PLAGIARISM DETECTION USING SUPERVISED MACHINE LEARNING ALGORITHM

[1]Shubhesh Amidwar
[1]Student,
[1]Computer Engineering,
[1]Rajarshi Shahu College of Engineering (Pune University), Pune, India

Abstract— Due to large amount of work on various topics of notable authors available online, the work of well-known authors is copied as it is which leads to the spread of plagiarism. Plagiarism has become a worrying problem to various authors. In this project, we attempt to find a solution for this issue by implementing a plagiarism detecting system for detecting whether the work of a notable author is copied or not. We implement this system by using modified version of supervised machine learning algorithm viz. Naïve Bayes. This system will let us know whether the text is plagiarized or not and will also tell us the name of the author whose work is being plagiarized.

Index Terms— Author Detection, Machine Learning, Naïve Bayes, Plagiarism

I. INTRODUCTION

Plagiarism means using someone's work without any acknowledgement of that person. Copying and Pasting work of various authors has led to Copyright Infringement. The plagiarism detection process usually starts with the training phase. In the training phase, used texts of known notable authors are selected from which Stop words are removed. The list of various stop-word is already mentioned in the code. Removing of stop-words from the text is a pre-processing phase. The next phase is the verification phase where the unattributed texts are compared with the previously computed text using a supervised machine learning algorithm. Then, the text is compared to check whether it is plagiarised or not. The system will also give the name of the author's whose work is being copied along the probability.

The proposed system makes use of supervised machine learning algorithm in contrast to the old algorithmic approach which is used in the existing system. This plagiarism system is flexible because the correctness of the machine learning algorithm is trained. This system can be made adaptable to various languages and not only English literature. By changing the set of stop-words in the given code according to any language we can detect plagiarism in that specific language. This system can be used by various people in various fields. These fields might include Journalism, Academics and many more.

II. PROPOSED SYSTEM

The architecture of the proposed plagiarism detection system is given as follows:

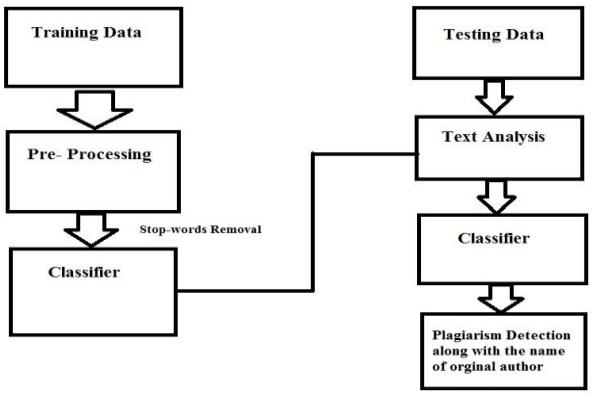

Fig.1

The data flow diagram (level 1) of our plagiarism detection system is given as follows:

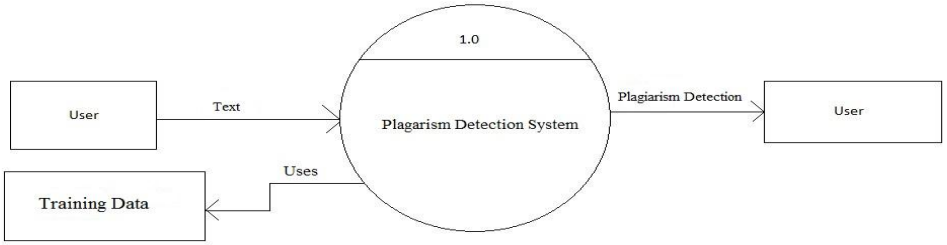

Fig.2

III. IMPLEMENTATION

We used a supervised machine algorithm namely Naïve Bayes. The Naïve Bayes Classifier is a simple probabilistic classifier which is based on applying Bayes rule with robust independence assumptions. A more expressive term for the original probability model would be "Independent feature model". The Bayesian Classification represents a statistical method as well as supervised learning method for classification.

Bayesian Classification provides a useful outlook for understanding and assessing many learning algorithms. Bayesian classification provides real-world learning algorithms and previous knowledge and practical data and be combined. It calculates obvious probabilities for hypothesis and it is vigorous to noise in input data.

In simple terms, a naïve Bayes classifier assumes that the occurrence of a specific feature of class is not related of any other feature. Even if the features depend upon the existence of the other features, a naïve Bayes classifier will consider all of the properties to autonomously contribute to the probability. Depending on the nature of the model, naïve Bayes classifiers can be trained proficiently in supervised learning setting.

