Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.11291621

Plagiarism Detection Using NLP

Keshav Sharma[1], Mr. Rohit Maheshwari[2]

School of Engineering & Technology,Career Point University, Kota
Email K19767@cpur.edu.in

Assistant Professor, School of Engineering & Technology, Career PointUniversity, Kota
Email rohit.maheshwari@cpur.edu.in

Abstracts:
Plagiarism detection is a crucial task in many fields, including academia, publishing, and journalism. It involves identifying instances of plagiarism, which is the act of copying someone else's work and passing it off as one's own.

One of the most promising approaches to plagiarism detection is using natural language processing (NLP). NLP is a field of computer science that deals with the interaction between computers and human language. NLP techniques can be used to analyze the text of a document and identify features that are indicative of plagiarism.

For example, NLP techniques can be used to identify the presence of unusual word combinations, repetitive phrases, and stylistic inconsistencies. These can be red flags that indicate that the document may be plagiarized.

Another way to use NLP for plagiarism detection is to compare the text of a document to a database of known plagiarized documents. If the document is found to be similar to any of the documents in the database, then it is likely that it is plagiarized.

NLP-based plagiarism detection systems are becoming increasingly sophisticated and accurate. However, they are still not perfect. One of the challenges of NLP-based plagiarism detection is that it can be difficult to distinguish between intentional plagiarism and unintentional plagiarism.

I Introduction:

Plagiarism is act of taking someone else’s work or ideas and passing them off as your own. It can be intentional or unintentional, but it is always wrong. Plagiarism can occur in any context, but it is especially common in academic setting.

There are many different ways to detect plagiarism, but some of the most common methods

involve using natural language processing(NLP). NLP is a field of computer science that deals with the interaction between computers and human (natural) languages. NLP techniques can be used to flag potential plagiarism.

One common NLP technique for plagiarism detection is text similarity analysis. This involves comparing two pieces of text are, the more likely it is that one was plagiarized from the other. Another common NLP technique for plagiarism detection is style analysis. This involves comparing the writing style of two pieces of text to see how similar they are. The more similar the writing styles of two pieces of text are, the more likely it is that one was plagiarized from the other.

There are a number of different NLP techniques that can be used for plagiarism detection. Someof the most common techniques include:

- Text similarity analysis: This involves comparing two pieces of text to see how similar they are. The more similar two pieces of text are, the more likely it is that one was plagiarized from the other. Text similarity analysis can be performed using a variety of different methods, such as n-gram matching, cosine similarity, and Jaccard similarity.
- Style analysis: This involves comparing the writing style of two pieces of text to see how similar they are. The more similar the writing styles of two pieces of text are, the more likely it is that one was plagiarized from the other. Style analysis can be performed using a variety of different methods, such as analyzing the use of words, phrases, and sentence structure.
- Deep learning: Deep learning is a type of machine learning that uses artificial neural networks to learn from data. Deep learning-based plagiarism detection systems can be trained on large datasets of plagiarized and non-plagiarized text. This allows them to learn the patterns and features that are associated with plagiarism. Deep learning-based plagiarism detection systems have been shown to be very effective at detecting plagiarism, even when it is disguised or paraphrased.

Conceptual Framework:

Plagiarism detection involves identifying similarities between two pieces of text to determine if one is a copy of the other. This process can be automated using natural language processing (NLP) techniques such as n-gram analysis, cosine similarity, and stylistic analysis. Effective plagiarism detection systems should be able to identify both direct copying and paraphrased

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.11291621

text.

II Review of Literature:

Plagiarism detection is the identification of stolen or unoriginal written work. This can include copying text without attribution, paraphrasing someone else's work without giving credit, or submitting someone else's work as your own.

There are a number of different approaches to plagiarism detection, including manual detection, electronic detection, and intrinsic detection. Manual detection is the most traditional method of plagiarism detection, and it involves having a human reviewer read and compare the suspected plagiarized work to the original source material. Electronic detection involves using software to scan the suspected plagiarized work for similarities to other sources. Intrinsic detection involves using statistical methods to identify unusual patterns in the writing style of a document that may suggest plagiarism.

