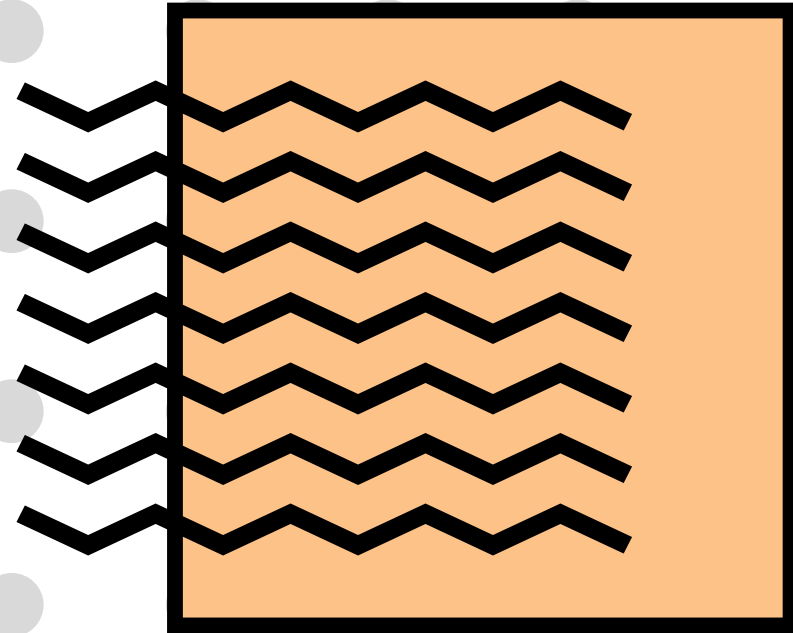


GROUP NO. 3



CASTAWAY ■ ■ ■ CORNER USING ML: PLAGIARISM DETECTION

Mentor:

Ms. Vandana Patil

Professor

OUR TEAM



Monik Kaole

Roll No. 9



Faustina Lazarus

Roll No. 10



Bennet Menezes

Roll No. 11



Dhruv Dave

Roll No. 12



OUTLINE



01

INTRODUCTION

03

METHODOLOGY

05

RESULTS AND
DISCUSSIONS

07

REFERENCES

02

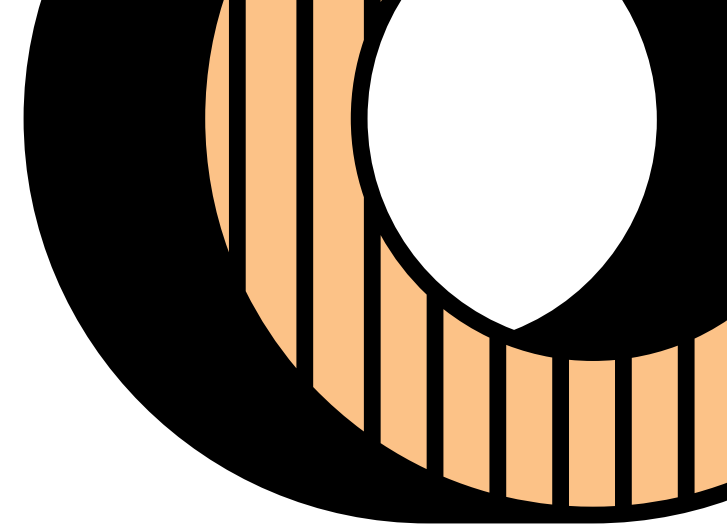
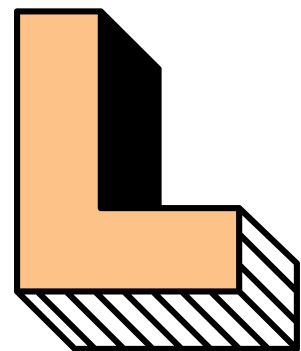
PROBLEM
STATEMENT

04

IMPLEMENTATION

06

CONCLUSION





INTRODUCTION

- **Plagiarism** is an increasingly common and rising problem in numerous sectors when it comes to authoring books or documents.
- Fraudsters employ a variety of plagiarism strategies, ranging from **simple synonym substitution and sentence structure alteration** to a more complicated process combining many forms of transformation.
- Our project provides an instrument for authors and writers to **determine** if their material is original or plagiarised.



