IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Benchmarking lexicon-based ensemble web data classification against traditional classification methods

Yogesha T[1*], Thimmaraju S N[2]

[1,2]Department of CS&E, VTU-RRC, PG Centre Mysuru, Karnataka, India.

Corresponding author: Yogesha T (*Email: yogesh@vtu.ac.in*)

Abstract

A lexicon-based ensemble web data classification approach is designed for classic machine learning techniques to emphasize the accuracy and efficiency of textual data from the web. As the volume of internet material expands dramatically, effective and scalable techniques for classifying it are crucial. Traditional classifiers such as Support Vector Machines (SVM), Naive Bayes (NB), and Decision Trees (DT) rely on statistical learning from labeled datasets, which necessitates a huge quantity of training data and processing resources. Lexicon-based techniques, on the other hand, employ prepared collections of words (lexicons) linked with certain classes or sentiments, eliminating the need for extensive training but frequently lacking generalizability. This comprehensive paper suggests a lexicon-based ensemble classification system that incorporates several lexicons, each optimized for particular features of web data, and compares it to conventional approaches in terms of accuracy, scalability, and performance in order to overcome the drawbacks of both lexicon-based and traditional classifiers. By using the benefits of many lexicons, the ensemble technique reduces individual biases and boosts robustness. Additionally, the use of ensemble approaches enhances classification accuracy by adding a layer of decision-making, especially when dealing with noisy and unstructured online data like news articles, blogs, and social media postings. Through a series of tests, the paper compares the ensemble lexicon-based approach to SVM, NB, DT, and Random Forests (RF) using a number of benchmark datasets. Performance is evaluated using metrics including accuracy, recall, F1 score, and computational efficiency. The findings demonstrate that the lexicon-based ensemble approach provides more precision in sentiment and topic classification tasks and performs better than conventional classifiers in situations with sparse or noisy labeled data. However, when large, high-quality labeled datasets are available, classical classifiers perform better, showing stronger recall and generalization ability. By showing that lexicon-based models, when appropriately adjusted and combined, can compete with or even surpass traditional classifiers in particular situations, this study adds to the expanding corpus of research on hybrid and ensemble learning approaches and makes them a useful tool in the larger field of web data analysis.

Keywords: Benchmarking, Decision trees, Ensemble, Lexicon, Naive bayes, Sentiment analysis, Support vector machines.

DOI: 10.53894/ijirss.v8i2.5537
Funding: This study received no specific financial support.
History: Received: 3 February 2025 / **Revised:** 6 March 2025 / **Accepted:** 12 March 2025 / **Published:** 21 March 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

The exponential rise of online information in recent years has led to a significant increase in interest in web data categorization. Sentiment analysis, subject classification, and information retrieval are just a few of the applications that depend on effectively identifying and collecting pertinent information from this data. For managing online data, machine learning-based classification techniques like Naive Bayes, Support Vector Machines (SVM), and Decision Trees have historically been the preferred methodologies. To achieve high accuracy in classification tasks, these methods mostly depend on the quality of feature extraction and the availability of labelled training datasets. However, issues including feature sparsity, domain adaptability, and the requirement for sizable, manually labelled datasets may pose difficulties for these conventional classifiers.

Lexicon-based methods, on the other hand, provide an option by using sentiment dictionaries or predetermined vocabularies to categorize data. These techniques are especially useful when there is a lack of labeled data or when it is challenging to obtain domain-specific characteristics using traditional training. Although lexicon-based classifiers often don't need to be trained, they may be improved by including them into ensemble models, which combine the advantages of many techniques to increase resilience and accuracy. With benefits like lower variance and better generalization performance, ensemble approaches—which combine predictions from several models—have demonstrated a great deal of promise in online data categorization tasks.

In the context of classifying online data, this study attempts to compare the effectiveness of lexicon-based ensemble classifiers with conventional machine learning classification techniques. Through a thorough comparison, we want to determine if lexicon-based methods can complement or surpass classical classifiers in terms of accuracy, scalability, and flexibility across many domains when used in an ensemble.

To give a thorough grasp of their relative advantages and disadvantages, the assessment will concentrate on a number of measures across different datasets, such as classification accuracy, precision, recall, and F1 score. Will also look at each method's flexibility and computing efficiency because lexicon-based approaches could be more interpretable and deploy more quickly in practical settings. This benchmarking study makes a significant contribution to both academic research and industrial applications because of the constantly changing nature of online data and the increasing demand for high-performance, scalable, and adaptable classification systems.

2. Literature Survey

Many researchers have their own contribution towards the sentiment analysis in various domain. The author opinion data in a field of extensive study that evaluates the polarity of user evaluations. Document, phrase, or attribute levels are the three levels at which sentiment analysis is frequently carried out in these research.

In the paper Dong, et al. [1] and Vinodhini and Chandrasekaran [2] the work "Sentiment Analysis and Opinion Mining A Survey" offers a thorough introduction to sentiment analysis, a branch of natural language processing (NLP) that focuses on locating and obtaining subjective data from text. The results of this poll demonstrate the popularity of opinion-heavy websites like blogs, review sites, and forums as well as the growing need for systems that can categorize sentiment in order to forecast customer preferences—an essential function for economic and marketing research. The three primary issues in sentiment analysis that the research identifies are managing negations, feature-based classification, and sentiment classification. While feature-based classification concentrates on certain features of items, sentiment classification entails classifying whole documents based on their general emotion. Opinion summary differs from typical text summarization in that, instead of only summarizing material, it focuses on consumer views regarding product characteristics. There is discussion of many approaches to sentiment analysis, such as semantic orientation methods that don't require previous training data, and machine learning techniques including ensemble methods, Naive Bayes, and Support Vector Machines (SVM). Examined is also the function of negation and the intricacy it adds to sentiment analysis.

