Bharath
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Declared as Deemed-to-be University under section 3 of UGC Act, 1956)
(Vide Notification No. F.9-5/2000 - U.3, Ministry of Human Resource Development, Govt. of India, dated 4th July 2002)
Accredited by NAAC
NATIONAL ASSESSMENT AND ACCREDITATION COUNCIL
Phone : 044-22290742 / 22290125 . Telefax : 044-22293886
Website : www.bharathuniv.ac.in
173, Agaram Road, Selaiyur, Tambaram,
Chennai - 600 073. Tamil Nadu.

Ref No.SMS-2018-O-20 Date: 22/07/2019

BHARATH INSTITUTE OF HIGHER EDUCATION & RESEARCH (Declared as Deemed to be University U/S 3 of UGC Act 1956) 22 JUL 2019 OFFICE OF THE DEAN-RESEARCH Selaiyur, Chennai-600 073. INDIA

TO
Mrs. Dr.M.Sriram,
Asst. Professor/IT,
BIHER.

Thro: Concern Head of the Department

Greetings!!!

We are happy to announce that the Research Advisory Committee has approved your proposal for Seed Money Scheme-2018 which was presented by you. You are requested to complete the proposal and send the progress report to the Dean Research in the prescribed time period.

Title of the Project: CONTENT BASED MULTI- LANGUAGE PLAGIARISM DETECTION TOOL USING BAYESIAN CLASSIFIER

Seed Money Amount: Rs.1, 00,000/- (Rupees One Lakh Only)

Approved on: 17/07/2019

Payment details:

Cheque No.375322 **Dated: 17/07/2019**

Bank Name: Indian Bank, Selaiyur, Chennai.

With Regards

Dean-Research

इंडियन बैंक
Indian Bank

सेलैयूर (तांबरम) शाखा, चेन्नई — 600 073
SELAIYUR (TAMBARAM) BRANCH. CHENNAI - 600 073
IFS Code: IDIB000S246

"VALID FOR THREE MONTHS ONLY"

1	7	0	7	2	0	1	9
D	D	M	M	Y	Y	Y	Y

Mr. M. Sriram. या धारक को OR BEARER

EES रुपये One Lakh Only

अदा करें ₹ 1,00,000/-

A/c. No. CA 6670628110

HMCIA
CBS Code: 02505

'AYABLE AT PAR AT ALL OUR BRANCHES"

Please sign above

⑈375322⑈ 600019250⑆ 29

PROPOSAL SUBMISSION

1. **Details of Principal Investigator**

Name	: Dr.M.Sriram
Designation	: Associate Professor
Highest Qualifications	: Ph.D.
Department	: Information Technology
E-mail	: sriramm.cse@bharathuniv.ac.in
Contact no	: 9952124912
Date of Joining	: 04.08.2014

2. **Details of Co - Principal Investigator**

Name	: Dr.G.Michael
Designation	: Associate Professor
Highest Qualifications	: Ph.D.
Department	: Computer Science and Engineering
E-mail	: micgeo270479@gmail.com
Contact no	: 9940284723
Date of Joining	: 07.07.2008

Technical details

1. Introduction

Nowadays, plagiarism is a really serious issue within the professional environment, or even within the education system. Since Internet is accessible to everyone, it is easy to use Internet as a source of information. However, copying documents from Internet can be considered as plagiarism: what can be found on Internet can come from a book, a research document or an article. It can even result to some legal problems, such as copyright infringement.

Although many of the laws and concepts are not new, the intellectual property concept is relatively recent, dating from the 19th century and this notion has been reassessed the last years with the creation of new online data sources such as Wikipedia, or even through the development of advanced search engines such as Google.

Since then, a new kind of software has emerged, the plagiarism detection softwares. There are several types of them; they can use databases (of thesis, books, or articles), Internet or comparison between files. This report is focusing on the development of an application, mostly using the extraction of data from Internet to check plagiarism and file to file comparisons. Because the project is a typical software development work, the synopsis has main focus on implementing this software; therefore it is mainly a technical report.

Cross-language plagiarism detection tries to automatically recognize and extract plagiarism amid documents in diverse languages. Plagiarized fragments can be translated. Exact replicas may have their structures altered to fleece the replication – this is identified as rephrasing and is far further challenging to identify. Online text publishing minimizes the difficulty of sharing and their reuse by other people. Some people copy text and reuse it without mentioning the authors. The huge amount of data that is provided by online internet resource networks maximizes the difficulty of detecting plagiarism effectively, as it requires more processing time. However, many types of data can be plagiarized, such as audio, text, images, and media clips.

