Multi-Class Text Classification using Machine Learning Techniques

Osamah Mohammed Alyasiri
School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia | Karbala Technical Institute, Al-Furat Al-Awsat Technical University, Karbala, Iraq
osama.alyasiri@atu.edu.iq

Yu-N Cheah
School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia
yncheah@usm.my (corresponding author)

Received: 20 December 2024 | Revised: 20 February 2025, 5 March 2025, and 11 March 2025 | Accepted: 17 March 2025

 DOI: https://doi.org/10.48084/etasr.9994

ABSTRACT

The exponential growth of the World Wide Web has led to an overwhelming flood of information from diverse sources. This stream in data underscores the critical need for automated Text Classification (TC) to effectively manage, organize, and facilitate information discovery. TC plays a pivotal role in various real-world applications, spanning society, academia, government, and industry, as it eliminates the reliance on manual data classification, which is both costly and time-intensive. Machine learning models have emerged as key enablers, enhancing TC, prediction accuracy, and efficiency. However, existing models often struggle with multi-class imbalanced TC, where uneven class distributions lead to biased predictions and suboptimal model performance. This issue is further compounded by the lack of comprehensive evaluations on diverse datasets, making it challenging to determine the most effective model under imbalanced conditions. To tackle these challenges, this study systematically evaluates five widely recognized supervised machine learning algorithms: Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), K-Nearest Neighbor (KNN), Decision Tree (DT), and Logistic Regression (LR) across 19 benchmark datasets. Based on the average performance across F1-score, Classification Accuracy, and statistical significance tests, LR achieved the highest rank, closely followed by SVM and MNB. In contrast, KNN and DT demonstrated comparatively inadequate performance.

Keywords-text classification; text categorization; multi-class dataset; machine learning models

I. INTRODUCTION

The increased pace of development in the Information Technology (IT) industry and the sheer volume of textual data available have changed approaches to decision-making, impacting economies, policies, and societal interactions. Existing applications, like social networks, location-based services, social media, and e-commerce, act as primary sources of creating, collecting, and disseminating vast amounts of textual data [1-3]. This unstructured data stream necessitates the development of efficient systems for organizing and classifying content, making TC a crucial research area in data mining and Natural Language Processing (NLP) [4-6]. TC assigns predefined categories to textual data, facilitating efficient information retrieval, storage, and organization [7,8]. A significant subfield, multi-class TC, aims to categorize texts into one of multiple overlapping classes, a challenge relevant to real-world applications, where textual data are diverse and multi-dimensional [5,9]. Machine learning has revolutionized this domain by replacing traditional rule-based approaches, which are labor-intensive, error-prone, and non-scalable [4,10]. The existing machine learning algorithms are categorized into supervised and unsupervised learning [11-13] and can learn new domains through analyzing immense amounts of data, pattern recognition, and accurate predictions. Supervised learning, which utilizes labeled data, has become the dominant paradigm for TC, outperforming unsupervised methods in applications, such as news categorization, sentiment analysis, and e-commerce content filtering.

However, despite advancements, several key challenges remain unaddressed. One major issue is the impact of class imbalance in multi-class TC, where the distribution of instances across classes is skewed, leading to biased predictions and degraded model performance. Existing methods struggle with this issue, particularly in real-world datasets. Another challenge arises from the inconsistency in classifier selection, which significantly influences predictive outcomes. Many studies inconsistently apply different classifiers across the same dataset, making it difficult to fairly assess techniques, such as Feature Selection (FS) for solving TC problems. For instance, studies on FS methods using the Reuters-21578 dataset have employed varying classifiers, including Artificial Neural Networks (ANN), KNN, DT, SVM,

and Naïve Bayes (NB) [14-18]. This variation introduces a notable source of bias, potentially masking the true impact of FS techniques. Additionally, while machine learning algorithms, such as SVM, MNB, DT, KNN, and LR, have been widely used in TC, there is no systematic evaluation of their performance across a diverse set of benchmark datasets under multi-class imbalance conditions. Since no algorithm meets all the requirements the circumstances could offer [19], a comprehensive study comparing these classifiers on multiple datasets is needed to determine the most suitable models for imbalanced multi-class TC.