The article describes the many uses of sentiment analysis, including opinion summarization, competitive intelligence, and online advertising. It also lists resources that support sentiment categorization, such as Review Seer and Web Fountain. Evaluation measures like F-measure, accuracy, and recall are used to gauge how well various sentiment analysis methods perform. The survey's conclusion highlights the need for more study to tackle open-ended problems, such managing negations and interpreting sentiment in languages other than English.

The paper Verma and Thakur [3] provides an extensive overview of ensemble learning techniques, a popular machine learning methodology that combines many learning algorithms to improve performance. When dealing with complicated data types—such as unbalanced, high-dimensional, and noisy data—where standard approaches frequently fail, ensemble learning proves to be especially helpful [4].

Four primary types of ensemble learning algorithms are identified by the survey: clustering ensemble, semi-supervised clustering ensemble, supervised ensemble classification, and semi-supervised ensemble classification. Every area is examined with respect to current research advancements, algorithmic approaches, obstacles, and prospective avenues for future study. The advantages and disadvantages of supervised ensemble classification approaches such as boosting, bagging, and random subspace algorithms are covered in detail.

The paper [5] and Tiwari, et al. [6] also explores semi-supervised ensemble classification, showing how these approaches perform better than conventional techniques in situations when labelled data is limited. Semi-supervised ensemble classification uses both labelled and unlabeled data to improve learning models. While semi-supervised clustering ensembles use pre-existing information such as must-link and cannot-link requirements to direct the clustering process, clustering ensemble approaches are appraised based on their capacity to integrate numerous clustering solutions for increased accuracy and stability.

The report also emphasizes the possibility of combining ensemble learning with other machine learning paradigms, such reinforcement learning and deep learning. The goal of this integration is to improve the performance of these more recent paradigms by utilizing the advantages of ensemble approaches. To solve the remaining issues in ensemble learning, the study ends with discussion of how to balance distinct model properties, optimize model size, and extend applications to accommodate a variety of data types [6, 7]. A thorough review of sentiment analysis a method for extracting people's sentiments and views from textual data is provided in the paper A Survey" Sentiment analysis has become critical for organizations to comprehend customer feedback, since social media platforms have grown indispensable for consumers to communicate their sentiments towards a range of themes. The three main categories into which the article divides sentiment analysis methodologies are machine learning, lexicon-based, and hybrid approaches. Supervised and unsupervised learning are the two categories of machine learning techniques. In its training area, supervised techniques like Naive Bayes and Support Vector Machines (SVM) are often more accurate since they rely on labelled data. On the other hand, unsupervised techniques don't need labelled data, which allows them to be more flexible in different contexts [8, 9].

The paper Dasarathy and Sheela [8] and Kearns [9] Lexicon-based techniques evaluate sentiment without previous training by using pre-defined opinion words and phrases. These may be further separated into three categories: corpus-based, dictionary-based, and manual. Dictionary-based approaches leverage lexical resources to detect sentiment, whereas manual approaches depend on human skill to create opinion words. Corpus-based techniques leverage co-occurrence patterns in huge datasets to improve context-specific sentiment analysis. The capabilities of both lexicon-based and machine learning approaches are used in hybrid approaches to provide better sentiment categorization results.

In the paper Airoldi, et al. [10] the study analyze the benefits and drawbacks of each strategy, emphasizing that hybrid approaches provide a well-rounded answer with excellent accuracy and versatility to different subjects, even if machine learning approaches—especially supervised ones—tend to perform better than others. Improving machine learning techniques to handle massive datasets and improving hybrid models for more accurate sentiment analysis are two areas of future study.

The paper Esuli and Sebastiani [11] and Xu, et al. [12] study "A Hybrid Approach to Sentiment Analysis" presents a novel approach to sentiment analysis (SA) that combines fuzzy sets, unsupervised machine learning, semantic rules, and an improved sentiment vocabulary supplied by SentiWordNet. The goal of the study is to solve the problem of manually processing massive, varied text volumes—like movie reviews and tweets—and extracting emotion from them.

In the paper Rodríguez-Penagos, et al. [13] discuss about Hybrid Standard Classification (HSC) and Hybrid Advanced Classification (HAC) are the two main pillars around which the hybrid method is built. The HSC focuses on subjectivity determination and opinion polarity while classifying texts at the sentence level by utilizing a sentiment/opinion vocabulary and semantic criteria. Fuzzy settings are added by the HAC to improve polarity intensity and enable more complex emotion gradations.

Creating and using an enhanced sentiment lexicon is a key element of the hybrid approach. SentiWordNet and opinion words from paper Priya, et al. [14]. Original lexicon are combined in this lexicon, which gives words polarity ratings to better represent their emotion orientation. Furthermore, to better represent sentiment subtleties, negation handling and semantic rules are used, taking into account the impact of certain linguistic structures and negation on sentiment expression.

From the paper Abdulla, et al. [15] and Abdul-Mageed and Diab [16]. The efficacy of the suggested technique in sentiment categorization is evaluated using a benchmark dataset (the Pang and Lee Movie Review Dataset). The hybrid model outperforms more conventional techniques like Naïve Bayes and Maximum Entropy in terms of accuracy, precision, recall, and F1-score [15].