The manual detection of plagiarism is inefficient for the large amounts of data that is published daily. Therefore, the automatic detection of plagiarism is necessary in order to protect the

copyright of original author's work. However, plagiarism detection is not easy and requires a great deal of effort to detect, analyses, and report plagiarism efficiently using expert processes. Therefore, the automatic detection of plagiarism should be intelligent enough to handle the processes of detection accurately. For example, people can rewrite original texts in many styles to avoid plagiarism detection using manual or electronic methods i.e., 25 can be written as 'twenty five'.The study mainly focuses on the design and implementation of an Arabic- English cross-language plagiarism detection method, which spontaneously detects the semantic relatedness among the words of two suspected and targeted documentations. A Linear Logistic Regression (LLR) approach is proposed as a classification approach that is responsible for detecting plagiarism based on two binary possibilities. The respite of this articles is prearranged as monitors: Section 2 designates work associated to cross-language plagiarism detection; Section 3 shows the projected technique; Section 4 explains the experimental results, and lastly, Section 5clarifies the decision.

Automatic plagiarism detection is mostly attentive, but not restrained to, hypothetical environments. Plagiarism approach consisting of alternative person's text inside the very specific canvases without the appropriate quotation (the clean get right of entry to the records via digital resources, such as the Web, constitute a high enticement to dedicate it). Plagiarism primarily based on exact reproduction is the perfect to come across. Conversely, when a plagiarism case suggests redrafting (altering phrases by means of synonyms or altering the edict of part of the text), the challenge turns into drastically tougher.

In plagiarism detection with regard, the doubtful text fragments are associated with a reference corpus with a purpose to novelty the probably source of the plagiarism instances. We have conceded out experimentations constructed totally at the thorough evaluation of reference and apprehensive word-level n-grams. The attained outcomes display that low values of n, besides n = 1 (unigrams), are the first-class desire to technique this undertaking.

2. Review of status of Research and Development in the subject

PLAGIARISM DETECTION METHODS

In both the textual document plagiarism and source code plagiarism, detection can be either:

- Manual detection:

Done manually by human, it is suitable for lectures and teachers in checking student's assignments but it is not effective and impractical for a large number of documents and not economical also need highly effort and wasting time.

- Automatic detection (Computer assisted detection):

There are many software and tools used in automatic plagiarism detection, like Turnitin, Urkund, PlagAware, PlagScan, Check for Plagiarism, iThenticate, PlagiarismDetection.org, Academic Plagiarism, The Plagiarism Checker, Docoloc and more.

FORMS OF PLAGIARISM

Plagiarism can take several distinct forms, including the following):

(1) Word-for-word plagiarism: direct copying of phrases or passages from a published text without quotation or Acknowledgement.

(2) Paraphrasing plagiarism: when words or syntax are Changed (rewritten), but the source text can still be recognised.

(3) Plagiarism of secondary sources: when original Sources are referenced or quoted, but obtained from a Secondary source text without looking up the original.

(4) Plagiarism of the form of a source: the structure of an argument in a source is copied (verbatim or Rewritten).

(5) Plagiarism of ideas: the reuse of an original thought from a source text without dependence on the words or form of the source.

(6) Plagiarism of authorship: the direct case of putting your own name to someone else's work

The easiest form of plagiarism to detect and prove is Verbatim or word-for-word text reuse (given a possible Source text to compare with). This can often be detected Using the simplest of automatic methods, but occurrences By students are often due to the fact that they are uncertain as to how to reuse source texts legitimately.

Other forms, such as paraphrasing and the reuse of structure can also be identified relatively easily, but get progressively harder as the plagiarist uses more complex rewrites or to hide the original text, or reuses only ideas and not the content. The extreme is ghost-writing: getting someone else to write the text for you. These forms of Plagiarism are not just harder to detect, but also harder to Prove.

Some methods were established as a way to discover original-plagiarised text combines on the idea of bendy seek techniques (capable of stumble on plagiarised fragments even supposing they're altered from their supply). If (suspicious & original) textual content fragments are near sufficient, it can be presumed that they may be a potential plagiarism case that prerequisites to be examined profounder. A easy choice is to transmit a contrast of textual content chunks based on word-stage n-grams. In Ferret, the situation and apprehensive texts are fragmented into tri-grams, combining double sets that are later associated. The quantity of commonplace tri-grams is taken into consideration with a view to locate ability plagiarism cases. An alternative is to split the file into chunks of text. PPChecker identifies plagiarized sentences on the idea of creating a repository of grammar and sentence parsing, to avoid tokens in text substrings that may be classified as a plagiarism vocabulary. Wordnet is a manual construct thesaurus that links words into a rigid synonym sets termed synsets. The synsets can be categorized and linked together to form an appropriate relationship. This relationship can then be linked again to create a class/subclass or "is-a" relationship for nouns. Wordnet forms the basis of using a controlled vocabulary for inverted indexing, thus eliminating redundancies. Our method is especially based on a mixture of the primary standards of PPChecker and Ferret. Though, as we designate inside the subsequent phase, the phrase-level n-grams evaluation is not done thinking about sentences or complete files, but in an uneven way (i.e., disbelieving sentence as opposed to orientation record.

