# Exploring Text Classification Algorithms: A Systematic Literature Review

Dr. Kavita Dhakad[1], Dr. Dipali Meher[2]
[1]Associate Professor, School of Information Technology, Indira University, Pune
[2]Assistant Professor, Department of Computer science, Modern College of Arts, Commerce and Science, Pune
doi.org/10.64643/IJIRTV12I12-202144-459

***Abstract - Text classification is an essential task in natural language processing with wide-ranging applications, including sentiment analysis, spams filtering, and document categorization. This study aims to investigate diverse text classification algorithms and assess their ability to accurately assign predefined categories to text documents. This Systematic Literature Review commences by providing an overview of various text classification techniques. In this research 31 research articles are considered. This includes traditional machine learning approaches such as Naive Bayes, Support Vector Machines (SVM), Decision Trees (DT) and Random Forest (RF) along with more advanced methods like deep learning models, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The study seeks to gain insights into their effectiveness for text classification tasks. It explores how each algorithm handles different types of data and the challenges they may encounter. The author will also identify application area of these algorithms. Understanding the performance characteristics of these algorithms is crucial for selecting the most appropriate approach for specific text classification scenarios. This study will give future directions to apply text classification in syllabus classification on large data.***

***Keywords: Text Classification, Text Classification Algorithm, Naive Bayes, Support Vector Machines, Deep Learning***

## I. INTRODUCTION

Every day, a large volume of data is generated by individuals and organizations, but a significant portion of this data, around 80%, is unstructured. Within this unstructured data, there exists a formless entity referred to as unstructured data, which lacks organization, search ability, and manageability. As a result, it remains dormant and lacks purpose or usefulness. This presents a major challenge for companies as they struggle to organize, categorize, and extract relevant information from raw data. To tackle this challenge, text mining has emerged as an essential tool.

In a business setting, unstructured text data encompasses various sources such as emails, social media posts, chats, support tickets, surveys, and more. Manual sorting of this information often leads to failure due to its time-consuming nature, high cost, and proneness to errors. In contrast, text mining has proven to be a dependable and cost-effective solution, offering accuracy, scalability, and rapid response times.

Text mining, also referred to as text analysis, encompasses the process of converting unstructured text into structured data that can be readily analyzed. It relies on natural language processing (NLP) to enable machines to understand human language and automatically process it. Text mining primarily employs two methods: text classification and text extraction. Text classification involves assigning tags or categories to texts based on their content. Automated text classification enables efficient tagging of large volumes of text data, producing favourable results in a short time without manual intervention. This approach finds exciting applications across various domains.

There are two common approaches to text classification: rule-based systems and machine learning-based systems.
Machine learning-based systems, learn from previous data and require training with relevant examples of text, known as training data, that have been appropriately tagged. The text data, transformed into vectors along with the expected predictions (tags), is fed into a machine learning algorithm, resulting in a classification model. This trained model can then extract relevant features from new, unseen text and make predictions accordingly. Several machine learning algorithms are commonly used for text classification: Naive Bayes family of algorithms (NB), Support Vector Machines (SVM), Decision

Tree(DT), Random Forest(RF) and Deep learning (DL).

To assess the performance of a text classifier, several parameters are measured: accuracy, precision, recall, and F1 score. These metrics provide a comprehensive understanding of how well a classifier model analyzes texts and enable an evaluation of its effectiveness. The purpose of this article is to examine the effectiveness of text classification algorithms in classifying text. The author aims to study these algorithms for its text classification.

## II. RESEARCH METHODOLOGY

The researcher employed the Systematic Literature Review (SLR) method, which is widely recognized as a prominent research approach for investigating concepts, techniques, and algorithms in the field of text classification. SLR provides a well-structured process for identifying, evaluating, and interpreting all relevant evidence from the available literature pertaining to a specific research topic. In this study, we adhere to the SLR procedure to ensure a comprehensive and systematic analysis of the subject matter.

