A Survey Paper on Plagiarism Detection Techniques

Dharmesh Namdev
Shri Vaishanav Institute of Technology & Science,
Road Dist. Indore M.P,
Post: Alwasa

Jayesh surana
Shri Vaishanav Institute of Technology & Science,
Road Dist. Indore M.P,
Post: Alwasa

ABSTRACT
In order to detect and forgeries in digital data different kind of approaches and techniques are recently proposed and implemented but most of the techniques are not able to deliver complete solution for such kinds of frauds and plagiarism therefore to optimize the existing solutions a review of different kind of plagiarism and there taxonomies are reported in this paper.

General Terms
Plagiarism, review article, detection techniques, taxonomies.

Keywords
Plagiarism, taxonomy, POS, feature, text, copy, paraphrasing

1. INTRODUCTION
Plagia rise (Play-ja-rise): To take and use another person's idea or writing or inventions as one's own. The word 'Plagiarism' has taken from Latin word 'Plagiare' which means "To kidnap"[13].

This paper gives an overview of automatic plagiarism recognition technology, with an emphasis on text-independent Recognition. copying is an perform of *scam*. It engages both **theft** someone else's work and **double-dealing** about it subsequently. Plagiarism is a "wrongful approach" and "purloining and publication" of another authors language, thoughts, ideas or expression" and this demonstrate of them as one's personal unique work[14]. The idea stays difficult with indistinct meanings and indistinct regulations. Plagiarism is considered academic dishonesty and a breach journalistic ethics, either it is a crime or in ethics this is a serious ethical offence it is copyright infringement. Usually plagiarism can be classified into two aspect:(1)literal plagiarism and (2) intelligent plagiarism.

After keen understanding of different linguistic patterns, the importance of sentences is determined based on statistical and linguistic features of sentences.We demeanor wide study of state-of-the-art methods for copying detection, contain CNG (character *n*-gram based), fuzzy-based (FUZZY), VEC (vector-based), stylo metric-based (STYLE), SYN (syntax-based), semantic-based (SEM), structural based (STRUC), and cross-lingual methods (CROSS).

2. BACKGROUND
According to the *Merriam-Webster OnLine Dictionary*, to "plagiarize" means:

1) To filch and palm off (the opinions otherwise phrase of different) as ones possess
2) To utilize (another's fabricate) with no praise the source
3) In the plagiarism forgeries are used broadsheet articles and had two tasks for a set of unique articles with intrinsic and extrinsic assessment [4]
4) To demonstrate as novel and unique a thought or invention derived from an existing origin.
5) To commit literary theft

As stated by the Oxford Dictionary, Plagiarism can be explained as taking someone else's effort and dissimulate it to be your individual attempt. Plagiarism is usually exposed by scholars and investigator during their continuous research work [13];

3. TYPE OF PLAGIARISM
There can be various kind of plagiarism. That are divided on the base of the amount to which the duplicate copy has been done. Firstly, there is theft exactly, as its own. This, in further words, is the exactly duplicate of another's work furthermore, there is "The Xerox copy" in which the person transcript significant portions of text straight from a unique source, without doing any changes. Thirdly the amanuensis offer to costume plagiarism by repetition from numerous dissimilar sources, changing few sentences and subsection now and there to create them healthy jointly as keep most of the original drafting. Another kind of copying is "The Self-Stealer" where the author "borrows" from his or her previous works, thus contravene strategy regarding the anticipation of innovation accept by most of the educational institutions. The common strand between the entire over categories is that in the entire over the basis of the unique work is not quoted [13].

There are rejection two beings, no issue what skills they use and how like notions they have, written accurately the similar passage. Thus, written passage, which is stemmed from dissimilar novelists, should be dissimilar, to some amount, excluding for cited segments. If suitable referencing is neglected, problems of plagiarism and conceptual property become apparent. The being of academic fraudulence problems has occurs, if not all, academic foundation and illuminator to set rule against the offence. Taken satisfied of any appearance necessitate straight or ultimately extract, in-text referencing, and referring the prime author in the list of references as shown in given figure [1][12].