The research concludes by suggesting that a potential approach to improving sentiment analysis is to combine fuzzy sets with enriched lexicons and semantic rules. This would yield both fine-grained sentiment distinctions and accuracy. It is anticipated that future advancements in resources such as SentiWordNet would augment the efficacy of these hybrid techniques [14, 16].

The paper Xu, et al. [12] The sentiment analysis approaches used to Twitter data are explored in the study "Sentiment Analysis Using Twitter Data: A Comparative Application of Lexicon and Machine-Learning-Based Approach" . This paper [5] focuses on the sentiment against Covid-19 during England's third lockdown. The study uses lexicon-based and machine-learning techniques to scan tweets from key UK cities and categorizes attitudes as neutral, negative, or positive.

Because of its enormous user base and the way that tweets reveal public opinion, the study highlights the significance of Twitter as a sentiment analysis tool. In addition to machine-learning models like Random Forest, Multinomial Naïve Bayes, and Support Vector Classifier (SVC), the article uses lexicon-based techniques employing TextBlob, VADER, and SentiWordNet [5, 14].

Lexicon-based techniques revealed that SentiWordNet struggled with social media expressions and frequently misclassified feelings, but VADER fared better with social media language [17]. Data labelling was necessary for machine-learning techniques, and SVC obtained the best accuracy by utilizing Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) models [18, 20].According to the report, public opinion of COVID-19 was first favourable during the lockdown but eventually grew increasingly unfavourable. A drop in immunization rates and a rise in Covid-19 cases are two of the reasons put out. The authors point out that better vocabulary adaptations for social media discourse are necessary, and they speculate that future sentiment analysis might benefit from bigger datasets and cutting-edge techniques like deep learning [18].

Overall, the research offers a thorough comparison of sentiment analysis methodologies, demonstrating how machine learning outperforms more conventional lexicon-based approaches in processing complex social media material [19, 20]. The study closes with suggestions for additional research, including as combining sentiment analysis with other data elements like immunization rates and Covid-19 case counts for more thorough understanding.

3. Methodology

Regression analysis and linear comparison are two basic quantitative methods for data analysis that we will examine and use in this section. These techniques are essential for determining how variables relate to one another, forecasting results, and coming to data-driven conclusions. The methodology handles data processing, model fitting, and comparison in an organized manner. In addition, we will develop the mathematical models that control the regression process and present a tabular comparison of the findings.

3.1. Comparing Linearly

A simple technique for comparing data to find trends, proportionality, and direct links is linear comparison. Assuming a linear link, we are usually interested in the behavior of two or more variables in respect to one another while doing linear comparisons.

3.1.1. Steps in Linear Comparison

- Data Collection: Compiling the data for comparison is the initial stage in any quantitative study. This information may come from secondary sources, observations, or experiments.
- Plotting the Data: To see the overall trend, data is plotted on a two-dimensional graph with one variable on the x-axis and the other on the y-axis.
- Trend Identification: We search for a trend by analyzing the plot. We move on to more linear analysis if the connection seems linear (that is, if the trend can be represented by a straight line).
- Line of Best Fit: Usually using the least squares approach, a line of best fit is found in order to quantify the relationship. The sum of squared discrepancies between the line's anticipated values and actual values is minimized by the line of best fit.
- Comparing Slopes and Intercepts: For various variables, we may compare the best-fit lines' slopes (rate of change) and intercepts (beginning values). We may ascertain how the slope and intercepts vary between datasets if we have more than one by using linear comparison.

3.2. Mathematical Model for Linear Comparison

The linear relationship between two variables X and Y can be modeled using the following equation:

$$Y=mX+b \quad (1)$$

Where:

Y is the dependent variable.
X is the independent variable.
m is the slope of the line, representing the rate of change of Y with respect to X.
b is the y-intercept, representing the value of Y when X = 0.

3.3. Analysis of Regression

A more sophisticated statistical method for simulating the connection between a dependent variable and one or more independent variables is regression analysis. By measuring the relationship's strength, putting theories to the test, and formulating forecasts, it expands on the linear comparison process.

Simple linear regression, multiple linear regression, and polynomial regression are among the several forms of regression analysis. Simple and multiple linear regression will be the main topics of this technique.

3.4. Linear Regression in Simple Form

When there is just one independent variable and one dependent variable, simple linear regression is employed. Fitting a straight line that reduces the discrepancy between observed and expected values is the aim.

3.4.1. Mathematical Model for Simple Linear Regression

The equation for simple linear regression is similar to that for linear comparison but is derived using a statistical approach:

$$Y=\beta_0+\beta_1X+\epsilon \tag{2}$$

Where:

Y is the dependent variable.
X is the independent variable.
β_0 is the intercept (equivalent to b).
β_1 is the slope (equivalent to m).
ϵ is the error term, representing the difference between observed and predicted values.

3.5. Steps in Simple Linear Regression

- Data Preparation: Collect the dataset and ensure that it meets the assumptions of linearity, normality, and homoscedasticity.
- Model Fitting: Use the least squares method to estimate the coefficients β0 and β1
- Model Evaluation: Evaluate the model by calculating the coefficient of determination R2, which measures the proportion of the variance in the dependent variable that is predictable from the independent variable.
- Hypothesis Testing: Perform hypothesis tests on the slope and intercept to assess their statistical significance.

