Plagiarism Detection and Text Classification Using Implicit Ascertain

Dr. Pradosh Chandra Patnaik[1], Dr. S. B. Kishor[2]

[1]Associate Professor and Head, Department of CSE,
Aurora's Scientific, Technological and Research Academy, Hyderabad.
[2]HoD, Department of Computer Science, S.P. College, Chandrapur

***Abstract*: Plagiarism is often regarded as a relatively minor offense with little impact beyond the plagiarist. The problem of plagiarism has increased recently due to the digital resources available on the World Wide Web. Detecting plagiarism in natural language with statistical methods or data. The purpose of any classification is to build a series of models to predict the class of different objects. Document Categorization is one of these applications and can be used in many classification tasks, for example, the new classification, language detection, Fatherhood performance, the type of text categorization, recommender systems, etc .In this paper we are going to take a tour of the concept of plagiarism detection and text classification method specially n-gram along with their types, functions and area of applicability.**

***Index Terms:* SVM, plagiarism, n-gram, text classification, kNN**

1. INTRODUCTION

Plagiarism is the failure of hardware cost, because almost common than we think. Plagiarism can be considered as the use of others. Plagiarism is often in fact, the decision process is considered an act. Often plagiarism is recognized relatively low guilty, a little, there a plagiarism appears [5]. Today more and more text collections accessible to the public through databases or literature large text documents. Recent developments show beyond a plagiarist, it is difficult to pinpoint when he plagiarized parts of a suitable piece of text that can be copied, it is very easy to find, as the system becomes increasingly important in view of plagiarism, a large volume of potential sources. Ohio University recently clarified by the recent case of plagiarism has been the impact, Athens. [1] Plagiarism and South Korea [2] as part of the Hwang case was reported. Recently, the problem of plagiarism, because the digital age has increased the resources available on the World Wide Web. Statistics or computer natural language plagiarism see [3], [4] for a classification of the various existing digital documents issued copy of plagiarism initiated research, which was begun in the 1990s "Distinction, for example, plagiarism process a semantically equivalent but different words and texts, change the organization, the program and see the description of the documents by a comma, and the ideas of others and be an important role in supporting a deeper understanding of language models" [5].

Plagiarism is usually divided into two categories: real-plagiarism /literal plagiarism and intellectual/intelligent plagiarism, a plagiarist, based on behavior (i.e., student or researcher to commit plagiarism).

1) *Literal Plagiarism*

Literal Plagiarism is a common practice and are important when making a copy does not spend a lot of time on the part of the academic crime is committed. For example, you can simply copy and paste the text from the Internet. [5]

2) *Intelligent Plagiarism*

Intelligent Plagiarism is a serious academic dishonesty in which the kidnappers trying to confuse readers by changing the contributions of others to appear as his own. Smart plagiarists are trying to hide, hide and modify the original work of several smart ways, including text manipulation, translation and the idea of adoption [5].

i) Text Manipulation: Plagiarism can be clouded by the manipulation of text and edit most of their appearance. Instead of words, synonyms / antonyms, short phrases are introduced to change the appearance, but not in the sense of the text. Synonyms for certain terms used in this Toc, although most of the clauses of the change and less left in the summary

ii) Translation: "Confusion makes translating text from one language to another without proper credit to the original source. Plagiarism includes reproduced machine translation (like Google translator) and version of the manual" [6]. "Back the translation plagiarism is (simplified) form of the correct default language text from one to the other and then retranslates back to the first. Clearly translated text can have bad English but plagiarists could spell checkers and other handling plagiarism obscure text"[5].

i) Idea Adoption: "The plagiarism consideration the gravity of the Idea that refers to the use of the views of others, such as performance, contributions, results and conclusions, without citing the source of original ideas" [7]. It is a serious crime to steal the ideas of others, which is a true learning problem. Borrow a few words, but there is no original ideas to improve the quality of English, especially foreign, and should not be construed as plagiarism [8]. Qualitative research showed that university professors blame or to see the different types of the idea of plagiarism to use their own experiences. But the solutions are necessary to see the concept of one mind, as it is important to check the quality of academic work in different ways, including, dissertations, articles, essays and assignments. Idea plagiarism can be divided into three types, but the feathers: the meaning of water in terms of the importance of the idea of home-based and depending on the context. "A close view of the idea of plagiarism can be water- based

description of the two texts, for example, the two pieces, which is the same idea, is expressed in different words. The water-based idea of plagiarism can be made use of, summarize and interpret the data" [5].

