TEXT AND IMAGE PLAGIARISM DETECTION

P. BHASKAR[1], BAJJURI NAGA NITHEESHA[2]

[1] Professor in the Department of CSE/MCA at QIS College of Engineering and Technology(Autonomous), Vengamukkapalem, Ongole-523272 Prakasam DT, AP

.[2] MCA Student in the Department of MCA at QIS College of Engineering and Technology(Autonomous), Vengamukkapalem, Ongole-523272, Prakasam DT, AP.

ABSTRACT_

Plagiarism is a major issue in academia and research, and it can affect all areas of the educational system. Plagiarism occurs in a variety of contexts among students, including their schoolwork, projects, essays, and so on. Students who plagiarise by copying assignments and receiving credit for work they haven't done lessen the learning experience they are supposed to be having. Plagiarism in research is being discussed more than ever before. There have been considerable harms to research as a result of web conditions and the ability to conduct complicated and intelligent searches in a short period of time. Text-focused plagiarism detection tools don't pay attention to visuals. An essential aspect of information sharing in articles and scientific research are photographs that communicate enormous amounts of information.. Flowcharts, because they include a lot of information, could be a source of plagiarism because of the wide range of images and the vast number of images contained in computer-generated texts. The goal of this research is to use the Histogram Model to determine how much of a paper contains photos that have been plagiarised.

1.INTRODUCTION

Plagiarism is a hot topic in educational circles all around the world. It refers to the practise of passing off someone else's work or ideas as your own without attribution. In essence, it's a repackaging of previously existent data. Precisely as S. Hannabuss defines it, plagiarism is the act of stealing from someone else's work and claiming it as one's own [5]. So many materials are now publicly available thanks to the enormous popularity of the

internet. The internet is now a vast resource for gathering information. People can simply obtain the information or data they need from the internet and then create a copy of it rather than authoring a text document themselves. As recent trends illustrate, plagiarism detection is becoming increasingly vital because it is so easy for a plagiarist to locate an adequate text fragment that can be copied. Since there are so many different sources, it becomes increasingly difficult to accurately detect plagiarised sections[7]. Plagiarism is a common occurrence in a variety of fields, including academia, journalism, science, and even politics. It is especially beneficial when there is no reference collection or all the possible copy sources are not present, thus document-to-document comparison methods can't be applied. Literal, intrinsic, extrinsic, exact copy and text modification are all forms of plagiarism [3]. In the same way, numerous methods of detecting plagiarism exist. As of right now, word manipulation technologies aren't accurate enough for real-world use. In order to detect plagiarism amongst text sets, we've developed a novel, simple method based on text identification via file transfer and machine learning. Comparing two files yields a percentage number based on our needed threshold value for detecting plagiarism and allowing us to obtain the plagiarised text series in question.

2.LITERATURE SURVEY

The main aim of this research work is to classify the emotional expression from the mouth region of the human face. As the initial task is to extract the mouth region from the facial image, a survey on various existing research works to segment the face expression images is reviewed and discussed.

2.1 P. Hurtik and P. Hodakova, "FTIP: A tool for an image plagiarism detection," 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR), 2015, pp. 42-47, doi: 10.1109/SOCPAR.2015.7492780.

Image plagiarism detection is the focus of this work. To be more explicit, we suggest a way to look for a plagiarised image in a database using our method. The most important criteria for a successful database search are speed and accuracy. S 0's proposed method is based on the F-transform technique, specifically on the F-transform. By reducing the domain dimension by a factor of ten, this method expedites the entire procedure. We demonstrate our method's speed and accuracy through a series of experiments and observations. In addition, we provide examples to illustrate how this approach

can be put to use in a wide range of situations.

2.2 M. Bouville, "Plagiarism: Words and ideas," Science and Engineering Ethics, vol. 14, no. 3, pp. 311–322, 2008.

Academic integrity is jeopardised by plagiarism. It misleads readers, harms the authors who have been plagiarised, and helps the plagiarist. Although these reasons demonstrate the wrongness of plagiarising the intellectual work of others, they are not applicable to the copying of words. Is it more important to steal someone else's ideas or to copy a few sentences that don't have a unique idea? To avoid confusion, the term "plagiarism" should not be applied to actions that are fundamentally different in nature and significance.

2.3 J. Kasprzak, M. Brandejs, and M. Kripac, "Finding plagiarism by evaluating document similarities," in Proc. SEPLN, vol. 9, 2009, pp. 24–28.

Plagiarism is becoming more and more prevalent as the Internet grows in popularity. Plagiarism can take many forms, from simple text copying to more complex concepts that are adopted without acknowledging the original source. For the most part, plagiarism detection research focuses on finding matches in strings. Intelligent plagiarism, in which the same content is conveyed in multiple ways, cannot be detected using this method. An approach to semantic text alignment based on sentence-level topic modelling is proposed in this research. Recalls and estimated plagdets from experiments with PAN corpora were much higher than the winning system in PAN2014. An example of intelligent plagiarism detection shows that topic modelling can be used to identify it.