Research Objectives:

1. To Explore text classification algorithms
2. To identify the practical applications of the algorithms.
3. To study Text Classification algorithms applied for Syllabi Text Classification

Search Term

TITLE-ABS-KEY ( ( *"Text Classification"* OR *"Text Mining"* ) AND ( *"Algorithm"*) )

Data Sources

We conducted a search query across various digital libraries, including Scopus, Web of Science, IEEE, and Google Scholar, to retrieve relevant papers for our study. These libraries serve as primary sources of literature containing potentially valuable studies on Text Mining and Text Classification. In our search, we utilized our predefined search terms targeting the title, keywords, and abstract of the papers. It is important to highlight that Google Scholar was chosen as a data source due to its inclusion of articles, research papers, and conference papers.

Data Selection

Researcher has selected 31 research papers which are covering the research objectives. The research papers cover the concepts, techniques, algorithms and application area of Text Mining and Text Classification. The research articles published since 2005 to 2022 are taken into consideration.

## III. DISCUSSION

After examining 31 research articles, the author has gained an understanding of the multitude of text classification algorithms employed by researchers for the classification of diverse documents. The findings and insights are elaborated as follows:

1. Naive Bayes family of algorithms (NB): This text classification algorithm is a probabilistic method used for categorizing texts into predefined categories or tags. It is based on Bayes' Theorem and the assumption of feature independence, which simplifies the calculations.
2. Support Vector Machines (SVM): this text classification algorithm is a supervised learning method used for categorizing texts into different predefined classes or categories. It works by creating a hyper plane that maximally separates the texts belonging to different classes in a high-dimensional feature space.
3. Decision Tree(DT): This text classification algorithm is a supervised learning method used for categorizing texts into different predefined classes or categories. It works by constructing a hierarchical tree-like structure that recursively partitions the feature space based on different textual features.
4. Random Forest (RF): text classification algorithm is an ensemble learning method used for categorizing texts into different predefined classes or categories. It combines multiple decision trees to create a robust and accurate classifier by leveraging the concept of ensemble learning.
5. Deep learning (DL): This text classification algorithms are a class of neural network-based models specifically designed to analyze and categorize texts. These algorithms leverage deep neural networks, which are capable of learning intricate patterns and representations from raw textual data.

These algorithms are studied for its features, limitation and applications. Study is summarised in table 1.

Table 1: Summary of algorithms and researcher contribution

| Algorithm | Papers | #Count |
|---|---|---|
| Naive Bayes[NB] | [S1], [S2], [S3], [S5], [S6], [S7], [S8], [S10], [S11], [S13], [S14], [S16], [S18], [S19], [S20], [S21], [S23], [S24], [S25], [S27], [S30], [S31] | 22 |
| Support Vector Machines[SVM] | [S1], [S2], [S3], [S5], [S6], [S7], [S8], [S10], [S11], [S12], [S13], [S15], [S16], [S18], [S19], [S21], [S22], [S23], [S24], [S25], [S27], [S28], [S29], [S31] | 24 |
| Decision Tree[DT] | [S1], [S3], [S5], [S6], [S7], [S8], [S11], [S13], [S14], [S19], [S23], [S24], [S27] | 13 |
| Random forest[RF] | [S1], [S2], [S5], [S7], [S9], [S10], [S18], [S19] | 8 |
| Deep Learning[DL] | [S1], [S2], [S3], [S4], [S17], [S19] | 6 |

Author has analysed out of 31 review articles that text classification algorithms features, limitations and applications are studied as : It is observed that SVM and NB are widely used for text classification of various types of documents.

Author has observed the evidences of applications of above mentioned text classification algorithms as: Syllabus classification, Job Posting Classification, News Classification, Social media documents, Health Care Domain, Research/Technical Paper and Document(General) Classification. The analysis of this is presented in Table 2. Rathod, N. et. al (2012), states in his article that SVM works well for classification of computer science syllabi with 81 features and also good at distinguishing syllabus of two subjects.

