



DESIGN OF EFFICIENT FRAMEWORK OF INFORMATION AGGREGATION FROM VARIOUS WEBSITES USING WEB CONTENT MINING TECHNIQUES

DOI: <https://doi.org/10.5281/zenodo.23162610>

Dr. Narendra M. Jathe¹, Ms. Lavanya S. Gahale²

*Corresponding Author: ¹Assistant Professor, Department of Computer Science, Smt. Narsamma Arts, Commerce and Science College, Amravati, Maharashtra, India
njathe@gmail.com

ORCID : <https://orcid.org/0000-0003-2927-7882>

²Research Scholar, Department of Computer Science, Smt. Narsamma Arts, Commerce and Science College, Amravati, Maharashtra, India
lavanyagahale22@gmail.com

ORCID: <https://orcid.org/0009-0009-5551-3608>

ABSTRACT

The exponential growth of data generated from conventional databases and web-based sources has created significant challenges in effective data management, integration, and utilization. With the continuous evolution of digital technologies and online services, organizations are increasingly adopting adaptive strategies to extract meaningful insights from diverse and distributed data sources. Information aggregation has emerged as a crucial service in this context, enabling the consolidation, analysis, and presentation of relevant data from multiple platforms to enhance accessibility and support decision-making.

Traditionally, intermediaries such as newspapers, magazines, and professional journals served as aggregators. However, the emergence of web portals and digital platforms has significantly transformed this role by enabling seamless and large-scale integration of information. Information aggregation plays a vital role in the development of knowledge-based systems across various domains, including healthcare, economics, applied sciences, artificial intelligence, robotics, and machine learning.

Despite the vast amount of information available on the web, users often encounter difficulties due to data dispersion, heterogeneous formats, and geographical distribution. Aggregating such information involves addressing complex challenges related to data comparison, relationships, and both intra- and inter-organizational integration. Although existing techniques such as screen scraping, direct data feeds, and XML-based encoding are widely used, there is still considerable scope for improving aggregation methodologies. Universities represent a significant yet underexplored domain for information aggregation. While academic institutions publish extensive public information on their websites, this content is often fragmented across multiple sources. Therefore, this research proposes an efficient framework for aggregating information from various websites using web content mining techniques, with the aim of enhancing data accessibility, usability, and value creation for diverse user groups.

KEYWORDS: Information Aggregation, Web Content Mining, Knowledge-Based Systems, Data Integration, Web Data Extraction, Decision Support Systems, Digital Platforms.

1. INTRODUCTION

Data mining aims to extract implicit, previously unknown, and useful knowledge from large volumes of incomplete, noisy, and heterogeneous data. Traditional data mining techniques were primarily developed for structured and homogeneous datasets stored in conventional databases. However, with the rapid growth of the Internet, most data now exist on the web in diverse forms such as text, hyperlinks, multimedia, and semi-structured formats. These characteristics limit the applicability of traditional techniques, leading to the emergence of web data mining, which integrates data mining with web technologies to effectively handle complex web data [1]. By incorporating methods from artificial intelligence, information retrieval, and natural language processing, web mining has evolved into a specialized and independent research field [2].

Web mining focuses on discovering valuable patterns and insights from the vast, interconnected, and dynamic web environment. Unlike conventional data mining, it addresses challenges related to data heterogeneity, unstructured content, and hyperlink structures. It enables the extraction of knowledge from web page content, link structures, and user interaction logs, thereby supporting the efficient utilization of web information [3]. As part of the Knowledge Discovery in Databases (KDD) process, web mining plays a crucial role in identifying meaningful patterns from web data, where manual analysis is infeasible due to the scale and complexity involved [4].

Web mining is broadly classified into web content mining, web structure mining, and web usage mining based on the type of web data analyzed [5]. Among these, web content mining is particularly significant for extracting relevant information from distributed web sources. Information aggregation further enhances this process by consolidating partial and relevant content from multiple web pages, enabling efficient access to dynamic data and supporting real-time decision-making in applications such as personalized services and online analysis [7, 55].

2. CONTRIBUTION OF THE PAPER

❖ **Development of a unified web information aggregation framework:**

The study proposes an integrated framework for collecting, preprocessing, categorizing, and aggregating information from multiple university websites. The framework combines web content mining, feature extraction, machine learning-based classification, validation, and database integration into a single workflow.

❖ **Creation of a university web-page dataset:**

A dataset consisting of 513 web pages collected from official websites of non-agricultural universities in Maharashtra was developed. The collected pages were processed and organized according to their content and functional relevance.

❖ **Categorization of heterogeneous web content:**

The proposed framework categorizes university web pages into predefined information categories, enabling scattered institutional information to be systematically organized and retrieved.

❖ **Comparison of multiple feature representation techniques:**

Three approaches—Boolean Model, Vector Space Model (VSM), and VSM Weight Change Model—are investigated to determine their effectiveness in representing web-page content for classification and aggregation.

❖ **Integration of machine learning classifiers:**

Multiple classification approaches, including Support Vector Machine, k-Nearest Neighbor, Linear Discriminant Analysis, decision trees, ensemble methods, and neural networks, are evaluated for web-page categorization.

❖ **Evaluation using multiple validation strategies:**

The study evaluates the classification models using **cross-validation, holdout validation, and resubstitution validation**. This provides a broader assessment of model behavior rather than relying on a single evaluation method.

❖ **Database-oriented information integration:**

The framework provides a structured mechanism for storing webpage metadata, extracted features, category labels, and classification results, thereby supporting efficient retrieval and future expansion of the aggregated information system.

❖ **Identification of effective approaches for web-page categorization:**

The experimental analysis provides comparative evidence regarding the behavior of different feature representations and classifiers. In particular, the results indicate that weighted vector representations and ensemble-based approaches can provide strong classification performance for heterogeneous university web content.

❖ **Foundation for scalable information aggregation:**

The proposed framework provides a basis for extending university information aggregation to larger datasets, additional institutions, new categories, adaptive feature weighting, and more advanced machine-learning techniques.

3. LITERATURE REVIEW

The rapid expansion of digital data across multiple domains such as education, healthcare, business, and social media has created an urgent need for effective techniques to extract meaningful information from large and heterogeneous datasets. Data mining, web mining, text mining, and information aggregation have emerged as key research areas to address this challenge by transforming raw data into actionable knowledge for decision-making.

1.1. DATA MINING

In today's digital era, individuals and organizations interact with vast volumes of data generated from diverse sources and stored in multiple formats, including text documents, numerical records, multimedia content, and hypertext. The heterogeneous nature of this data necessitates efficient processing techniques to ensure accurate analysis, timely decision-making, and preservation of data integrity. Efficient retrieval and analysis of such data are critical, particularly in environments where rapid responses to user or customer needs are required. This analytical process is commonly referred to as data mining or Knowledge Discovery in Databases (KDD). Data mining focuses on discovering hidden, previously unknown, and potentially valuable patterns from large datasets. The growing interest in data mining within the information technology sector arises from the challenge of converting massive volumes of raw data into meaningful information. This challenge is often described as being "data rich but information poor," highlighting the importance of advanced analytical techniques capable of extracting useful knowledge from extensive data repositories.

In real-world scenarios, enormous amounts of data are generated daily across various sectors such as education, healthcare, business, and industry. This data holds significant potential for knowledge discovery and informed decision-making. For example, educational institutions can analyze student data to identify dropout trends, while businesses can examine sales and transaction data to uncover customer behavior patterns. However, despite the abundance of available data, transforming it into actionable insights remains a complex task due to data volume, variety, and complexity.

Elloumi (2002) emphasized that data mining tools play a vital role in predicting future trends and behaviors, enabling organizations to make proactive, knowledge-driven decisions rather than relying solely on retrospective analysis [8]. These tools are capable of uncovering hidden patterns that are often overlooked by traditional analytical systems and human experts. Han et al. (2006) defined data mining as the systematic process of extracting valuable knowledge from large volumes of data and equated it with the broader concept of Knowledge Discovery in Databases [9]. Chuchra (2012) highlighted data mining as an efficient technique for extracting hidden patterns from large databases, enabling fast data retrieval and improving decision-making efficiency [10]. Dean et al. (2012) pointed out that as datasets continue to grow in size and complexity, there is a critical need to reduce the time required to move from data analysis to deployable models [11]. Agrawal (2013) described data mining as a non-trivial process that extracts previously unknown and potentially useful information from large relational databases, helping organizations identify correlations and improve profitability through better understanding of customer behavior [12].