An initial probability is called as a prior probability which we get before any additional information is obtained. The probability is called as a posterior probability value which we get or revised after any additional information is obtained.

$$P(h \mid D) = \frac{P(D \mid h) * P(h)}{P(D)}$$

where,
P(h) : Independent probability of h: Prior Probability
P(D) : Independent probability of D
P(h | D) : Conditional Probability of D given h: Likelihood
P(D | h) : Conditional Probability of h given D: Posterior Probability

Snapshot of the User Interface of Plagiarism Detection System:

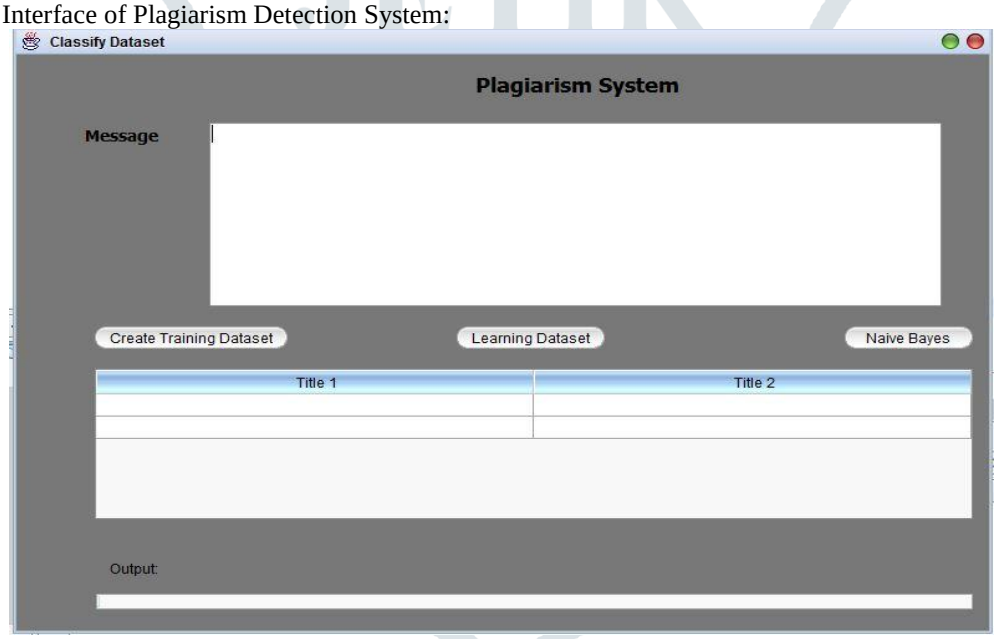

Fig.3

Snapshot of UI of Training Dataset:

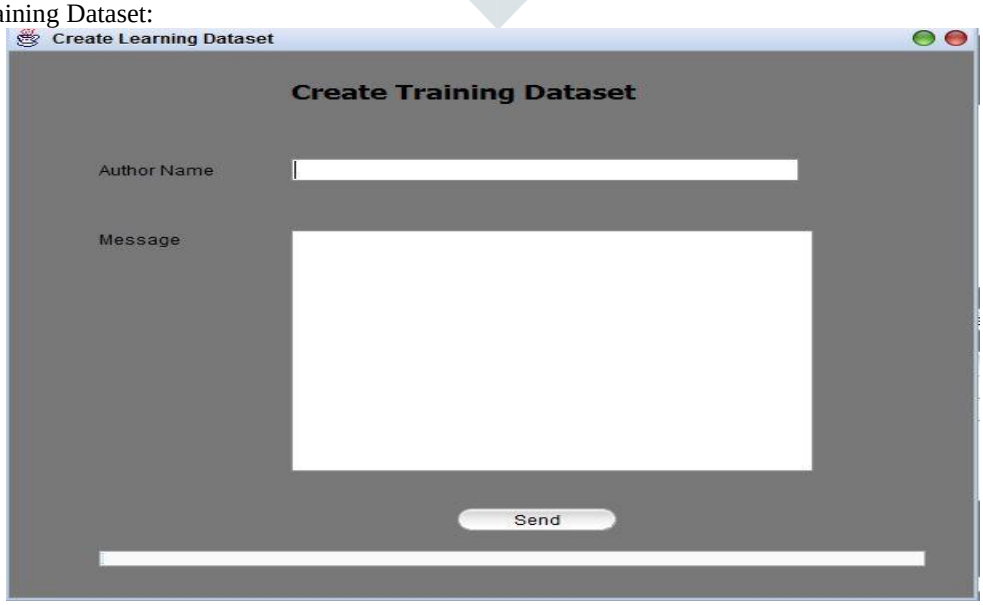

Fig.4

IV. EXPERIMENTAL RESULTS

We now give a text as an input. The following algorithm will give the appropriate result.

I. Results when the text is plagiarized along with the name of the author whose text is copied. The author's name along with the probability is displayed.