This section provides an overview of related works that deal with the uncovering of cross-language plagiarism. Under this topic, Baroni and Bernardini. Conducted experiments within a

domain-specific corpus that consisted of English, Arabic, French, Spanish and Russian texts that were translated into Italian.

They employed the SVM classifier on lemmatized words and POS sequences and obtained the best accuracy through a mishmash of structures including 1-gram word with tf-idf increment, and 2-grams and 3-grams POS tags. They concluded that the task is dependent on the dispersal of ngrams of purpose words and morpho-syntactic structures.

In a related study, Pouliquen et al. illustrated a statistical method that mapped multilingual documents into a language-independent document representation that gauged the similarity between mono and cross-lingual documents. Moreover, Anguita et al. introduced a cross-language plagiarism system for English-translated copies of Spanish document's detection. Their system was comprised of three stages; namely translation detection, internet search and report generation. They classified text paragraphs with the help of supervised learning techniques (i.e., Support Vector Machines) as originally written in a specific language (N for Native language and F for Foreign language).

Furthermore, in a study conducted by BarronCedeno et al. statistical techniques were used to detect cross-lingual plagiarism. Specifically, they made consumption of the IBM Model 1 alliance model, fitted with a statistical bilingual glossary, for the analysis of plagiarism in a similar corpus. Initial studies in English and Spanish text fragments obtained acceptable outcomes, but other experimentations required a cross-lingual corpus for the evaluation phase.

In an extension of the work by Pinto et al., English versus Spanish and English versus Italian leaflets were tested using the IBM Model 1 alignment model based on a bilingual statistical dictionary.

The system unswervingly pinpointed the simultaneous words across different languages. The above revisions indicate that association could be crucial to retrieval tasks involving cross-language information.

Also in the same line of study, was the work by Shiraz and Yaghmaee. They introduced a method based on the overall dependence of textual contents that provided and employed the Vector Space Model (VSM). The method automatically detected bilingual plagiarism from

English Persian. In the context of Indonesian-English cross language plagiarism detection, Alfikri and Purwarianti proposed a method consisting of three primary components, known as pre-processing, heuristic retrieval and detailed analysis. In a recent study, Omar et al. demonstrated a plagiarism detection algorithm using both Arabic and English languages using the 'Bing' search engine. The system supported both languages and used fingerprint and content comparison containing string-matching and tree-matching algorithms. The English publications obtained precision values of 80% while the Arabic publications obtained 90% precision.

Lastly, the pioneering Arabic-English cross language plagiarism recognition, using the Winnowing algorithm, was proposed by Aljohani et al. to detect Arabic verdicts converted from English sources without giving credence to the unique authors.

NAME OF TOOLS/ REFERENCES	Uses	Languages supported
EPHORUS HTTP://WWW.EPHORUS.COM/HHOM	Ephorus is composed of three services: Ephorus internet compares ith documents on the internet, Ephorus group with documents of parallel student groups, and Ephorus database with documents handed in before or at other educational institutes with an Ephorus account	English, Spanish, Portuguese, German, Finnish, Swedish, Norwegian, Danish, Dutch, French, Italian, Polish, Russian, Turkish, Greek, Croatian, Serbian Bosnian, Czech, Arabic
PLAGIARISMSCANNER HTTP://WWW.PLAGIARISMSCANNER.COM/	It is a commercial online plagiarism detecting application which runs against internet resources, that is websites, digital databases and online libraries such as questia or proquest.	
SAFE ASSIGN HTTP://SAFEASSIGN.COM	Safe assign is a plagiarism prevention service which is not independent, but offered at no additional cost as a part of blackboard products (blackboard sells solutions in virtual learning environments).	English, Arabic , Chinese ,Dutch French , German, Japanese, Spanish.
TURNITIN HTTP://TURNITIN.COM/STATIC/INDEX.PHP	Turnitin is commercial anti plagiarism most popularly used system. Turnitin stores and computes unique fingerprint for a given document. It computes detailed document similarities for a	Turnitin supports 19 languages: English, Arabic, Chinese (traditional and