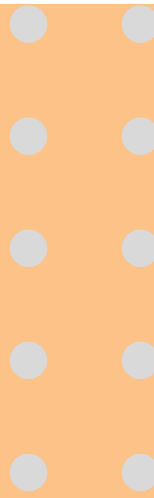
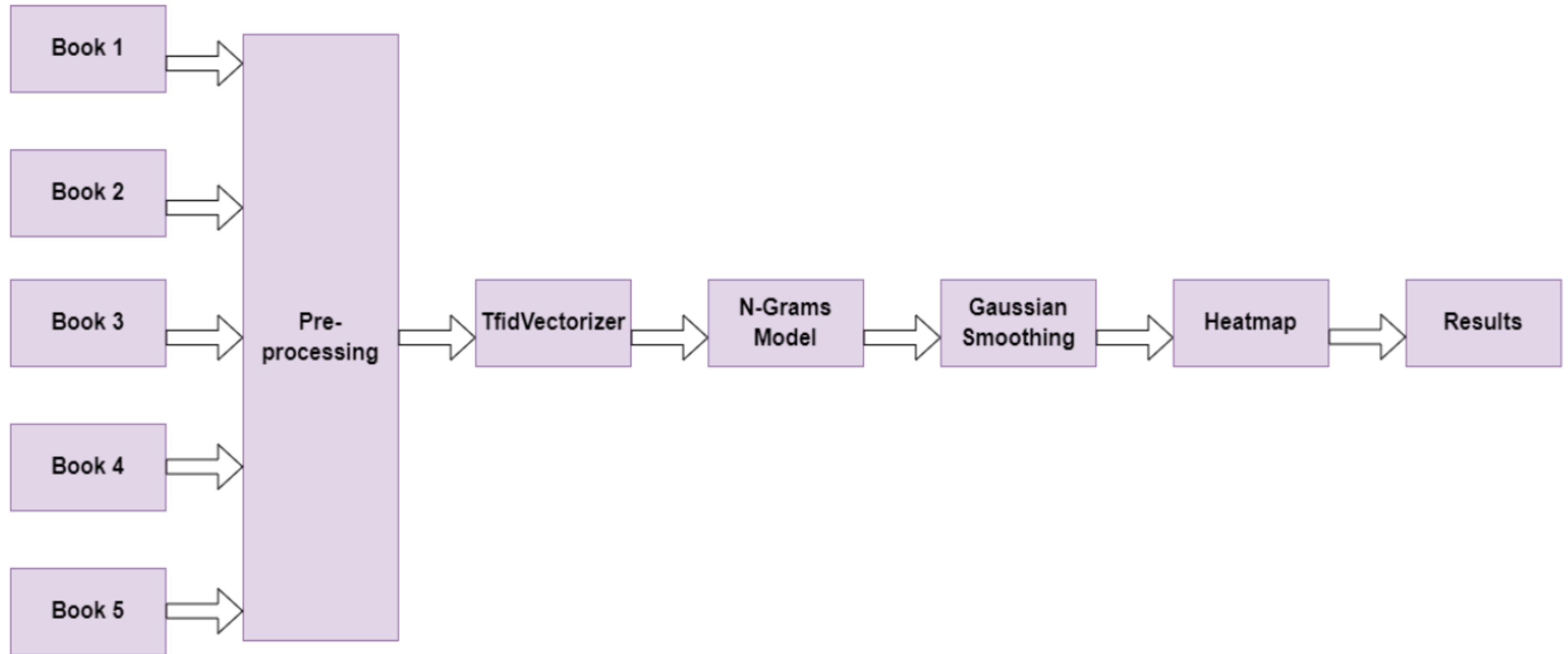
PROBLEM STATEMENT

- Keeping **precise** track of every language and remark may be a real pain at times.
- As a result, the story appears uninteresting and unfinished, and the author loses out on a potentially **lucrative opportunity**.
- The abuse of sources is a problem associated with plagiarism. Writers should be careful not to modify or misrepresent the original text material while utilizing and crediting **sources' ideas** in their essays.