Previous studies have demonstrated the effectiveness of machine learning models in TC across various domains, including finance [20], tourism [21], healthcare [22], and online news analysis [23]. Additionally, multiple models have been developed for multilingual TC, covering languages, such as Urdu, English, Arabic, French, and Chinese [24-30]. Survey papers in this field have extensively discussed the advantages and disadvantages of traditional approaches, highlighting emerging challenges and future directions for TC [1, 2, 5, 6, 31-33]. A major concern in text mining is the complexity of classification algorithms when applied to large-scale datasets without human intervention. Several studies have explored different machine learning techniques to develop adequate classification models. Authors in [34] conducted a comparative study of five machine learning classifiers: SVM, KNN, MNB, LR, and Random Forest (RF), using the IMDB and SPAM datasets. Their findings revealed that KNN performed best on the SPAM dataset (98.5% accuracy), whereas LR was more effective for the IMDB dataset (85.8% accuracy). However, this study focused on binary classification problems and did not examine the challenges of multi-class TC, especially under class-imbalanced conditions. Similarly, authors in [25] compared SVM, KNN, and NB for English TC using datasets from the UCI library. The study concluded that SVM outperformed the other classifiers in terms of Precision, Recall, and F1-score, making it more suitable for large datasets. Although this research provides valuable insights, it lacks an in-depth exploration of multi-class scenarios and class imbalance issues, which are crucial in real-world applications. In the domain of medical TC, authors in [35] evaluated six machine learning models for predicting medical conditions based on user-generated drug reviews. Their results showed that the Linear Support Vector Classifier (LSVC) achieved the highest F1-score (0.88), demonstrating the potential of machine learning models in healthcare text analysis. However, the study was domain-specific, making its findings less generalizable to broader multi-class TC problems. Authors in [36] benchmarked six machine learning and five deep learning models on the 20 Newsgroups dataset, revealing that LR (82.74% accuracy) outperformed other machine learning models, while a bi-channel Convolutional Neural Network (CNN) performed best among deep learning models (80.27%). Their research highlights the efficiency of simpler models, like LR and CNN, but does not examine the effect of classifier consistency in multi-class TC. Authors in [37] compared LR, RF, and KNN using a BBC news dataset, demonstrating that LR paired with the TF-IDF vectorizer achieved the highest Accuracy (97%), followed by RF (93%) and KNN (92%). Although their study reinforces the stability of LR for small datasets, it lacks a systematic evaluation across multiple datasets to determine generalizability.

Despite extensive research in TC, several limitations remain unaddressed. One major issue is the challenge of multi-class classification under imbalanced conditions, as many studies focus on binary classification tasks rather than the challenges of multi-class classification, particularly when dealing with imbalanced datasets. Another limitation is the inconsistency in classifier selection, as prior research lacks a standardized approach to classifier evaluation, leading to inconsistencies in FS and performance assessments across different datasets. Additionally, most existing studies analyze classifiers using one or two datasets, making their conclusions dataset-dependent and limiting their generalizability. To address these gaps, a large-scale comparative evaluation is conducted on five widely used machine learning classifiers: SVM, MNB, KNN, DT, and LR, across 19 benchmark datasets. Unlike previous studies that focus on a single dataset, classifier performance is assessed across multiple datasets, with considerations for class imbalance and statistical significance tests to ensure a fair and unbiased evaluation. Through this rigorous large-scale comparative analysis, the gap in multi-class TC research is bridged, guiding future applications in academia and industry.

II. PROPOSED MODEL AND METHODOLOGY

This study employs machine learning-based classification models, including MNB, SVM, KNN, DT, and LR, to evaluate performance across 19 benchmark datasets. These datasets, varying in size and class distribution, provide a comprehensive basis for comparison. Performance was measured using the weighted average of the F1-score and Classification Accuracy for each model. To ensure an unbiased evaluation, each dataset was split into a training set (80%) and a test set (20%). A three-fold subdivision was adopted, where the training set was exclusively used for training and validation, while the test set remained reserved for final performance assessment. Additionally, the training set was further divided into 10 subsets using StratifiedKFold cross-validation, preserving the original class distribution, a crucial factor in multi-class classification tasks [38]. The final evaluation was conducted on the unseen test set to estimate the model's generalization capability accurately. All experiments were implemented in a Python environment, utilizing its extensive machine learning and data analysis libraries. Figure 1 illustrates the proposed model architecture.

A. The Benchmark Datasets

This study's experiments were conducted on 19 widely recognized multi-class text datasets, available on the official WEKA platform website [39]. These datasets span various domains and data characteristics, originating from prominent sources, including TREC, OHSUMED, Reuters-21578, and the WebACE project. Specifically, they vary significantly in terms of the number of documents, classes, document length, and category distribution, thus providing a diverse benchmark to robustly evaluate TC methods, as demonstrated in [40]. The data were initially transformed into word counts in [41].

Specifically, the dataset "fbis" is derived from the Foreign Broadcast Information Service data of TREC-5, while datasets "la1s" and "la2s" originate from the Los Angeles Times data of TREC-5. Datasets, such as "tr11", "tr12", "tr21", "tr23", "tr31", "tr41", "tr45", and "new3s", were created from collections in TREC-5, TREC-6, and TREC-7. Additionally, datasets "oh0", "oh5", "oh10", "oh15", and "ohscal" belong to a subset of the OHSUMED collection extracted from the MEDLINE database. The datasets "re0" and "re1" are sourced from the Reuters-21578 TC test collection, while the dataset "wap" originates from the WebACE project. Detailed descriptions of these 19 benchmark text datasets are provided in Table I.