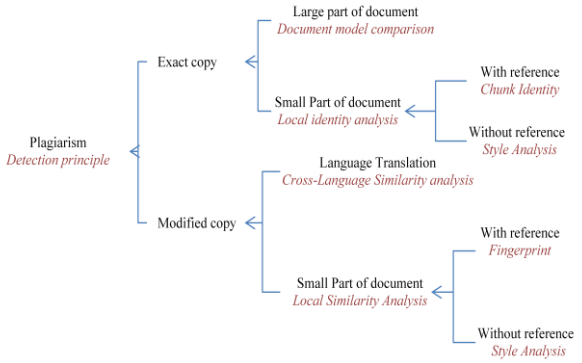

Fig.1. Plagiarism types with some related detection principles

As exposed in Fig. 2. The classification splits copying into two usual types: *literal plagiarism* and *clever plagiarism*, based on the *plagiarist's behavior* (i.e., student's or researcher's way of committing plagiarism).

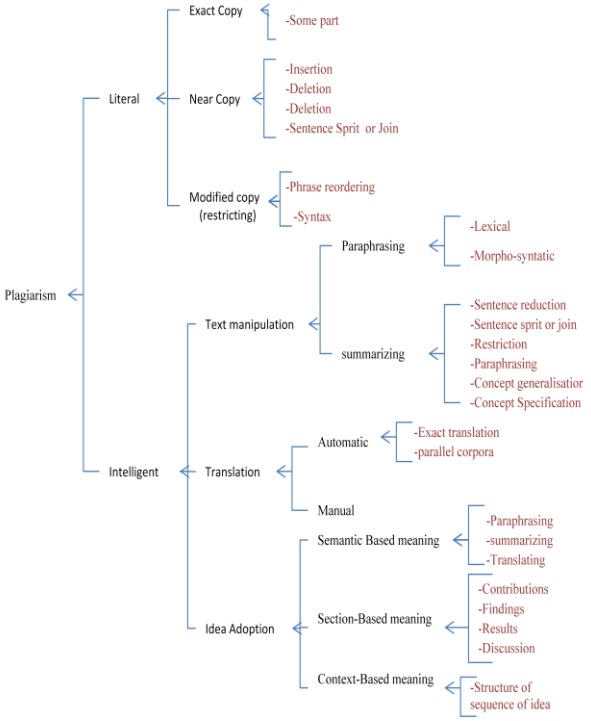

Fig.2. Taxonomy of plagiarism

3.1 Literal Plagiarism

They just copy and paste the passage as of the Internet. A side ofmodifications in the unique passage (noticeable as underscored) [1].

3.2 Intelligent Plagiarism

Intelligent plagiarists try to hide, and change the original work in different clever methods, contain text treatment, conversion, and thought acceptance [1]

3.2.1 Text Manipulation

Copying can be difficult to understand by influence the passage and altering most of its aspects. Be an example of syntactical and lexical summarize, where underscore words are restored with synonyms/antonyms, and small expressions are insert to modify the manifestation, but not the thought, of the passage. Summarize while preserve the semantic.

3.2.2 Translation

Obfuscation can as well be complete by interpreting the passage from one word to one more with no suitable referencing to the unique basis. Interpret copying contains mechanical conversion (e.g., Google translator) and physical conversion (e.g., by people who speak both languages).

3.2.3 Idea Adoption

Idea acceptance is the majority solemn copying that refers to the employ of other's thoughts, such as consequences, assistance, findings, and conclusions, without citing the unique basis of thoughts. It is a main offence to take thoughts of others, which is a real educational difficulty that requires to be examine[1].

4. RESOURCES

There are many resources available around us to create plagiarism. Some of them possibilities are trying to explain here in which plagiarism of any data or document can be occurs.

1. The growth of the Internet and the resources available[5]
2. The web continues to grow at an increasing rate; this provides staff and students with a vast array of material of varying quality[5]
3. The scale of the web has meant it is impossible for most staff to be familiar with all the relevant text in the way they may have been with paper based material[5]
4. Material can be accessed from outside the Institution's campus, thus allowing students to carry out research on the Internet regardless of location[5]
5. As the opportunity of commercial gain from the Internet is recognized, there is a growth in the number of 'Cheat sites' providing either pre-written or customized essays[5]
6. The web provides a quicker way of plagiarizing other people's text by eliminating the need to retype text copied, text can simply be cut and pasted into a word processor document.[5]
7. The use of newspaper articles and has two tasks (single- and multi-document summarization) for two types of intrinsic evaluations [4].
8. Direct read and write from books. These books can be famous or not.
9. And from the social networking sites such as "twitter" or 'face book' and others, person can be easily theft or stealing material.

5. TECHNIQUES TO PRESERVE DATA

It is very necessary to secure data or document after creating or design them. Some techniques are discussed below by which data and communication can be more secure. And plagiarism of documents can be prevented.