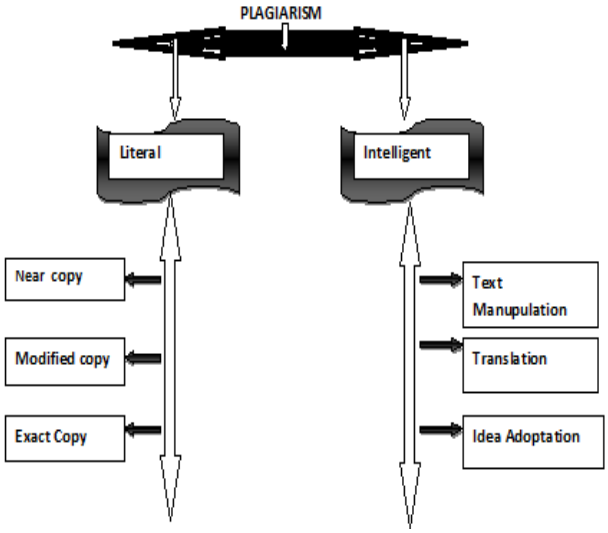

Fig.1.classification of Plagiarism

A holistic approach to the idea of plagiarism can be seen through the context-based adaptation, where the author's structure of ideas (eg, section, subsection, and the logical sequence of ideas), but not necessarily the exact content has been plagiarized from the source. Although the author writes and paraphrases much of the text, while retaining the logical sequence of ideas, this practice is considered plagiarism idea in certain fields of research. [7].

2. PLAGIARISM DETECTION SYSTEMS

Today more text documents available to the public through a large collection of literary documents or database. As recent events show, the result of plagiarism in such programs is very important, as it is very easy to plagiarist to find the correct piece of text that can be copied, on the other hand, is it is difficult to correctly identify the steps fictional account of a lot of possible sources. Two main approaches to identify plagiarism in text books known [9] external/extrinsic and intrinsic algorithms, algorithms when compared with the external document given a set of infinite source of such documents on the web the world of the accused, and intrinsic only to examine the suspicious document. Application techniques often used external methods include the grams [10] or the n-gram [11] comparison, all standard IR techniques as traditional techniques subsequences [12] or machine learning [13]. On the other hand, basic needs approach to understand the writing style of the author in some way and use other services such as the frequency of words and predefined classes [10], the complexity of the analysis [15] or the -gram [14] and the fictional categories. Although most of the external algorithms remarkably better than algorithms using a large set of data intrinsic value received from the Internet, the inherent practical ways in which such a set of available data. For example, scientific documents using information available publications digitally, proof of authenticity is almost impossible to program a computer to do it. In addition, authors can modify the source data so that even advanced text comparison algorithms, fault-tolerant, as the longest common subsequence [16] cannot find a match. In addition, the intrinsic are used as art to help reduce the set of source documents for CPU and / or memory-intensive external processes.

Plagiarism is the process document, analyze its contents, revealing plagiarized parts, and the same original documents, if available [5]. While people have the potential to suspect plagiarism in the same store or look at writing style, "it requires a lot of effort to achieve the potential sources of the material was solid evidence against offenders to provide" [17] . The need for computer systems to detect plagiarism could be due to a failure to process large documents and retrieve all the parts of the original sources suspect.

Compare extrinsic plagiarism is the suspect document against a set of spring collection features a range of data used to the idea of another suspected [18]. Intrinsic plagiarism, strengthen and restore Attribution creativity through still works the same with different end goals. Overall, the style of writing, testing and / or analysis of the complex dimension. They are

i) Intrinsic plagiarism detection doubt;

ii) To check that the data obtained from a particular author or a secure back;

iii) Say that the text writers and writer Imaging

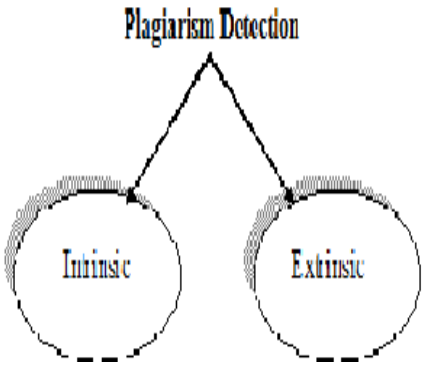

Fig.2: Plagiarism Detection system

3. TECHNIQUES FOR PLAGIARISM

It is not possible to classify all the possible ways the system can be transformed into another of the same (or similar) to work. However, two changes in the general strategies identified.