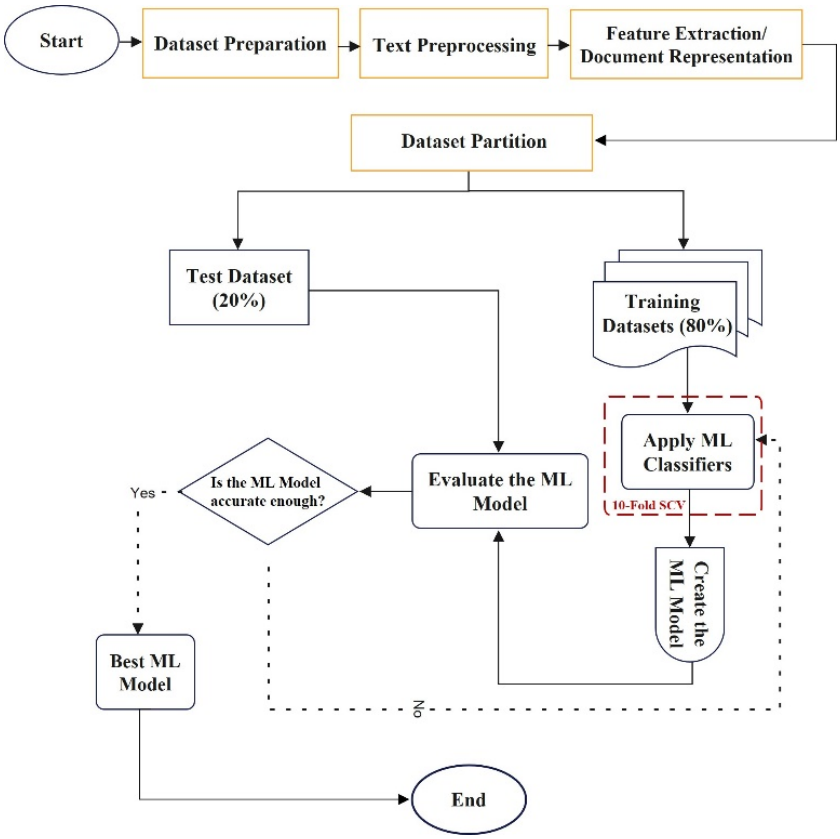

Fig. 1. The proposed model architecture.

TABLE I. THE 19 BENCHMARK DATASETS USED IN EXPERIMENTS

Dataset	#Documents	#Features	#Classes	#min class	#max class	#avg class	Document types
fbis	2463	2000	17	38	506	144.9	News articles
la1s	3204	13196	6	273	943	534.0	
la2s	3075	12433	6	248	905	512.5	
new3s	9558	26832	44	104	696	217.2	
oh0	1003	3182	10	51	194	100.3	Medical abstracts
oh10	1050	3238	10	52	165	105.0	
oh15	913	3100	10	53	157	91.3	
oh5	918	3012	10	59	149	91.8	
ohscal	11162	11465	10	709	1621	1116.2	
re0	1504	2886	13	11	608	115.7	News articles
re1	1657	3758	25	10	371	66.3	
tr11	414	6429	9	6	132	46.0	TREC documents
tr12	313	5804	8	9	93	39.1	
tr21	336	7902	6	4	231	56.0	
tr23	204	5832	6	6	91	34.0	
tr31	927	10128	7	2	352	132.4	
tr41	878	7454	10	9	243	87.8	
tr45	690	8261	10	14	160	69.0	
wap	1560	8460	20	5	341	78.0	Web pages

B. The Multi-classifiers

To ensure a comprehensive and equitable comparison, five prominent classifiers, MNB, SVM, KNN, DT, and LR, were selected.

- MNB was chosen for its simplicity and computational efficiency, making it a dominant approach in TC tasks, where the assumption of feature independence approximately holds. MNB was used with its default settings.
- SVM is particularly well-suited for high-dimensional feature spaces. By constructing linear or nonlinear hyperplanes, SVM effectively separates samples into different classes. In this study, a linear SVM classifier was implemented using LibLINEAR with L2-regularized L2-loss support vector classification (dual formulation).
- KNN was selected due to its simplicity, effectiveness, and widespread application in TC. It assigns class labels based on the majority vote of the KNN (where k=3 in this study)

using Euclidean distance. KNN is particularly beneficial in cases where local neighborhood structures play a crucial role in classification.

- DT was included for its interpretability and ability to handle both categorical and continuous features. DT recursively partitions datasets based on feature thresholds, forming a tree-like structure that is easy to understand and implement.
- LR is widely used in TC as it directly models class membership probabilities without making strong distributional assumptions. Unlike linear regression, which does not constrain probabilities within a valid range, LR uses the sigmoid function to map predicted values to probabilities between 0 and 1. In this study, LIBLINEAR with L2-regularized LR (dual formulation) was employed to optimize LR performance.