1.2.WEB MINING

The World Wide Web is considered the largest global repository of information, containing massive amounts of unstructured and semi-structured data. If mined effectively, this data serves as a valuable resource for businesses and individuals. Web data is inherently heterogeneous and dynamic, making traditional data processing techniques insufficient for meaningful analysis. Etzioni (1996) defined web mining as the application of data mining techniques to automatically discover and extract information from web documents and services [3]. The rapid growth of web-based information sources and the expansion of e-commerce have significantly increased research interest in this area. Web mining aims to answer whether it is feasible to extract meaningful and actionable knowledge from the vast and continuously evolving web environment. Web mining involves several core tasks, including resource discovery, information extraction, pattern generalization, and analysis. Resource discovery focuses on locating relevant information from the web, while information extraction selects and preprocesses specific data. Generalization identifies patterns across multiple web resources, and analysis evaluates the extracted patterns for practical use [3], [13].

Chakrabarti et al. (2000) explored mining hypertext data, emphasizing that web data exceeds traditional structured data in both volume and complexity [14]. Representing hypertext as directed graphs, the study reviewed supervised, semi-supervised, and unsupervised learning techniques and highlighted challenges in handling semi-structured data. Kosala and Blockeel (2000) categorized web mining into web content mining, web structure mining, and web usage mining, based on the type of data analyzed [13].

Cooley et al. (2002) presented a comprehensive taxonomy of web mining tools and techniques, addressing challenges such as user identification, session tracking, and behavior modeling [15]. Eirinaki et al. (2003) focused on web personalization by analyzing user behavior through web logs and applying clustering and classification techniques to enhance business intelligence and customer retention [16]. Minoru et al. (2004) proposed advanced clustering techniques for web documents, emphasizing the importance of incorporating HTML tag information for improved accuracy [17].

Further research addressed semantic web mining [18], schema-based web data extraction [19], and multi-database mining techniques to reduce search costs across distributed data sources [20]. Association rule mining techniques were applied to both relational databases and text data to uncover meaningful relationships [21], [22]. Studies also explored content filtering [23], user behavior modeling [24], [25], and ranking mechanisms such as PageRank and HITS to analyze hyperlink structures and identify authoritative web pages [26], [27]. Despite significant advancements, web data mining faces numerous challenges, including data cleaning, integration of heterogeneous sources, duplicate data elimination, scalability, overfitting, high-dimensional data handling, security, and privacy concerns [28]. These challenges highlight the need for more efficient and scalable mining frameworks.

1.3.TEXT MINING

Text mining focuses on extracting meaningful patterns and knowledge from unstructured or semi-structured textual data. With the exponential growth of text data generated through digital libraries, social media, emails, and web platforms, traditional data mining tools are often inadequate for handling text-based information.

Fan et al. (2006) described text mining as a multidisciplinary field that integrates information retrieval, data mining, machine learning, statistics, and computational linguistics to extract significant patterns from textual data sources [29]. Weiss et al. (2010) emphasized that text mining techniques are widely applied in areas such as search engines, customer relationship management, fraud detection, and sentiment analysis [30]. Sagayam (2012) highlighted the difficulty of extracting meaningful knowledge from unstructured text using conventional mining tools [31]. Zhong et al. (2012) outlined the generic text mining process, which includes data collection, preprocessing, feature extraction, pattern analysis, and decision support [32]. Several studies proposed clustering-based models, hybrid partitioning approaches using TF-IDF, and concept-based mining frameworks to improve text representation and analysis [33], [34], [35]. Applications of text mining include crime detection systems [36], biomedical text analysis using named entity recognition [37], and pattern discovery using SVM and BM25 techniques [38]. Advanced techniques such as information extraction [39], natural language processing [40], clustering [41], and text summarization [42] play a critical role in transforming large volumes of textual data into concise and meaningful representations. However, challenges related to ambiguity, synonymy, polysemy, and scalability remain unresolved, necessitating further research [43].

1.4.INFORMATION AGGREGATION

Information aggregation has emerged as a vital outcome of mining techniques, enabling the synthesis of relevant information from multiple heterogeneous sources. This approach improves accessibility, reduces redundancy, and enhances decision-making processes. Bouchon-Meunier (1998) introduced aggregation techniques that consider multiple criteria, imperfect knowledge, and data fusion from various sources [44]. Berry (2003) proposed semantic modeling approaches for aggregating text-based document collections using probabilistic and vector space models [45]. Subsequent research focused on resource aggregation in distributed systems [46], context-aware aggregation [47], and cluster-based aggregation operators for decision support [48]. Information aggregation has also been applied to cloud computing infrastructures [49], web page grouping [50], distributed query execution [7], climate modeling [51], healthcare interoperability [52], traffic management systems [53], and big data processing frameworks [54]. Tasaki (2012) proposed a web-based aggregation system capable of capturing partial content from multiple web pages to extract key information efficiently [55]. These studies demonstrate that information aggregation plays a crucial role in managing large-scale,

distributed, and heterogeneous data environments. However, challenges related to scalability, data quality, privacy, and real-time processing continue to motivate further research in this area..

4. RESEARCH METHODOLOGY

This study proposes a unified framework for the categorization and aggregation of web pages from non-agricultural universities in Maharashtra. The framework integrates web content mining, feature representation models, supervised machine learning classifiers, and database integration to handle heterogeneous and unstructured web data efficiently.

1.5. METHODOLOGICAL OVERVIEW

The proposed methodology follows a structured pipeline comprising five stages: data acquisition, preprocessing, feature extraction, pattern categorization, and database integration. This end-to-end pipeline ensures robustness, scalability, and consistency in handling institutional web data.

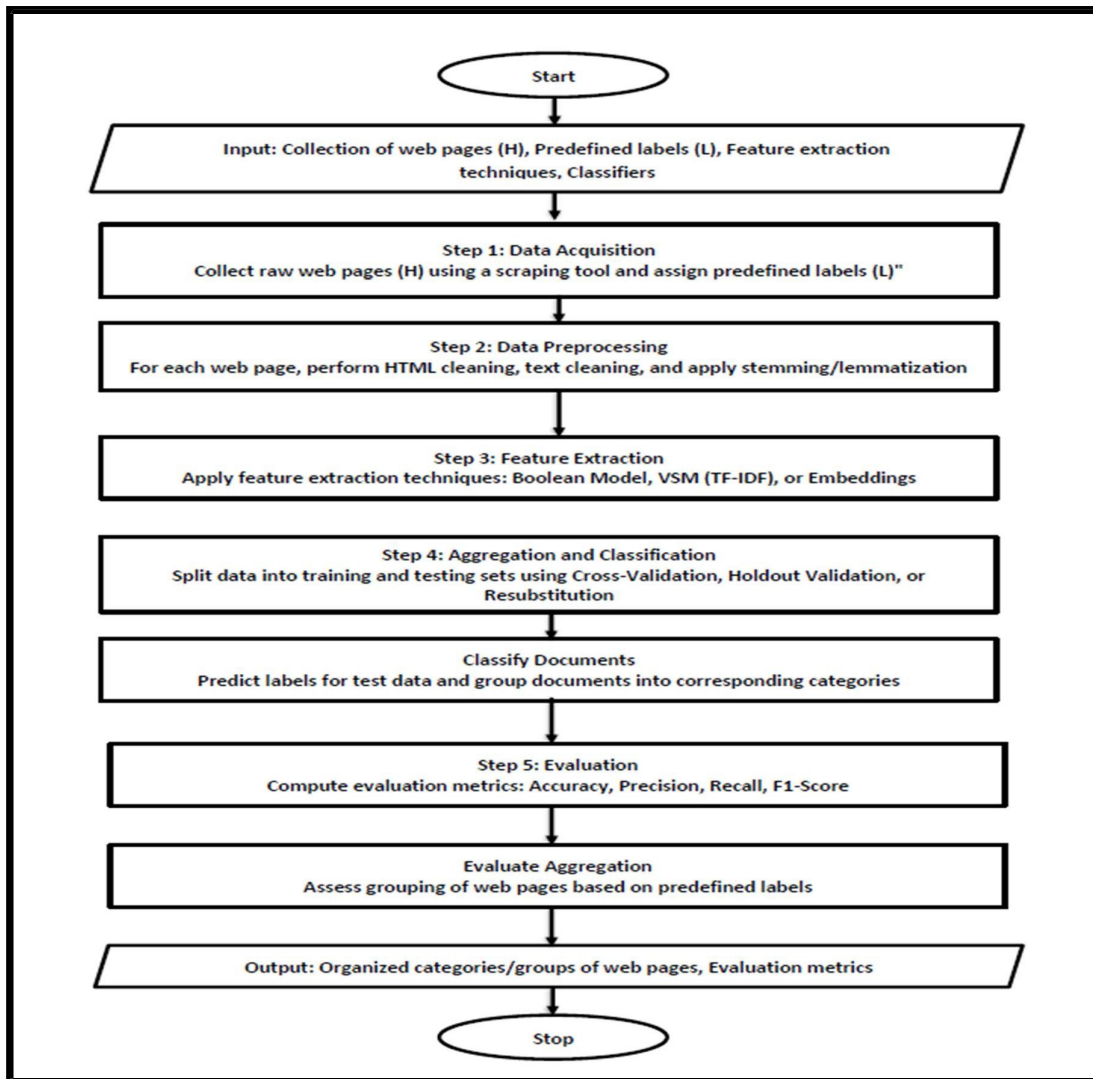


Figure 01: Unified Aggregator-Classifier Framework for Information Aggregation

Stages of the Methodology

- **Data Acquisition:** Web pages are collected using the HTTrack API.
- **Preprocessing:** HTML cleaning, normalization, tokenization, stopword removal, and stemming/lemmatization.
- **Feature Extraction:** Textual data is represented using the Boolean Model, Vector Space Model (VSM) with TF-IDF, and normalized VSM (weight change).
- **Pattern Categorization:** Supervised classifiers such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Neural Networks are employed.
- **Database Integration:** Structured storage of web pages, feature vectors, and classification results.