Changes to realize them in principle are carried out by sophisticated text editor. They do not have sufficient knowledge of the language to be analyzed. General methods are, for example, reformulations add or remove comments; change the formatting; names and identifiers repair. A structural change requires the kind of information that would be needed to analyze it and the "language of the subject. Few examples are cycles (eg while...do to REPEAT...UNTIL or vice versa); nested if statements could be the cause or switch statements: In case of a decision of some of the statements that can be changed without changing the meaning of the program; Calls to subroutines can be tilted, and deciding on operands to be replaced (eg, x <y can be y> x). Our current solution does not deal with structural change. Many of these techniques can circumvent by simply removing all comments and blank tokenizing source application. Process of creating tokens can for example replace all the names of identifiers with the same reasons. A simple method has been very successful in making [19, 20].

4. TEXT CLASSIFICATION

Text Categorization (TC) is a technique that is often used as the basis for applications in document processing and visualization, web mining, surveillance technology, patent analysis, etc. Evaluation of different methods of experimentation, the basis for the choice of a classifier, a solution for a specific problem with. Not a single classifier is always better [21], so that for practical purposes, we need to develop a methodology for the efficient operation. Text Categorization, which is also known as text classification, relates to the problem of automatically assigning given text passages (paragraphs or documents) into predefined categories. Task of text categorization is to classify documents based in predefined classes, their content automatically.

The widespread and increasing availability of text documents in electronic form increases the importance of the use of automatic methods for analyzing the content of text documents. The method of using experts in the field to identify new text documents and maps them well-defined categories is time consuming, expensive and has its limitations. As a result, the identification and classification of text documents based on their content is becoming imperative. A series of statistical learning techniques and machines developed for the classification of text, including the regression model, the k-nearest neighbor, decision tree, Naive Baye, Support Vector Machines, using n-grams and many others. Such techniques are used in many areas of the English language as the language of identification, proof of plagiarism, of authors, the type of text categorization, news categorization, recommendation systems, spam filtering, etc.

i) Support Vector Machines(SVM)

SVM is a new learning method. They are well based computation learning theory and very open to theoretical understanding and analysis. Moreover, in contrast to traditional text classification methods SVM is to be very robust, eliminating the need for expensive parameters. Support vector machines are basically the structure risk minimization [9] calculation of learning theory. The idea of structural risk minimization, it is, and a hypothesis h that we guarantee the lowest true error. SVM is very universal breadwinner. In its basic form, SVM learning linear threshold function. But through a simple plug-in "an appropriate kernel function, it can be used for polynomial classifiers, radial basis function (RBF) network and three-layer sigmoid neural network learning.

ii) k-nearest neighbour

k is The most important parameter in a text categorization system based on k-nearest neighbour algorithm (KNN). In the classification, k nearest documents to test in a training set initially set. Then predication according to the category distribution among these k nearest neighbours can be. In general, the class distribution of the training set is uneven. Some classes may have more samples than others. Therefore, it is very sensitive to the choice of parameter k system performance. Many researchers have found that the ANN algorithm is a very good performance in their experiments on different data sets. The idea behind the K-nearest neighbour algorithm is quite simple. To classify a new document, the system finds the k nearest neighbours among the training documents, and uses the categories of the k nearest neighbours, category candidates [22] mass. One of the drawbacks of KNN algorithm are its efficiency, because it must compare a test document with all samples in the training set.

iii) n-grams

An n-gram is a subsequence of n objects in a specific order. Models in Computational Linguistics n-gram characters (n-grams Level) are used for various applications often predict words (word level n-gram) or predictions. N-gram is a string of length n is

extracted from a document. Typically n a large text is set in a given collection of documents and queries in this corpus, which is the Corpus.

To produce the n-gram vector for the documents, a window length character, moves through the text, a forward sliding through a fixed number of characters (usually one) at one time. The character sequence in the window at each window position is detected.

Many current classifier keywords (which are usually a single word) from the documents. In many classification approaches that keyword, the representative of a different concept or semantic unit is adopted. However, the reality is different: One word can represent different meanings and different words may refer to the same meaning. These are the problems of synonymy. For example, the word bank, part of the memory of a computer, a bank, a steep slope, a collection of some sort, or even a pool shot. The words of the same root word derived tend to produce many of the n-grams to a question with another form of Word documents contain different forms because of the word, help, recovered. Door approach allows us n-gram equivalent expressed as detection word pairs. The N-gram "CO" for example, the first n-grams in the term "natural".