A number of different plagiarism detection tools are available, both commercial and open-source. Some popular tools include Turnitin, SafeAssign, and CopyCatch. These tools can be used to check for plagiarism in a variety of different formats, including text documents, images, and audio files.

The use of plagiarism detection tools has become increasingly common in recent years, as the amount of online information has grown and it has become easier to copy and paste text without attribution. However, plagiarism detection tools are not foolproof, and there are a number of ways to bypass them. For example, plagiarists may try to avoid detection by using synonyms or paraphrasing text, or by using machine translation to translate text from one language to another.

Despite these limitations, plagiarism detection tools can be a valuable tool for preventing plagiarism. They can be used to identify potential plagiarism cases that can then be investigated further. They can also be used to educate students about the importance of academic integrity.

In addition to using plagiarism detection tools, there are a number of other things that

can bedone to prevent plagiarism. These include:

- Providing clear guidelines on plagiarism: Providing students with clear guidelines on what constitutes plagiarism can help to prevent them from plagiarizing unintentionally.
- Teaching students about proper citation: Teaching students how to properly cite sources can help to ensure that they give credit to the original authors of their work.
- Encouraging students to use original sources: Encouraging students to use their own words and ideas can help to prevent them from plagiarizing.
- Using a variety of assessment methods: Using a variety of assessment methods, such as essays, presentations, and projects, can make it more difficult for students to plagiarize.
- Creating a culture of academic integrity: Creating a culture of academic integrity, where students understand the importance of honesty and originality, can help to prevent plagiarism.

III Methodology:

The research methodology for plagiarism detection involves a combination of theoretical and empirical approaches. The theoretical approach involves reviewing existing literature on plagiarism detection techniques, framework, and challenges. This helps to establish a foundation for understanding the current state of the field and identity areas for further research.

The empirical approach involves conducting experiments and studies to evaluate the effectiveness of different plagiarism detection techniques and develop new methods. This often involves collecting and analyzing large datasets of text documents, both plagiarized and non-plagiarized, to train and test machine learning models.

Here is a more detailed breakdown of the research methodology for plagiarism detection:

1. Literature Review: Conduct a comprehensive review of existing literature on plagiarism detection techniques, frameworks, and challenges. This involves identifying relevant academic papers, conference proceedings, and technical reports.
2. Problem Formulation: Clearly define the research problem and objectives. This involves specifying the specific challenges or limitations of existing plagiarism detection methods and the goals of the research project.
3. Data Collection: Collect a large and diverse dataset of text documents, both plagiarized and

non-plagiarized. This may involve gathering data from online sources, academic repositories, or conducting controlled experiments.

4. Feature Engineering: Extract relevant features from the text data. This may involve using natural language processing (NLP) techniques to identify features such as n-grams, word similarity, and syntactic structure.
5. Model Development: Develop machine learning models for plagiarism detection. This may involve using supervised learning techniques, such as support vector machines (SVMs) or random forests, or unsupervised learning techniques, such as topic modeling or anomaly detection.
6. Model Evaluation: Evaluate the performance of the developed models using a held-out test set. This involves calculating metrics such as accuracy, precision, recall, and F1-score to assess the models' ability to correctly identify plagiarized and non-plagiarized documents.
7. Result Analysis: Analyze the results of the experiments and studies to identify patterns, trends, and insights. This may involve statistical analysis, visualization techniques, and comparative analysis with existing methods.
8. Conclusion and Future Directions: Draw conclusions from the research findings and discuss potential future directions for research. This may involve proposing new research questions, suggesting improvements to existing methods, or identifying promising areas for further exploration.

IV Result and Analysis

Research in plagiarism detection has yielded significant findings that have contributed to the development of more effective and accurate plagiarism detection tools. These findings have helped to address various challenges in plagiarism detection, including identifying paraphrased text, detecting improper citations, and handling machine-translated plagiarism.