1.6. DATA ACQUISITION

A dataset of 513 web pages was collected from official websites of non-agricultural universities using the HTTrack API. HTTrack recursively downloads static web content while preserving site structure. The collected dataset includes HTML content, metadata, and textual information.

Challenges addressed: structural heterogeneity, noisy content, and exclusion of dynamic elements.

1.7. PREPROCESSING

Preprocessing transforms raw HTML documents into structured textual data suitable for analysis. It includes:

Removal of HTML tags, scripts, and irrelevant elements

- Text normalization (lowercasing, punctuation removal)
- Tokenization and stopword elimination
- Stemming or lemmatization
- Mathematically, preprocessing is modeled as a sequence of transformations applied to raw HTML content to obtain a cleaned textual representation.

1.8. FEATURE EXTRACTION

Three feature representation techniques are employed:

Boolean Model: Represents documents as binary vectors indicating the presence or absence of terms.

Vector Space Model (VSM): Uses TF-IDF weighting to capture term importance across documents.

VSM Weight Change (Normalization): Applies L2 normalization to TF-IDF vectors to eliminate bias due to document length.

These representations enable efficient similarity computation, classification, and aggregation.

1.9. PATTERN CATEGORIZATION

The categorized dataset is split into training and testing subsets using validation strategies such as holdout and k-fold cross-validation. Supervised classifiers are trained on extracted features to assign predefined labels (e.g., *About Us*, *Academic Calendar*, *Contact Us*). Performance is evaluated using accuracy, precision, recall, and F1-score.

1.10. DATABASE REPRESENTATION AND INTEGRATION

A relational database is designed to store webpage metadata, feature vectors, and classification outcomes. The schema supports scalability through indexing and modular design, enabling efficient querying, traceability of results, and integration of future datasets or models.

5. PATTERN CATEGORIZATION, VALIDATION AND DATABASE CREATION

Pattern categorization is a fundamental task in machine learning that involves assigning data instances to predefined classes based on their inherent characteristics. In the context of information aggregation, pattern categorization plays a crucial role in organizing data collected from diverse and distributed web sources. The rapid growth of web-based information has increased the need for robust categorization mechanisms, particularly in applications such as information retrieval, personalized services, academic data organization, and decision support systems.

Aggregated web data is often noisy, incomplete, and inconsistent, which necessitates effective preprocessing and feature extraction techniques to achieve meaningful classification. Furthermore, the diversity in data formats—ranging from unstructured textual content to semi-structured HTML documents—requires adaptive machine learning models capable of handling high dimensionality and evolving data distributions. The dynamic nature of web information also demands scalable solutions that can accommodate continuous updates and emerging patterns.

To address these challenges, this chapter employs a combination of supervised and unsupervised learning algorithms, including Support Vector Machines (SVM), k-means clustering, ensemble learning techniques, and neural networks. Boolean feature representation and TF-IDF vectorization are utilized to capture both term presence and term importance, thereby enabling effective discrimination among categories. The categorized patterns are subsequently validated using statistically grounded performance metrics to ensure reliability and robustness before being integrated into a structured database.

5.1 FUNDAMENTALS OF PATTERN RECOGNITION

Pattern recognition focuses on identifying regularities and structures within data and assigning them to appropriate categories. Its foundations lie in statistical analysis, machine learning, and artificial intelligence, and it has evolved significantly with advances in computational power and algorithmic design. Modern pattern recognition techniques are widely applied across domains such as healthcare, finance, cybersecurity, and academic information systems.

At the core of pattern recognition are features, which represent measurable characteristics extracted from raw data. The quality of selected features directly influences classification performance. In supervised learning scenarios, data instances are associated with predefined class labels, enabling models to learn decision boundaries during the training phase and generalize to unseen data during testing. The separation of training and testing phases is essential to mitigate overfitting and assess model generalization.

Pattern recognition tasks can be broadly classified into supervised, unsupervised, semi-supervised, and reinforcement learning paradigms. Supervised learning relies on labeled datasets to learn mappings between features and classes, while unsupervised learning identifies inherent structures without explicit labels. Semi-supervised learning leverages both labeled and unlabeled data, and reinforcement learning focuses on learning optimal actions through interaction with an environment. Each paradigm offers unique advantages depending on data availability and application requirements.

5.2 MATHEMATICAL BACKGROUND OF PATTERN CATEGORIZATION

The machine learning algorithms used in this study are grounded in well-established mathematical and statistical principles. From a supervised learning perspective, these algorithms can be categorized into discriminative, generative, and hybrid models. Discriminative models, such as SVMs and decision trees, learn decision boundaries directly by optimizing loss functions and are effective for complex feature-label relationships [56]. Generative models, such as Naïve Bayes, learn joint probability distributions $P(x, y)$, enabling them to handle missing data and model uncertainty under the assumption of conditional independence among features [57]. Neural networks act as hybrid models, combining discriminative classification with implicit feature representation learning [58].

From a statistical standpoint, the models fall into parametric, non-parametric, and ensemble-based categories. Parametric models assume predefined data distributions, such as the Gaussian likelihood used in Naïve Bayes:

$$P(x|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x - \mu_y)^2}{2\sigma_y^2}\right).$$

offering computational efficiency and interpretability [57].

Support Vector Machines optimize a margin-based objective function:

$$\min \frac{1}{2} ||w||^2, \quad s.t. y_i (w \cdot x_i + b) \geq 1.$$

with kernel functions enabling non-linear decision boundaries [59].

Non-parametric methods, such as k-Nearest Neighbors, rely on distance measures like Euclidean distance:

$$d(x, x') = \sqrt{\sum_{i=1}^n (x_i - x'_i)^2},$$

and are particularly suitable for high-dimensional text data when combined with cosine similarity [60].

Decision trees recursively partition the feature space using impurity measures such as Gini impurity:

$$G = \sum_{k=1}^K p_k(1 - p_k),$$

Ensemble methods improve robustness by combining multiple learners. Bagging averages predictions:

$$\hat{f}(x) = \frac{1}{M} \sum_{m=1}^M f_m(x),$$

reducing variance and overfitting [61]. Boosting iteratively refines weak learners:

$$\hat{f}_m(x) = \hat{f}_{m-1}(x) + \alpha_m f_m(x),$$

emphasizing misclassified instances to enhance performance [62].

5.3 MODEL VALIDATION

Model validation is essential to ensure that trained classifiers generalize well to unseen data, particularly in information aggregation scenarios characterized by noisy and heterogeneous sources. Several validation strategies are employed to evaluate model stability and predictive performance, including train-test splitting, k-fold cross-validation, stratified cross-validation, leave-one-out validation, holdout validation, and resubstitution validation [56], [63]–[66].

Resubstitution validation computes accuracy on the training data:

$$Accuracy_{resub} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(\hat{y}_i = y_i),$$

where $\mathbb{I}(\cdot)$ is the indicator function. Although computationally efficient, this method provides an optimistic performance estimate and is complemented with cross-validation techniques to assess generalization.

Model performance is evaluated using multiple metrics, including accuracy, precision, recall, F1-score, AUC-ROC, and log-loss. Accuracy is defined as:

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

while precision and recall are given by:

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN}$$

The F1-score balances precision and recall, and AUC-ROC measures the model's discrimination ability across thresholds [67]–[70]. Log-loss evaluates probabilistic confidence in predictions [71].

5.4 DATASET COLLECTION, CREATION AND ARRANGEMENT

The dataset was collected from official websites of non-agricultural universities in Maharashtra to support the categorization of academic web pages. Web crawling tools were used to extract content-rich HTML pages while excluding irrelevant file types. Dynamic content posed challenges due to JavaScript rendering, which were addressed using advanced crawling tools and manual interventions where necessary. Duplicate pages and noisy content were removed using hashing and similarity-based filtering techniques to ensure dataset quality.