C. Evaluation Criteria

In TC, particularly topic classification tasks involving multi-class imbalanced datasets, selecting appropriate evaluation metrics is substantial. Relying solely on classification accuracy may be insufficient since accuracy does not adequately reflect performance across imbalanced classes [5]. Consequently, metrics like the F1-score, which integrate both Precision (P) and Recall (R), are more suitable. The F1-score effectively balances False Positives (FP) and False Negatives (FN), thus providing a more reliable assessment for imbalanced scenarios [5]. In this research, the proposed models were evaluated using a weighted average of the F1-score alongside Classification Accuracy, computed from standard confusion matrix components: True Positives (TP), True Negatives (TN), FP, and FN:

$$P = \frac{TP}{TP+FP} \quad (1)$$

$$R = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F1-score} = \frac{2(P \cdot R)}{P + R} \quad (3)$$

$$\text{Classification Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

III. EXPERIMENTAL RESULTS AND DISCUSSION

Tables II and III compare the performance of MNB, SVM, KNN, DT, and LR, across 19 benchmark datasets, respectively. The comparison is based on two key performance metrics: the weighted average of F1-score, as displayed in Table II, and Classification Accuracy, as evidenced in Table III, using Term Frequency (TF) as the feature representation method, with the best results being highlighted in bold. The averages presented at the end of Tables II and III for each classifier across all datasets serve as a comprehensive indicator of relative performance. These results are further complemented by detailed statistical analyses, which are discussed in the next section. Additionally, Figure 2 provides a visualization of the performance of the multi-classifier evaluated across all 19 benchmark datasets, offering a holistic view of its efficacy.

The experimental results presented in Tables II and III show that dataset characteristics, including dimensionality, class distribution, the number of classes, and dataset size, highly influence the performance of machine learning classifiers. Datasets with a high number of classes and imbalanced distributions tend to challenge most classifiers, particularly those sensitive to uneven class distributions, such as KNN and DT. In contrast, LR and SVM exhibit greater adaptability due to their ability to optimize decision boundaries in high-dimensional spaces. LR consistently ranks as the top performing classifier, followed closely by SVM and MNB. The superior performance of LR can be attributed to three main factors. First, LR benefits from built-in L2 regularization (Ridge regression), which prevents overfitting in high-dimensional datasets by penalizing large coefficients, making it particularly effective in feature-rich environments, like 'new3s' with 26,832 features and 'la2s' with 12,433 features. Second, LR is computationally efficient compared to SVM, which often requires expensive kernel transformations, rendering LR more scalable for large datasets. Third, LR's assumption of a linear decision boundary allows it to model feature dependencies more effectively than MNB, which assumes feature independence. While SVM also performs well, it exhibits slightly lower scores in datasets with large feature spaces due to the computational overhead of finding optimal hyperplanes. MNB, despite its efficiency in handling textual data, is inherently limited by its feature independence assumption, which leads to suboptimal performance in datasets where word dependencies play a crucial role, such as 'tr21' and 'ohscal'.

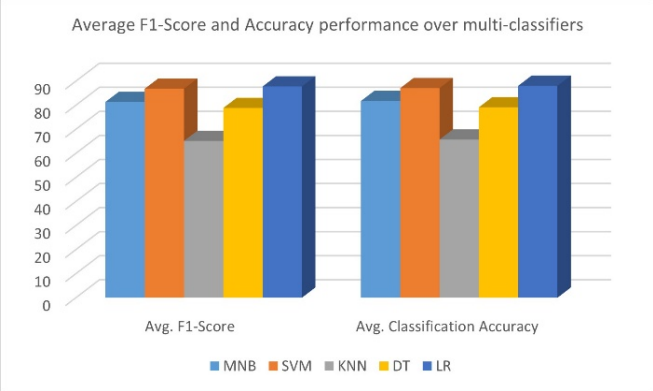

Fig. 2. The performances of multi-classifiers for all 19 benchmark datasets.

KNN and DT exhibit noticeable drops in performance due to inherent weaknesses in handling high-dimensional data and complex decision boundaries. KNN struggles with large feature spaces due to the curse of dimensionality, which makes distance-based calculations less meaningful, as seen in datasets like 'new3s' and 'ohscal,' where its F1-scores are relatively low, that is, at 68.58 and 47.72, respectively. DT, on the other hand, tends to overfit small and imbalanced datasets, as it relies on greedy heuristics that do not generalize well across multi-class scenarios. The number of classes in each dataset also significantly impacts classifier performance. Datasets with a small number of dominant classes tend to be easier to classify, while datasets with a large number of classes increase classification complexity and misclassification rates. LR and SVM manage multi-class classification effectively due to their

optimization techniques, followed closely by MNB, while KNN and DT struggle as the number of classes increases, leading to noticeable performance degradation in datasets like 'tr21' and 'ohscal'.

IV. STATISTICAL ANALYSIS

After presenting the experimental results to test the proposed methods, a statistical test is performed to compare the performance of the five classifiers considered in this study. A statistical significance test is employed using the Friedman and Nemenyi statistical significance tests [42], as proposed in [43], to compare multiple classifiers across multiple datasets. The Friedman test serves as a non-parametric alternative to the Analysis of Variance (ANOVA), while the Nemenyi post-hoc test evaluates pairwise differences in the classifiers' average ranks. If the Friedman test rejects the null hypothesis of equal performance, the present work proceeds with the Nemenyi post-hoc test at a significance level of 0.05 to determine whether rank differences are statistically significant. The classifiers' average rankings, derived from the Friedman test, are summarized in Tables II and III.

