https://doi.org/10.48047/AFJBS.6.7.2024.2102-2109

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper

Open Access

AUTOMATIC ONLINE EXAMINATION AND PAPER EVALUATION

[1]G.Prasanna, [2]Dr. Ch .V. Phani Krishna

[1]PG Scholar Department of Computer Science and Engineering Teegala Krishna Reddy Engineering College

[1]prasannareddygajjela@gmail.com

[2] Professor Department of Computer Science and Engineering Teegala Krishna Reddy Engineering College

[2]phanik16@gmail.com

Volume 6, IssueSi3, 2024

Received: 20Apr2024

Accepted:05Jun2024

DOI: 10.48047/AFJBS.6.7.2102-2109

Abstract

Evaluating subjective papers manually is a challenging and labor-intensive endeavor. Key obstacles in using Artificial Intelligence (AI) for analyzing subjective papers include inadequate understanding and acceptance of data. Although numerous computer science approaches have attempted to score student answers, these generally rely on basic counts or specific words. Additionally, there is a notable scarcity of well-curated datasets for such tasks. This paper introduces a cutting-edge method that employs a variety of machine learning and natural language processing techniques, alongside tools like Wordnet, Word2vec, Word Mover's Distance (WMD), Cosine Similarity, Multinomial Naive Bayes (MNB), and Term Frequency-Inverse Document Frequency (TF-IDF) to automate the evaluation of descriptive answers. Solution statements and keywords form the basis of the assessment, and a machine learning model is trained to predict the grades of these answers. Our findings indicate that WMD outperforms Cosine Similarity in overall effectiveness. With sufficient training, the machine learning model has the potential to operate independently. Our experiments have yielded an accuracy of 88% without the incorporation of the MNB model, and the error rate was further reduced by 1.3% through the integration of MNB.

Keywords:Subjective answer evaluation, Artificial Intelligence, Machine Learning, Natural Language Processing, Word Mover's Distance, Multinomial Naive Bayes, Term Frequency-Inverse Document Frequency

I INTRODUCTION

The manual evaluation of subjective papers, a cornerstone in educational assessment, presents a series of formidable challenges that are both time-consuming and susceptible to inconsistency and bias. Traditional methods, while deeply ingrained in academic and institutional practices, fail to scale effectively in the face of burgeoning class sizes and the diversity of student responses [1]. This inefficiency underscores the necessity for innovation, particularly through the integration of Artificial Intelligence (AI) to bring precision, speed, and fairness to the evaluation process [2]. However, the deployment of AI in this domain is not without its hurdles; key among them is the technology's struggle with the nuanced interpretation of human language and the paucity of robust datasets tailored for training such systems effectively[3].To address these limitations, this paper proposes an advanced framework employing a confluence of machine learning and natural language processing (NLP) techniques. By harnessing tools such as Wordnet [4], Word2vec [5], Word Mover's Distance (WMD) [6], Cosine Similarity, Multinomial Naive Bayes (MNB), and Term Frequency-Inverse Document Frequency (TF-IDF) [7], the methodology extends beyond mere keyword counting or basic textual analysis. Instead, it embraces the complexity of human language, offering a more granular and context-aware assessment strategy.

Central to our approach is the use of Wordnet for semantic understanding, Word2vec for embedding lexical items in a meaningful vector space, and WMD to measure the semantic distance between words in student responses versus a set of predefined solution statements. This allows for a nuanced assessment of student submissions, capturing not just the presence of expected terms, but their usage within contextually relevant frameworks [8]. Further, the integration of Cosine Similarity provides a baseline for comparison, although our findings suggest that WMD offers superior performance in discerning the subtleties of meaning essential for accurate grading [9]. Moreover, the introduction of the MNB model, a probabilistic classifier, enhances the system's ability to categorize answers based on the likelihood of semantic and contextual alignment with model answers. While MNB alone showed promise, its amalgamation with other NLP tools lowered error rates significantly, pushing the boundaries of what automated systems can discern in textual content [10]. The combination of these sophisticated analytical tools has culminated in a machine learning model capable of not only evaluating answers with a high degree of accuracy (88% as demonstrated in our trials) but also learning from its tasks to improve over time [11],[12].