	selected set of documents with similar fingerprint. Internal document storage is composed of archived student papers, journals, periodicals and books. The document storage is being enlarged by automatic web page crawling.	simplified), Dutch, Finnish, French, German, Italian, Japanese, Korean, Polish, Portuguese, Romanian, Russian, Spanish, Swedish, Turkish and Vietnamese.
URKUND HTTP://WWW.URKUND.COM/INT/EN/	Ukund is an automated online plagiarism detection system. Its system is easier to use than previous ones, since the entire process is automated by email sending (no need to access to a site or login),and therefore it only requires that you know how to send and read emails	
NOPLAGIAT.COM HTTP://WWW.NOPLAGIAT.COM/	Noplagiat.com is a French online detection tool with a very simple interface. It can be interesting if you are looking for something very simple to use, or for an occasional use, with no subscription. The user sends his documents to the site through a form, then he launches the analysis and the engine checks for similarities with contents found on the internet or also in an internal database.	
COMPILATIO.NET HTTP://WWW.COMPILATIO.NET/EN/	Compilatio.net is an interesting, Antiplagiarism solution, since it offers a different point of view from other tools.	
POMPOTRON.COM HTTP://WWW.POMPOTRON.COM/	The user sends his document and the plagiarism detection is launched. Many formats are supported for the detection. Once the analysis done, the user gives his mail address and is directed to the payment step.	
PLAGIARISMDETECT HTTP://WWW.PLAGIARISMDETECT.COM/	The fact that plagiarism detect emphasizes this plugin reflects that the concept of saving time by doing the plagiarism detecting task directly in an editor is as interesting as an online solution for some people. However, it limits the format of documents to Ms word, which is not very practical	
PLAGIARISMDETECT	Plagiarism detector is a standalone computer	

OR HTTP://PLAGIARISMDETECTOR.COM/	desktop application for plagiarism detection, which runs only on windows.	
EVE2 HTTP://WWW.CANEXUS.COM/	Eve2 is another commercial anti plagiarism system. For an input document it returns links to web pages from which an author could plagiarized.eve2 uses" advanced searching tools" to locate suspect sites. It compares both the given and the found document and highlights" plagiarism" in red.	These systems were developed only for English, while other programs were adapted to deal with French, German and Chinese languages
COPY CATCH HTTP://WWW.CFLSOFTWARE.COM/	A Uk system which concentrates on comparison within a group of students. The software compares text from work collected by email or on disk using a similarity threshold that will detect essays which are very similar or dissimilar to other class essays by communality of words and phrases	
FREE ONLINE		
PLAGIUM HTTP://WWW.PLAGIUM.COM/	Plagium is a very simple online plagiarism detection tool. You just have to paste your original text, and Plagium will search for redundancies over the web. There are many free, online tools, but most of them look like Plagium, meaning they are very simple, with just a copy- paste system.	
SEE SOURCES HTTP://WWW.PLAGSCAN.COM/SEESOURCES/	Seesources.com is also an online plagiarism detection tool. It resembles to Plagium and many other free online tools, but here you can also load documents in Ms word, html and text format.	
COPY SCAPE HTTP://WWW.COPYSCAPE.COM/	Copy scape is another variant of free online tool; nevertheless its particularity is that it is designed for checking plagiarism of web pages only.	
PLAGISERVE HTTP://WWW.P	The service is based in Ukraine.	

LAGISERVE.COM/	
DUPLI CHECKER HTTP://WWW.DUPLICHECKER.COM/	Dupli checker just automates a process that the user could do himself.
PLAGIARISM CHECKER HTTP://WWW.PLAGIARISMCHECKER.COM/	Plagiarism checker is also a free online plagiarism detection tool. Automatically adds the quotation marks and special operators for you