TABLE II. WEIGHTED AVERAGE OF F1-SCORE COMPARISONS FOR TF, ACROSS FIVE PROMINENT MACHINE LEARNING CLASSIFIERS

Dataset	MNB	SVM	KNN	DT	LR
fbis	75.64	83.69	67.87	72.37	84.00
la1s	**89.53**	87.08	58.00	75.45	88.27
la2s	91.00	91.76	64.97	74.11	**92.32**
new3s	78.30	87.02	68.58	70.35	**88.27**
oh0	87.14	87.18	51.10	82.83	**88.62**
oh10	76.00	77.49	48.69	78.65	78.44
oh15	**87.13**	84.64	53.92	75.09	85.67
oh5	84.10	87.93	49.40	79.18	**88.99**
ohscal	73.10	75.31	47.72	67.63	77.69
re0	80.92	84.19	76.50	76.39	**88.36**
re1	81.64	83.50	62.74	80.34	**87.49**
tr11	81.13	**89.05**	77.17	77.96	**89.05**
tr12	77.09	88.57	72.89	75.94	**88.57**
tr21	63.44	**86.09**	79.45	79.53	**86.09**
tr23	70.60	**87.58**	72.88	92.76	**87.58**
tr31	91.40	97.77	81.70	95.17	**98.30**
tr41	95.32	95.29	86.43	96.11	95.29
tr45	83.22	91.83	71.75	82.69	**91.90**
wap	78.76	**83.85**	46.19	65.02	82.71
Average	81.34	**86.83**	65.16	78.82	**87.77**
Ranking	3.0526	2.1316	4.8421	3.4737	1.5

The critical diagram representation proposed in [43] provides a visual method for comparing the results. In this diagram, the horizontal axis represents the classifiers' average ranks, with lower-ranked (better-performing) methods positioned to the left and higher-ranked (worse-performing) methods to the right. Classifiers that do not differ significantly are connected by a bold horizontal line, while significant performance differences occur if the rank difference between two classifiers exceeds the Critical Difference (CD). At the top of the diagram, the CD value calculated by the Nemenyi test is shown. These tests enhance the credibility and accuracy of the experimental findings by mitigating family-wise error and offering a comprehensive comparison of group performances [42, 43].

TABLE III. CLASSIFICATION ACCURACY COMPARISONS FOR TF, ACROSS FIVE PROMINENT MACHINE LEARNING CLASSIFIERS

Dataset	MNB	SVM	KNN	DT	LR
fbis	75.86	83.77	67.34	72.41	**84.18**
la1s	**89.70**	87.05	56.94	75.51	88.30
la2s	91.06	91.87	62.76	74.15	**92.36**
new3s	78.87	86.98	66.74	70.45	**88.28**
oh0	87.06	87.06	51.24	82.59	**88.56**
oh10	76.67	77.62	48.10	**79.05**	78.57
oh15	**87.43**	84.70	55.74	75.41	85.79
oh5	84.24	88.04	50.54	79.35	**89.13**
ohscal	73.26	75.41	49.04	67.58	**77.79**
re0	80.73	84.05	76.41	76.74	**88.37**
re1	84.04	84.04	66.27	81.33	**88.25**
tr11	80.72	**89.16**	77.11	78.31	**89.16**
tr12	77.78	**88.89**	73.02	77.78	**88.89**
tr21	63.24	**88.24**	83.82	79.41	**88.24**
tr23	71.17	87.80	75.61	**92.68**	87.80
tr31	91.40	97.85	81.72	95.16	**98.39**
tr41	95.45	95.45	86.36	**96.59**	95.45
tr45	83.33	**92.03**	73.91	83.33	**92.03**
wap	80.77	**84.29**	45.83	65.06	83.33
Average	81.74	**87.07**	65.71	79.10	**88.05**
Ranking	3.1053	2.1316	4.8421	3.4211	1.5

The p-values for both the F1-score and classification accuracy are 0.00001, leading the Friedman test to strongly reject the null hypothesis. This result confirms statistically significant differences among the classifiers, necessitating a post-hoc analysis via the Nemenyi test. The results of these post-hoc comparisons, along with the corresponding CD values, are illustrated in Figures 3 and 4. It is observed that the CD value of 1.3993 defines the threshold for significant differences: if the rank difference between two classifiers exceeds 1.3993, their performances are significantly different.

According to Figures 3 and 4, the classifiers LR and SVM are connected, indicating that their performances are not significantly different. Likewise, SVM and MNB are not significantly different, nor are KNN and DT. However, classifiers, like LR and DT, are disconnected, demonstrating significant differences in performance. Furthermore, in Figure 4, the accuracy-based rankings display that the rank difference between DT and KNN (4.8421 - 3.4211 = 1.421) exceeds the CD (1.3993), confirming a statistically significant performance gap.

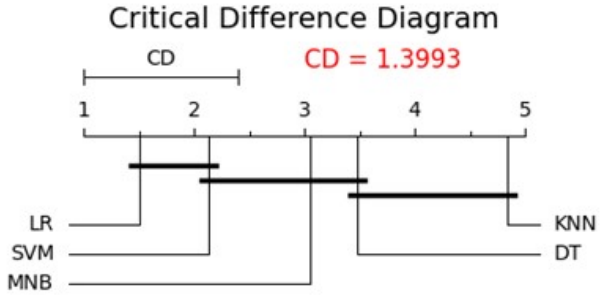

Fig. 3. Nemenyi test's CD diagrams for F1-Score metric across all 19 benchmark datasets.