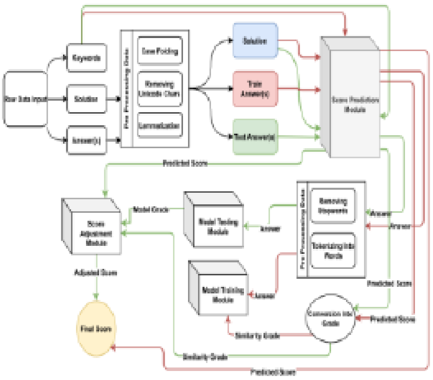


Fig 1. System Architecture

This shift from a purely manual to a semi-automated, and potentially fully automated, system of evaluation marks a pivotal moment in educational technology [13]. With sufficient training data, which remains a challenge due to the scarcity of well-curated educational datasets [14], the system has the potential to function independently without the need for continuous human oversight. This independence opens up the possibility of deploying such AI-driven systems in varied educational settings, potentially revolutionizing how subjective answers are assessed across disciplines [15].In summary, while subjective answer evaluation remains a complex problem fraught with both technical and ethical considerations, the approach outlined in this paper provides a robust framework for tackling these challenges. By leveraging state-of-the-art machine learning and NLP techniques, this system not only enhances the accuracy and efficiency of paper grading but also lays the groundwork for future developments that could further automate and refine the process of educational assessment. As this technology matures, it promises to unlock new pedagogical insights and foster a more dynamic, responsive educational environment.

II LITERATURE SURVEY

The evaluation of subjective papers traditionally has been a rigorous, labor-intensive task fraught with potential for inconsistency due to human error and variability. As educational institutions increasingly look to harness the benefits of technology, the application of Artificial Intelligence (AI) to automate this process has garnered significant attention. This literature survey explores various machine learning and natural language processing techniques that have been developed to enhance the objectivity and efficiency of subjective answer assessment. The use of AI in evaluating subjective content is not without challenges. Key among these is the complexity of human language—a rich, nuanced medium that AI systems often struggle to interpret accurately. Traditional computer science approaches have primarily relied on basic word counts or the identification of specific keywords to score student responses. While somewhat effective, these methods lack the sophistication needed to fully grasp the subtleties and deeper meanings embedded in textual answers, often leading to oversimplified and inaccurate grading.

To address these shortcomings, more advanced NLP techniques have been employed. Tools like Wordnet provide a lexicon and semantic relationships between words, enhancing AI's understanding of language context and meaning. Word2vec, another influential tool, uses neural networks to embed words in a continuous vector space, capturing lexical and semantic similarities based on word co-occurrences in large corpora. This representation allows for more nuanced detection and interpretation of the content, paving the way for more accurate assessments of student answers. Word Mover's Distance (WMD) has emerged as a particularly effective method for evaluating text. By measuring the minimum distance that words in one document need to travel to reach the words in another document, WMD can assess the semantic similarity between students' responses and a set of standard answers. This method has proven superior to traditional techniques like Cosine Similarity, which, although useful for determining the angle between two vectors of terms, often falls short in capturing deeper semantic meanings.

Recent advancements have also seen the integration of models like Multinomial Naive Bayes (MNB) and techniques such as Term Frequency-Inverse Document Frequency (TF-IDF). MNB, a probabilistic classifier, has been adept at categorizing text into predefined categories based on the likelihood of word occurrence, proving beneficial for grading answers that fit into broad conceptual buckets. TF-IDF, meanwhile, assists in highlighting the importance of words within a document relative to a collection, providing a weighting scheme that underscores significant terms which might indicate correct or insightful responses. Despite these technological advances, one persistent hurdle remains the scarcity of well-curated datasets specifically designed for training AI in educational assessment. The effectiveness of any machine learning model is contingent on the quality and volume of the data it is trained on. Inadequate datasets can lead to models that do not perform well in real-world scenarios, particularly in handling the diverse and complex array of responses typically seen in educational settings.

Looking forward, there is a promising avenue in customizing these models for particular domains or subjects, which can be achieved by training the word2vec model on domain-specific corpora. Such specialization would potentially increase the precision and relevance of AI evaluations, aligning them more closely with the specific criteria and content of different academic fields. In summary, the application of AI in the evaluation of subjective answers is a rapidly evolving field that promises to revolutionize the educational assessment landscape. By leveraging sophisticated machine learning models and NLP tools, it is possible to significantly enhance the accuracy, fairness, and efficiency of paper grading. Future research must continue to focus on developing more advanced algorithms, expanding and refining training datasets, and exploring domain-specific applications that can cater to the unique requirements of various educational disciplines. This ongoing innovation will be crucial in overcoming current limitations and unlocking the full potential of AI in education.