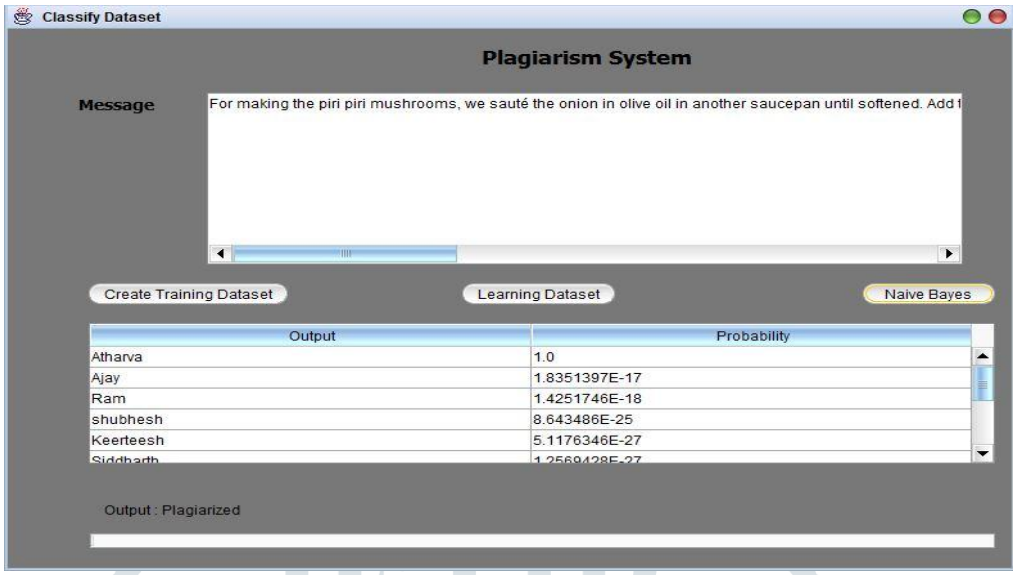

Fig.5

II. Results when the given text is not plagiarized.

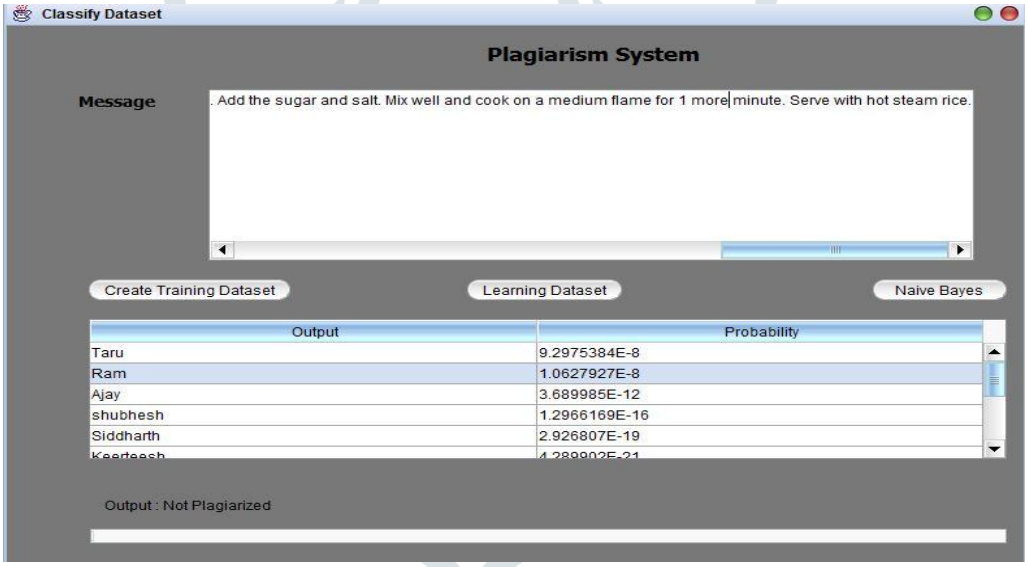

Fig.6

V. CONCLUSION

The project 'Plagiarism Detection using Supervised Machine Learning Algorithm' is to build a system which identifies whether the given text is plagiarized or not along with the name of the author whose work is being copied.

The biggest challenge here is reading of PDF files. Hence, text is entered manually. The future work shall include reading of various types of documents in various formats. We will try to improve the accuracy by using various features. The use of other supervised machine learning algorithms can also be considered in the distant future.

REFERENCES

[1] Xinhao Wang, Keelan Evanini, James Bruno, Matthew Mulholland "Automatic plagiarism detection for spoken responses in an assessment of English language proficiency", IEEE Spoken Language Technology Workshop (SLT), 2016.

[2] Sathiamoorthy Manoharan, "Personalized Assessment as a Means to Mitigate Plagiarism", IEEE Transactions on Education, Year: 2017, Volume: 60, Issue: 2.

[3] Cosmin Strileţchi; Mircea Vaida; Ligia Chiorean; Sorin Popa, "A cross-platform solution for software plagiarism detection", 2016 12th IEEE International Symposium on Electronics and Telecommunications (ISETC), 2016.