Table 2: application of Text Classification Algorithm

| Application | Papers Addressed | #Count |
|---|---|---|
| Syllabus Classification | [S4], [S22], [S27], [S28], [S29], [S30] | 6 |
| Job Posting Classification | [S7], [S10] | 2 |
| News Classification | [S9], [S14], [S15], [S20], [S21] | 5 |
| Social media documents | [S12], [S15],[S18], [S20] | 4 |
| Health Care Domain | [S23] | 1 |
| Research/Technical Paper | [S25], [S26] | 2 |
| Document(General) Classification | [S13], [S15], [S20] | 3 |

## IV. CONCLUSION

This systematic literature review initiates by providing a comprehensive overview of various concepts, techniques, algorithms, and applications related to text classification. The research encompasses the analysis of 31 research articles, which cover a wide range of approaches. These include traditional machine learning methods such as Naive Bayes, Support Vector Machines (SVM), Decision Trees (DT), and Random Forest (RF), as well as more advanced techniques like Deep Learning (DL) models, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Through this examination, the researcher gained valuable insights into the effectiveness of these algorithms for text classification tasks. The study explores how each algorithm handles different types of data and discusses the challenges they may encounter. The author specifically identified several application areas where these algorithms have been successfully employed, including Syllabus Classification, Job Posting Classification, News Classification, Social Media Documents, Health Care Domain, Research/Technical Paper, and Document (General) Classification. Notably, the research findings highlight that Naive Bayes and Support Vector Machines (SVM) performed well in Syllabus Classification. The study also provides valuable guidance for future research directions in applying text classification to syllabus classification tasks, particularly when dealing with large datasets.

## REFERENCES

[1] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., ... & He, L. (2022). A survey on text classification: From traditional to deep learning. ACM Transactions on Intelligent Systems and Technology (TIST), 13(2), 1-41.

[2] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160.

[3] Bhavani, A., & Kumar, B. S. (2021). A review of state art of text classification algorithms. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1484-1490). IEEE.

[4] Yasukawa, M., Yokouchi, H., & Yamazaki, K. (2020). Syllabus Mining for Analysis of Searchable Information. International Journal of Institutional Research and Management, 4(1), 46-65.

[5] Maleki, F., Ovens, K., Najafian, K., Forghani, B., Reinhold, C., & Forghani, R. (2020). Overview of machine learning part 1: fundamentals and classic approaches. Neuroimaging Clinics, 30(4), e17-e32.

[6] Gupta, R. (2020). A survey on machine learning approaches and its techniques. In 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1-6). IEEE.

[7] Nasser, I., & Alzaanin, A. H. (2020). Machine learning and job posting classification: A comparative study. International Journal of Engineering and Information Systems (IJEAIS) ISSN, 6-14.

[8] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

[9] Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5, 1-16.

[10] Mittal, S., Gupta, S., Shamma, A., Sahni, I., & Thakur, D. N. (2020). A performance comparisons of machine learning classification techniques for job titles using job descriptions.

[11] Ray, S. (2019). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE.

[12] VM, N., & Kumar R, D. A. (2019). Implementation on text classification using bag of words model. In Proceedings of the second international conference on emerging trends in science & technologies for engineering systems (ICETSE-2019).

[13] Naresh, E., Vijaya, K. B., Pruthvi, V. S., Anusha, K., & Akshatha, V. (2019). Survey on Classification and Summarization of Documents. International Journal of Research in Advent Technology, 7(6S).

[14] Ababneh, J. (2019). Application of Naïve Bayes, decision tree, and K-nearest neighbors for automated text classification. Modern Applied Science, 13(11), 31.

[15] Chatterjee, S., Jose, P. G., & Datta, D. (2019). Text classification using SVM enhanced by multithreading and CUDA. International Journal of Modern Education and Computer Science, 12(1), 11.

[16] Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. Artificial Intelligence Review, 52(1), 273-292.