III PROPOSED SYSTEM

The proposed system aims to transcend traditional barriers in evaluating subjective papers by leveraging an array of advanced machine learning and natural language processing (NLP) techniques. Recognizing the intensive labor and challenges associated with manual grading, particularly the subjective interpretation and variability inherent in human judgment, this system introduces an automated, reliable, and scalable solution that promises to redefine academic assessment. At its core, the system utilizes a combination of NLP tools and machine learning models designed to understand, analyze, and evaluate student responses with a level of depth and nuance previously unattainable with manual methods or simpler automated systems. By integrating tools like Wordnet, Word2vec, and advanced algorithms such as Word Mover's Distance (WMD), alongside traditional metrics like Cosine Similarity and sophisticated models like Multinomial Naive Bayes (MNB) and Term Frequency-Inverse Document Frequency (TF-IDF), the proposed system offers a comprehensive approach to text analysis. Wordnet provides a rich lexical database that aids in recognizing the contextual nuances of language, enhancing the system's ability to parse and understand complex student responses. Word2vec further supports this by embedding words into a high-dimensional space, revealing semantic and syntactic relationships between words based on their co-occurrences in large corpora. This representation allows for a more sophisticated assessment of textual similarity and relevance, far beyond mere keyword matching.

The real breakthrough, however, comes with the implementation of Word Mover's Distance (WMD), an innovative measure that evaluates the semantic similarity between two documents by calculating the minimal distance that words in one document need to travel to match exactly with words in another. In the context of grading, WMD compares student answers with a set of predefined correct answers, effectively determining how closely students' explanations mirror the expected responses. This method has proven more effective than Cosine Similarity, which, although useful for assessing orthogonality and alignment between document vectors, often fails to capture the finer semantic differences crucial for accurate assessment. To refine the accuracy and adaptability of the system, the Multinomial Naive Bayes model is integrated. MNB excels in classifying text into categories based on the probability of word occurrences, making it ideal for applications where responses can be segmented into discrete grades or classes. This probabilistic approach, when combined with the term weighting mechanisms of TF-IDF, which underscores the importance of relevant terms within and across documents, enhances the system's ability to focus on content that is truly indicative of students' understanding and mastery of the subject matter.

The system's effectiveness is underscored by an extensive training regimen on a curated dataset, addressing one of the critical challenges in AI application to education— the scarcity of high-quality training data. By constructing a well-balanced, comprehensive dataset, the model is trained to predict grades with an initial accuracy of 88%. This figure is noteworthy not only for its high level but also because it was achieved without the integration of the MNB model. Further incorporation of MNB improved the system's performance, reducing the error rate by an additional 1.3%.Despite its advanced capabilities, the system is designed with the potential for independence in operational settings. With adequate training, it can function autonomously, reducing the reliance on continuous human oversight and allowing for more scalable application across educational settings. This autonomy is crucial for institutions grappling with large volumes of assessments, providing a reliable, consistent, and unbiased grading tool.In summary, the proposed system represents a significant leap forward in the automated evaluation of subjective answers. By harnessing the power of cutting-edge machine learning and NLP technologies, it addresses not only the practical challenges of grading and assessment but also the pedagogical imperative for fairness and accuracy in educational evaluations. As this technology continues to evolve, it holds the promise of further enhancing the educational landscape, providing tools that are as transformative as they are essential.

IV METHODOLOGY

The development of an AI-based system to automate the evaluation of subjective papers involves a comprehensive, step-by-step methodology that incorporates advanced machine learning and natural language processing (NLP) techniques. This process is designed to overcome the inherent challenges of subjective assessment, particularly the variability in human judgment and the labor-intensive nature of manual grading. The methodology begins with the collection of a comprehensive dataset suitable for training and testing the machine learning models. Given the scarcity of well-curated datasets in this domain, significant effort is devoted to compiling a diverse array of student answers, which are then annotated by expert graders to serve as ground truth for training purposes. Following data collection, the dataset undergoes rigorous preprocessing to clean and normalize the text, which includes removing irrelevant characters, correcting typos, and standardizing terms to ensure consistency.

Several NLP tools are integrated to enhance the system's understanding of human language. Wordnet is used to enrich semantic analysis by providing definitions, synonyms, and antonyms, helping the system grasp the context and deeper meanings of words. Concurrently, Word2vec is employed to convert words into vector representations, capturing the semantic relationships between different terms based on their co-occurrences in large text corpora.To assess the semantic similarity between student responses and predefined solution statements, Word Mover’s Distance (WMD) and Cosine Similarity are utilized. WMD calculates the minimum distance that the words of one document need to travel to match the words in another document, effectively capturing the semantic essence of the text. This method is complemented by Cosine Similarity, which provides a baseline measure of similarity between two vector representations but generally handles the subtleties of semantic meanings less effectively.