Detecting Paraphrased Plagiarism:

- Paraphrase Detection Models: The use of machine learning models trained on large datasets of paraphrased and non-paraphrased text has shown promise in identifying paraphrased plagiarism. These models can learn to recognize subtle changes in wording and semantic similarity to detect paraphrased content.

- Syntactic Analysis: Analyzing the syntactic structure of text, such as sentence structure and grammatical patterns, can provide additional clues for detecting paraphrased plagiarism. Changes in syntactic patterns can be indicative of paraphrasing attempts.

Identifying Improper Citations:

- Citation Pattern Analysis: Analyzing citation patterns, such as the frequency and consistency of citations, can help identify potential cases of improper citations. Unusual citation patterns may indicate that sources are not being properly referenced.
- Citation Similarity Assessment: Assessing the similarity between cited sources and the text in question can help identify cases where sources are being cited but not accurately incorporated. This can be done using text similarity algorithms to compare the content.

Handling Machine-Translated Plagiarism:

- Language Identification: Identifying the language of the text can help detect machine-translated plagiarism. Statistical methods can be used to determine the original language of the text before translation.
- Machine Translation Detection: Developing algorithms that can recognize patterns and anomalies in machine-translated text can help identify instances of plagiarism involving machine translation. These algorithms can analyze stylistic features and language usage to detect translated content.

Addressing Multimodal Plagiarism:

- Cross-Media Analysis: Developing techniques to analyze and compare content across different media formats, such as text, images, and videos, can help detect plagiarism involving multiple media types. This requires the development of algorithms that can extract meaningful features from different media and compare them effectively.
- Multimedia Plagiarism Detection Models: Training machine learning models on datasets of multimedia content, including plagiarized and non-plagiarized examples across different media formats, can help develop more robust plagiarism detection tools for multimedia content.

Understanding Cultural Differences in Plagiarism:

- Comparative Studies: Conducting comparative studies across cultures to understand different perspectives on plagiarism can help develop culturally sensitive plagiarism detection tools. These studies can identify cultural norms and expectations regarding plagiarism and inform the design of detection algorithms.
- Multilingual Plagiarism Detection: Developing multilingual plagiarism detection tools that can handle different languages and cultural contexts can help address the challenges of detecting plagiarism in diverse settings.

Evaluating Effectiveness in Real-World Scenarios:

- Field Studies: Conducting field studies in actual educational and professional settings can provide valuable insights into the effectiveness of plagiarism detection tools in real-world contexts. These studies can identify challenges, limitations, and areas for improvement.
- User Feedback and Evaluation: Gathering feedback from users, such as instructors, students, and professionals, can help evaluate the usability, effectiveness, and acceptance of plagiarism detection tools in real-world scenarios.

Developing Adaptive and Evolving Systems:

- Machine Learning with Continuous Learning: Implementing machine learning algorithms with continuous learning capabilities can enable plagiarism detection systems to adapt to new forms of plagiarism and improve over time. This can involve incorporating new data, identifying emerging patterns, and updating models accordingly.
- Human-AI Collaboration: Exploring ways to combine human expertise and AI capabilities can enhance plagiarism detection. Human judgment and feedback can be integrated into AI-powered systems to improve their accuracy and adaptability.

Ensuring User Privacy and Ethical Considerations:

- Data Privacy Protection: Implementing robust data privacy measures, such as anonymization,

encryption, and access control, can protect user data and ensure ethical use of plagiarism detection tools.

- Transparency and Bias Mitigation: Developing transparent and unbiased plagiarism detection systems is crucial to avoid unfair judgments and ensure fair treatment of users. This involves explaining the algorithms' decision-making processes and mitigating biases that may arise from training data.

Promoting Responsible Use of Detection Tools:

- Guidelines and Education: Providing clear guidelines and educational resources for the responsible use of plagiarism detection tools can help ensure that they are used effectively and ethically.
- Promoting Academic Integrity: Integrating plagiarism detection tools into broader efforts to promote academic integrity can help create a culture that values original work, proper citation, and ethical practices in education and research.