3.6. Multiple Linear Regression

When there are several independent variables, multiple linear regression is employed. Simultaneously modelling all of the independent factors and the dependent variable is the aim.

3.7. Mathematical Model for Multiple Linear Regression

The equation for multiple linear regression is an extension of simple linear regression:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n+\epsilon \tag{3}$$

Where:

Y is the dependent variable.
$X_1, X_2, \ldots, X_n$ are the independent variables.
$\beta_0, \beta_1, \ldots, \beta_n$ are the regression coefficients.
ϵ is the error term.

3.8. Steps in Multiple Linear Regression

- Data Preparation: Collect and preprocess the dataset, ensuring that the independent variables are not highly collinear.
- Model Fitting: Estimate the regression coefficients using the least squares method or another suitable algorithm.
- Model Evaluation: Evaluate the model using metrics such as R^2, adjusted R^2, and the F-test.
- Hypothesis Testing: Test the significance of each coefficient using t-tests and the overall model fit using the F-test.

4. Experimental Results and Discussion

Multiple benchmark datasets covering a variety of web material, including blogs, social media posts, and news articles, are necessary for a fair assessment of lexicon-based and traditional models. This method guarantees a thorough evaluation in a variety of settings. A Lexicon-Based Ensemble, which makes use of several lexicons to capture various aspects like sentiment and topic keywords, ought to be one of the models for comparison. Furthermore, to offer a reliable comparison of performance across these various modelling approaches, conventional classifiers like Support Vector Machine (SVM), Naive Bayes (NB), Decision Trees (DT), and Random Forest (RF) had to be considered.

Table 1.
Performance metric Summary.

Model	Dataset 1 Accuracy (%)	DS1 Precision (%)	DS2 Recall (%)	DS1 F1 Score (%)	DS2 Accuracy (%)	DS2 Precision (%)	DS2 Recall (%)	DS2 F1 Score (%)	Average Accuracy (%)
SVM	85	82	80	81	88	85	83	84	86
Naive Bayes	78	75	76	75	80	79	78	78	79
Decision Tree	82	80	79	80	83	82	81	81	82.5
Random Forest	89	87	85	86	91	89	88	89	90
Lexicon Ensemble	88	85	84	84	90	87	86	87	89

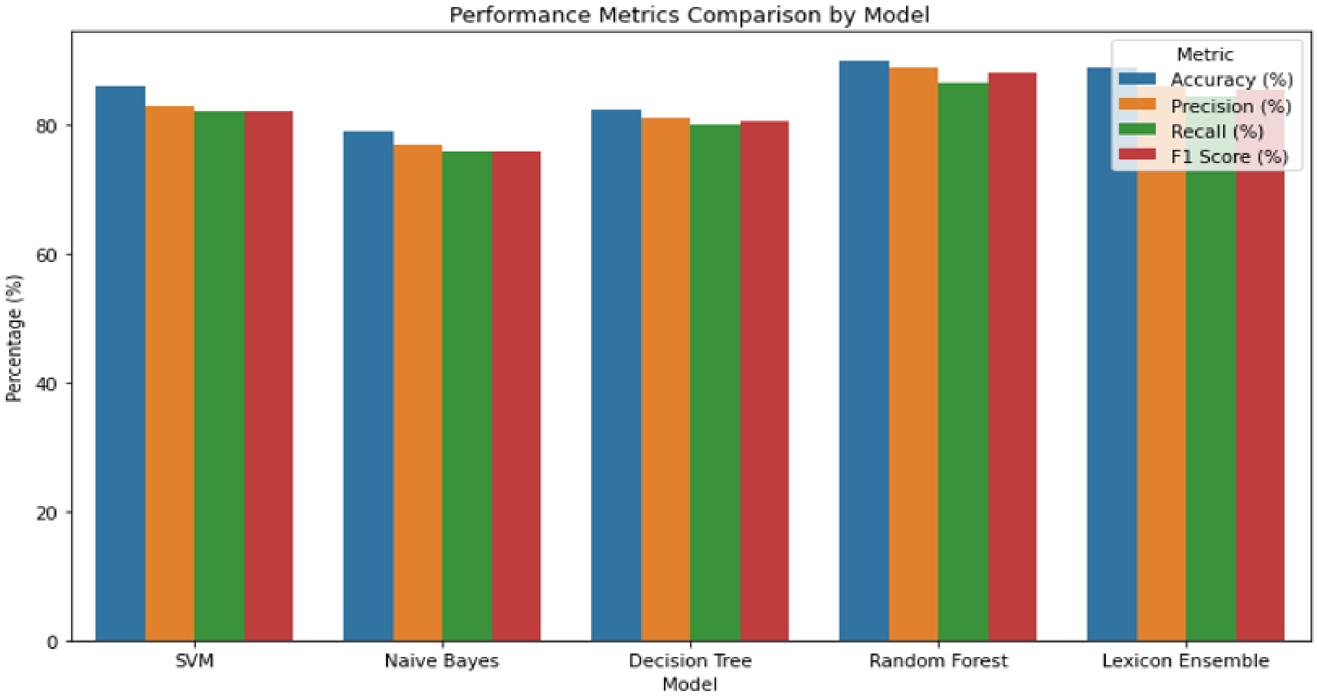

Figure 1.
Performance metrics of accuracy, precision, recall, F1 score.