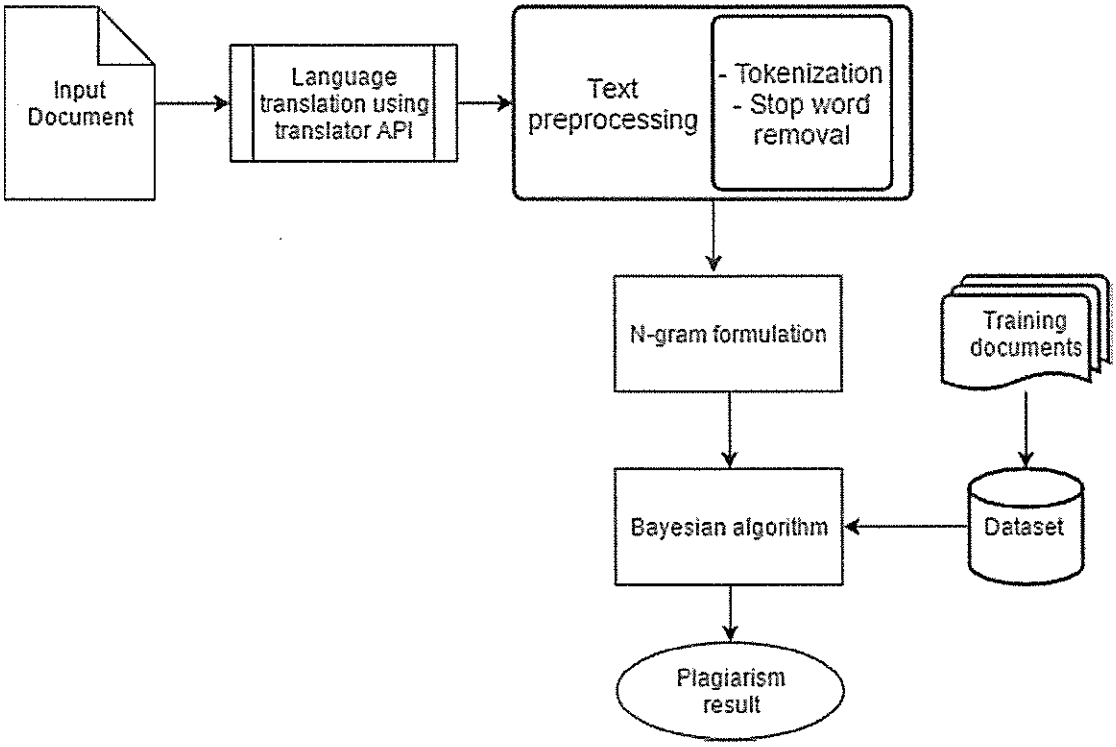


Figure: 1 Block Diagram

2.1 International Status: Nil

2.2 National Status: Nil

3. Progress/achievement so far,

1. The given document which may be in any language is converted into English using goslate language API.
2. The language corpus is considered as a dataset which will have large volume of document samples. Higher the size of corpus, more accurate the prediction will be.
3. Then the given document is tokenized, means the whole document is splitted as individual words.
4. 1-gram, 2-gram and 3-gram models are created.
5. The created model is compared against the models of corpus to predict plagiarism probability.
6. Higher the probability value, more the document suffers from plagiarism.

4. Work Plan:

4.1. LANGUAGE-ENGLISH PLAGIARISM DETECTION

The framework of the proposed Language-English plagiarism detection technique, which illustrates all stages of the process, can be seen in Figure 1.The proposed method consists of six major phases as follows:

Text pre-processing: There are two main processes of the Text preprocessing phase; (1) tokenization, (2) stop word removal. The main aim of this phase is to prepare the original document's dataset for similarity comparisons with extra texts.

i. Tokenization: Tokenization is the technique of infringement up a circulation of text into phrases, phrases, symbols, or different significant elements; known as tokens. The listing of tokens develops the center for auxiliary processing, which includes parsing or textual content mining. Tokenization is beneficial, both in semantics (wherein it is a shape of textual content dissection) and in laptop technology (in which it paperwork a part of lexical evaluation). Classically, tokenization happens at the phrase level. However, it is once in a while hard to describe what is destined by using a "phrase." A tokenizer frequently trusts on artless heuristics, as a model:

i. Punctuation and whitespace may or may not be comprised in the resultant slope of tokens.

ii. All connecting sequences of alphabetic charms are part of one token; similarly with numbers.

iii. Tokens are unglued by whitespace appeals, such as a space or line break, or by punctuation charms.

In languages that use inter-word spaces (including most languages that use the Latin alphabet, and maximum programming languages), this tactic within reason truthful.

Conversely, even right here there are numerous facet cases, including contractions, hyphenated phrases, emoticons, and large constructs, consisting of URIs (which for some functions may also substance as single tokens).

In the proposed method, the tokenizer breaks up a stream of input text of the Language file into words, phrases, symbols, or other tokens Moreover, the English text resulting from the translation of the Language text is also broken up into tokens by using the same tokenization process.

ii. Stop words removal

In natural language processing, forestall phrases are phrases that are sifted out earlier than or after the dispensation of natural language facts (text). There is no any single regular slant of sojourn words used by all natural language processing gear; and undeniably, not all utensils even use this kind of listing. Some equipment specially avoid removing those stop words to support phrase searches. Any institution of words may be chosen as prevent words for a given reason. For plagiarism, a number of the maximum commonplace prevent phrases are brief function words, including, such as 'in', or in Tamil as 'இல்'.In this instance, stop words, if not filtered, can basis glitches when searching for words that comprise them; principally words such as ' இருந்து ' ,'அது,' or 'மீது' which mean "from", "it" or "on", respectively. In addition, and regarding English textual content, after the tokenization step has divided the person sequence, the subsequent step is to do away with the forestall phrases, which include preposition question and auxiliary verbs. The list of the English stop words that has been used in this study is a default English stop words list, and is a well-known list used by many researchers, including. In addition, the Langugae stop words list is taken from a study. After the tokenization has been

done and the words, stop words and special characters have been identified, the stop words removal is applied.