[17] Luan, Y., & Lin, S. (2019). Research on text classification based on CNN and LSTM. In 2019 IEEE international conference on artificial intelligence and computer applications (ICAICA) (pp. 352-355). IEEE.

[18] Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. International Journal of Research in Marketing, 36(1), 20-38.

[19] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[20] Panda, M. (2018). Developing an efficient text pre-processing method with sparse generative Naive Bayes for text mining. International Journal of Modern Education and Computer Science, 11(9), 11.

[21] Miao, F., Zhang, P., Jin, L., & Wu, H. (2018). Chinese news text classification based on machine learning algorithm. In 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) (Vol. 2, pp. 48-51). IEEE.

[22] Orellana, G., Orellana, M., Saquicela, V., Baculima, F., & Piedra, N. (2018). A text

mining methodology to discover syllabi similarities among higher education institutions. In 2018 International Conference on Information Systems and Computer Science (INCISCOS) (pp. 261-268). IEEE.

[23] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919.

[24] Desai, A. (2015). A review on knowledge discovery using text classification techniques in text mining. International Journal of Computer Applications, 111(6).

[25] Adeva, J. G., Atxa, J. P., Carrillo, M. U., & Zengotitabengoa, E. A. (2014). Automatic text classification to support systematic reviews in medicine. Expert Systems with Applications, 41(4), 1498-1508.

[26] Nguyen, T. H., & Shirai, K. (2013). Text classification of technical papers based on text segmentation. In Natural Language Processing and Information Systems: 18th International Conference on Applications of Natural Language to Information Systems, NLDB 2013, Salford, UK, June 19-21, 2013. Proceedings 18 (pp. 278-284). Springer Berlin Heidelberg.

[27] Rathod, N., & Cassel, L. N. (2012). Machine learning in building a collection of computer science course syllabi. In Theory and Practice of Digital Libraries: Second International Conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings 2 (pp. 357-362). Springer Berlin Heidelberg.

[28] Ota, S., & Mima, H. (2011). Machine learning-based syllabus classification toward automatic organization of issue-oriented interdisciplinary curricula. Procedia-Social and Behavioral Sciences, 27, 241-247.

[29] Yu, X., Tungare, M., Yuan, W., Yuan, Y., Pérez-Quinones, M., & Fox, E. A. (2009). Automatic syllabus classification using support vector machines. In Handbook of Research on Text and Web Mining Technologies (pp. 61-74). IGI Global.

[30] Joorabchi, A., & Mahdi, A. E. (2009). An automated syllabus digital library system for higher education in Ireland. The Electronic Library, 27(4), 640-658.

[31] Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974.

APPENDIX

Coding of Research Articles:

| Code | Citation |
|---|---|
| S1 | Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., ... & He, L. (2022). A survey on text classification: From traditional to deep learning. ACM Transactions on Intelligent Systems and Technology (TIST), 13(2), 1-41. |
| S2 | Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160. |
| S3 | Bhavani, A., & Kumar, B. S. (2021). A review of state art of text classification algorithms. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1484-1490). IEEE. |
| S4 | Yasukawa, M., Yokouchi, H., & Yamazaki, K. (2020). Syllabus Mining for Analysis of Searchable Information. International Journal of Institutional Research and Management, 4(1), 46-65. |
| S5 | Maleki, F., Ovens, K., Najafian, K., Forghani, B., Reinhold, C., & Forghani, R. (2020). Overview of machine learning part 1: fundamentals and classic approaches. Neuroimaging Clinics, 30(4), e17-e32. |
| S6 | Gupta, R. (2020). A survey on machine learning approaches and its techniques. In 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1-6). IEEE. |