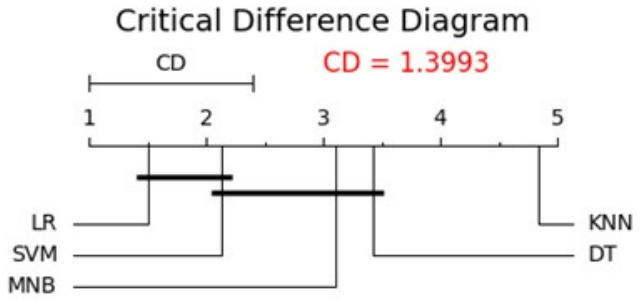

Fig. 4. Nemenyi test's CD diagrams for classification accuracy metric across all 19 benchmark datasets.

Thus, this study concludes with the key insight that LR achieves the best rank, followed closely by SVM and MNB. In contrast, KNN and DT perform poorly compared to the other classifiers. Consequently, there are clusters of classifiers (i.e., LR, SVM, MNB) whose performances are statistically indistinguishable.

V. CONCLUSION AND FUTURE WORK

Text Classification (TC) is one of the critical tasks of text mining that has drawn a lot of interest in recent years because of the increasing availability of digital documents and the need for efficient, adaptive search capabilities. Current TC mainly relies on machine learning paradigms, where models are trained on labeled document sets through inductive learning. In this investigation, five well-known classifiers were assessed, including Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), and Logistic Regression (LR) on 19 different benchmark datasets. The results demonstrate that LR achieved the highest performance, with an average F1-score of 87.77 and classification accuracy of 88.05, followed closely by SVM (average F1-score: 86.83, average accuracy: 87.07) and MNB (average F1-score: 81.34, average accuracy: 81.74). In contrast, KNN (average F1-score: 65.16, average accuracy: 65.71) and DT (average F1-score: 78.82, average accuracy: 79.10) showed relatively worse performance. Statistical analysis, using the Friedman and Nemenyi statistical significance tests and the critical diagram representation, proposed in [43], confirms that LR, SVM, and MNB form a group of classifiers with statistically similar accuracy, significantly outperforming both KNN and DT.

For future work, the authors intend to incorporate other forms of text representation that are more sophisticated than the use of binary term-frequency vectors, including the word embedding techniques and term weighting schemes that are supervised in nature, such as the TF-ICF scheme. This is expected to enhance the identification of benefits that can improve the efficiency of existing classifiers by employing these complex approaches. Moreover, it is planned to explore more sophisticated techniques of optimization and Feature Selection (FS) to make a more accurate term selection; it can be regarded as another significant line of future research. Such efforts are hoped to enhance the specificity of classification, capacity of scaling up, and flexibility of the model across different TC domains.

ACKNOWLEDGMENT

This work was supported by the Fundamental Research Grant Scheme (FRGS), Ministry of Higher Education Malaysia, under Grant FRGS/1/2023/ICT02/USM/02/4.

REFERENCES

[1] O. M. Alyasiri, Y.-N. Cheah, A. K. Abasi, and O. M. Al-Janabi, "Wrapper and Hybrid Feature Selection Methods Using Metaheuristic Algorithms for English Text Classification: A Systematic Review," *IEEE Access*, vol. 10, pp. 39833–39852, 2022, https://doi.org/10.1109/ACCESS.2022.3165814..

[2] V. Dogra *et al.*, "A Complete Process of Text Classification System Using State-of-the-Art NLP Models," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–26, Jun. 2022, https://doi.org/10.1155/2022/1883698.

[3] W. Q. A. Saif, M. K. Alshammari, B. A. Mohammed, and A. A. Sallam, "Enhancing Emotion Detection in Textual Data: A Comparative Analysis of Machine Learning Models and Feature Extraction Techniques," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16471–16477, Oct. 2024, https://doi.org/10.48084/etasr.7806.

[4] A. Palanivinayagam, C. Z. El-Bayeh, and R. Damaševičius, "Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review," *Algorithms*, vol. 16, no. 5, Apr. 2023, Art. no. 236, https://doi.org/10.3390/a16050236.

[5] O. M. Alyasiri, Y.-N. Cheah, H. Zhang, O. M. Al-Janabi, and A. K. Abasi, "Text classification based on optimization feature selection methods: a review and future directions," *Multimedia Tools and Applications*, Jul. 2024, https://doi.org/10.1007/s11042-024-19769-6.

[6] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A Survey on Text Classification Algorithms: From Text to Predictions," *Information*, vol. 13, no. 2, Feb. 2022, Art. no. 83, https://doi.org/10.3390/info13020083.

[7] A. Wahdan, M. Al-Emran, and K. Shaalan, "A systematic review of Arabic text classification: areas, applications, and future directions," *Soft Computing*, vol. 28, no. 2, pp. 1545–1566, Jan. 2024, https://doi.org/10.1007/s00500-023-08384-6.