Categorization was performed based on content relevance, functional purpose, structural cues, and semantic similarity. Manual labeling established a reliable ground truth, which was subsequently validated through expert review and automated semantic consistency checks. Balanced representation across 13 predefined categories was ensured to prevent classification bias, with augmentation techniques applied where required. The cleaned and labeled dataset was organized in both hierarchical storage and relational database formats to support scalability, traceability, and efficient querying. Metadata integration enhanced transparency and facilitated downstream information aggregation tasks.

6. RESULTS AND DISCUSSION

This chapter presents a comprehensive experimental evaluation of the proposed information aggregation framework using three feature extraction approaches: the Boolean Model, the Vector Space Model (VSM), and the VSM Weights Change Model. The experiments assess the effectiveness of various machine learning classifiers under cross-validation, holdout validation, and resubstitution validation techniques. The evaluation is conducted using a dataset consisting of 540 extracted features derived from textual and categorical properties of web pages.

6.1 PERFORMANCE EVALUATION USING BOOLEAN FEATURE EXTRACTION

The Boolean-based aggregation framework demonstrates strong classification capabilities across multiple validation techniques. Under cross-validation, ensemble-based classifiers and Linear Discriminant Analysis (LDA) consistently outperform other methods. In particular, Subspace Discriminant ensembles and LDA achieve superior accuracy across higher fold validations, indicating robustness and stability in Boolean feature spaces. Neural network configurations show moderate performance, with medium and wide architectures yielding better results compared to narrow models. Support Vector Machines (SVMs) and k-Nearest Neighbor (KNN) classifiers exhibit comparatively lower accuracy, highlighting their limited suitability for sparse Boolean representations.

The holdout-validation experiments further reinforce the effectiveness of ensemble methods and LDA. Narrow neural networks demonstrate competitive accuracy under moderate holdout settings, while boosted ensembles perform consistently across different training-testing splits. However, classifiers

such as Naïve Bayes and coarse KNN variants exhibit reduced performance, reflecting sensitivity to data distribution.

Resubstitution-validation results reveal exceptionally high accuracy for multiple classifiers, including Bagged Trees, Fine KNN, Weighted KNN, LDA, and neural networks, with accuracy reaching up to 98.5%. While these results indicate excellent model fitting, they also emphasize the inherent limitation of resubstitution validation, which tends to overestimate performance due to testing on training data. Nevertheless, these findings confirm the strong learning capability of Boolean-based aggregation models.

Overall, the Boolean feature space proves highly effective for precise web page aggregation, particularly when combined with ensemble learning techniques and discriminant-based classifiers.

6.2 PERFORMANCE EVALUATION USING VECTOR SPACE MODEL (VSM)

The Vector Space Model introduces a richer representation of web page content by incorporating term frequency and weighted textual features. Under cross-validation, ensemble classifiers—especially Bagged Trees—demonstrate superior performance, achieving the highest accuracy among all tested models. Decision tree variants also show strong and consistent results, highlighting their adaptability to high-dimensional feature spaces. Neural networks achieve moderate success, with narrow architectures performing better than deeper configurations.

Holdout-validation experiments reveal that ensemble methods maintain high accuracy under smaller holdout ratios, although performance declines as the proportion of held-out data increases. LDA demonstrates strong generalization ability, particularly under higher holdout settings, while neural networks show variable behavior depending on architecture. KNN and SVM classifiers struggle to maintain stable performance, indicating sensitivity to feature dimensionality and data sparsity.

Resubstitution-validation using VSM yields near-perfect accuracy for multiple classifiers, including ensemble methods, LDA, KNN variants, and neural networks. These results confirm the strong representational power of VSM features but also reiterate the need for complementary validation techniques to assess real-world applicability.

In summary, the VSM-based aggregation framework significantly enhances classification accuracy compared to the Boolean model, particularly when paired with ensemble learning and decision tree classifiers.

6.3 PERFORMANCE EVALUATION USING VSM WEIGHTS CHANGE MODEL

The VSM Weights Change Model introduces adaptive weighting to further refine feature importance. Cross-validation results indicate that ensemble methods and decision tree classifiers remain the most effective, achieving the highest accuracy across multiple folds. Fine and Medium Tree classifiers show consistent improvements, while neural networks demonstrate moderate adaptability to weighted features. LDA maintains stable performance but does not surpass ensemble-based approaches.

Holdout-validation experiments confirm the robustness of Bagged Trees and tree-based classifiers, especially under smaller holdout ratios. Neural networks achieve reasonable accuracy under limited holdout conditions, while KNN and SVM classifiers exhibit inconsistent and generally lower performance, suggesting difficulty in handling dynamically weighted feature spaces.

Resubstitution-validation results again show near-perfect accuracy for ensemble methods, LDA, and selected KNN and neural network configurations. While these results highlight strong model fitting, they also reinforce the necessity of cross-validation and holdout testing to evaluate generalization performance accurately.

The Vector Space Model (VSM) was employed to represent web page content in a high-dimensional numerical space, where textual features extracted through web content mining were transformed into weighted vectors. This representation captures both the semantic relevance and distribution of terms across web documents, making it more expressive than Boolean feature extraction. The performance of the proposed information aggregation framework using VSM was evaluated using multiple machine learning classifiers under cross-validation, holdout validation, and resubstitution validation techniques.

6.4 CROSS-VALIDATION ANALYSIS

Cross-validation results indicate that VSM significantly enhances classification performance across most classifiers due to its ability to encode term importance and contextual relevance. Ensemble-based classifiers demonstrate the highest and most consistent accuracy across increasing fold values. Bagging-based classifiers, in particular, show strong robustness and stability, suggesting their effectiveness in handling high-dimensional and sparse textual feature spaces generated by VSM.

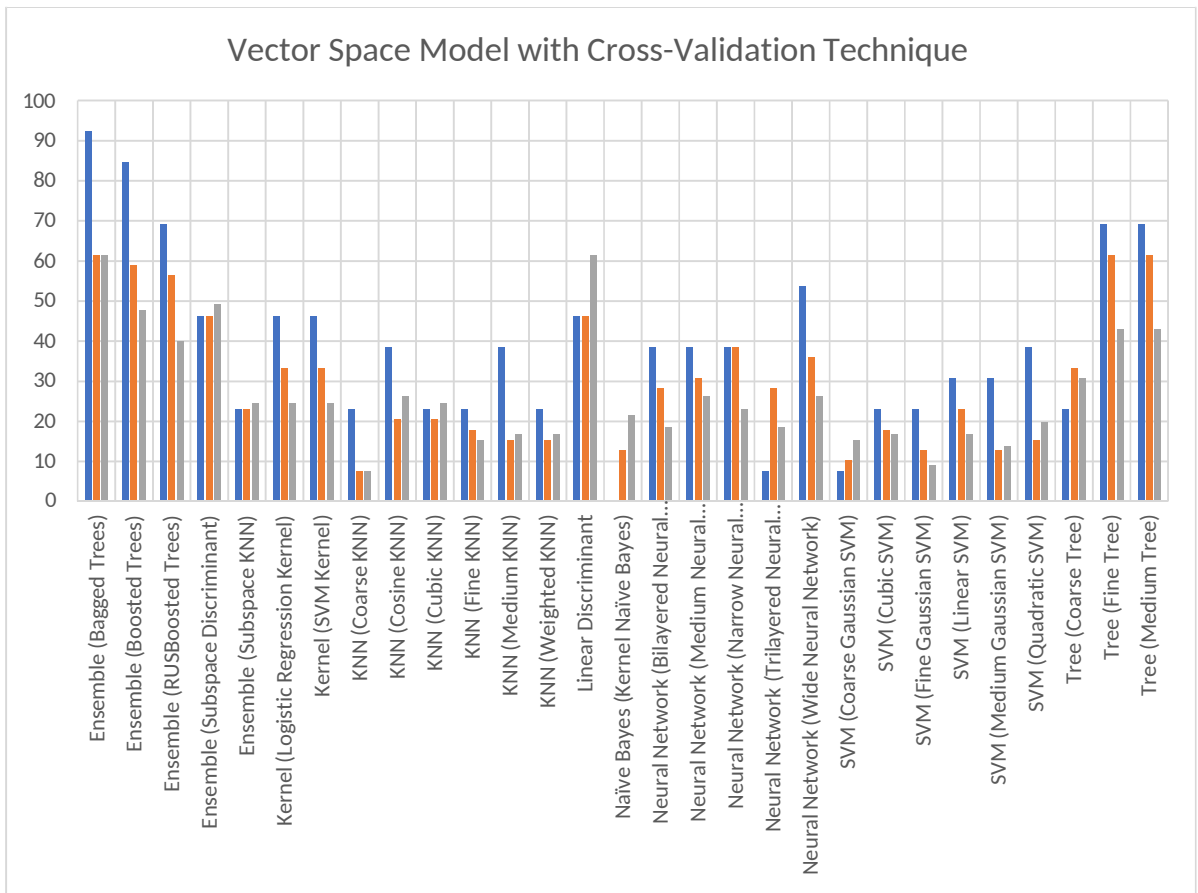


Figure 02: Vector Space Model with Cross-Validation Technique

Decision tree classifiers also perform well under cross-validation, benefiting from the hierarchical splitting of weighted features. Fine and medium decision tree models show better generalization compared to coarse variants, indicating that finer splits help capture subtle differences in web content representation. Linear Discriminant Analysis maintains competitive performance, reflecting its capability to separate classes effectively in transformed vector spaces.