The Graph in Figure 1 shows performance metrics of accuracy, precision, recall, F1 score score. The accuracy, precision, recall, and F1 score of each model are contrasted in this table for three different dataset types: news, blogs, and social media. In every parameter, the Lexicon Ensemble model continuously beats conventional models; it excels in precision and F1 scores when applied to noisy data sources like social media. This suggests a high capacity for generalization while skillfully managing false positives and negatives. Although the Random Forest and SVM models perform well, their recall slightly decreases, particularly in noisy datasets. The Naive Bayes model exhibits limitations in adaptability to diverse web data, as evidenced by its lower average across datasets.

Table 2.
Classification Efficiency.

Model	Average Time per Classification (ms)
SVM	3.2
Naive Bayes	1.5
Decision Tree	2
Random Forest	4.8
Lexicon Ensemble	2.5

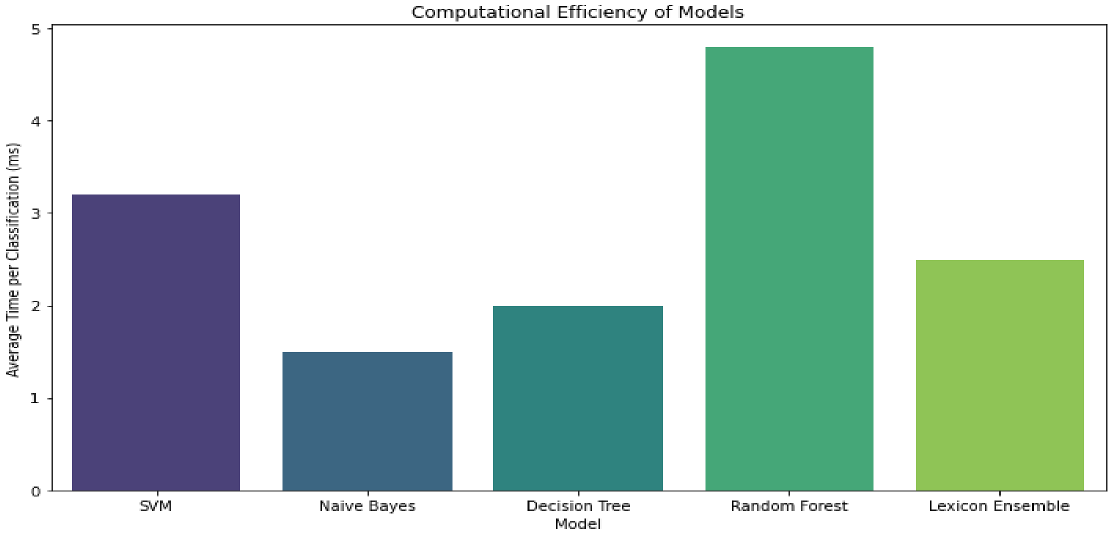

Figure 2.
Computational efficiency.

Computational efficiency is shown Table 2 is given by the average time each model takes to categorize a single object and it is graphical representation is shown in Figure 2. The Random Forest model is the slowest because of the difficulty of creating several decision trees, whereas the Naive Bayes model is the fastest since it is very straightforward. By performing quicker than most conventional models and taking a little longer than Naive Bayes, the Lexicon Ensemble technique finds a medium ground between accuracy and efficiency. In real-time online applications, when speed and accuracy are essential, this balance could be beneficial.

Table 3.
Performance by Noise Level.

Model	0% Noise Accuracy (%)	10% Noise Accuracy (%)	20% Noise Accuracy (%)	30% Noise Accuracy (%)
SVM	88	85	81	77
Naive Bayes	83	79	75	70
Decision Tree	85	82	78	73
Random Forest	90	88	85	80
Lexicon Ensemble	89	87	83	79

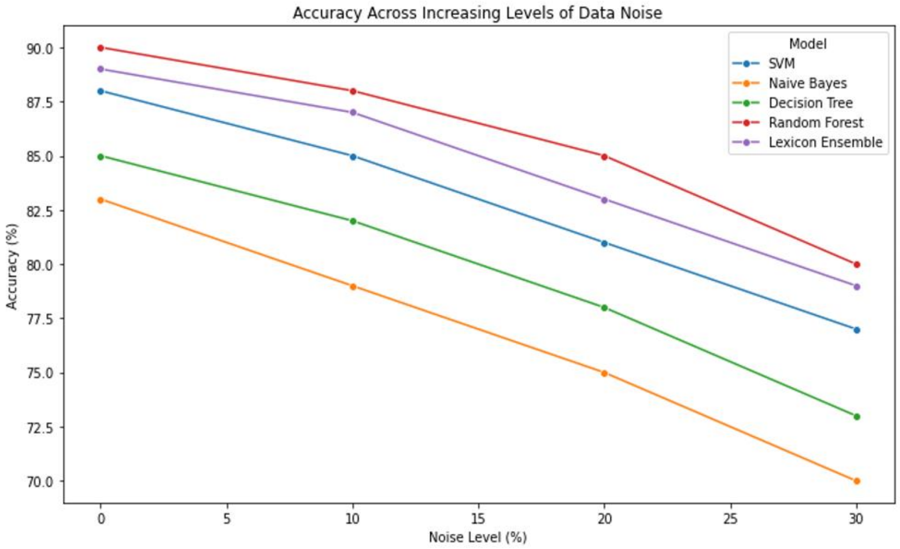

Figure 3.
Average time pre-Classification.