5.1 Copy Right

In the combined States and numerous other countries, the appearance of innovative thoughts is thinker intellectual property, and is secluded by patent laws, just like innovative development. Almost all forms of appearance fall beneath patent defense as long as they are record in some medium (such as a book or a processor file) [3].

5.2 Water Marking Of Document

The watermarking is a technique to attain the patent defense of multimedia satisfied. Because the multimedia symbolizes numerous dissimilar media such as text, image, video, audio, and graphic objects, and they disclose extremely dissimilar features in defeat information within them, dissimilar watermarking algorithms suitable to every of them should be residential[7].

5.3 Authentication

Water mark is worn to give authentication. It may be intended in such a method that, any probable modification in the container data either destroyed the watermark or generates a

variance among the substance and the water mark that can easily be detected. [6]

5.4 Copy Control

Watermark may content knowledge necessary by the substance possessor that determined the strategy of repetition the digital contented. The knowledge included by the watermark specifies 'content may not be copied' or 'only one copy' etc. then, the appliance used for repetition the content may be necessary by rule to content water mark detector, which follows directive given by the content owner[6]

5.5 Digital signature

Watermark may be used to identify the owners and by having of this information user may contact the owner of obtain the authorized copy or using the contented [6].

5.6 Finger Printing

Watermark is used to identifying the buyers. This can help to trace illegal copies. When digital media is distributed it can contain hidden information about user thus the licensed copy belonging to a specific user. This also resolves the possible conflict about the ownership. This thing is referred to as "Fingerprint" [6].

5.7 Broadcast Monitoring

Automatic identification of ownership of data is required and used in responsible for monitoring the radio and the television broadcast. This may decide the royalty payments. [6]

5.8 Hash Function

With help of cryptographic hash functions to generate digital fingerprints of so-called text chunks, which are evaluate a database of innovative text route fingerprints. While cryptographic fingerprints recognize a text chunk precisely, the value of these outlookskeepfaithon offset and volume of chunks inside mutually plagiarized and original text [12].

5.9 Secrete Communication

This technique is used to transformation secretly from source to destination in the hidden way. This method is referred to stenography [6]

6. HOW TO DETECT PLAGIARISM

Fig. 3.shows that black-box structure for a copying discovery scheme. It has one basis input that is articulated as a query/distrustful text *dq*, and a further elective input, which is the orientation compilation *D*, such as the mesh. The output is the doubtful fragments/ segment (e.g., paragraphs, statements, etc.), if establish with basis of copying, if obtainable. The subsequent section evaluation a lot of characteristics within the black-box contain tasks, skills, techniques, and evaluation.[1]

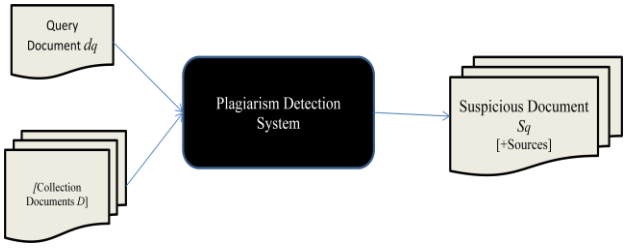

Fig.3. Black-box drawing for plagiarism detection scheme.

7. PLAGIARISM DETECTION TASKS

Copying discovery is split into two official jobs: extrinsicand inherent extrinsic plagiarism detection.

Extrinsic Plagiarism Discovery: Extrinsic copying discovery is a technique of comparinga doubtful document against a set of origin collection whereby various text features are used to doubtful plagiarism.

*Intrinsic Plagiarism Detection:*Intrinsic plagiarism detection is the three similar tasks which are authorship verification, and authorshipattribution so far with different termination aims. In all of them, writing method is compute and/or characteristicdifficulty is investigated.

The dissimilar end aims of these assignments are:

1) To uncertain plagiarism in the intrinsic plagiarism detection.
2) To verify whether the text stems from a precise creator or not in the authorship confirmation; and
3) To element the passage to creators in the authorship ascription.

Text detection process can be divided into two steps:

1) Pre Processing step and
2) Processing step.

Pre Processing is structured representation of the original text. It usually includes:

1) Sentences boundary identification. In English, sentence boundary is identified with presence of dot at the end of sentence.
2) Stop-Word Elimination—Common words with no semantics and which do not aggregate relevant information to the task are eliminated.[2] After that create a list of stop word considered, and weighted term distributed to all words and part of speech (pos) tagging done.