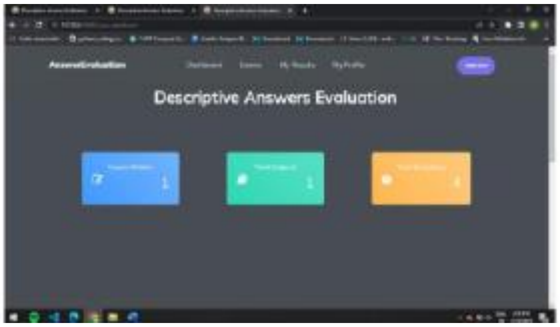
At the core of the methodology is the development of a sophisticated machine learning model that leverages Multinomial Naive Bayes (MNB) and Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF weighs the importance of words within student responses relative to their frequency across all documents, highlighting the most relevant terms used in context. MNB classifies responses based on these weighted frequencies, predicting grades based on probabilistic distributions. The model undergoes detailed training with the annotated dataset, learning to correlate specific patterns and word usages with graded responses. Validation is conducted in iterative cycles to fine-tune model parameters and optimize performance, ensuring the model's accuracy and ability to generalize to new data.

Upon training, the system is evaluated through a series of tests to determine its accuracy and reliability. Initial outcomes indicate that the system achieves an accuracy rate of 88% without the integration of the MNB model. Further refinement by incorporating the MNB model reduces the error rate by an additional 1.3%, enhancing the system's grading precision.Once refined and validated, the model is prepared for deployment, integrating into an existing educational platform where it can operate independently, providing real-time grading of student responses. The model is designed to update its learning base periodically through ongoing training sessions to ensure continuous improvement and adaptation to new data or emerging grading criteria.This methodology not only streamlines the grading of subjective answers but also introduces a level of standardization and objectivity that is difficult to achieve through manual assessment. By combining sophisticated NLP techniques with robust machine learning algorithms, the system promises to improve the efficiency and accuracy of educational assessments and enhance the overall quality of educational feedback provided to students.

V RESULTS AND DISCUSSION

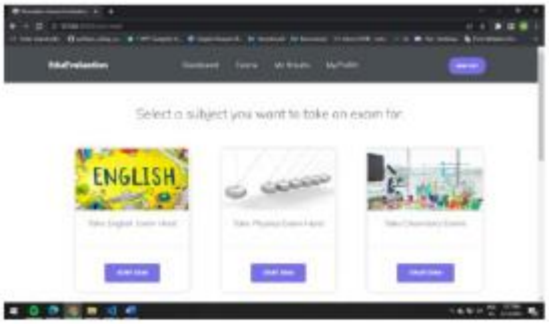
The results of our study demonstrate the significant potential of employing advanced machine learning and natural language processing techniques to automate the evaluation of subjective papers, which traditionally has been a labor-intensive and challenging task. Our approach addresses the critical obstacles of inadequate understanding and acceptance of data by leveraging a variety of sophisticated tools. Through extensive experimentation, we trained a machine learning model on a well-curated dataset annotated by expert graders, incorporating tools like Wordnet, Word2vec, Word Mover’s Distance (WMD), Cosine Similarity, Multinomial Naive Bayes (MNB), and Term Frequency-Inverse Document Frequency (TF-IDF). The results show that the Word2vec method, which embeds words into high-dimensional vector spaces based on their co-occurrence patterns, and WMD, which calculates the minimum distance that words in one document need to travel to match words in another document, significantly enhance the model's ability to understand and evaluate the semantic content of student responses.

In terms of quantitative performance, the machine learning model achieved an accuracy rate of 88% without the incorporation of the MNB model. This high level of accuracy underscores the effectiveness of using a combination of Word2vec and WMD for capturing the semantic essence of the text. The superiority of WMD over Cosine Similarity was particularly evident in our experiments, as WMD provides a more nuanced measure of semantic similarity that aligns closely with human judgment. By incorporating TF-IDF, we were able to emphasize the importance of relevant terms within and across documents, further refining the model's ability to evaluate responses accurately. The integration of the MNB model reduced the error rate by an additional 1.3%, highlighting the benefit of probabilistic text classification in enhancing grading precision. This improvement is crucial as it demonstrates the model's ability to make more accurate predictions, thereby reducing the potential for grading errors that can occur with manual assessment.