[8] O. M. Alyasiri, Y.-N. Cheah, and A. K. Abasi, "Hybrid Filter-Wrapper Text Feature Selection Technique for Text Classification," in *2021 International Conference on Communication & Information Technology (ICICT)*, Basrah, Iraq, Jun. 2021, pp. 80–86, https://doi.org/10.1109/ICICT52195.2021.9567898.

[9] S. Anitha, E. Kavi Varshini, N. Haritha Mahalakshmi, and S. Jishnu, "Optimizing Multi-Class Text Classification Models for Imbalanced News Data," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, Jun. 2024, pp. 1–6, https://doi.org/10.1109/ICCCNT61001.2024.10724277.

[10] R. Li, M. Liu, D. Xu, J. Gao, F. Wu, and L. Zhu, "A Review of Machine Learning Algorithms for Text Classification," in *Cyber Security*, vol. 1506, Eds. Singapore: Springer Nature Singapore, 2022, pp. 226–234.

[11] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science," in *Supervised and Unsupervised Learning for Data Science*, Eds. Cham: Springer International Publishing, 2020, pp. 3–21.

[12] A. Ali and W. K. Mashwani, "A Supervised Machine Learning Algorithms: Applications, Challenges, and Recommendations," *Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences*, vol. 60, no. 4, Dec. 2023, https://doi.org/10.53560/PPASA(60-4)831.

[13] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, "An Unsupervised Machine Learning Algorithms: Comprehensive Review," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 911–921, Apr. 2023, https://doi.org/10.12785/ijcds/130172.

[14] M. Asif, A. A. Nagra, M. B. Ahmad, and K. Masood, "Feature Selection Empowered by Self-Inertia Weight Adaptive Particle Swarm

Optimization for Text Classification," *Applied Artificial Intelligence*, vol. 36, no. 1, Dec. 2022, Art. no. 2004345, https://doi.org/10.1080/08839514.2021.2004345.

[15] P. Grover and S. Chawla, "Text Feature Space Optimization Using Artificial Bee Colony," in *Soft Computing for Problem Solving*, vol. 1057, Eds. Singapore: Springer Singapore, 2020, pp. 691–703.

[16] R. Janani and S. Vijayarani, "Text Classification Using K-Nearest Neighbor Algorithm and Firefly Algorithm for Text Feature Selection," in *Advances in Electrical and Computer Technologies*, vol. 672, Eds. Singapore: Springer Singapore, 2020, pp. 527–539.

[17] R. Joseph Manoj, M. D. Anto Praveena, and K. Vijayakumar, "An ACO–ANN based feature selection algorithm for big data," *Cluster Computing*, vol. 22, no. S2, pp. 3953–3960, Mar. 2019, https://doi.org/10.1007/s10586-018-2550-z.

[18] A. Singh and A. Kumar, "Text document classification using a hybrid approach of ACOGA for feature selection," *International Journal of Advanced Intelligence Paradigms*, vol. 20, no. 1-2, 2021, Art. no. 158, https://doi.org/10.1504/IJAIP.2021.117613.

[19] B. Mahesh, "Machine Learning Algorithms - A Review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, Jan. 2020, https://doi.org/10.21275/ART20203995.

[20] M. N. Ashtiani and B. Raahemi, "News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review," *Expert Systems with Applications*, vol. 217, May 2023, Art. no 119509, https://doi.org/10.1016/j.eswa.2023.119509.

[21] Q. Li, S. Li, S. Zhang, J. Hu, and J. Hu, "A Review of Text Corpus-Based Tourism Big Data Mining," *Applied Sciences*, vol. 9, no. 16, Aug. 2019, Art. no. 3300, https://doi.org/10.3390/app9163300.

[22] A. Salau, N. Agwu Nwojo, M. Mahamat Boukar, and O. Usen, "Advancing Preauthorization Task in Healthcare: An Application of Deep Active Incremental Learning for Medical Text Classification," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12205–12210, Dec. 2023, https://doi.org/10.48084/etasr.6332.

[23] S. G. Tesfagergish, R. Damaševičius, and J. Kapočiūtė-Dzikienė, "Deep Fake Recognition in Tweets Using Text Augmentation, Word Embeddings and Deep Learning," in *Computational Science and Its Applications – ICCSA 2021*, vol. 12954, Eds. Cham: Springer International Publishing, 2021, pp. 523–538.

[24] M. N. Asim, M. U. Ghani, M. A. Ibrahim, W. Mahmood, A. Dengel, and S. Ahmed, "Benchmarking performance of machine and deep learning-based methodologies for Urdu text document classification," *Neural Computing and Applications*, vol. 33, no. 11, pp. 5437–5469, Jun. 2021, https://doi.org/10.1007/s00521-020-05321-8.

[25] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alexandria Engineering Journal*, vol. 60, no. 3, pp. 3401–3409, Jun. 2021, https://doi.org/10.1016/j.aej.2021.02.009.

[26] N. Aljedani, R. Alotaibi, and M. Taileb, "HMATC: Hierarchical multi-label Arabic text classification model using machine learning," Egyptian Informatics Journal, vol. 22, no. 3, pp. 225–237, Sep. 2021, https://doi.org/10.1016/j.eij.2020.08.004.