V Conclusion:

Plagiarism detection has become increasingly important in today's digital age, where information is readily available and easily copied. Effective plagiarism detection tools are crucial for upholding academic integrity, promoting original work, and ensuring fair evaluation in educational and professional settings.

Research in plagiarism detection has yielded significant advancements in recent years, leading to the development of more accurate, efficient, and versatile tools. Machine learning techniques, particularly deep learning, have played a pivotal role in enhancing the capabilities of plagiarism detection systems. These techniques can identify subtle patterns and semantic similarities in text, enabling them to detect paraphrased plagiarism and machine-translated content.

Despite these advancements, challenges remain in addressing the ever-evolving nature of plagiarism. Researchers continue to explore new methods for detecting plagiarism in multimedia formats, cross-lingual contexts, and emerging forms of academic dishonesty.

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.11291621

Future Directions for Plagiarism Detection

As technology advances and plagiarism techniques become more sophisticated, the field of plagiarism detection will continue to evolve. Here are some key areas for future research:

1. Addressing Multimodal Plagiarism: Developing robust methods for detecting plagiarism across different media formats, such as images, videos, and audio, is essential to combat the increasing prevalence of multimedia plagiarism.
2. Cross-lingual Plagiarism Detection: Enhancing the ability of plagiarism detection systems to handle different languages and cultural contexts is crucial for addressing the global nature of plagiarism.
3. Human-AI Collaboration: Exploring ways to integrate human expertise and AI capabilities can further improve the accuracy and adaptability of plagiarism detection systems. Human judgment and feedback can provide valuable insights for refining algorithms and identifying emerging forms of plagiarism.
4. Continuous Learning and Adaptation: Implementing machine learning algorithms with continuous learning capabilities can enable plagiarism detection systems to adapt to new forms of plagiarism and improve over time. This involves incorporating new data, identifying emerging patterns, and updating models accordingly.
5. Addressing Ethical Considerations: Ensuring user privacy, transparency, and fairness in plagiarism detection systems is paramount. Researchers should prioritize data privacy protection, mitigate biases in algorithms, and develop transparent decision-making processes.
6. Promoting Responsible Tool Usage: Educating users about the proper and ethical use of plagiarism detection tools is crucial to maximize their benefits and minimize potential misuse.
7. Collaboration and Knowledge Sharing: Fostering collaboration and knowledge sharing among researchers, educators, and technology developers can accelerate progress in plagiarism detection. This can involve open-source initiatives, data sharing agreements, and regular conferences or workshops.
8. Exploring New Techniques: Investigating emerging technologies, such as natural language generation (NLG) and natural language understanding (NLU), can lead to innovative approaches for detecting plagiarism and identifying potential misuse of AI tools for

generating plagiarized content.

Suggestion & Recommendations / Future Scope:

Suggestions for Future Research in Plagiarism Detection

1. Addressing Multimodal Plagiarism: Develop robust methods for detecting plagiarism zcross different media formats, such as images, videos, and audio. This could involve using image recognition, video analysis, and audio fingerprinting techniques.
2. Cross-lingual Plagiarism Detection: Enhance the ability of plagiarism detection systems to handle different languages and cultural contexts. This could involve developing multilingual corpora, training models on cross-lingual data, and incorporating cultural awareness into detection algorithms.
3. Human-AI Collaboration: Explore ways to integrate human expertise and AI capabilities to improve plagiarism detection. This could involve developing interactive systems that allow human reviewers to provide feedback and refine algorithms, or creating hybrid systems that combine human and AI strengths.
4. Continuous Learning and Adaptation: Implement machine learning algorithms with continuous learning capabilities to enable plagiarism detection systems to adapt to new forms of plagiarism and improve over time. This could involve incorporating new data streams, identifying emerging patterns, and updating models dynamically.
5. Addressing Ethical Considerations: Prioritize data privacy protection, mitigate biases in algorithms, and develop transparent decision-making processes in plagiarism detection systems. This could involve anonymizing data, using fairness-aware algorithms, and providing clear explanations for detection decisions.
6. Promoting Responsible Tool Usage: Educate users about the proper and ethical use of plagiarism detection tools to maximize their benefits and minimize potential misuse. This could involve developing guidelines, providing tutorials, and incorporating responsible use practices into educational curricula.
7. Collaboration and Knowledge Sharing: Foster collaboration and knowledge sharing among researchers, educators, and technology developers to accelerate progress in plagiarism detection. This could involve open-source initiatives, data sharing agreements, and regular conferences or workshops.
8. Exploring New Techniques: Investigate emerging technologies, such as natural language