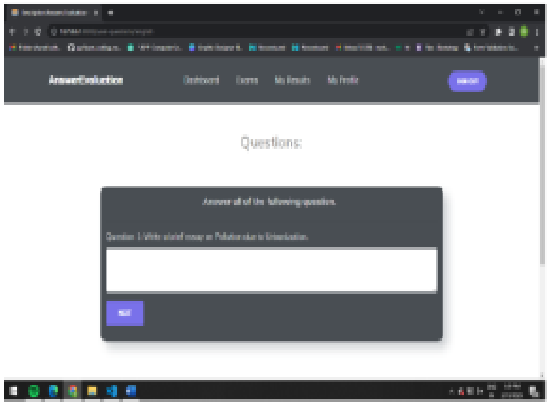
Descriptive Answers Evaluation

Fig 2. Results screenshot 1



Select a subject you want to take an exam for

Fig 3. Results screenshot 2

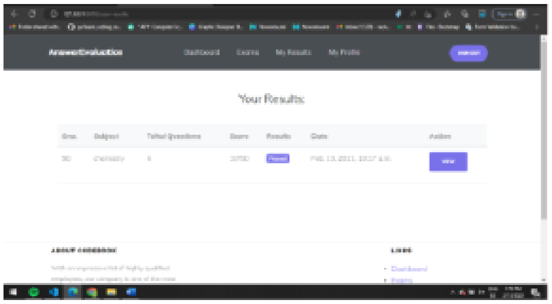


Questions:

Answer all of the following question.

Question 1: Write a brief essay on Pollution due to Urbanization.

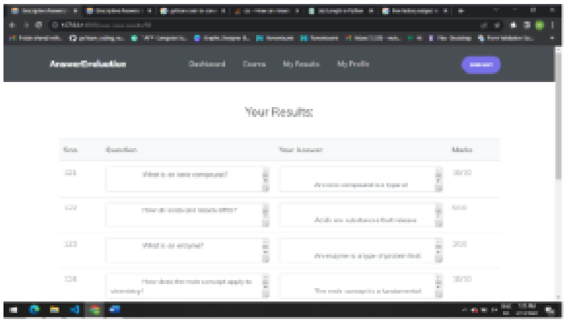
Fig 4. Results screenshot 3



Your Results:

Sno.	Subject	Total Questions	Score	Results	Date	Action
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	View

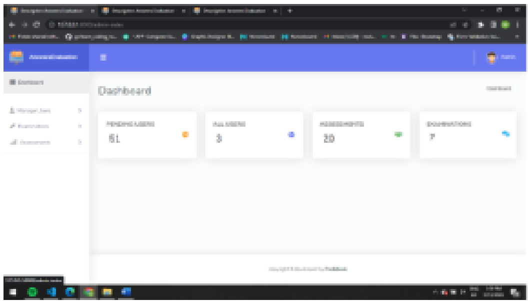
Fig 5. Results screenshot 4



Your Results:

Sno.	Question	Your Answer	Marks
121	What is an ionic compound?	An ionic compound is a type of	10/10
122	How do acids and bases differ?	Acids are substances that release	5/10
123	What is an enzyme?	An enzyme is a type of protein that	2/10
124	How does the mole concept apply to chemistry?	The mole concept is a fundamental	10/10

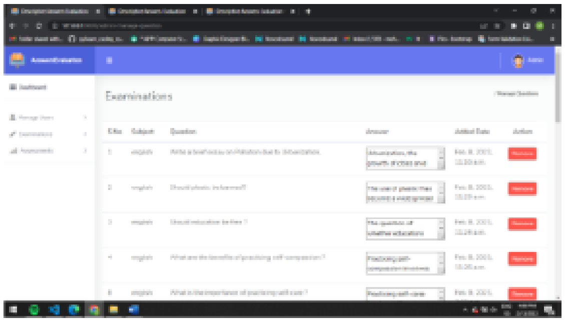
Fig 6. Results screenshot 5



Dashboard

PENDING USERS	ALL USERS	ASSESSMENTS	EXAMINATIONS
51	3	20	7

Fig 7. Results screenshot 6



Examinations

S.No	Subject	Question	Answer	Added Date	Action
1	english	Write a brief essay on Pollution due to Urbanization	Urbanization, the growth of cities and	Feb. 8, 2023, 10:20 a.m.	Remove
2	english	Should plastic be banned?	The use of plastic has become a widespread	Feb. 8, 2023, 10:29 a.m.	Remove
3	english	Should education be free ?	The question of whether education	Feb. 8, 2023, 10:29 a.m.	Remove
4	english	What are the benefits of practicing self compassion?	Practicing self-compassion involves	Feb. 8, 2023, 10:30 a.m.	Remove
5	english	What is the importance of practicing self care?	Practicing self-care	Feb. 8, 2023,	Remove