iii. n-gram model

N-gram that compares substring by means of substring to decide the number of comparable substrings that exist in both sentence1 (s1) and sentence2 (s2). Where c is the number of commonplace substrings among each is the total number of substrings in sentence1 (s1) and sentence2 (s2).

$$2*c/(|S_1|+|S_2|)$$

Hence, the most number of substrings of some unique length is the quantity of phrases within the sentence. Therefore, the number of words of the pointer desk is enough for dealing with substrings.

Probability Calculation

A plagiarism detection machine can be assessed as a type system; wherein each verdict goes to certainly one of training: plagiarized or unique. In this study, 3 assessment techniques are used; precision, take into account and F-Measure. The consequences of plagiarism detection can be disbursed as 4 kinds: actual high-quality, authentic negative, false fantastic and fake poor [16]. TP- True Positive is a fixed of plagiarized quantities previously noticed by way of the machine. TN- True Negative is a hard and fast of non-plagiarized parts and the system chooses them as such. FP- False Positive is a set of non-plagiarized parts, however the machine spotted them as plagiarized. FN - False Negative is a hard and fast of plagiarized components, however the machine did now not detect them. In phrases of these four units, consider may be described as follows: the bear in mind degree is distinct as the ratio of pertinent plagiarized quantities perceived by the system. Recall is a fragment of effectively categorized check instances divided with the aid of the quantity of check cases manually classified as similar. The second overall performance metric is precision. The exactness metric is recycled to degree the accurateness of the plagiarism uncovering system. The precision is defined as:

$$\text{Precision} = TP/(TP+FP)$$

4.2 ALGORITHM

Basic Considerations

Let $d_1, d_2, d_3 \ldots\ldots d_n$ be the number of documents in the document corpus(D)

Let S_n be the number of sentences

$S_1, S_2, \ldots\ldots\ldots S_n \quad \sum_n \quad Dn$

Let w_n be the number of words in each sentence

$w_1, w_2, \ldots\ldots\ldots w_n \quad \sum_n \quad Sn$

To calculate the percentage of plagiarism in document (d) in corpus (D) then-gram model is proposed

The given document 'd' is spitted into n-grams using Bayesian Principle

The suspicious documents is divided into sentences (si) si is divided into word n -grams: The set of n –grams signifies the sentence:

- a document d is not divided into condemnations, but merely into word n-grams.
- Each verdict SI 2 S is explored single to N over the reference documents:

Bayesian classifier rule

Using Bayes theorem, the conditional probability can be given as,

$$P(ck|x) = \frac{P(ck).P(x|ck)}{P(x)} \quad (1)$$

From 1, the bayesian classifier can be applied for plagiarism detection by modifying the equation as

$$P(si|D) = \frac{P(si).P(d|si)}{P(D)} \quad (2)$$

where

P(si|D)is the probability of a sentence si in the document corpus D:

P(si)is the probability of occurrence of a sentence si in the document d:

P(d|si) = $\sum$ Pi P(wi|Dcorpus) is the probability of n-grams of the documents in the document corpus D :

The joint probability model of the proposed algorithm is equivalent to

$$P(S_i', g_1, g_2, g_3, \ldots\ldots, g_n) \quad (3)$$

Where

S_i' – n grams of a sentence, $g_1, g_2, g_3, \ldots, g_n$ – n-gram in the document corpus

Using the chain rule,

$$P(S_i', g_1, g_2, g_3, \ldots, g_n) = P(g_1, g_2, \ldots g_n, S_i')$$

$$= P(g_1 | g_1, \ldots g_n, S_i') P(g_2, \ldots g_n, S_i')$$

$$= P(g_1 | g_1, \ldots g_n, S_i') P(g_2 | g_3, \ldots g_n, S_i') P(g_3, \ldots g_n, S_i')$$

$$= P(g_1 | g_1, \ldots g_n, S_i') P(g_2 | g_3, \ldots g_n, S_i') \ldots\ldots P(g_{n-1} \ldots g_n, S_i') P(S_n | S_i'). P(S_i') \quad (4)$$

According to navie's conditional independence assumptions the n-grams in the corpus is characterized as,

g_i is independent of g; for $i \neq j$, means the n-grams in the corpus are conditionally independent.

