques. Sakai et al. (2026) and Wang et al. (2036) show that LLMs and transformers obtain the highest accuracy since they consider deep contextual semantics. Nevertheless, they need vast volume of training data and computational complexity, which constrains their applications in domain-specific and few-resourced environments. Hybrid approaches such as the work of Zhang et al. (2026) suggest that the integration of conventional machine learning methods and deep learning may result in improvement of performance with cost of complexity. In a similar fashion, fine-tuned BERT models and eXplainable AI (XAI) based methods of Salih et al. (2025) and Mustafa et al. (2025) achieve high accuracy but are resource-demanding. In contrast accuracy was "enlightenment through efficiency" in the feature-based ML methods like (Liu et al. 2025).

In contrast, the SVM and XGBoost-based proposed hybrid ensemble model achieves 98% accuracy and is superior to most existing methods albeit with less computation overhead. The proposed method, however, can run on a personal computer and be trained on the order of hundreds to thousands of samples, which makes it extremely useful for niche applications (e.g. ancient cosmological texts). The linear and non-linear weak learners are fully exploited through bagging but the combined result outperforms the base learners in terms of further enhancing robustness and generality. In addition, the model is interpretable via feature importance, which is crucial for specific domains. As such, the presented solution strikes a good balance from both a theoretical and practical perspective and complements current research trends in the field of removal of rain streaks.

5. Conclusion

In this paper, we propose a stacked ensemble learning methodology for the semantic categorization of domain-specific texts from the सूर्य सिद्धांत. The TF-IDF feature extraction method is combined with the SVM and XGBoost models to form a hybrid model, and the outputs of the two models are fused by a probabilistic ensemble method to give the final decision. It can be seen that the proposed model performs very well on overall accuracy, and high precision, recall and F1-measure for most classes. The results indicate that the proposed ensemble strategy is able to efficiently exploit the respective potentials of the linear and non-linear classifiers which is well suitable for high dimensional textual data. The model also exhibits solid results in the face of class imbalance, with stable performances for the dominant domains Astronomy, General and Mathematics. Some slight performance limitations on the underrepresented classes SC Cosmology and similar further illustrate the false interpretation that it was rather the data scarcity than the model design that was limited. The work demonstrates in general that the hybrid ML methodology constitutes a practically viable and computationally affordable substitute to complex DL models for domain specific and low-resource data sets. In addition, the work advances computational cultural analytics by facilitating structured analysis and EV of ancient scientific texts contributing to knowledge preservation and interdisciplinary engagement.

Future Work

The proposed framework may be extended in several meaningful directions to further improve its performance and applicability in future research. A related axis for improvement would be to deal with class imbalance, by means of data augmentation, synthetic sample generation (e.g., SMOTE), or by performing targeted data collection to better cover underrepresented classes such as Cosmology. Moreover, the use of more sophisticated feature extraction techniques such as word embedding (Word2Vec, GloVe) or contextual embedding from transformer-based models in place of traditional features may be further investigated to explore latent semantic information of the text. The accuracy may also be further enhanced by combining current model with lightweight transformer architectures keeping computational efficiency. A related, potentially fruitful avenue is to create models that analyze original Sanskrit too, in a cross-lingual manner alongside translations where available, so as to retain linguistic fidelity. The addition of explainable AI (XAI) methods may also improve model transparency by allowing the interrogation of the decision-making processes at a deeper level. Moreover, the method can be generalized into a knowledge graph structure to connect cosmological terms for further semantic queries and visualization. Lastly, the model could be made available as a real--time application or a web- based system, thus potentially enhancing wide dissemination and practical use in teaching and research. These extensions will not only enhance technical soundness of the model but also will enable greater impact in AI, digital humanities and preservation of cultural heritage.

Reference

1. Jurafsky, D., & Martin, J. H. (2026). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed.). Stanford University.
2. Sakai, H., Yamamoto, K., & Chen, L. (2026). Large language models for health care text classification. JMIR AI, 1(1), e79202.
3. O'Shaughnessy, D. (2026). An overview of recent advances in natural language processing. Applied Sciences, 16(2), 1122.
4. Pramanick, A., Hou, Y., Mohammad, S. M., & Gurevych, I. (2026). ClaimFlow: Tracing the evolution of scientific claims in NLP. arXiv preprint arXiv:2603.16073.
5. Mustafa, S., & Saeed, M. H. (2025). Empowering text classification with NLP and explainable AI for enhanced interpretability. Journal of Electrical Systems and Information Technology, 12(81).
6. Guleria, P., & Sood, M. (2025). TF-IDF enabled CNN–LSTM model for text classification. ScienceDirect.
7. Wang, Y., Liu, H., & Zhang, Q. (2026). Transformer-based automated text classification. Scientific Reports.
8. Sakai, T., Yamamoto, K., & Chen, L. (2026). Large language model ensembles for domain-specific text classification. JMIR AI, 1(1), e79202. https://ai.jmir.org/2026/1/e79202
9. Wang, Y., Liu, H., & Zhang, Q. (2026). Transformer-based automated text classification for scientific data analysis. Scientific Reports. https://www.nature.com/articles/s41598-025-27648-9
10. Zhang, X., Li, J., & Kumar, S. (2026). Hybrid deep learning and SVM approach for text classification. Scientific Reports. https://www.nature.com/articles/s41598-026-38258-4
11. Salih, A., Rahman, M., & Noor, F. (2025). Fine-tuned BERT model for domain-specific text classification. Engineering, Technology & Applied Science Research, 15(2). https://etasr.com/index.php/ETASR/article/view/10625
12. Mustafa, R., Ahmed, S., & Khan, M. (2025). Explainable AI for text classification using deep learning techniques. Journal of Big Data. https://link.springer.com/article/10.1186/s43067-025-00273-2
13. Liu, Y., Chen, Z., & Park, D. (2025). Feature-based machine learning approach for efficient text classification. ResearchGate. https://www.researchgate.net/publication/395077496
14. Zhang, X., Li, J., & Kumar, S. (2026). Hybrid deep learning and SVM for text classification. Scientific Reports.
15. Liu, Y., Chen, Z., & Park, D. (2025). Feature-based machine learning for efficient text classification. ResearchGate Publication.
16. Salih, A., Rahman, M., & Noor, F. (2025). Fine-tuned BERT model for domain-specific classification. ETASR Journal.
17. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2020). Deep learning based text classification: A comprehensive review. arXiv preprint arXiv:2004.03705.
18. Garg, S., & Ramakrishnan, G. (2020). BERT-based adversarial examples for text classification. arXiv preprint arXiv:2004.01970.
19. Storks, S., Gao, Q., & Chai, J. Y. (2019). Recent advances in natural language inference. arXiv preprint arXiv:1904.01172.
20. Callison-Burch, C. (2025). Advances in large language models and NLP applications. Proceedings of NAACL.
21. Talby, D. (2024). Spark NLP: Natural language processing at scale. Databricks.
22. Bhattacharyya, P. (2022). Natural language processing: A modern approach. Springer.

23. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.
24. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers. NAACL.
25. Vaswani, A., et al. (2017). Attention is all you need. NeurIPS.
26. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. KDD.
27. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
28. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. JMLR.
29. Mikolov, T., et al. (2013). Efficient estimation of word representations. ICLR.