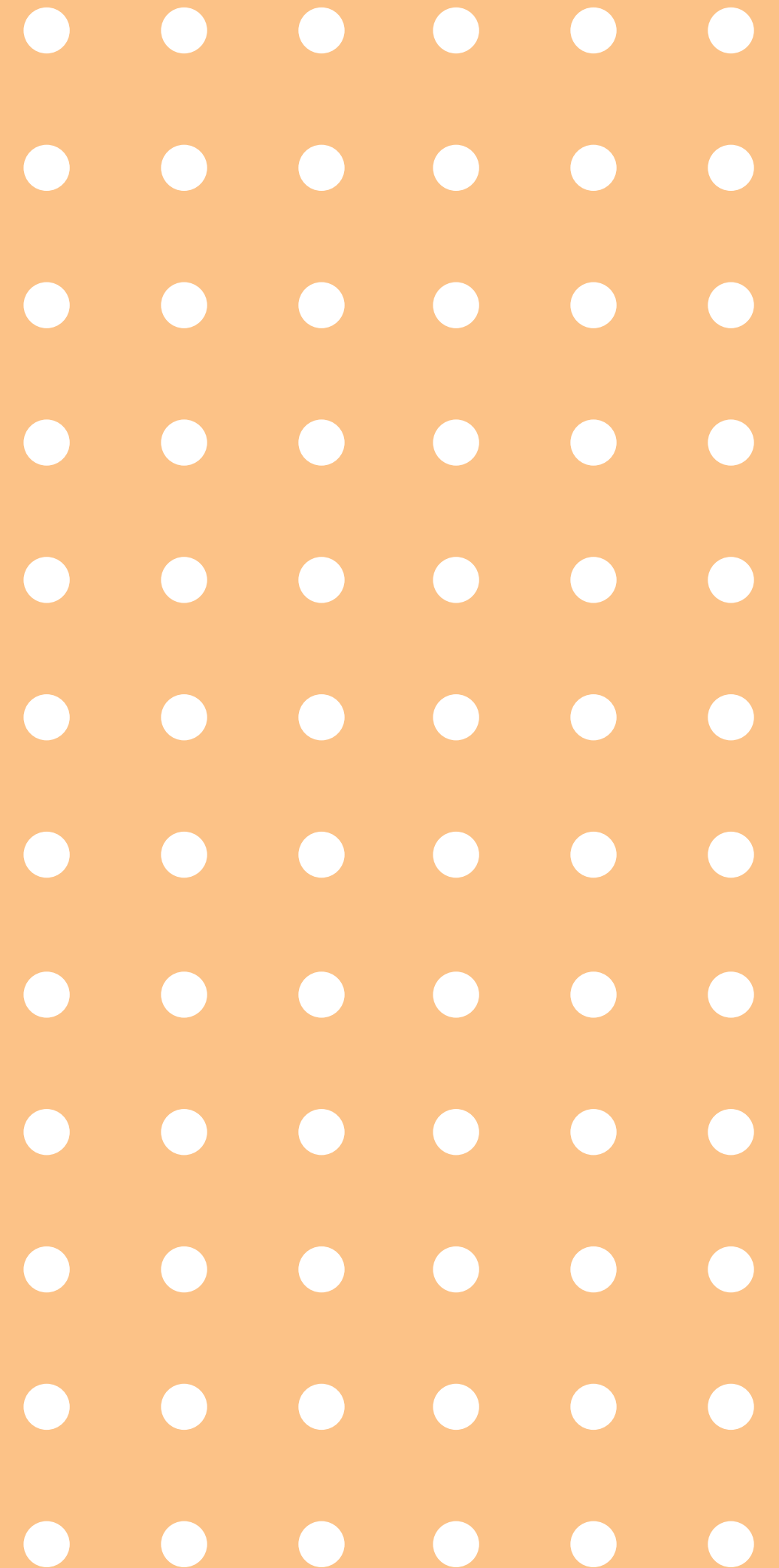
METHODOLOGY





METHODOLOGY

- The **dataset** contains 4 different books written by different authors.
- We will then convert those dataset into **txt format** and will feed into the code or model which we have trained.
- After applying the vectorize and similarity algorithms in the dataset, we get back results in float format which if we multiply by 100 we get it into the percentage the dataset has been **plagiarized**.