The joint model becomes

$$P(S_i' | g_1, g_2, g_3, \ldots, g_n) \alpha P(S_i', g_1, g_2, g_3, \ldots, g_n)$$

$$\alpha P(S_i') \; P(g_1|S_i') P(g_2|S_i') P(g_3|S_i') \quad (5)$$

$$= P(S_i') \prod_{i=1}^{n} P(g_i|S_i') \quad (6)$$

In the above equation, the n-grams used can be 1-gram, 2-grams, 3-grams.

Investigational Outcomes

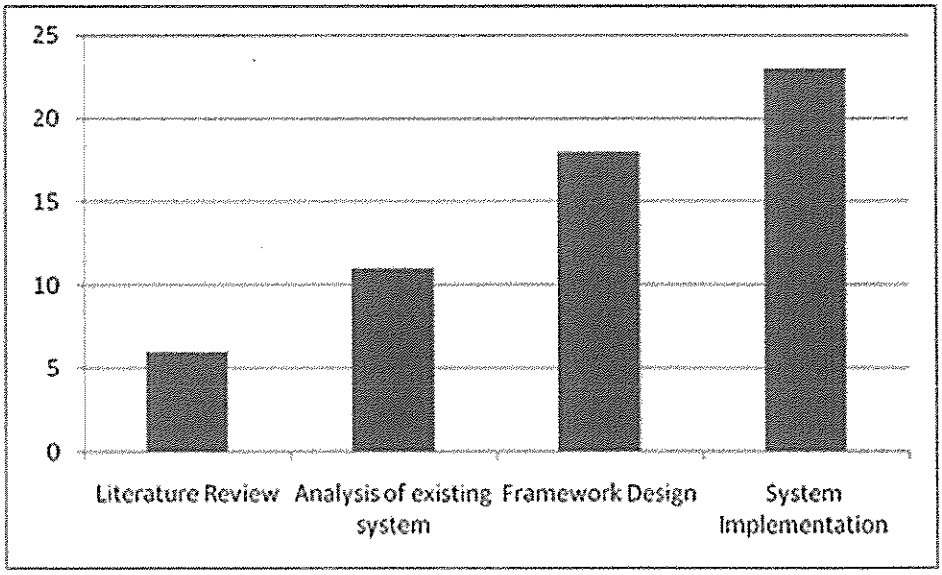
The objective of our research is to express the best n-gram glossary to detect plagiarism belongings. We have demonstrated n-gram levels in the range [1, ⋯ , 11]. The proposed model of this study was programmed with Python programming language. The objective of the projected model is to solve the problem of plagiarism in any language text that may be copied from English text. Thus, several experiments are carried out to find the best setting (within our research scope and objectives) for Any Language-English text system. In this research, a total of

318 different language files are used for both training and test. Different Language files are divided into paragraphs; then, going through the pre-process steps, translated into English. Next, extracting key phrases is done. In addition, all English files were used for the comparison of both training and testing stages.

4.3 Time Schedule of activities giving milestones through BAR diagram.

Work plan (including detailed methodology and time schedule)

Sl. No.	Activity / Milestone	1st Year		2nd Year	
1.	Literature Review	1-6			
2.	Analysis of existing system		7-12		
3.	Framework Design			13-18	
4.	System Implementation				19-24



4.4. Expected outcome within the time period of Seed Money Scheme

- The proposed system predicts the amount of plagiarism from specific file with accurate percentage.
- The process of tokenization makes the n-gram model more efficient thereby the

plagiarism prediction will be accurate.

- Removal of stop words, means the most common terms of language (say Idioms) are excluded from the n-gram modeling so that the wrong predictions are avoided.
- N-gram model has semantics. (language rules so that the model will be executed fast and accurate)
- The usage of goslate API for language translation ensures that the document given in any language can be converted to English without explicitly.