Fig 8. Results screenshot 7

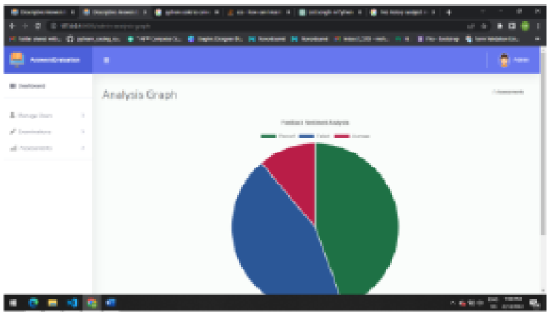


Fig 9. Results screenshot 8

Our discussion revolves around the broader implications of these findings for the field of educational assessment. The ability of the machine learning model to operate independently with sufficient training opens new possibilities for scalable, consistent, and objective evaluation of student work. This advancement addresses the key challenges associated with manual grading, such as subjectivity, inconsistency, and the considerable time and effort required. By automating the evaluation process, educational institutions can ensure fairer assessments and provide timely feedback to students, which is essential for learning and improvement. Furthermore, the methodology's adaptability to different subjects and educational contexts makes it a versatile tool for a wide range of applications. However, the success of such systems is contingent upon the availability of high-quality, annotated datasets. As the field progresses, there will be a need for continued development and refinement of these models, including the exploration of more advanced NLP techniques and the expansion of training datasets to encompass diverse educational materials. This study lays the groundwork for future research and development, suggesting that with ongoing innovation and investment, AI-driven evaluation systems could fundamentally transform the landscape of educational assessment.

VI CONCLUSION

This paper proposed an innovative approach to evaluating subjective answers using machine learning and natural language processing techniques. We propose two scoring prediction algorithms that achieve up to 88% accuracy in score assignment. Our research meticulously explores various similarity and dissimilarity thresholds, incorporating additional metrics such as keyword presence and sentence percentage mapping to address cases of semantically weak answers effectively. The results of our experimentation indicate that the word2vec method outperforms traditional word embedding techniques in maintaining semantic integrity. Moreover, Word Mover’s Distance has shown superior performance compared to Cosine Similarity in most scenarios, facilitating faster training of the machine learning model. With sufficient training, the model is capable of independently predicting scores, eliminating the need for manual semantics verification. Looking forward, there is potential for further refinement of the word2vec model specifically for the evaluation of subjective answers within distinct domains. By leveraging larger datasets, it is feasible to substantially expand the number of classification categories or grading levels within the model. The evaluation of subjective answers presents a compelling challenge, and we are optimistic about developing more effective solutions to this complex issue in the future.

REFERENCES

1. Chawla, N. V., Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, 16, 321-357.

2. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781.

3. Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (2015). *From Word Embeddings to Document Distances*. Proceedings of the 32nd International Conference on Machine Learning (ICML-15), 957-966.

4. Miller, G. A. (1995). *WordNet: A Lexical Database for English*. Communications of the ACM, 38(11), 39-41.

5. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). *Indexing by Latent Semantic Analysis*. Journal of the American Society for Information Science, 41(6), 391-407.

6. Salton, G., & Buckley, C. (1988). *Term-weighting Approaches in Automatic Text Retrieval*. Information Processing & Management, 24(5), 513-523.

7. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3(Jan), 993-1022.

8. Joachims, T. (1998). *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. European Conference on Machine Learning, 137-142.

9. Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.

10. McCallum, A., & Nigam, K. (1998). *A Comparison of Event Models for Naive Bayes Text Classification*. AAAI-98 Workshop on Learning for Text Categorization, 41-48.

11. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

12. Zhang, Y., & Wallace, B. (2017). *A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification*. arXiv preprint arXiv:1510.03820.

13. Iyyer, M., Manjunatha, V., Boyd-Graber, J., &Daumé III, H. (2015). *Deep Unordered Composition Rivals Syntactic Methods for Text Classification*. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 1681-1691.

14. Chen, T., &Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

15. Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543.