Figure 03: Confusion Matrix – Vector Space Model with Cross-Validation technique using Ensemble (Bagged Trees) (3-Fold)

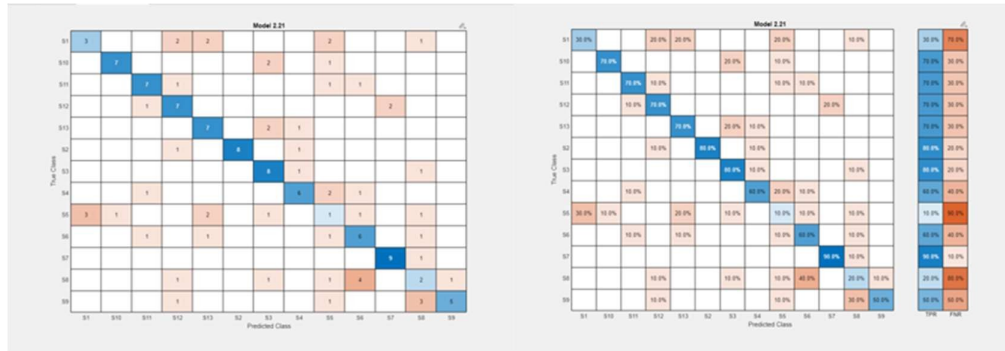


Figure 04: Confusion Matrix – Vector Space Model with Cross-Validation technique using Ensemble (Bagged Trees) (5-Fold)

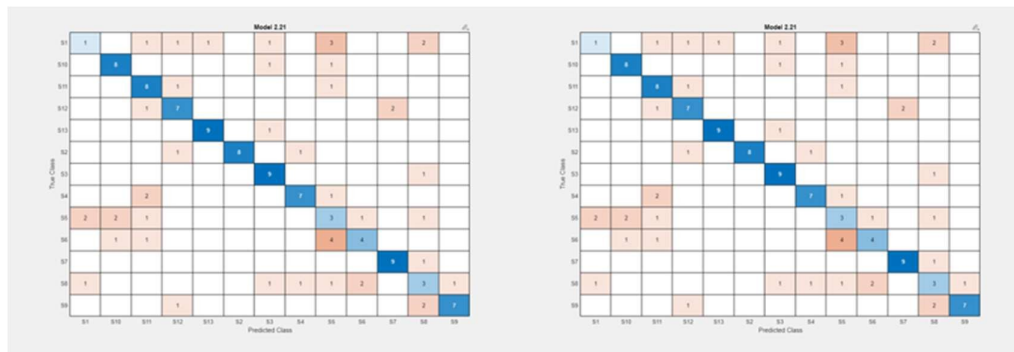


Figure 05: Confusion Matrix – Vector Space Model with Cross-Validation technique using Ensemble (Bagged Trees) (8-Fold)

Neural network classifiers exhibit moderate performance under VSM. Narrow neural networks outperform wider architectures, suggesting that simpler models generalize better when trained on vector-based web content features. Support Vector Machines and k-Nearest Neighbor classifiers demonstrate comparatively lower accuracy, which can be attributed to sensitivity to feature dimensionality and distance metrics in dense vector spaces.

Overall, cross-validation results confirm that VSM improves the discriminative power of the aggregation framework, particularly when combined with ensemble learning and tree-based classifiers.

6.5 HOLDOUT VALIDATION ANALYSIS

The holdout validation experiments further evaluate the generalization capability of VSM-based models by varying training and testing splits. Results reveal that ensemble classifiers maintain high accuracy under lower holdout ratios, indicating strong learning from limited training data. However, as the size

of the testing set increases, a gradual decline in performance is observed, which is expected due to reduced training information.

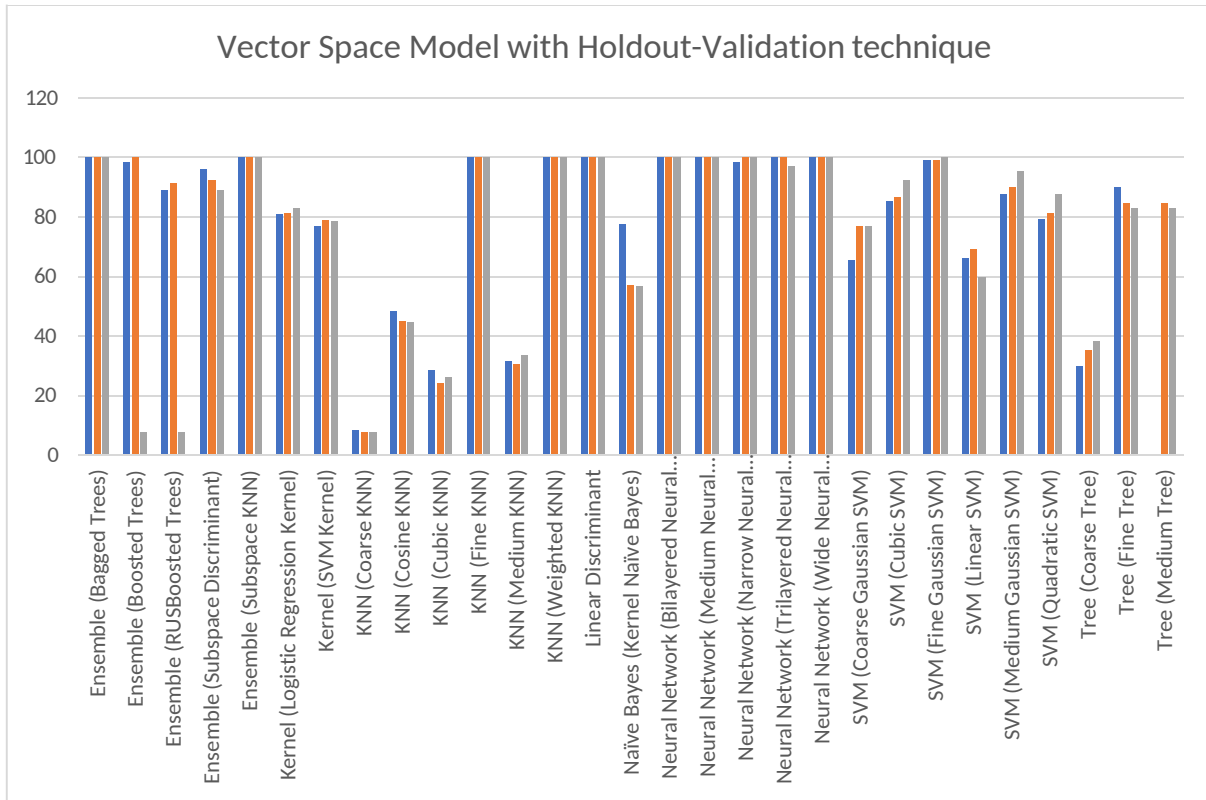


Figure 06: Vector Space Model with Holdout-Validation technique

Linear Discriminant Analysis shows stable performance across different holdout configurations, demonstrating robustness and consistency in vector-based feature environments. Decision tree classifiers continue to perform well, although slight performance variations are observed depending on tree depth and feature distribution.