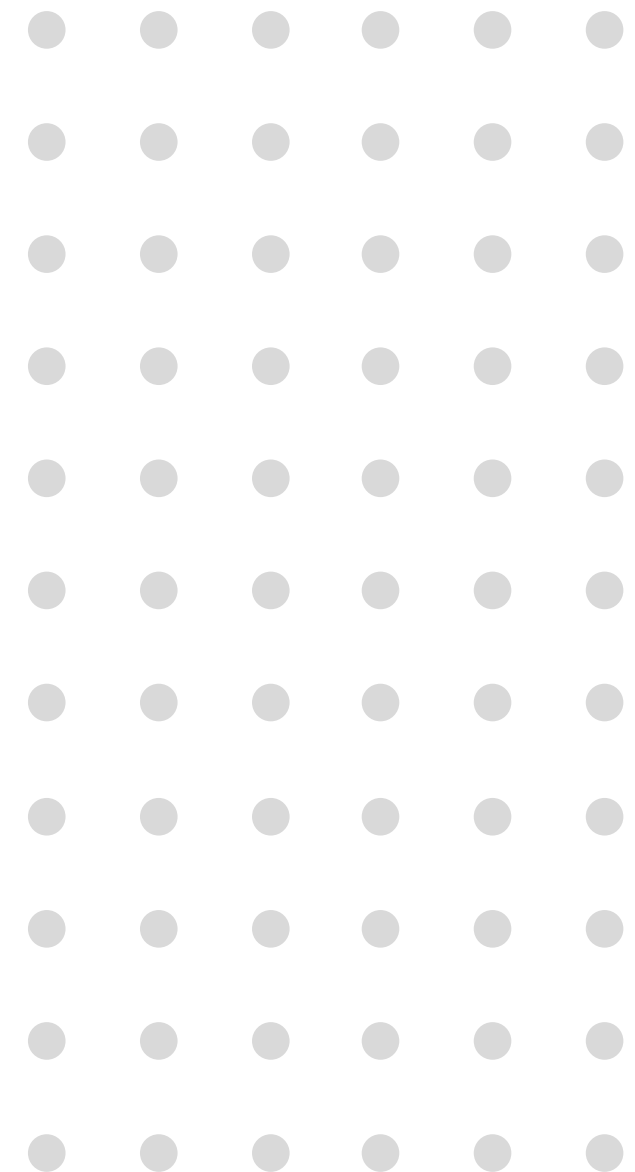


IMPLEMENTATION



The algorithm we are using are:

- **Vector embeddings** are central to many NLP (Natural Language Processing), recommendation, and search algorithms.
- **N-gram Language Model:** An N-gram language model scores words based on the preceding window of context. For plagiarism, however, the emphasis is on copied sequences of words, not on similarities at an abstract level. A paraphrasing should not set off an alarm, but a direct copying should.
- **Visualization:** We can represent a book as a heatmap image where each pixel corresponds to the score of one word. This allows us to quickly gauge if plagiarism is likely, and which parts of a text were most likely to have been plagiarized.