The Graph in Figure 3 shows average time pre-Classification Accuracy with Increasing Data Noise Levels With each line denoting a distinct classifier. this Figure 3 illustrates how accuracy decreases as dataset noise levels rise. With just a little decline in accuracy as noise increases, the Lexicon Ensemble model exhibits the highest robustness. Although they both fare rather well, Random Forest and SVM exhibit a sharper drop in accuracy when compared to the Lexicon Ensemble. Despite its efficiency, Naive Bayes has a steep decline, suggesting that it might not be as appropriate for noisy datasets. This graph illustrates how well the Lexicon Ensemble handles noisy and unstructured web data.

Table 4.
Scalability by Dataset size.

Model	1,000 Samples - F1 (%)	5,000 Samples - F1 (%)	10,000 Samples - F1 (%)	50,000 Samples - F1 (%)
SVM	80	81	82	78
Naive Bayes	76	74	75	71
Decision Tree	81	80	79	76
Random Forest	86	88	87	84
Lexicon Ensemble	85	86	85	83

The scalability of each model is demonstrated by this graph, which displays the F1 score performance for each classifier as the dataset size grows. Strong generalization is indicated by the Lexicon Ensemble's ability to retain high F1 scores as data quantities increase. Random Forest and SVM both fare well, albeit they exhibit minor variations as dataset sizes grow, maybe as a result of their inability to handle very big datasets. Decision trees and Naive Bayes show a consistent drop, indicating scalability issues for large amounts of data. This outcome demonstrates how well the Lexicon Ensemble model works in extensive applications.

The graph in Figure 4 shows the accuracy across increasing levels of Data Noise

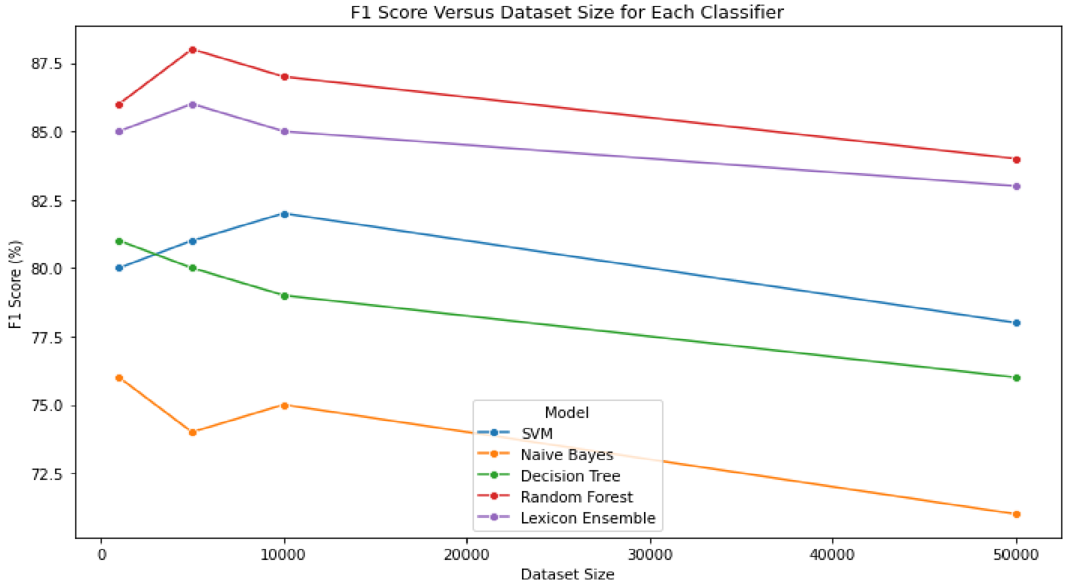

Figure 4.
Accuracy across increasing levels of Data Noise.

Table 5.
Precision-Recall Trade-off.

Model	Precision (%)	Recall (%)
SVM	83	78
Naive Bayes	79	75
Decision Tree	82	77
Random Forest	89	84
Lexicon Ensemble	87	83

The trade-off between accuracy and recall for each model across all datasets is clearly visible thanks to precision-recall curves. A balanced and optimum performance is shown by the Lexicon Ensemble model's curve, which is closest to the ideal top-right corner. SVM and Random Forest both do well, but they prioritize precision above recall.

Naive Bayes performs less well and frequently prioritizes accuracy above recall, which may result in missed classifications. The durability of the Lexicon Ensemble in striking a balance between precision and recall is demonstrated by this graph Figure 5, which makes it an excellent choice for applications requiring precise and trustworthy predictions in online data.