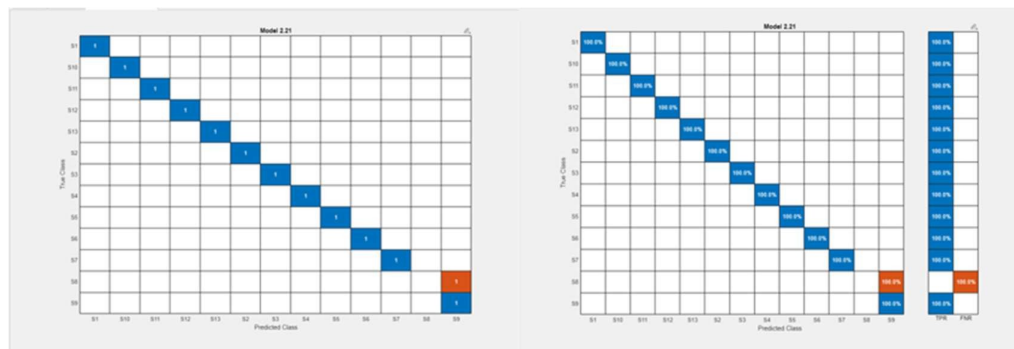


Figure 07: Confusion Matrix – Vector Space Model with Holdout-Validation technique using Ensemble (Bagged Trees) (10-Holdout)

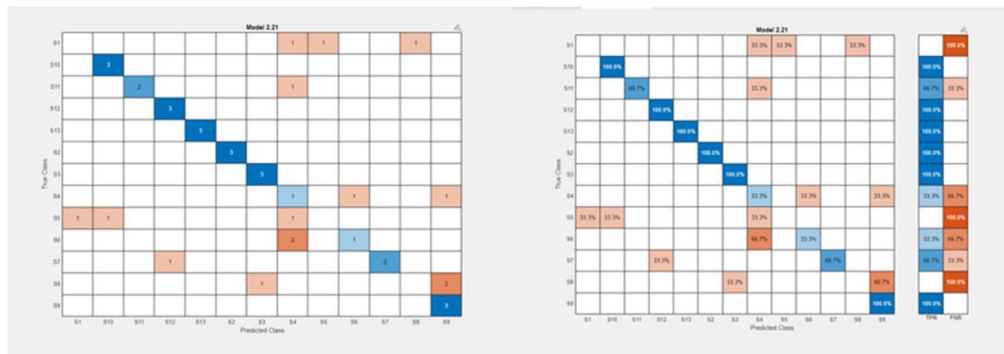


Figure 08: Confusion Matrix – Vector Space Model with Holdout-Validation technique using Ensemble (Bagged Trees) (30-Holdout)

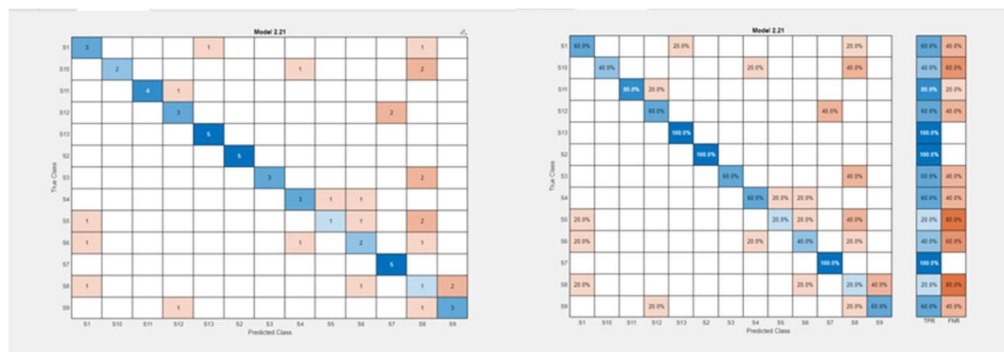


Figure 09: Table **Error! No text of specified style in document..1**: Confusion Matrix – Vector Space Model with Holdout-Validation technique using Ensemble (Bagged Trees) (50-Holdout)

Neural network classifiers display fluctuating accuracy across different holdout ratios, suggesting sensitivity to training data size. Narrow networks remain more stable compared to deeper architectures. KNN and SVM classifiers again show limited adaptability to VSM features, particularly under higher holdout conditions, reflecting challenges in distance-based classification of high-dimensional text vectors.

These findings highlight that VSM-based aggregation systems benefit from classifiers capable of managing complexity, feature weighting, and dimensionality effectively.

6.6 RESUBSTITUTION VALIDATION ANALYSIS

Resubstitution validation yields near-perfect classification accuracy for most classifiers using VSM features. Ensemble methods, decision trees, Linear Discriminant Analysis, KNN variants, and neural networks achieve exceptionally high accuracy, demonstrating their strong ability to learn and fit vector-based representations.

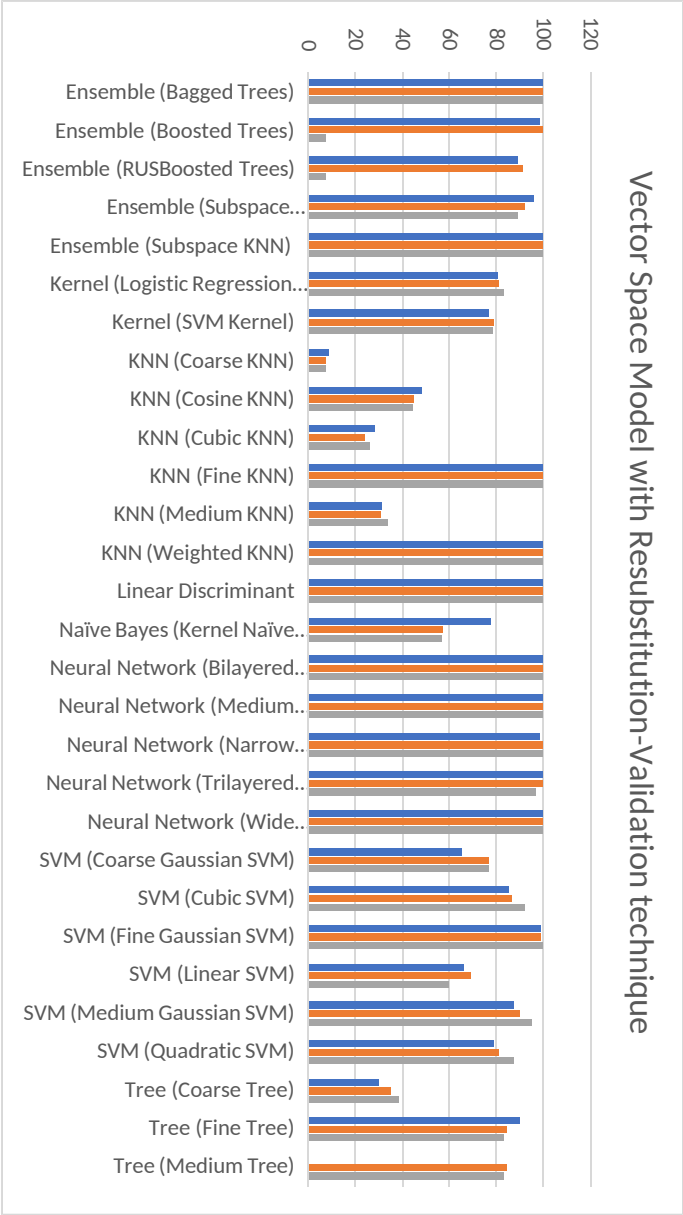


Figure 10: Vector Space Model with Resubstitution-Validation technique

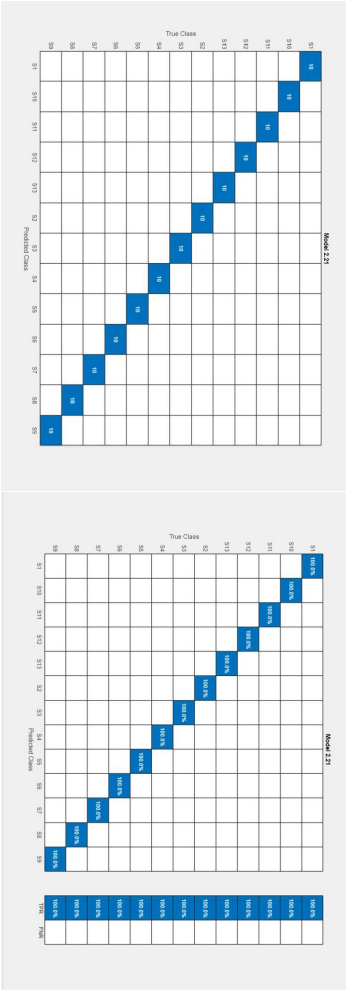


Figure 11: Confusion Matrix – Vector Space Model with Resubstitution-Validation technique using Ensemble (Bagged Trees) (0% of SetAside)

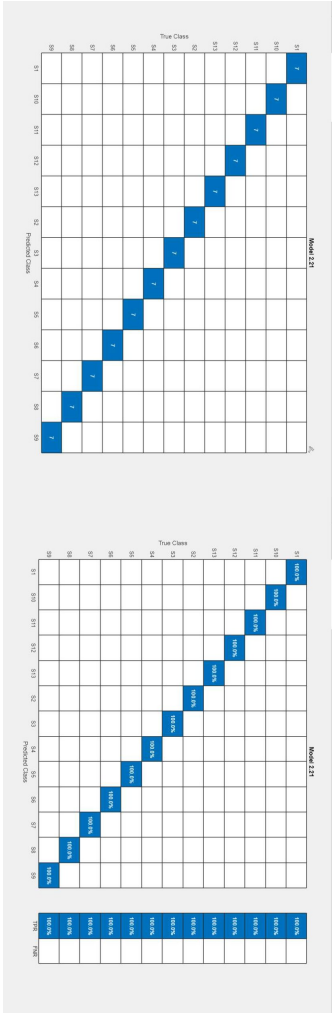


Figure 12: Confusion Matrix – Vector Space Model with Resubstitution-Validation technique using Ensemble (Bagged Trees) (30% of SetAside)