A section of character N-gram is a string-N long. Even in the literature, the term, the term co-occurring set of characters in a character string (for example, an N-gram of the first and third characters of a word)

Dice only process incoming N-gram text genre and counting the occurrences of all n-grams. To do this, the system performs the following steps:

a) Spilled text in separate letters and apostrophes tokens.
b) Scanning of each symbol of the generation of all possible n-grams of N = 1 to the fifth
c) The hash to find a table, counters N-grams, and raises it. The hash table is used, common mechanism to ensure that each n-gram has its own disk
d) When you are ready, give all n-grams and their bills
e) Sort in reverse order with the number of occurrences. Keeping only the N-gram.

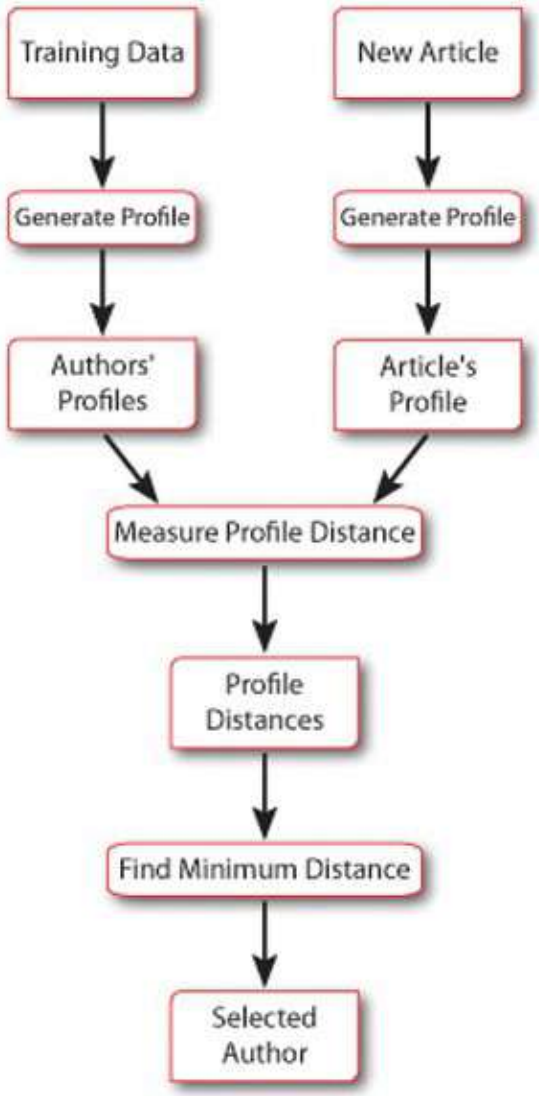

Fig.3.N-gram based Text Classification

N-grams is not a classifier, it is a probabilistic language model, modelling sequences of base units, these basic units may be words, phonemes, letters, etc., the N-gram is basically a probability distribution over sequences of length n and may use when building a

representation of a text. N-gram-based text categorization is also among the methods used in the English language for text categorization, which has good performance.

Parameters	KNN	SVM	N-Gram
Performance	good	better	better
Understanding	easy	moderate	Very simple
Error Detection	medium	large	large

Table.1.Discussion of text classification

5. CONCLUSIONS

This paper discusses different themes in the field of text classification techniques which can be helpful for detecting plagiarism. When the pursuit of time to improve the efficiency of the calculation is to reduce mechanical and various proposals are introduced and researchers working on the same challenge. This work takes us to the distribution of information technology short of plagiarism and text also aims to summarize some of the recent and effective for classification, with new strategies. The task of text categorization is to classify documents predefined classes based on their content automatically. N-gram showed better performance for text categorization.

6. ACKNOWLEDGMENT

We would like to thank to the researchers who motivated us to move ahead in the field of plagiarism and gave guidance that how percentage of plagiarism can be improved.