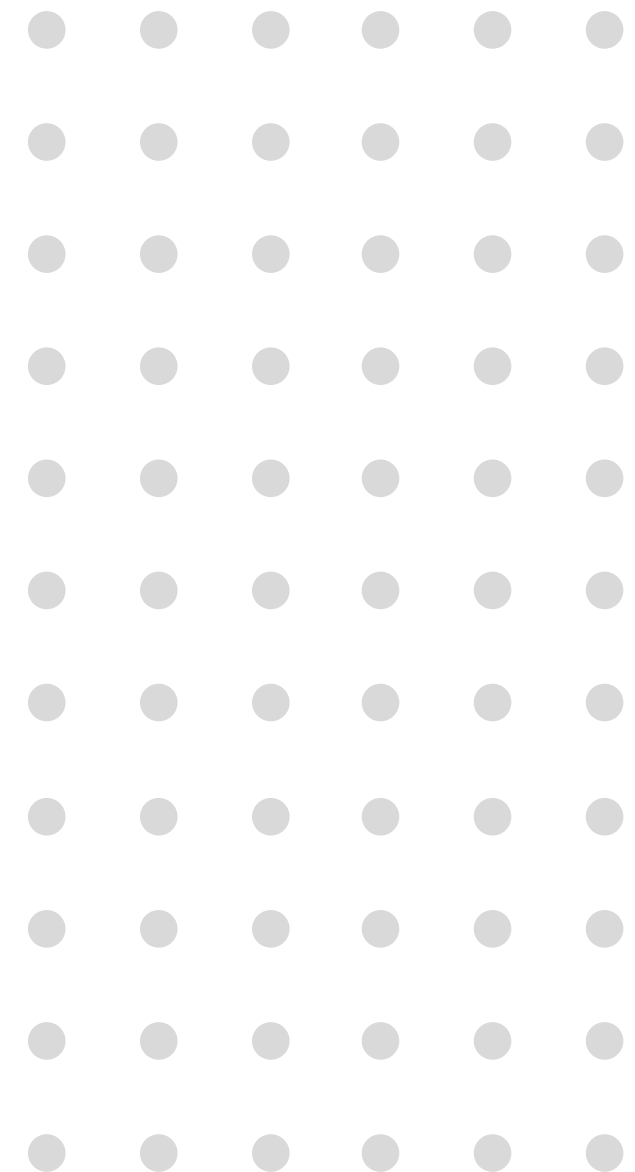




PERFORMANCE METRICS



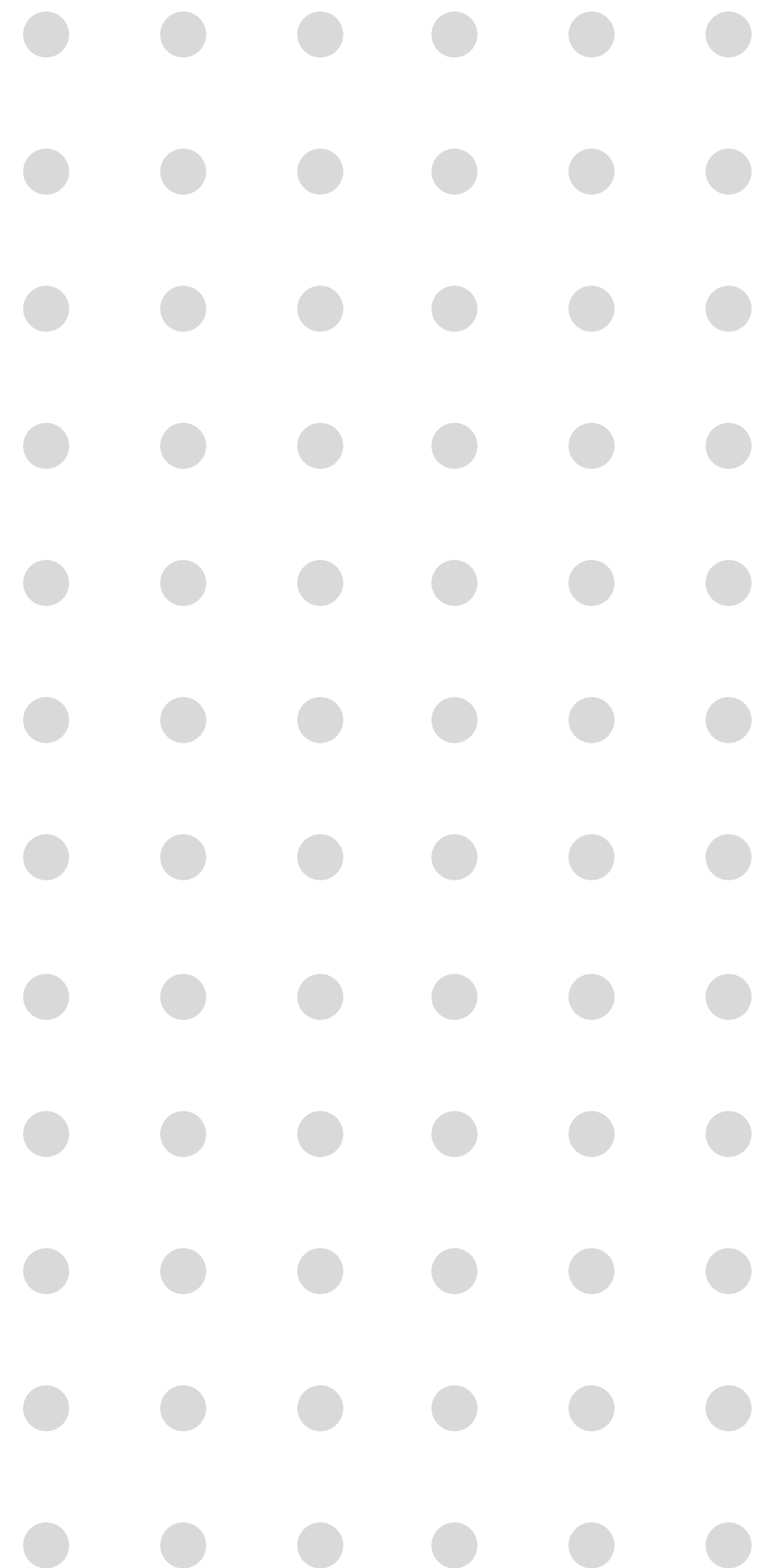
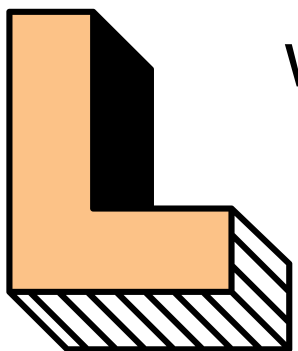
- The accuracy of our Plagiarism model is 89 % calculated using $(TP+TN)/(TP+FP+TN+FN)$.
- Our model has a precision percentage of 97 % calculated using $TP/(TP+FP)$.
- The recall percentage is 54 % calculated using $TP/(TP+FN)$.