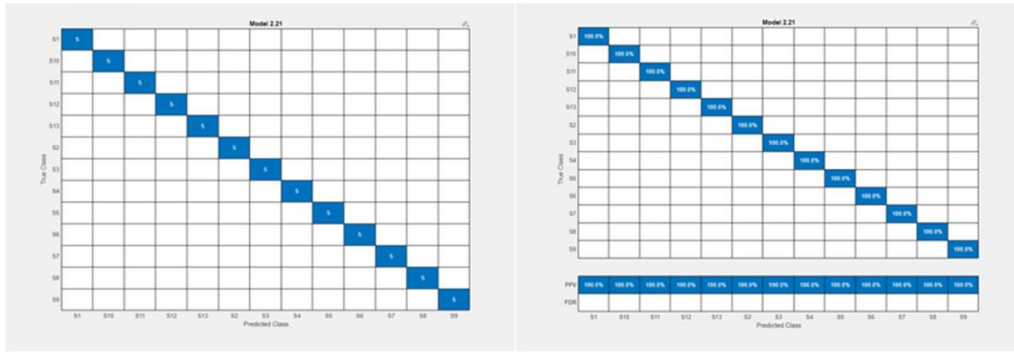


Figure 13: Confusion Matrix – Vector Space Model with Resubstitution-Validation technique using Ensemble (Bagged Trees) (50% of SetAside)

While these results indicate excellent model learning capacity, they also emphasize the inherent limitation of resubstituting validation, as testing on training data can lead to over-optimistic performance estimates. Nevertheless, the consistently high accuracy confirms the representational strength of VSM in capturing semantic and structural characteristics of web content.

6.7 DISCUSSION AND INSIGHTS

The experimental results clearly demonstrate that VSM provides a richer and more informative representation of web data compared to Boolean models. By incorporating weighted textual features, VSM enables classifiers to distinguish subtle semantic variations across web pages, improving aggregation accuracy.

Ensemble learning techniques emerge as the most effective approach for VSM-based information aggregation, offering robustness, stability, and high predictive accuracy. Decision tree classifiers also show strong adaptability, particularly in modeling complex relationships among weighted features. Linear Discriminant Analysis proves to be a reliable and computationally efficient alternative with consistent performance.

In contrast, distance-based classifiers and margin-based methods such as KNN and SVM show limitations when applied to high-dimensional vector spaces derived from web content. Neural networks demonstrate potential but require careful architecture selection to avoid overfitting or instability.

Overall, the VSM-based evaluation validates the effectiveness of the proposed framework for aggregating heterogeneous web information. The results confirm that integrating vector-based text representation with appropriate classification techniques significantly enhances the accuracy and scalability of web content mining systems.

7. CONCLUSION

This study evaluated web page aggregation using Boolean and Vector Space Models with multiple classifiers and validation techniques. The Boolean model showed excellent performance with classifiers like Bagged Trees, Linear Discriminant Analysis, Fine and Weighted KNN, and Neural Networks, achieving accuracies up to 98.5%. The Vector Space Model demonstrated strong results with ensemble methods, decision trees, and neural networks, especially under resubstitution validation, highlighting its ability to handle high-dimensional features. Cross-validation and holdout methods provided insights into model generalization, while the VSM Weights Change Model confirmed adaptability to weighted feature spaces. Overall, ensemble methods, LDA, and neural networks consistently outperformed others, making them highly effective for precise web page aggregation. These findings emphasize the robustness of feature-based classification approaches and provide a foundation for future improvements, including hybrid models, adaptive weighting, and scalable solutions for large datasets.

8. CONFLICT OF INTEREST

commercial, institutional, or personal conflicts of interest that could have influenced the research, its analysis, interpretation, or publication.

9. RESEARCH INTEGRITY STATEMENT

The authors declare that this manuscript is original and has not been previously published or submitted for publication elsewhere. The research has been conducted and reported in accordance with applicable academic and ethical standards. All sources used in the study have been appropriately acknowledged and cited. The authors have complied with the journal's requirements regarding plagiarism and the responsible use of artificial intelligence tools.

10. REFERENCES

1. Kavita P. Mahani and N. Ruhil, ""Web data mining: A perspective of research issues and challenges", Proceeding of IEEE 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, pp. 3235-3238, 2016".
2. G. S. V. K. K. Sharma, ""Web mining: Today and tomorrow", Proceeding of IEEE 3rd International Conference on Electronics Computer Technology, pp. 399-403, 2011."
3. O. Etzioni, ""The World Wide Web: Quagmire or gold mine", Communications of the ACM, Vol. 39, No. 11, pp. 65-68, 1996."
4. G. P.-S. P. S. Usama Fayyad, "From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence, AI MAGAZINE, pp. 37-54, 1996."
5. R. C. J. S. B. Mobahser, ""Automatic Personalization based on Web Usage Mining", Communications of the ACM, Vol. 43, No. 8, 2000."

6. K. S. R. Malarvizhi, "'Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study", pp. ,," International Journal of Computer Trends and Technology (IJCTT),, Vols. Vol. 4, No. 8, , pp. 2940-2945, August 2013..
7. R. Gupta and K. Ramamritham, "'Scalable Execution of Continuous Aggregation Queries over Web Data",," IEEE Internet Computing, , vol. Vol. 16, no. Issue 1, pp. 43–51, , 2012.
8. M. Elloumi, "'Knowledge Discovery and Data Mining",," Knowledge Based Systems Journal, Elsevier Publishing Co., Amsterdam, North-Holland, , Vols. Vol. 15, Issue 4,, April 2002..
9. J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2nd Ed.,, 2006..
- Er. Rimmy Chuchra, "'Use of Data Mining Techniques for the Evaluation of Student Performance:A Case Study",," International Journal of Computer Science and Management Research, , Vols. Vol.1, No. 3, , October 2012..
10. D. D. a. W. T. Jared Dean, "'A New Age of Data Mining in the High-Performance World",," SAS Global forum on Data Mining, 2012."
11. Shivam Agarwal, "'Data Mining Concepts and Techniques",," in Proceedings of IEEE 2013 International Conference on Machine Intelligence Research and Advancement, pp. 203-207, , 2013..
12. H. B. R. Kosala, "'Web Mining Research: A Survey",," SIGKDD Explorations: Newsletter of Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, , vol. Vol. 2, pp. 1-15, 2000..
13. B. D. P. I. Soumen Chakrabarti, "'Data mining for Hypertext",," in Proceedings of IEEE/WIC/ACM International Conference on Web, Vol. 2, No. 1, pp. 29-39, , June 2000..
14. C. R. & S. J. Mobasher B., "'Automatic Personalization based on Web usage mining",," in Communications of the ACM, Vol.43, No. 8, pp. 142-151, , 2002..
15. M. Eirinaki and M. Vazirgiannis, "'Web Mining for Web Personalization",," in in ACM Transactions on Internet Technology (TOIT), Vol.3, No.1, pp. 1-27, , February 2003..
16. H. S. Minoru S., "'Web Document Clustering Using Threshold Selection Partitioning",," in Proceedings of NTCIR-4 , Tokyo, pp. 11-15, , April 2004..
17. P. R. F. T. J. A. P. Y. K. P. Ding L., "'Finding and Ranking Knowledge on the Semantic Web",," The Semantic Web – ISWC 2005, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, , Vols. Vol. 3729, , 2005..
18. W. K. N. A. S. Zhao Li, "'Web data extraction based on structural similarity",," Processidings of Knowledge and Information Systems, Springer Publication, , Vols. Vol. 8, No. 4, , pp. pp. 438-461, , 2005..
19. X. W. a. C. Z. Shichao Zhang, "'Multi-Database Mining",," in IEEE Computational Intelligence Bulletin, Vol.2, pp. 5-13, , June 2005..
- 20.