[27] X. Liu *et al.*, "Adapting Feature Selection Algorithms for the Classification of Chinese Texts," *Systems*, vol. 11, no. 9, Sep. 2023, Art. no. 483, https://doi.org/10.3390/systems11090483.

[28] M. F. Ibrahim, M. A. Alhakeem, and N. A. Fadhil, "Evaluation of Naïve Bayes Classification in Arabic Short Text Classification," *Al-Mustansiriyah Journal of Science*, vol. 32, no. 4, pp. 42–50, Nov. 2021, https://doi.org/10.23851/mjs.v32i4.994.

[29] M. F. Ibrahim and A. Al-Taei, "Title-Based Document Classification for Arabic Theses and Dissertations," in *Advances in Data and Information Sciences*, vol. 318, Eds. Singapore: Springer Singapore, 2022, pp. 189–203.

[30] H. Alshammary, M. F. Ibrahim, and H. A. Hussein, "Evaluating The Impact of Feature Extraction Techniques on Arabic Reviews Classification," *InfoTech Spectrum: Iraqi Journal of Data Science*, vol. 1, no. 1, pp. 42–54, Jun. 2024, https://doi.org/10.51173/ijds.v1i1.10.

[31] Q. Li *et al.*, "A Survey on Text Classification: From Traditional to Deep Learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 13, no. 2, pp. 1–41, Apr. 2022, https://doi.org/10.1145/3495162.

[32] M. Thangaraj and M. Sivakami, "Text Classification Techniques: A Literature Review," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 13, pp. 117–135, 2018, https://doi.org/10.28945/4066.

[33] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 3, May 2021, Art. no. 160, https://doi.org/10.1007/s42979-021-00592-x.

[34] S. U. Hassan, J. Ahamed, and K. Ahmad, "Analytics of machine learning-based algorithms for text classification," *Sustainable Operations and Computers*, vol. 3, pp. 238–248, 2022, https://doi.org/10.1016/j.susoc.2022.03.001.

[35] S. Joshi and E. Abdelfattah, "Multi-Class Text Classification Using Machine Learning Models for Online Drug Reviews," in *2021 IEEE World AI IoT Congress (AIIoT)*, Seattle, WA, USA, May 2021, pp. 0262–0267, https://doi.org/10.1109/AIIoT52608.2021.9454250.

[36] C. M. Suneera and J. Prakash, "Performance Analysis of Machine Learning and Deep Learning Models for Text Classification," in 2020 IEEE 17th India Council International Conference (INDICON), New Delhi, India, Dec. 2020, pp. 1–6, https://doi.org/10.1109/INDICON49873.2020.9342208.

[37] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augmented Human Research*, vol. 5, no. 1, Dec. 2020, Art. no. 12, https://doi.org/10.1007/s41133-020-00032-0.

[38] S. Szeghalmy and A. Fazekas, "A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning," *Sensors*, vol. 23, no. 4, Feb. 2023, Art. no. 2333, https://doi.org/10.3390/s23042333.

[39] *19MclassTextWc dataset*, 2006, G. Forman, Accessed: 5, Mar. 2025. [Online]. Available: https://sourceforge.net/projects/weka/files/datasets/text-datasets/19MclassTextWc.zip/download.

[40] R. G. Rossi, R. M. Marcacini, and S. O. Rezende, "Benchmarking text collections for classification and clustering tasks," *Institute of Mathematics and Computer Sciences*, Nov. 2013.

[41] E. H. Han and G. Karypis, "Centroid-Based Document Classification: Analysis and Experimental Results," in *Principles of Data Mining and Knowledge Discovery*, vol. 1910, Springer Berlin Heidelberg, 2000, pp. 424–431.

[42] D. G. Pereira, A. Afonso, and F. M. Medeiros, "Overview of Friedman's Test and Post-hoc Analysis," *Communications in Statistics - Simulation and Computation*, vol. 44, no. 10, pp. 2636–2653, Nov. 2015, https://doi.org/10.1080/03610918.2014.931971.

[43] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1-30, 2006.

AUTHORS PROFILE

Osamah Mohammed Alyasiri received a B.Sc. degree in computer science from Mustansiriyah University, Iraq, in 2009 and an M.Sc. degree in computer science from Dr Babasaheb Ambedkar Marathwada University (Dr BAMU), India, in 2013. He is currently pursuing a Ph.D. degree with the School of Computer Sciences, Universiti Sains Malaysia. He is also a Lecturer at Al-Furat Al-Awsat Technical University, Karbala Technical Institute, Department of Computer Network and Software Techniques, Iraq. His research interests include Artificial Intelligence, Text Mining, Text Classification, Information Retrieval, Machine Learning, Optimization, Pattern Recognition, Feature Selection, and AI Chatbots.

Yu-N Cheah received his B.Comp.Sc. (Hons.) and Ph.D. degrees from Universiti Sains Malaysia in 1998 and 2002, respectively. He is currently an associate professor at the School of Computer Sciences, Universiti Sains Malaysia. His research interests include sentiment analysis, semantic technologies, knowledge management, intelligent systems, and health informatics