The automated *part of speech (POS)* tagging is a well known problem, In POS tagging, a fortunate label is allocate to every utterance of the verdict. POS labeling is generally worn for lexical passage investigation. In POS labeling, it obtain a verdict as contribution and allocates an exclusive label to every expression in the verdict. it comes to index term extraction[8]. Examples: number of adjectives or pronouns [12]. Index term is a term that catches the essence of the sentence. A subtask of POS Tagging is noun either proper or common is identifying in a document. Nouns are used as a most important feature to express the meaning of the text in natural language processing applications [8]. Study of analysis of two lines or sentences meaning can be define and compare. This investigation provides a permanent design to the files and the sentence include the biased conditions will be stock up in a database. These are the summarized sentences [8] SID (Software Integrity examination technique) thatcalculate this metric through a heuristic pressure algorithm.[9].

Stemming—the purpose of stemming is to obtain the stem or radix of each word, which emphasize its semantics. In Processing step, features influencing the relevance of sentences are decided and calculated and then weights are assigned to these features using weight learning method. Final score of each sentence is determined using Feature-weight

equation. Top ranked sentences are selected for final summary[2].

To detect the plagiarism we have to also consider text summarization of the documents. Text summarization can be classified keen on two- mining and concept. Mining means to choose the slogan or verdicts having the uppermost achieve beginning the innovative passage and mutual to acquire the new shorter passage without altering the basis passage. Concept signifies to probe and interpret the text by using linguistic methods. Extraction method is used to produce the summary in automated text summarization system.[8][1]

8. FEATURES OF (POS) PART OF SPEECH TAGGING

Most stylo metric features are based on the following semiotic features [12]:

1. Text statistics, which operate at the character level. Examples: word distance end to end, question marks, number of commas.
2. Syntactic characteristics, which compute writing manner at the sentence level.Examples: use of function words, sentence lengths
3. Part-of-speech characteristics to compute the exploit of words classes.Examples: amount of adjectives or pronouns
4. Closed-class word puts to analyze extraordinary words.Examples: number of stop words, foreign words, \di_cult" words
5. Structural features, which behave text association.Examples: paragraph distance end to end portion, extent

Several program plagiarism exposure methods have previously been developed. Depended on which attribute assets they employ to evaluate two programs, these method can be generally faction into two groups: attribute-counting systems and structure metric systems. A easy attribute-counting scheme only calculate the number of different machinists, distinct operands, full quantity of operators of all kinds, and total number of operands of all types, and then create a profile with these statistics for each program.[9]

9. OTHER PLAGIARISM DETECTION TECHNIQUES

There are many methods are founded to stop and detect Plagiarism. Some of Important methods are disrobe here:

9.1 Character-Based Methods

When automatic plagiarism detection is taking away examining a position body, a doubtful text is evaluate to a set of actual action in series to relate the plagiarized texts pieces to their potential origin. One of the great complexities in this job is to discover plagiarized particlesthat have been converted (by rephrasing, addition or removal, for instance) from the origin text. The common of copying discovery algorithms rely on quality-based lexical characteristics,word-based lexical characteristics, and syntax characteristics, such as verdicts,to compare the inquiry text *dq*with every applicanttext *dx* ∈*Dx*. Matching series in this background canbe *precise* or *estimated* [1][10].

9.2 Vector-Based Methods

Lexical and grammar characteristics may be contrast as vectors of expressions/tokens rather than string. The resemblance can be compute by via vector resemblance coefficients, i.e., word *n*-gram is represent as a vector of *n* expressions/tokens, verdicts and masses are resent as also term vectors or nature *n*-grams vectors; then, the resemblance can be appraise by using corresponding, Jaccard (or Tanimoto), Dice's overlie (or repression), Cosine, Euclidean, or Manhattan coefficients. A table should generate that explains these vector resemblance metrics with arithmetical symbol and behind illustration [1].

9.3 Syntax-Based Methods

Some research works have used syntactical features to gauge the text resemblance and copying discovery. In fresh study, Elhadi and Al-Tobi and Elhadi and Al-Tobi used POS labels characteristics followed by other series resemblance metrics in the investigation and computation of resemblance among passages. This is based on the perception that like (precise copies) texts would have alike (precise) syntactical construction (series of POS Tags) [1].

9.4 Semantic-Based Methods

A verdict can be delighted as a collection of words prearranged in an exacting order. Two verdicts can be semantically the similar but be different in their organization, e.g., by using the vigorous against inactive voice, or conflicting in their word option. Semantic looms appear to contain had fewer notice in copying discovery, which could be owing to the complexity of representing semantics, and the difficulty of delegate techniques. Li *et al.* and Bao*et al.* worn semantic characteristics for resemblance investigation and obfuscated copying discovery. A technique to calculate the semantic relationship among small ways of verdict span is suggested based on the knowledge removed from a prepared lexical database and quantity figures [1].