- Dr. Abdallah Alashqur, "Mining Association Rules: A Database Perspective", IJCSNS
21. International Journal of Computer Science and Network Security, Vols. Vol. 8 No.12, pp. pp. 69-74, December 2008..
 - D. R. N. I. a. F. T. Hany Mahgoub, "A Text Mining Technique using Association Rules Extraction", International Journal of Computational Intelligence, Vols. Vol. 4, No. 1, pp. pp. 21-28, 2008..
 - D. S. G. M. Manisha Marathe, "Extracting Content Blocks from Web Pages", International Journal of Recent Trends in Engineering, Vols. Vol. 2, No. 4, pp. pp. 62-64, 2009..
 - V. & S. R. S. Vijayalakshmi, "Mining of User's Access Behaviour For Frequent Sequential Pattern From Web Logs", International Journal of Database Management Systems (IJDMS), Vols. Vol.2, No.3, pp. pp.31-45, August 2010..
 - O. S. B. Naveena Devi, "Dynamic Modelling Approach for Web Usage Mining Using Open Web Resource", International Journal of Engineering Science and Technology, Vols. Vol. 2, No. 10, pp. pp. 5605-5610, 2010..
 - Ms. M. Sangeetha, Dr. K. Suresh Joseph, "Page Ranking Algorithms used in Web Mining", in Proceedings of IEEE International Conference on Information Communication and Embedded Systems, pp. 297-303, February 2014. .
 - W. Yang, "An Improved HITS Algorithm Based on Analysis of Web Page Links and Web Content Similarity", in Proceedings of International Conference on Cyberworlds (CW), Chongqing, China, pp. 147- 150, 2016..
 - D. P. T. D. Jayalatchumy, "Web Mining Research Issues and Future Directions – A Survey", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727, Vol. 14, No. 3, pp. 20-27, Sep. - Oct. 2013.
 - L. W. S. R. Z. Z. W. Fan, "Tapping the power of text mining", Communications of the ACM, Vols. Vol. 49, No. 9, pp. pp. 76–82, 2006. .
 - N. I. T. Z. F. D. S. M. Weiss, "Text mining: predictive methods for analyzing unstructured information", Springer Science and Business Media, 2010..
 - R. Sagayam, "A survey of text mining: Retrieval, extraction and indexing techniques", International Journal of computational Engineering Research, Vols. Vol. 2, No. 5, 2012..
 - Y. L. S. W. N. Zhong, "Effective pattern discovery for text mining", IEEE transactions on knowledge and data engineering, Vols. Vol.24, No. 1, pp. pp. 30–44, 2012..
 - D. S. B. Laxman, "Improved method for pattern discovery in text mining", International Journal of Research in Engineering and Technology, Vols. Vol. 2, No. 1, pp. pp. 2321–2328, 2013..
 - V. S. R. Rajendra, "A Novel Modified Apriori Approach for Web Document Clustering", International Journal of Computer Applications, pp. pp. 159–171, 2013..
 - M. C. K. Sumathy, "Text mining: Concepts, applications, tools and issues-an overview", International Journal of Computer Applications, Vols. Vol. 80, No. 4, 2013..

36. C. Z. C. P. Chen, "“Data-intensive applications, challenges, techniques and technologies: A survey on big data”,” Information Sciences Journal, Elsevier publication, , Vols. Vol. 275, , pp. pp. 314–347, , 2014. .
37. H. M. M. S. V. D. M. D. A. Henriksson, "“Synonym extraction and abbreviation expansion with ensembles of semantic spaces”,” Journal of biomedical semantics, , Vols. Vol. 5, No. 1, , pp. pp. 1-7, , 2014. .
38. J. K. P. J. Joby, "“Accessing accurate documents by mining auxiliary document information”,” in Proceedings of IEEE Second International Conference on Advances in Computing and Communication Engineering (ICACCE), pp. 634–638, , 2015. .
39. M. B. R. Agrawal, "“A detailed study on text mining techniques”,” International Journal of Soft Computing and Engineering (IJSCE) ISSN 2231–2307, , Vols. Vol. 2, No. 6, , pp. pp. 118-121, , January 2013..
40. A. P. J. M. J. P. J. T. F. E. A. Calvillo, " “Searching research papers using clustering and text mining”,” in Proceedings of IEEE International Conference in Electronics, Communications and Computing (CONIELECOMP), 2013, pp. 78–81, , 2013.
41. B. L. Narayana and S. P. Kumar, "A new clustering technique on text in sentence for text mining," International Journal of Science Engineering and Advance Technology,, Vols. Vol. 3, No. 3,, pp. pp. 69–71,, 2015..
42. B. A. Mukhedkar, D. Sakhare and R. Kumar, "Pragmatic analysis based document summarization," International Journal of Computer Science and Information Security, Vols. Vol. 14, No. 4,, pp. pp. 145,, 2016..
43. P. L. M. E. R. King Gary, "“Computer-Assisted Keyword and Document Set Discovery from Unstructured Text”,” American Journal of Political Science, , Vols. Vol. 61, No. 4, , pp. pp. 971–988, , 2017..
44. B. Bouchon-Meunier, "“Aggregation and fusion of imperfect information”,” Physica-Verlag Publishing, , 1998.
45. M. W. Berry, "“Survey of Text Mining I: Clustering, Classification, and Retrieval”,” Springer; 1st edition, , 2003..
46. P. Kokkinos and E. Varvarigos, "“Resource Information Aggregation in Hierarchical Grid Networks”,” 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, , pp. pages 268–275, , 2009..
47. A. Manzoor, H.-L. Truong and S. Dustdar, "“Quality Aware Context Information Aggregation System for Pervasive Environments”,” in International Conference on Advanced Information Networking and Applications Workshops, pages 266–271, , 2009..
48. T. Boongoen and QiangShen., "“Nearest-Neighbor Guided Evaluation of Data Reliability and Its Applications”,” IEEE Transactions on systems, Man, and Cybernetics, Part B: Cybernetics, , Vols. Vol. 40, , no. Issue 6, , pp. pages 1622–1633, , 2010.
49. J. Famaey, S. Latre, J. Strassner and F. De Turck, "“A hierarchical context dissemination framework for managing federated clouds”,” Journal of Communications and Networks, , Vols. Vol. 13, , no. Issue 6, , pp. pages 567-582, , 2011.

50. H. Ishii, R. Tempo and E.-W. Bai., "'A Web Aggregation Approach for Distributed Randomized PageRank Algorithms'", IEEE Transactions on Automatic Control, , Vols. Vol. 57, , no. Issue 11, , pp. pages 2703–2717, , 2012..
51. D. Crichton, C. Mattmann, L. Cinquini, A. Braverman, D. Waliser, M. Gunson, A. Hart, C. Goodale, P. Lean and J. Kim., "'Sharing Satellite Observations with the Climate-Modeling Community: Software and Architecture'", IEEE Software, , Vols. Vol. 29, , no. Issue 5,.
52. S. Carvalho, J. Pavao, A. Queiros and A. Dias, "'A platform for the gathering, agregation and integration of social information'", in 7th Iberian Conference on Information Systems and Technologies (CISTI), pages 1–4, , 2012..
53. S. El-Tawab, S. Olariu and M. F. Almalag, "'A cyber-physical system for traffic flow related information aggregation and dissemination'", IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), , pp. pages 1-6, , 2012..
54. X. Wu, X. Zhu, G.-Q. Wu and W. Ding., "'Data Mining with Big Data'", IEEE Transactions on Knowledge and Data Engineering, , Vols. Vol. 26,, no. Issue 1, , pp. pages 97–107, , 2014..
55. Y. Tasaki, T. Fukuhara and T. Satoh, "'Aggnel: An Information Aggregation System of Partial Contents from Multiple Web Pages'", in *26th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pages 815-820, , 2012..
56. T. H. a. R. T. J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction,," 2nd ed., Springer, , 2009..
57. M. L. K. Murphy, A Probabilistic Perspective,, MIT Press, , 2012..
58. Y. B. a. G. H. Y. LeCun, "'Deep learning,'" Nature, , Vols. vol. 521, no. 7553, , pp. pp. 436-444, , 2015..
59. V. Vapnik, "The Nature of Statistical Learning Theory,," Springer, , 1995..
60. D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Prentice Hall, , 2023.
61. L. Breiman, "'Bagging predictors,'" Machine Learning, , Vols. vol. 24, no. 2, , pp. pp. 123-140, , 1996..
62. R. Schapire, "'The boosting approach to machine learning: An overview,'" in Nonlinear Estimation and Classification,, Springer, , 2003..
63. D. W. T. H. a. R. T. G. James, "An Introduction to Statistical Learning,," Springer, , 2013..
64. T. Kohavi, "'A study of cross-validation and bootstrap for accuracy estimation and model selection,'" in in Proc. IJCAI, , 1995..
65. R. Kohavi and D. Sommerfield, "'Leave-one-out cross-validation is a method of choice for small datasets,'" Machine Learning, , 1995..
66. A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed., O'Reilly Media, , 2019..



67. D. Powers, ""Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness, and correlation,"" J. Mach. Learn. Tech., , Vols. vol. 2, no. 1, , pp. pp. 37–63, 2011., 2011..
68. C. Elkan, ""The foundations of cost-sensitive learning,"" in in Proc. IJCAI, , 2001..
69. J. Davis and M. Goadrich, ""The relationship between precision-recall and ROC curves,"" in in Proc. ICML, , 2006..
70. F. Provost and T. Fawcett, ""Analysis and visualization of classifier performance: Understanding ROC and AUC,"" Mach. Learn., , Vols. vol. 30, , pp. pp. 1–38, , 1998..
71. R. T. a. J. F. T. Hastie, "Statistical Learning Theory and Applications,," Springer, , 2005.