3.PROPOSED SYSTEM

Training and testing are the two aspects of the proposed system. They are viewed as using the Histogram in the learning phase and the modelling done by this network in the testing phase for the recognition stage during the train phase. Based on correlation rates between query photos and images in database, the data analysis approach selects the images with the most similar correlations to the query image. At this point, the correlation values are reported as the tested image plagiarism, and the expert is responsible for the ultimate interpretation.

3.1 IMPLEMENTATION

1. Upload text dataset: we collect the text dataset from Kaggle website. We used that dataset to compare with other files.

2. Upload image dataset: we collect the image dataset to compare this images from the uploaded images.

3. Use lcs for text plagiarism: we are using the LCS method to check the uploaded file is plagiarism or not. For this we are using NLP algorithm.

4. Use histogram values for image plagiarism: we converted all the image dataset into hist values. Now we are converting the uploaded image into hist values and we compare the dataset hist values with uploaded image hist values.

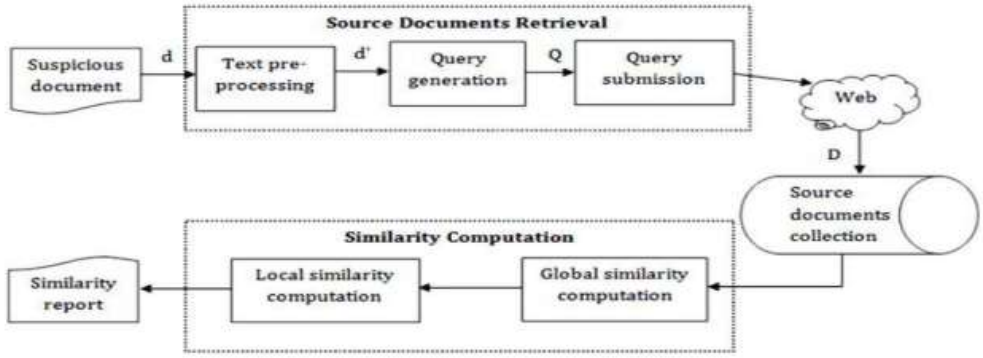

Fig 1:Work Flow

3.2 ALGORITHM DETAILS

It is a branch of computer science, human language, and artificial intelligence called Natural Language Processing (NLP). It is the technology used by machines to understand, analyse, manipulate, and interpret human languages. Translating and summarising texts, recognising and extracting relationships between words, and segmenting topics are just some of the activities that may be accomplished using this tool.

- In a matter of seconds, a user can ask a question about any topic and receive a quick answer thanks to NLP technology.
- In the case of NLP, this means that it provides only the information that is relevant to the question at hand.
- NLP enables computers to converse with people in their own language. a
- It saves a lot of time.
- The majority of businesses use NLP to increase the speed and accuracy of their documentation processes, as well as to locate specific data in massive databases.

3.3 DATASET INFORMATION

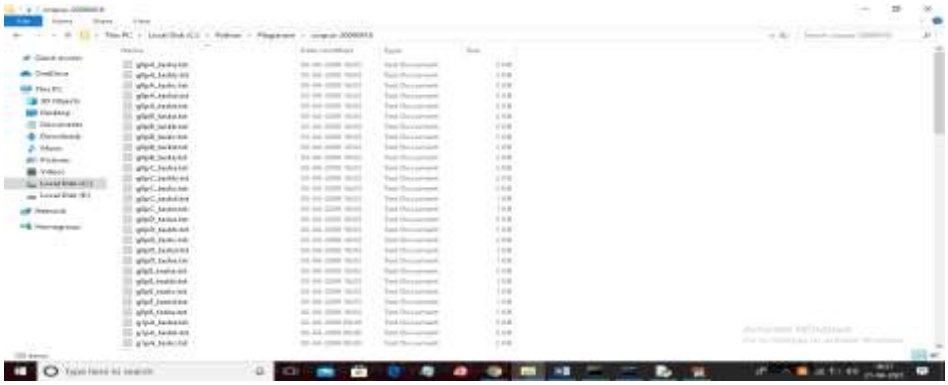

Fig 2:We are using below images to build histogram model and if any suspicious image similarity finds with this histogram then plagiarism will be detected. See below images used to build histogram model