Career Point International Journal of Research(CPIJR)
©2022 CPIJR | Volume 4 | Issue 2 | ISSN : 2583-1895
December-2023 | DOI: https://doi.org/10.5281/zenodo.11291621

generation (NLG) and natural language understanding (NLU), to identify potential applications for detecting plagiarism and misuse of AI tools for generating plagiarized content.

Recommendations for Future Research in Plagiarism Detection

1. Develop a comprehensive framework for evaluating the effectiveness of plagiarism detection systems in real-world scenarios. This framework should consider various factors such as accuracy, precision, recall, F1-score, robustness to different types of plagiarism, and user acceptance.
2. Conduct longitudinal studies to investigate the impact of plagiarism detection tools on academic integrity and student learning outcomes. This could involve analyzing changes in plagiarism rates, student perceptions, and overall academic performance.
3. Explore the potential of using plagiarism detection tools to promote formative assessmentand feedback in educational settings. This could involve using detection results to identify areas for improvement and provide students with personalized guidance.
4. Investigate the ethical implications of using plagiarism detection tools in different contexts, such as employment screening and intellectual property protection. This could involve developing ethical guidelines and ensuring fair and transparent use of these tools.
5. Promote the development of open-source plagiarism detection tools and datasets to facilitate collaboration and innovation in the field. This could involve creating public repositories, organizing hackathons, and providing funding for open-source projects.
6. Encourage the integration of plagiarism detection tools into educational software and learning management systems to provide seamless and integrated support for academic integrity. This could involve developing standardized interfaces, data exchange protocols,and compatibility across different platforms.

References

1. Alzubaidi, L., Zhang, J., Humaidi, A. J., Duan, Y., Santamaría, J., Fadhel, M. A., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 1-74. https://doi.org/10.1186/s40537-021-00444-8
2. Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00444-8
3. Yu, J., & De Antonio, A. (2022). Deep Learning (CNN, RNN) Applications for Smart Homes: A Systematic Review. *Computers*, *11*(2), 26. https://doi.org/10.3390/computers11020026
4. Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN Framework for Crop

Yield Prediction. *Frontiers in Plant Science*, *10*, 492736. https://doi.org/10.3389/fpls.2019.01750

5. Kotappa, Y. G., Krushika, M., Ravichandra, M., & Pranitha, Mrs. (2022). A Review Paper on Computer Vision and Image Processing. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 2(2), 68.
6. Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306. https://doi.org/10.1016/j.physd.2019.132306
7. P. Garg and A. Sharma, "A distributed algorithm for local decision of cluster heads in wireless sensor networks," *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, Chennai, India, 2017, pp. 2411-2415, doi: 10.1109/ICPCSI.2017.8392150.
8. A. Sharma and A. Sharma, "KNN-DBSCAN: Using k-nearest neighbor information for parameter-free density based clustering," *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, Kerala, India, 2017, pp. 787-792, doi: 10.1109/ICICICT1.2017.8342664.
9. Sherstinsky, A. (2018). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *ArXiv*. https://doi.org/10.1016/j.physd.2019.132306
10. Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, *60*, 102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383
11. Kaur, P. (2023). Artificial Intelligence. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 11(X), 597. https://doi.org/10.22214/ijraset.2022.44306
12. Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*(1), 1-27. https://doi.org/10.1186/s41239-019-0171-0