A Method for Text Plagiarism Classification Using Deep Learning

Mumthaz Beegum.M[1], Aji S [2]

Department of Computer Science, University of Kerala, Kerala,

[1]mumthazbeegumm@gmail.com,[2]aji@keralauniversity.ac.in

***Abstract*— This work proposes a new method for detection and classification of plagiarism in the text using deep learning strategy. The linguistic features that can give more information in respect to the context of the document are used in the experiments. It is found in the experiments that, the shallow syntactic features, Parts of Speech tags and Chunks, are more effective and reliable features in the text plagiarism classification. The irrelevant features in the collection of feature set are eliminated through different selection procedures. The results obtained in the experiments show that the proposed method is consistent and performed well compared to other existing methods.**

***Keywords*— Text Plagiarism, POS Features, Chunk Features, Feature Selection, Classification, Deep Learning.**

I. INTRODUCTION

The plagiarism detection, a binary classification process, begins with transforming the suspicions-source document pair into a collection of features. These features are representatives of their corresponding suspicious -source document pair in a discriminative high-dimensional space. According to Merriam Webster the Plagiarism is "the act of using another person's words or ideas without giving credit to that person". In text plagiarism, the content is obfuscated or manipulated in different ways to create a document which seems to be a new work. Normally the plagiarism can happen through different activities vary from simple copy-paste to higher level sentence restructuring. Plagiarism detection task can be divided into two categories - extrinsic and intrinsic detection [1]. In the former, the suspicious document is compared against a collection of the source document which can be done either offline or online. In Intrinsic detection, the text contents in the documents are verified with the help of structural distributions and stylometric information within the documents. Literal or copy-paste and intelligent plagiarisms are the two types of text plagiarisms. In this work, the syntax-based linguistic features are used for plagiarism classification. The features are extracted using Parts of Speech (POS) tagging and Chunking the text. The best features are selected using correlation-based ranking and these features are used for classification in the proposed Deep Learning Strategy. The Short Answer Corpus (PSA) [2], a collection of manually plagiarised documents, is used in the experiments to evaluate the proposed method.

The coming sections of the paper is organized as follows. A review of the related literature is discussed in section 2. The method and working of the proposed method are described in section 3. Section 4 and 5 are used to explain the experimental results, analysis of the result and conclusion of the work.

II. LITERATURE REVIEW

This section briefly discusses the works already came out in the domain to detect text plagiarism based on document level, passage level detections and context-based approaches. The plagiarism detection process consists of different stages and there are many notable research contributions in all these processes as well.

Lancaster et al. [3] uses different types of corpora in the experiments and the classification has done on the basis of metric complexity. Alzahrani et al. introduced a classification method based on document, structure, paragraph, line, statement, word and character. The main approaches used in document-level plagiarism detection are N-gram and Vector space Models (VSM). The similarity matrices like Jaccard coefficient, dice coefficient, kullback-Liebler distance metrices are used in the bigram and trigram-based comparisons [5, 6,7].

Ravi et al. [8] use the unsupervised machine learning clustering techniques like K-Means and Fuzzy C-Means methods. The proposals evolved in PAN 2012-2015 are mainly used Vector space Models (VSM) approaches with term frequency-inverse document-Frequency weighting (tf-idf) schemes. Some of the contributors have explored the NLP techniques namely Parts of Speech (POS) tagging and Chunking.

Rafiei et al. used chunk-based queries by extracting the words with the highest tf-idf for each chunk [9]. Raviet al.used POS tagging for feature extraction and tf-idf for sorting keywords[10,11]. The POS tagging is also used in query formation [12, 13,14]. Suchomel et al., [15] derived the features based on a set of parameters such as key-word , paragraph and phrase in their method. Zubarev et al. used the important noun features for query formulation [16]. Many researchers used syntax or semantic

based NLP techniques for passage level plagiarism detections [4, 17, 18, 19, 20]. The cross lingual text similarity detection uses the distribution of words or word embedding [21, 22]. In reuse detection [23,24], the semantic similarity measures using the word embeddings are played a key role. Some other existing works [25,26], classifies the suspicious-source document pair either as plagiarised or non-plagiarised using the features which include similarity scores computed from the N -gram representations, language models, longest common sub sequences and dependency relations.

In the proposed work, shallow linguistic features -POS and chunk features -are used for classifying a given suspicious-source document pair.