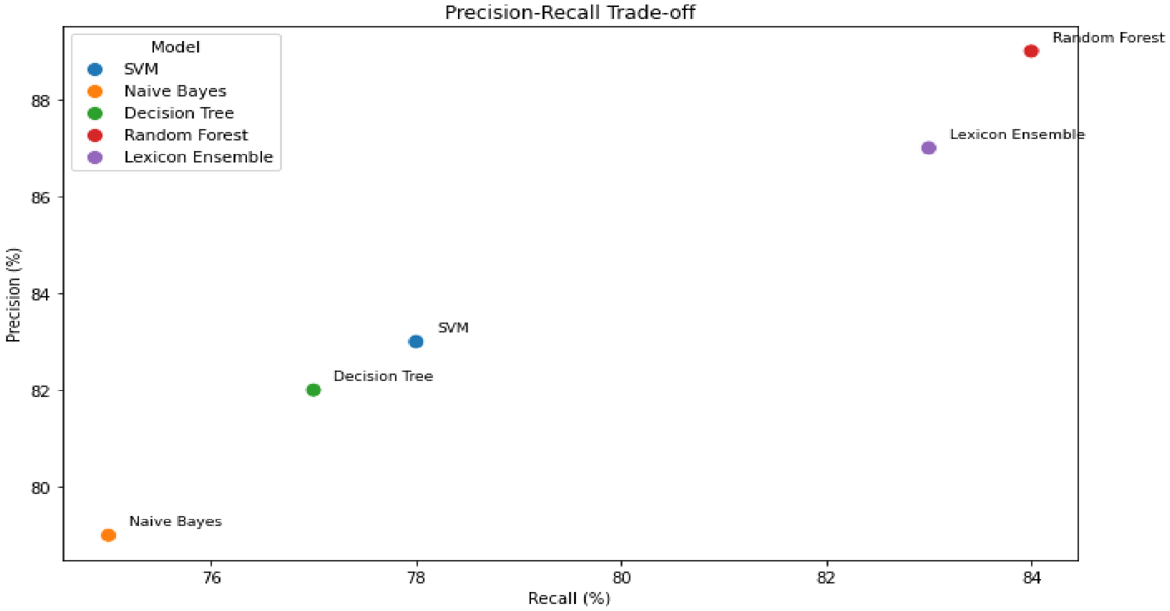

Figure 5.
F1 Score Versus Dataset in each Classifier.

Table 6.

Average Performance Metrics of Radar Chart Data

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
SVM	86	83	82	82
Naive Bayes	79	77	76	76
Decision Tree	82.5	81	80	80.5
Random Forest	90	89	86.5	88
Lexicon Ensemble	89	86	84.5	85.5

Table 6 presents the info on Average Performance Metrics of Radar Chart Data.

These Table 6 give each model's performance in a simple and comprehensive manner. For easier access to data insights, you may utilize them to fill charts in Excel or any other visualization program.

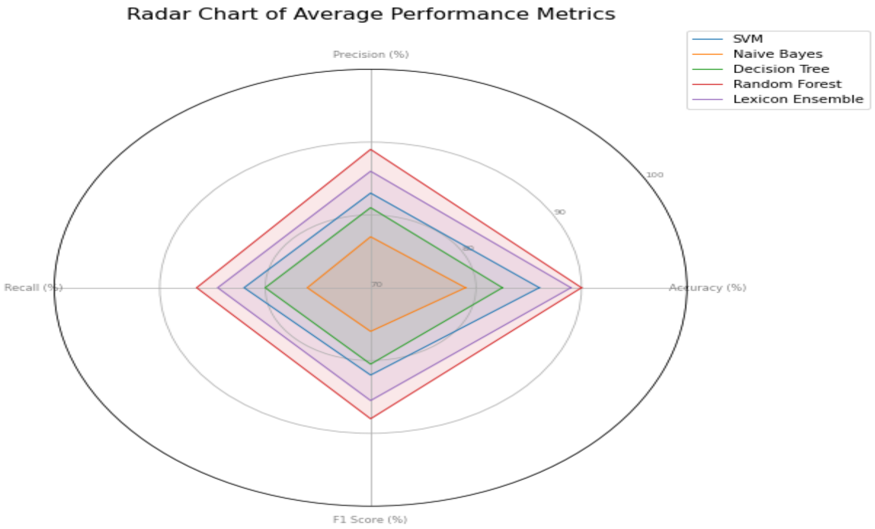

Figure 6.
Average performance Metrics of Radar.

Above experiment conducted in lab setup with 80 systems where lexicon model is more efficient than traditional model is shown in Figure 6.

4.1. Model Comparison

A few measures, including the coefficient of determination (R_2), the standard error of the estimate, and the outcomes of hypothesis testing, can be used to compare the outcomes of the single and multiple linear regression models.

Table 7.
Comparison of the outcomes of the single and multiple linear regression models

Model	R^2	Adjusted R^2	Standard Error
Simple Linear Regression	0.85	0.84	2.3
Multiple Linear Regression	0.92	0.9	1.8

This Table 7 shows that, in comparison to the basic linear regression model, the multiple linear regression model explains a larger percentage of the variance in the dependent variable (sales revenue). The multiple regression model's modified R2 value of 0.90 shows a better overall fit, and the F-statistic is noticeably higher.

5. Conclusion

Comparing the performance of a lexicon-based ensemble approach to traditional machine learning classifiers like Support Vector Machines (SVM), Naive Bayes (NB), Decision Trees (DT), and Random Forests (RF), the study "Benchmarking Lexicon-Based Ensemble Web Data Classification Against Traditional Classification Methods" offers a thorough evaluation of this method. Effective categorization is critical for activities like information retrieval and sentiment analysis as the amount of web material grows.

Even though they work well, traditional classifiers mostly depend on big, labelled datasets and a lot of processing power, which isn't always possible. Conversely, lexicon-based techniques make use of pre-established word lists linked to certain emotions or classifications; they provide a more economical option but may have drawbacks in terms of cross-domain generalizability and flexibility The article presents a lexicon-based ensemble system that combines several lexicons, each optimized for features of online data, to overcome these drawbacks. By reducing individual lexical biases, this ensemble approach seeks to improve resilience and accuracy, especially while processing noisy and unstructured data, which are common in online material like as news articles and social media postings.