RESULT & DISCUSSIONS

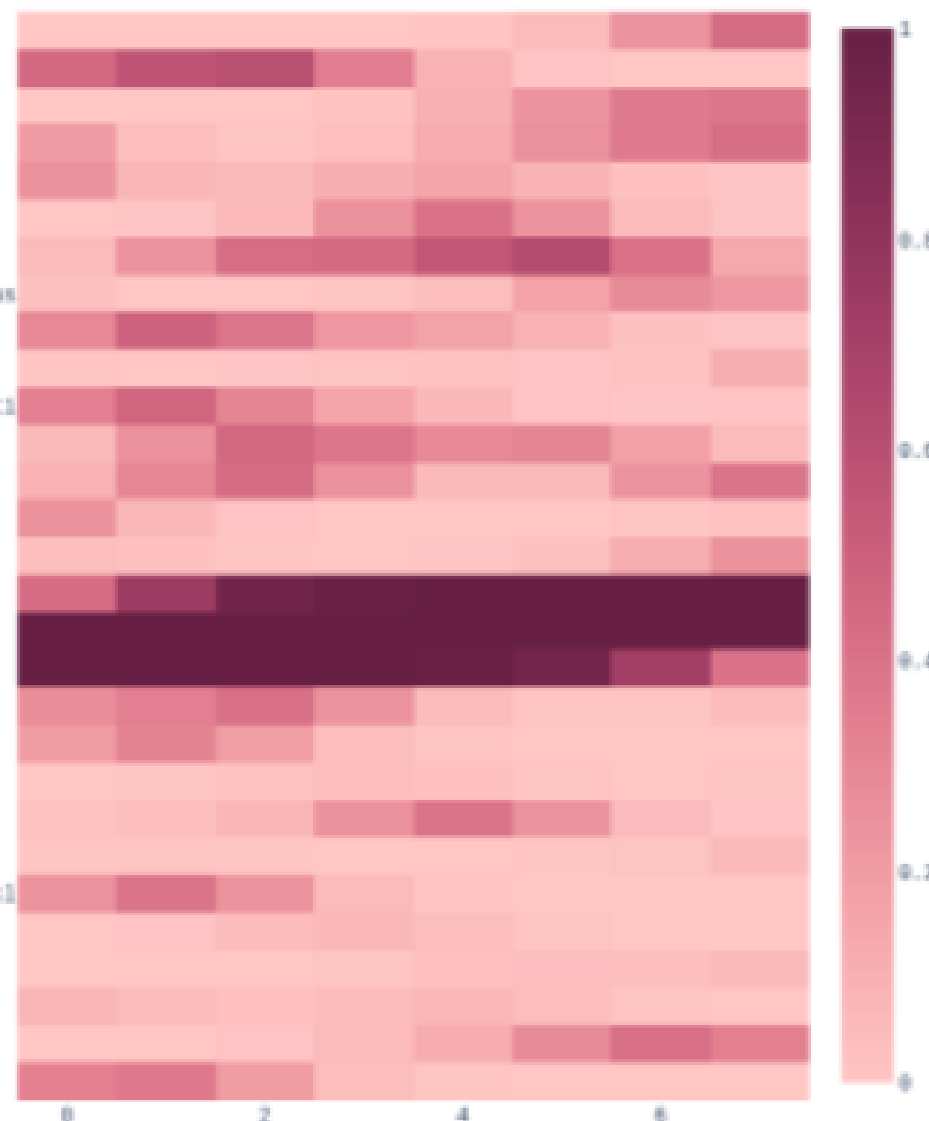
- As the input for the task of **plagiarism detection** is passage-level text, the sentence-level paraphrase recognition system has been modified to handle passages.
- The source and suspicious passages are split into sentences. In order to determine the **closest matching source sentence** for the suspicious passage sentences, the extent of unigram overlap is computed between the sentences in both the passages.
- For every sentence in the suspicious passage, the source sentence, which has the **highest word overlap**, is paired with it.



