International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Novel Features for plagiarism detection in Marathi Language

Ramesh R. Naik, Maheshkumar B. Landge, C. Namrata Mahender
Department of CS and IT, Dr.B.A.M.University, Aurangabad, Maharashtra, India

ABSTRACT

Plagiarism is stealing information or idea from someone without giving proper acknowledgement. Currently plagiarism is increasing in different fields like education, industry. There is need to prevent plagiarism. There are four types of stylometric features available namely: lexical, syntactic, semantic, content specific. In this paper we have added three new features for detecting plagiarism namely noun, adjective and rhyming words. For calculating these features, we have used our own Marathi text corpus. These features will be useful for detecting plagiarism and linguistic researchers.

Keywords : Plagiarism Detection, Feature Extraction, Stylometric Features.

I. INTRODUCTION

Plagiarism is increasing day by day. It has become a worldwide problem this problem is getting worse mainly because of the increase in the volume of online publications. Relying only on exact word or phrase matching for plagiarism detection is not sufficient now. The act of plagiarizing is when you use someone else's words and ideas without giving due credits to the original writer and regard the work as your own [1].
The text document in digital form has drastically increased worldwide.

II. LITERATURE SURVEY

We have studied stylometric Features for plagiarism detection, which are following:

There are four types of stylometric features as following: Lexical features, Structural features, Syntactic features and content specific features:

Lexical features: lexical features are used to the lexical structure of the text which operate the word level of the document in order to trace the plagiarism in the suspicious document [2].lexical features are divided into two types i)character based feature and ii) word based feature.
Character based features, which are following:
Characters count, total number of alphabet character, total number of uppercase characters and total number of digit characters. Frequency of letters and special characters.

Word based features, which are following:
Total number of words, Total number of characters in words, average word length, average sentence length.

Structural features: Structural features conversely, take into account the way the words are distributed throughout the document. Very few plagiarism detection approaches have been developed to handle structural or tree features [3].

Syntactic features: Syntactic features identify plagiarism by dividing the text into
Chunks, sentences and phrases to aid efficient comparison of the original and
Suspicious documents [4].

Content specific features: content specific features are used to characterise certain activities. like frequency of content specific keywords. [4].

III. PREPROCESSING

A. Tokenization:

Tokenization is the process of separating tokens from input text. Each word is separated from sentence by

white space or punctuation marks and treat as single token and then deal with each word individually. [5]

B. Stopword Remove:

Remove functional words (articles, pronouns prepositions, complementisers, and determiners)

C. Stemming:

A stemming is a process of converting morphologically identical words to root word affixes without applying morphological analysis of that term. The purpose of using a stemmer in plagiarism detection is to increase the possibility of detecting rewrites of a given word [6].

IV. FEATURE EXTRACTION

Feature extraction can be defined as a process of extracting a set of new features from the features set that is generated in feature selection stage. Feature extraction is a basic and fundamental pre-processing step to pattern Recognition and machine learning problem. There is no text corpus available for Marathi language. We have created our own Marathi text corpus using three Marathi poems. We have taken summary of Marathi poems written by different authors. We have considered 18 Marathi text files in our text corpus.

We have calculated three new features using text corpus namely: noun, adjective, rhyming words

I. **Noun:** This feature contains noun words present in file.
II. **Adjective:** This feature contains adjective words present in file.
III. **Rhyming words:** This feature contains Rhyming words present in file.

Table 1. The below table shows extracted features.

Sr. No.	Name of File	Noun	used as it is noun	Adjective	used as it is adjective	Rhyming Words	used as it is Rhyming Words
1	3_Mahesh1.txt	कविता	yes	व्यक्त	No	पात	Yes
		कुसुमाग्रजा	yes	निळनिळ	Yes	डोलत	No
		निसर्ग	no	छोटासा	Yes	हसत	No
		गोष्ट	no	प्रसन्न	No	म्हणत	No
		भाव	no	छोटासा	No	गात	No
		गवताचं	yes			पावसात	No
		पाणी	yes			बरसात	Yes
		पाखरू	yes			खेळतो	No
		मोती	yes			म्हणतो	No
		मनीच	yes			लोळत	No
		पिल्लू	yes				
		उपमा	no				
		निसर्गाचे	no				
		वर्णन	no				
2	3_Rameshh2.txt	कवी	no	सहभागी	no	पात	Yes
		कुसुमाग्रज	yes			डोलत	No
		कवितेत	no			हसत	No
		निसर्गाचे	no			म्हणत	No
		वर्णन	no			गात	No
		निसर्गाशी	no			पावसात	No
		संवाद	no			बरसात	No
		प्रयत्न	no			खेळतो	No
		गवताचे	no			म्हणतो	No
		पक्षी	no			लोळत	No
		मोती	yes				
		मांजरीच	yes				
		पिल्लू	yes				
		कार्य	yes				

V. CONCLUSION

The plagiarism detection is a serious problem in academics. There are number of tool available in English language. However, there is no tool available to detect the plagiarism for Marathi language. To get good accuracy in plagiarism detection features play important role. This paper covers four types of features like lexical, syntactic, semantic, content specific, there are three new feature added namely:noun,adjective,rhyming words of poem. These three features are useful for Marathi plagiarism detection.

VI. REFERENCES

[1]. Chris Park. Rebels without a Clause: Towards an Institutional Framework for Dealing with Plagiarism by Students. Journal of Further and Higher Education Vol. 28, No. 3, August 2004.

[2]. Alzahrani, S.M., Salim, N. and Abraham, A. (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, pp. 133-149.

[3]. Chow, T.W.S. and Rahman, M.K.M. (2009), "Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection", IEEE Transactions on Neural Networks, Vol. 20 No. 9, pp. 1385-1402.

[4]. Ramya, L. and Venkatalakshmi, R. (2013), "Intelligent plagiarism detection", International Journal of Research in Engineering & Advanced Technology, Vol. 1 No. 1, pp. 171-174.
[5]. K. Leilei, Q. Haoliang, W. Shuai, D. Cuixia, W. Suhong, and H. Yong, "Approaches for candidate document retrieval and detailed comparison of plagiarism detection," Notebook for PAN at CLEF 2012.
[6]. M. Sanchez-Perez, G. Sidorov, and A. Gelbukh, "A winning approach to text alignment for text reuse detection at pan 2014," Notebook for PAN at CLEF, pp. 1004–1011, 2014.