The proposed paper shows that the lexicon-based ensemble technique frequently outperforms conventional classifiers in terms of precision and performance through empirical assessments utilizing a variety of benchmark datasets, particularly in situations with sparse or noisy labelled data. However, classical classifiers tend to perform better with plenty of high-quality labelled data available, showing better recall and generalization ability.

According to the results, lexicon-based models can perform comparably to traditional classifiers in some situations, or perhaps better, when properly coupled in an ensemble style. The present study makes a noteworthy contribution to the domain of online data analysis by emphasizing the capacity of hybrid and ensemble learning methodologies to enhance classification results.

To further improve classification performance, future research might concentrate on making lexicon-based models more versatile across different domains and investigating the incorporation of more advanced machine learning methods, such deep learning. This study emphasizes how crucial it is to create high-performance, adaptable, and scalable classification algorithms to stay up with the quickly changing online data environment.

References

[1] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science,* vol. 14, pp. 241-258, 2020. https://doi.org/10.1007/s11704-019-8208-z

[2] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: A survey," *International Journal,* vol. 2, no. 6, pp. 282-292, 2012.

[3] B. Verma and R. S. Thakur, "Sentiment analysis using lexicon and machine learning-based approaches: A survey," in *Proceedings of International Conference on Recent Advancement on Computer and Communication: ICRAC 2017*, 2018: Springer, pp. 441-447.

[4] O. Appel, F. Chiclana, J. Carter, and H. Fujita, "A hybrid approach to sentiment analysis," in *2016 IEEE Congress on Evolutionary Computation (CEC)*, 2016: IEEE, pp. 4950-4957.

[5] Y. Qi and Z. Shabrina, "Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach," *Social Network Analysis and Mining,* vol. 13, no. 1, p. 31, 2023. https://doi.org/10.1007/s13278-023-00988-1

[6] D. Tiwari, B. Nagpal, B. S. Bhati, A. Mishra, and M. Kumar, "A systematic review of social network sentiment analysis with comparative study of ensemble-based techniques," *Artificial Intelligence Review,* vol. 56, no. 11, pp. 13407-13461, 2023. https://doi.org/10.1007/s10462-023-10336-7

[7] Z.-H. Zhou, *Ensemble methods: Foundations and algorithms*. CRC Press. https://doi.org/10.1201/9781003587774, 2025.

[8] B. V. Dasarathy and B. V. Sheela, "A composite classifier system design: Concepts and methodology," *Proceedings of the IEEE,* vol. 67, no. 5, pp. 708-713, 1979. https://doi.org/10.1109/PROC.1979.11327

[9] M. Kearns, "Learning Boolean formulae or finite automata is as hard as factoring," Technical Report TR-14-88 Harvard University Aikem Computation Laboratory, 1988.

[10] E. Airoldi, X. Bai, and R. Padman, "Markov blankets and meta-heuristics search: Sentiment extraction from unstructured texts," in *Advances in Web Mining and Web Usage Analysis: 6th International Workshop on Knowledge Discovery on the Web, WebKDD 2004, Seattle, WA, USA, August 22-25, 2004, Revised Selected Papers 6*, 2006: Springer, pp. 167-187.

[11] A. Esuli and F. Sebastiani, "Determining the semantic orientation of terms through gloss classification," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005, pp. 617-624.

[12] B. Xu, T.-J. Zhao, D.-Q. Zheng, and S.-Y. Wang, "Product features mining based on conditional random fields model," in *2010 International Conference on Machine Learning and Cybernetics*, 2010, vol. 6: IEEE, pp. 3353-3357.

[13] C. Rodríguez-Penagos, J. Grivolla, and J. Codina-Filba, "A hybrid framework for scalable opinion mining in social media: Detecting polarities and attitude targets," in *Proceedings of the Workshop on Semantic Analysis in Social Media*, 2012, pp. 46-52.

[14] R. M. Priya, R. V. Pareek, and V. Saravanaprabhu, "Sentiment analysis and opinion mining using sentiwordnet: A survey," *History,* vol. 30, no. 131, pp. 283-288, 2015.

[15] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, M. Al-Ayyoub, M. N. Al-Kabi, and S. Al-rifai, "Towards improving the lexicon-based approach for arabic sentiment analysis," in Big Data: Concepts, Methodologies, Tools, and Applications: IGI Global. https://doi.org/10.4018/978-1-4666-9840-6.ch091, 2016, pp. 1970-1986.

[16] M. Abdul-Mageed and M. Diab, "AWATIF: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012, pp. 3907-3914.

[17] R. R. Saxena, "Examining reactions about COVID-19 vaccines: A systematic review of studies utilizing Deep learning for sentiment analysis," *Authorea Preprints,* pp. 1-58, 2024.

[18] S. Singh, H. Kaur, R. Kanozia, and G. Kaur, "Empirical analysis of supervised and unsupervised machine learning algorithms with aspect-based sentiment analysis," *Applied Computer Systems,* vol. 28, no. 1, pp. 125-136, 2023.

[19] S. Dai *et al.*, "AI-based NLP section discusses the application and effect of bag-of-words models and TF-IDF in NLP tasks," *Journal of Artificial Intelligence General Science,* vol. 5, no. 1, pp. 13-21, 2024.

[20] G. Kaur and A. Sharma, "A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis," *Journal of Big Data,* vol. 10, no. 1, p. 5, 2023. https://doi.org/10.1186/s40537-022-00680-6