9.5 Fuzzy-Based Methods

In fuzzy-based techniques, corresponding remains of text, such as verdicts, befall estimated or indistinct, and realizes a variety of resemblance values that series from one (precisely coordinated) to zero (totally dissimilar). The notion "fuzzy" in copying discovery can be modeled by consider that every word in a text is connected with a unclear set that includes words with similar sense, and present is a amount of resemblance among expressions in a text and the downy set.[1]. FL supplies a easy way to appear at a exact conclusion based winning unclear, vague, indefinite, noisy, or absent input information. FL's loom to manage difficulties mimics how a being would make decision, only much earlier.

9.6 Structural-Based Methods

It is worth noting that all the aforementioned techniques use *flat* characteristics demonstration. In fact, flat characteristic demonstration use lexical, syntactic, and semantic characteristics of the passage in the paper, but do not obtain keen on description contextual resemblance, which are based on the methods the words are used during the essay, i.e., segments and subsections.

10. CURRENT POLICIES

Conference call-for-paper announcements and journal submission guidelines usually have a short statement about the use of previously published results. The ACM policy on prior publication and simultaneous submission allows the submission of "papers that appeared previously in refereed

publications" provided that: "the paper has been substantially revised. Similarly, the IEEE policy expressly states that plagiarism, self-plagiarism, fabrication, and falsification are "unacceptable". Both policies give substantial leeway to the journal editor or program chair to decide when a submitted work meets minimum novelty standards. Both policies emphasize novelty of the new result as an important criterion, and ACM puts a number to it: "at least 25% of the document is thing not published before. though, this is a rather individual obligation that is left up to every publication to elucidate [11].

11. CONCLUSION

Copyright act are promises to preserve the author's wrights of authenticity of originality but due to alteration , modification , insertion , deletion and other kind of operation perform on the digitized data can completely change the mean of concept and on addition there are not a single methodology is available to protect originality of document completely. In near future more literature is collected to find the solution of plagiarism kinds of frauds detection.

12. ACKNOWLEDGMENT

The reported survey is intended to find essential contain for solving over academic issues that is not feasible for industrial used so please verify contain before use in industrial projects.

13. REFERENCES

[1] Salha M. Alzahrani, NaomieSalim, and Ajith Abraham, MARCH 2012. Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 42, NO. 2,

[2] Vishal Gupta, Gurpreet Singh Lehal, AUGUST 2010. A Survey of Text Summarization Extractive Techniques, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3,

[3] What is Plagiarism?Plagiarism Detection and Prevention,Final Report on the JISC Electronic Plagiarism Detection Project Gill Chester 1 August 2001

[4] Text Summarization Challenge 2, Text summarization evaluation,Manabu Okumura, Takahiro Fukusima, Hidetsugu Nanba

[5] Dr. AshishBansal and Sarita Singh Bhadauriya, 2006. Digital Watermark Techniques and Principles An Evolutionary Approach, ASIAN JOURNAL OF INFORMATION TECHNOLOGY 5(10): 1082-1087, MEDWELL ONLINE

[6] Young-Won Kim, Kyung-Ae Moon, and Il-Seok Oh 2003. A Text Watermarking Algorithm based on Word Classification and Inter-word Space Statistics, IEEE Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003) 0-7695-1960-1/03

[7] MANISHA PRABHAKAR, NIDHI CHANDRA,AUTOMATIC TEXT SUMMARIZATION BASED ON PRAGMATIC ANALYSIS. International Journal of Scientific and Research Publications, Volume 2, Issue 5, May 2012 1 ISSN 2250-3153.

[8] Xin Chen, Brent Francia, Ming Li, Member, JULY 2004,Shared Information and Program Plagiarism Detection, IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 50, NO. 7.

[9] Alberto Barr´on-Cede˜no and Paolo Rosso, On Automatic Plagiarism Detection Based on n-Grams Comparison, Springer-Verlag Berlin Heidelberg 2009.

[10] Christian Collberg and Stephen Kobourov, SELF-PLAGIARISM IN COMPUTER SCIENCE, No. 4 COMMUNICATIONS OF THE ACM, April 2005/Vol. 48,

[11] Sven Meyer zuEissen, Benno Stein, and Marion KuligPlagiarism Detection without Reference***Collections.Springer 2007.***

[12] http://www.legalserviceindia.com/article/l222-Plagiarism.html

[13] http://en.wikipedia.org/wiki/Plagiarism