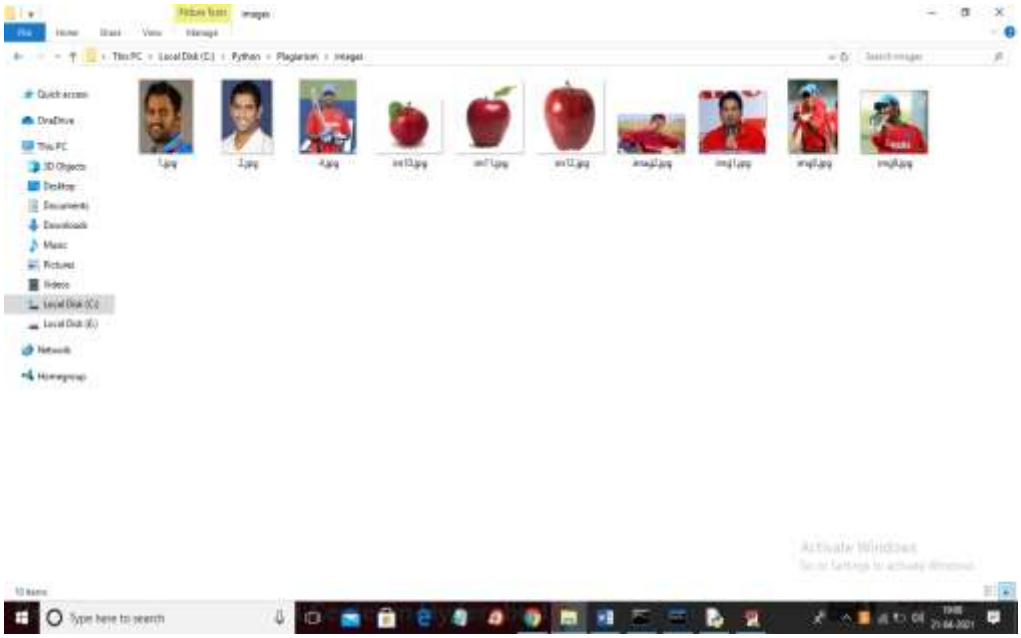

Fig 3:Above images are available inside "images" folder

4.RESULTS AND DISCUSSION

Fig4 :In above screen we can see for database image and uploaded image we generated histogram and we can see there is no match in histogram so no plagiarism will be detected and now close above graph to get below result

Fig 5:In above screen histogram pixel matching score is 15173 out of 40000 pixels so image is not plagiarised and now upload image from "images" folder and see result

Fig 6:In above screen histogram matching score is 40000 which means all pixels matched so plagiarism is detected in above result.

5.CONCLUSION

It is critical to detect plagiarism in order to protect one's work. Students, faculty, and administrators must be aware of plagiarism and anti-plagiarism techniques, according to this study [6]. In this research, we present a simple technique for detecting images and text plagiarised from student assignments. As well as being able to check a huge number of assignments fast and efficiently, it has a high detection rate.

REFERENCES

[1] Imam Much IbnuSubroto and Ali Selamat, "Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine," TELKOMNIKA, Vol.12, No.1, March 2014, pp. 209-218.

[2] UpulBandara and GaminiWijayrathna ,"Detection of Source Code Plagiarism Using Machine Learning Approach," International Journal of Computer Theory and Engineering, Vol. 4, No. 5, October 2012, pp.674-678.

[3] SalhaAlzahrani, Naomie Salim, Ajith Abraham, and Vasile Palade," iPlag: Intelligent Plagiarism Reasoner in Scientific Publications," IEEE World Congress on Information and Communication Technologies, 2011.

[4] BarrónCedeño, A., & Rosso, "On automatic plagiarism detection based on n-grams comparison," In Advances in Information Retrieval, Vol. 5478. Lecture Notes in Computer Science, pp. 696–700, Springer.

[5] Ahmad Gull Liaqat and Aijaz Ahmad, "Plagiarism Detection in Java Code," Degree Project, Linnaeus University, June 2011, pp. 1-7.

[6] A Selamat, IMI Subroto and Choon-Ching Ng, "Arabic Script Web Page Language Identification Using HybridKNN Method," International Journal of Computational Intelligence and Applications, 2009, pp. 315-343.

[7] Michael Tschuggnall and Gunther Specht , "Detecting Plagiarism in Text Documents through Grammar-Analysis of Authors," pp. 241-255.

[8] Bill B. Wang, R I. (Bob) McKay, Hussein A. Abbass and Michael Barlow, "Learning Text Classifier using the Domain Concept Hierarchy," ACT 2600, pp. 1-5.

[9] Francisco R., Antonio G., Santiago R., Jose L., Pedraza M.,

and Manuel N., ―Detection of Plagiarism in Programming Assignments,‖ IEEE Transactions on Education, vol. 51, No.2, pp.174-183, 2008.

AUTHOR PROFILE

Dr. P. Bhaskar Professor in The Department of CSE/MCA at QIS College of Engineering and Technology. (Autonomous), Vengamukkapalem, Prakasam (Dt.) He is having 20 years of Teaching Experience and 13 years of Research Experience and Published more than 20 Research Publications & His area of interest is Artificial intelligence and image processing &Biometric Systems

B. Naga Nitheesha PG Scholar in The Department of MCA QIS College of Engineering and Technology (Autonomous), Vengamukkapalem, Prakasam (Dt). Areas of Interests Image Processing and Machine Learning.