| | |
|---|---|
| S7 | Nasser, I., & Alzaanin, A. H. (2020). Machine learning and job posting classification: A comparative study. International Journal of Engineering and Information Systems (IJEAIS) ISSN, 6-14. |
| S8 | Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386. |
| S9 | Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5, 1-16. |
| S10 | Mittal, S., Gupta, S., Shamma, A., Sahni, I., & Thakur, D. N. (2020). A performance comparisons of machine learning classification techniques for job titles using job descriptions. |
| S11 | Ray, S. (2019). A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 35-39). IEEE. |
| S12 | VM, N., & Kumar R, D. A. (2019). Implementation on text classification using bag of words model. In Proceedings of the second international conference on emerging trends in science & technologies for engineering systems (ICETSE-2019). |
| S13 | Naresh, E., Vijaya, K. B., Pruthvi, V. S., Anusha, K., & Akshatha, V. (2019). Survey on Classification and Summarization of Documents. International Journal of Research in Advent Technology, 7(6S). |
| S14 | Ababneh, J. (2019). Application of Naïve Bayes, decision tree, and K-nearest neighbors for automated text classification. Modern Applied Science, 13(11), 31. |
| S15 | Chatterjee, S., Jose, P. G., & Datta, D. (2019). Text classification using SVM enhanced by multithreading and CUDA. International Journal of Modern Education and Computer Science, 12(1), 11. |
| S16 | Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. Artificial Intelligence Review, 52(1), 273-292. |
| S17 | Luan, Y., & Lin, S. (2019). Research on text classification based on CNN and LSTM. In 2019 IEEE international conference on artificial intelligence and computer applications (ICAICA) (pp. 352-355). IEEE. |
| S18 | Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. International Journal of Research in Marketing, 36(1), 20-38. |
| S19 | Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150. |
| S20 | Panda, M. (2018). Developing an efficient text pre-processing method with sparse generative Naive Bayes for text mining. International Journal of Modern Education and Computer Science, 11(9), 11. |
| S21 | Miao, F., Zhang, P., Jin, L., & Wu, H. (2018). Chinese news text classification based on machine learning algorithm. In 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) (Vol. 2, pp. 48-51). IEEE. |
| S22 | Orellana, G., Orellana, M., Saquicela, V., Baculima, F., & Piedra, N. (2018). A text mining methodology to discover syllabi similarities among higher education institutions. In 2018 International Conference on Information Systems and Computer Science (INCISCOS) (pp. 261-268). IEEE. |
| S23 | Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919. |
| S24 | Desai, A. (2015). A review on knowledge discovery using text classification techniques in text mining. International Journal of Computer Applications, 111(6). |

| | |
|---|---|
| S25 | Adeva, J. G., Atxa, J. P., Carrillo, M. U., & Zengotitabengoa, E. A. (2014). Automatic text classification to support systematic reviews in medicine. Expert Systems with Applications, 41(4), 1498-1508. |
| S26 | Nguyen, T. H., & Shirai, K. (2013). Text classification of technical papers based on text segmentation. In Natural Language Processing and Information Systems: 18th International Conference on Applications of Natural Language to Information Systems, NLDB 2013, Salford, UK, June 19-21, 2013. Proceedings 18 (pp. 278-284). Springer Berlin Heidelberg. |
| S27 | Rathod, N., & Cassel, L. N. (2012). Machine learning in building a collection of computer science course syllabi. In Theory and Practice of Digital Libraries: Second International Conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings 2 (pp. 357-362). Springer Berlin Heidelberg. |
| S28 | Ota, S., & Mima, H. (2011). Machine learning-based syllabus classification toward automatic organization of issue-oriented interdisciplinary curricula. Procedia-Social and Behavioral Sciences, 27, 241-247. |
| S29 | Yu, X., Tungare, M., Yuan, W., Yuan, Y., Pérez-Quinones, M., & Fox, E. A. (2009). Automatic syllabus classification using support vector machines. In Handbook of Research on Text and Web Mining Technologies (pp. 61-74). IGI Global. |
| S30 | Joorabchi, A., & Mahdi, A. E. (2009). An automated syllabus digital library system for higher education in Ireland. The Electronic Library, 27(4), 640-658. |
| S31 | Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974. |