5. **Suggested Plan of action stating the name of funding agency where the project will be communicated for financial support within the time period of project.**

6. **Bibliography: Nil**

7. **List of Projects submitted/implemented by the Investigators (Separate for Pi and Co-PI) : NIL**

7.1 Details of Projects submitted to various funding agencies:

Sl. No.	Title	Cost in lakhs	Month of submission	Role as PI/ Co-	Agency	Status
	NA	NA	NA	NA	NA	NA

7.2 Details of Projects under implementation

Sl. No.	Title	Cost in lakhs	Duration	Role as PI/ Co-PI	Agency
	NA NA	NA	NA	NA	NA

7.3 Details of Projects completed during the last 5 years

Sl. No.	Title	Cost in lakhs	Duration	Role as PI/ Co-PI	Agency
	NA	NA NA	NA	NA	NA

8 List of publications published by the Investigators, if any:

a) Co - Principal Investigator

Sl.No	Details
1.	Michael, G., Priya, N., Pothumani, S" Decoupling the partition table from access points in DHTs" International Journal of Engineering and Advanced Technology, 2019, 8(6 Special Issue 2), pp. 105–108
2	Michael, G., Priya, N., Pothumani, S." Sequential analysis of hierarchical databases for efficient handling International Journal of Engineering and Advanced Technology, 2019, 9(1), pp. 7193–7196
3	Michael, G., Nalini, C., Pothumani, S." QoS based enhanced system determination plan for 4G frameworks" International Journal of Engineering and Advanced Technology, 2019, 8(6 Special Issue 2), pp. 283–286
4	Theivasigamani, S., Jeyapriya, D., Michael, G." Secure controlling system for cross layer reliability for ultra huge scale framework" International Journal of Engineering and Advanced Technology, 2019, 8(6 Special Issue 2), pp. 229–231

b) Principal Investigator

Sl.No	Details
1	M. Sriram, R.M.Suresh" Comparing Expert Systems And Randomized Algorithms Using SAC" International Journal of Pharmacy & Technology, Vol. 8 | Issue No.3 | 17245-17251.
2	M. Sriram, R.M.Suresh" ANN -Conten Based Victimization – Review" International Journal of Pure and Applied Mathematics, Volume 116 No. 20 2017, ISSN: 1311-8080.
3	M. Sriram, R.M.Suresh" AI- Techniques Used For Contentbased Victimisation" International Journal of Pure and Applied Mathematics, Volume 116 No. 20 2017, ISSN: 1311-8080.
4	M. Sriram, R.M.Suresh" Ontology Matching Techniques For Information Retrieval" International Journal of Pure and Applied Mathematics, Volume 116 No. 20 2017, ISSN: 1311-8080

9 **Budget**

Sl. No.	Equipment /Soft ware	Quantity	Amount in INR
1.	Software cost and Design Implementation	1	50,000/-
2.	Consumables		20,000/-
3.	Travel support for the purpose of research work.	---	10,000/-
4.	Contingency	---	10,000/-
5.	Others	---	10,000/-
	Total		1,00,000/-

10 **Name of at least two subject experts from the Institute and one from the outside Institute with their contact details:**

a) Dr.M.Senthil – Professor, Department of CSE, SKP Engineering College, Tiruvannamalai -601 108.

b) Dr.G.Ayyappan – Professor, Dept. of CSE, BIHER, Chennai.

CERTIFICATE FROM THE INVESTIGATOR

Project Title:"CONTENT BASED MULTI-LANGUAGE PLAGIARISM DETECTION TOOL USING BAYESIANCLASSIFIER"

It is certified that

1. I do hereby agree to submit a complete proposal for financial support to the external funding agency within the time period of SMS-2018
2. I undertake that spare time on equipment procured in the project will be made available to other users.
3. I agree to submit a certificate from Institutional Biosafety Committee, if the project involves the utilization of genetically engineered organisms. I also declare that while conducting experiments, the Biosafety Guidelines of Department of Biotechnology, Department of Health Research, GOI would be followed in to.
4. I agree to submit ethical clearance certificate from the concerned ethical committee, if the project involves field trails/experiments/exchange of specimens, human & animal materials etc.
5. I agree to abide by the terms and conditions of SMS-2018, BIHER, and Chennai.

Name and signature of
Principal Investigator

Name and signature of
Co-Principal Investigator

Date: 27.02.2019
Place: Chennai - 73

Forwarded by Head of the Department

Signature of the Head

PROJECT EVALUATION FORMAT

Recommendation Sheet

Name of the Principal Investigator	M.Sriram
Name of the Co-Investigator	Dr.G.Michael
Name of the Department	IT
Title of project	"CONTENT BASED MULTI-LANGUAGE PLAGIARISM DETECTION TOOL USING BAYESIAN CLASSIFIER"
Recommendation of the evaluation committee	– Recommended –
Financial allocation recommended	Rs. 1,00,000 –

Sl. No.	Equipment /Soft ware	Quantity	Amount in INR
1.	Software cost and Design implementation		60,000/-
2.	Consumables		15,000/-
3.	Travel support for the purpose of research work.	---	5,000/-
4.	Contingency	---	10,000/-
5.	Others	---	10,000/-
	Total		1,00,000/-

Name and Signature of the Research Advisory Committee members with date.

(Dr. P. Kaveichandran)