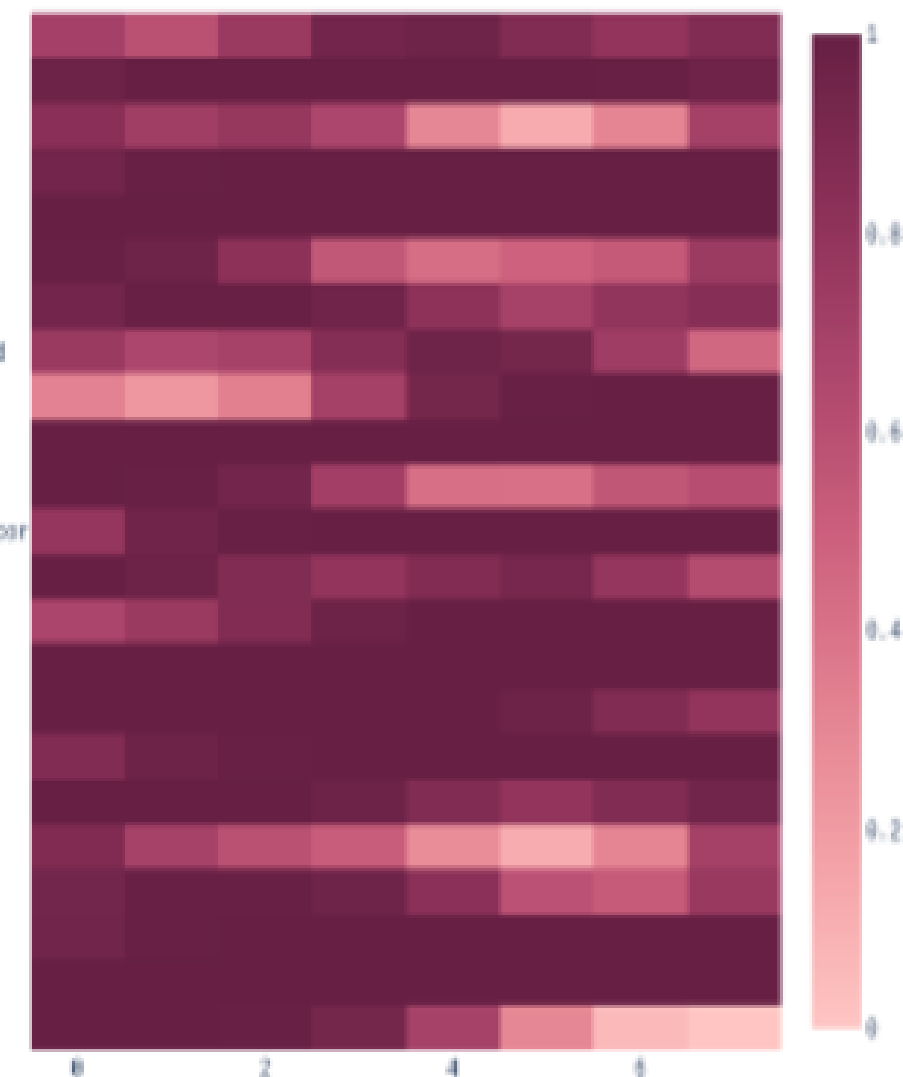


RESULT & DISCUSSIONS

In 1909 Rutherford disproved Sir J.J. Thomson's model of the atom as a uniformly distributed substance because only very few of the alpha particles in his beam were scattered by large angles after striking the gold foil while most passed completely through. Rutherford knew that the gold atoms' mass must be concentrated in a tiny dense nucleus. Encyclopædia Britannica Inc. Most alpha particles pass straight through the gold foil, which implied that atoms are mostly composed of open space. Some alpha particles were deflected slightly, suggesting interactions with other positively charged particles within the atom. Still, other alpha particles were scattered at large angles, while a very few even bounced back toward the source. Rutherford famously said later it was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you. Only a positively charged and relatively heavy target particle, such as the proposed nucleus, could account for such strong repulsion. The negative electrons that balanced electrically the positive nuclear charge were regarded as traveling in circular orbits about the nucleus. The electrostatic force of attraction between electrons and nuclei was likened to the gravitational force of attraction between the revolving planets and the sun. Most of this planetary atom was open space and offered no resistance to the passage of the alpha particles.