REFERENCES

[1] R. Tomsho, "Student plagiarism stirs controversy at Ohio University," *Wall Str. J.*, vol. CCXLVIII, no. 38, pp. A1–A10, 2006.
[2] Anonymous, "Misconduct: Dissertation blues," *Science*, vol. 314, p.415, Oct. 2006
[3] S. Brin, J. Davis, and H. Garcia-Molina, "Copy detection mechanisms for digital documents," in Proc. ACMSIGMOD Int. Conf.Manage. Data, New York, 1995, pp. 398–409.
[4] N.Shivakumar and H. Garcia-Molina, "SCAM: A copy detection mechanism for digital documents," in D-Lib Mag., 1995
[5] Salha M. Alzahrani, Naomie Salim, and Ajith Abraham, Senior Member," Understanding Plagiarism Linguistic Patterns, IEEE Trans, Applications And Reviews, Vol. 42, No. 2, March 2012,pp.133-149.
[6] M. Jones, "Back-translation: The latest form of plagiarism," presented at the 4th Asia Pacific Conf. Edu Integr., Wollongong, Australia,2009.
[7] M. Roig, Avoiding Plagiarism, Self-Plagiarism, and Other uestionable Writing Practices: A Guide to EthicalWriting.NewYork: St. Johns niv. Press, 2006.
[8] M. Bouville, "Plagiarism: Words and ideas," Sci. Eng. Ethics, vol. 14, pp. 311–322, 2008.
[9] Martin Potthast, Andreas Eiselt, Alberto Barr´on-Cede˜no, Benno Stein, and Paolo Rosso. Overview of the 3rd International Competition on Plagiarism Detection. In Vivien Petras, Pamela Forner, and Paul Clough, editors, Notebook Papers of CLEF 11 Labs and Workshops, 2011
[10] Gabriel Oberreuter, Gaston L'Huillier, Sebasti´an A. R´ıos, and Juan D. Vel´asquez. Fast- Docode: Finding Approximated Segments of N-Grams for Document Copy Detection. In Martin Braschler, Donna Harman, and Emanuele Pianta, editors, Notebook Papers CLEF 10 Labs and Workshops, 2010
[11] Enrique Vall´es Balaguer. Putting Ourselves in SME's Shoes: Automatic Detection of Plagiarism by the WCopyFind tool. In Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse, pages 34–35, 2009.
[12] Thomas Gottron. External Plagiarism Detection Based on Standard IR Technology and Fast Recognition of Common Subsequences - Lab Report for PAN at CLEF 2010.In Martin Braschler, Donna Harman, and EmanuelePianta, editors, CLEF (Notebook) Papers/LABs/Workshops), 2010.
[13] Jun-Peng Bao, Jun-Yi Shen, Xiao-Dong Liu, Hai-Yan Liu, and Xiao-Di Zhang. Semantic Sequence Kin: AMethod of Document Copy Detection. In Honghua Dai, Ramakrishnan Srikant, and Chengqi Zhang, editors, Advances in Knowledge Discovery and Data Mining, volume 3056, pages 529–538. Springer Berlin, Heidelberg, 2004.
[14] Mike Kestemont, Kim Luyckx, and Walter Daelemans. Intrinsic Plagiarism Detection Using Character Trigram Distance Scores. In V. Petras, P. Forner, and P. Clough, editors, CLEF 2011 Labs and Workshop, Notebook Papers, Amsterdam, The Netherlands,2011.
[15] Leanne Seaward and Stan Matwin. Intrinsic Plagiarism Detection using Complexity Analysis. In CLEF (Notebook Papers/Labs/Workshop), 2009.

[16] L. Bergroth, H. Hakonen, and T. Raita. A Survey of Longest Common Subsequence Algorithms. In Proceedings of the Seventh International Symposium on String Processing Information Retrieval (SPIRE'00), SPIRE '00, pages 39–48, Washington, DC,USA, 2000. IEEE Computer Society.
[17] M. Potthast, A. Barr´on-Cede˜no, B. Stein, and P. Rosso, "Cross-language plagiarism detection," *Language Resources & Evaluation*, pp. 1–18,2010.
[18] M. Potthast, B. Stein, A. Eiselt, A. Barr´on-Cede˜no, and P. Rosso,"Overview of the 1st international competition on plagiarism detection,"in Proc. SEPLN, Donostia, Spain, pp. 1–9.
[19] L. Prechelt, G. Malpohl, and M. Philippsen. JPlag: Finding plagiarisms among a set of programs. Technical report 2000-1, Fakultat fur Informatik, UniversitatKarlsruhe, Germany, 2000.
[20] M. S. Joy and M. Luck. Plagiarism in programming assignments. IEEE Transactions on Education, 42(2):129{133, 1999.
[21] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, 2002
[22] Manning C. D. and Schutze H., 1999. Foundations of Statistical Natural Language Processing [M]. Cambridge: MIT Press.