The popular theory of atomic structure at the time of Rutherford's experiment was the plum pudding model. This model was invented by Lord Kelvin and further developed by J.J. Thomson. Thomson was the scientist who discovered the electron and that it was a part of every atom. Thomson believed the atom was a sphere of positive charge throughout which the electrons were spread out a bit like raisins in a Christmas pudding. The existence of protons and neutrons was unknown at this time. An alpha particle is a submicroscopic positively charged particle of matter according to Thomson's model. If an alpha particle were to collide with an atom, it would just fly straight through its path being deflected by at most a fraction of a degree. At the atomic scale, the concept of solid matter is meaningless. The Thomson atom is a sphere of positive electrical charge tethered in place by its mass. Thus, the alpha particle would not bounce off the atom like a ball but might pass right through if the atom's electric fields are weak enough to allow it.





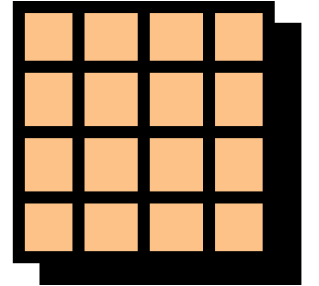
CONCLUSION

- We have developed a new system for the detection of plagiarism based on machine learning methods.
- It's interest is the extraction of characteristics without losing the sense of the document by using vector word embedding technique.
- The proposed system has the ability to detect not only that there is plagiarism but also the probabilities of the existence of each type of plagiarism.
- In future, we plan to further develop our website and add this detection system on our website like buttons, and add a grammatical checker and error checker to increase our resources to a greater extent, in order to help authors and developers to the best of our abilities.

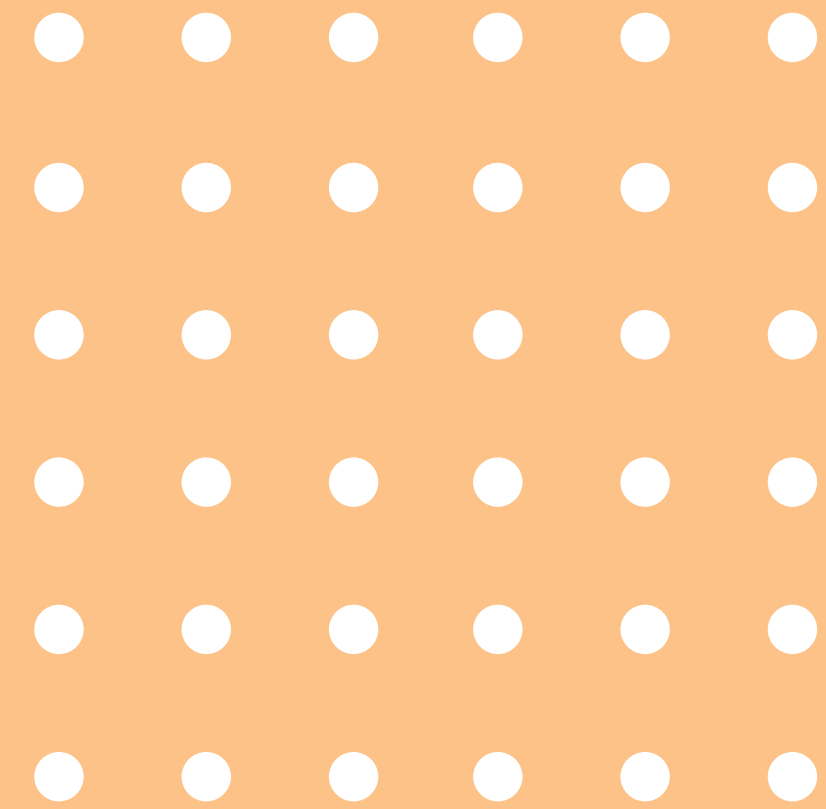




REFERENCES



- [1] S. Autade and A. Suryawanshi, “A Systematic Literature survey on Plagiarism detection Tools” International Journal of Application or Innovation in Engineering & Management (IJAIEEM), Volume 7, Issue 2, February 2018, Pune, India.
- [2] J. Y. B. Katta, “Machine Learning for Source-code Plagiarism Detection”, Master of Science in Computer Science and Engineering, International Institute of Information Technology, July 2018, Hyderabad, India.
- [3] E. M. Hambi and F. Benabbou, “A New Online Plagiarism Detection System based on Deep Learning”, (IJACSA) International Journal of Advanced Computer Science and Applications, University Hassan II, Casablanca, Morocco, Vol. 11, No. 9, 2020



THANK YOU