III. PROPOSED APPROACH

In the proposed method, the given suspicious-source document pair is categorised according to the degree of plagiarism. Each Suspicious-Source document is represented as features. Consider a set of n documents pair D, (DP_1, DP_2,…....,DP_n).A document pair DP_i have the suspicious document (DPsusp) and source document (DPsrc) as components. The Fig 1 show the work flow of the proposed approach.

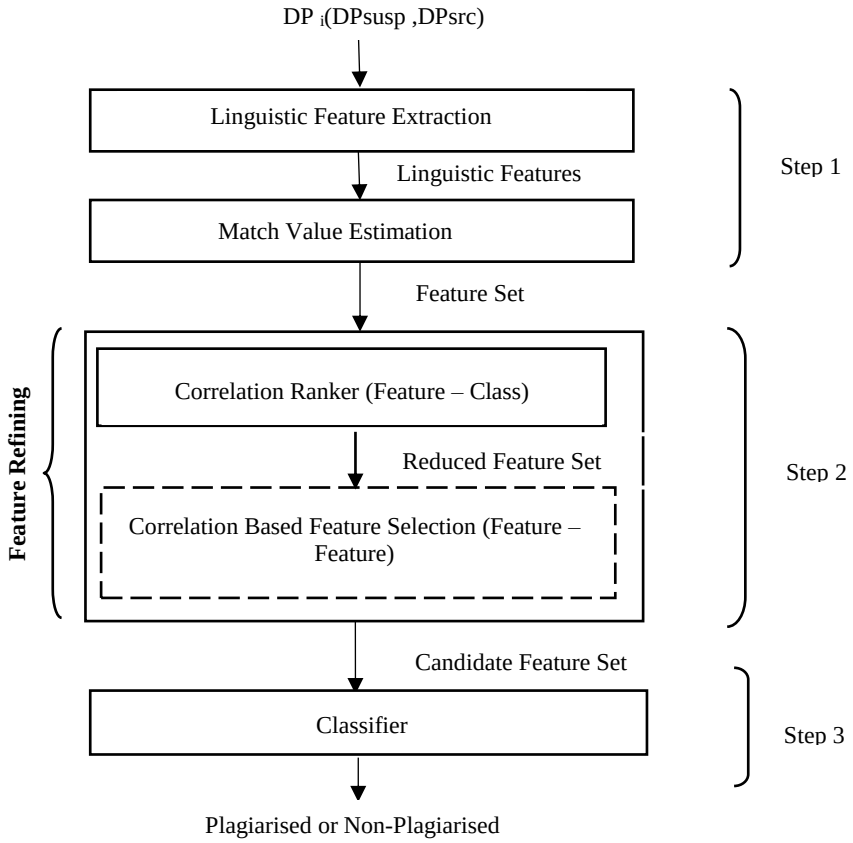

Fig.1.Work Flow of the Proposed Approach

A. Linguistic Feature Extraction

The linguistic features are extracted from the suspicious – source documents as the first step of feature extraction. Part of Speech (POS) and Chunk are the main features extracted using the Shallow NLP techniques namely POS Tagging and Chunking respectively.

a. POS (Parts of Speech) Features

Tokenization of the text document is the initial step in the POS feature extraction. After the tokenization process stop word removal, regular expression removal, lower casing and lemmatization are done. The lemmatization is the process of grouping words to their dictionary forms called as lemma. The POS tagging is applied in these tokens to classify into word lasses such as Noun, Verb, Adjective, Adverb, Preposition and conjunctions etc. In this work Noun (FN), Verb (FV), Adjective (FADJ) and Adverb (FADV) are extracted as POS features.

Consider a sentence “In the coming month ,Veena planning to go Paris along with the her father suresh”.

Then extracted POS features are [[('in', 'IN'), (u'come', 'JJ'), ('month', 'NN'), ('veena', 'NN'), (u'plan', 'NN'), ('go', 'VBP'), ('paris', 'NN'), ('along', 'RB'), ('father', 'RB'), ('suresh', 'NN')]].

b. Chunk Features

To extract the chunk features the document is chunked into a group of words. The stop words, the words that frequently occur within a document usually semantically irrelevant, are eliminated from these chunks. The final chunk features are obtained after the lower casing and lemmatization. The chunks are extracted in two ways according to the length of the chunk-length, l>=1 and l>=2.

B. Match Value Computation

Match value is the content similarity measure of suspicious -source document pair and it is computed from the features extracted in the previous phase. The match value has to be computed for every feature in the document pair. The normalized, with the length of the largest document, the match value is computed as

$$Macth(DP_i) = (Match(Dsusp, Dsrc) = \frac{|\text{PDsus\_fi}| \cap |\text{PDsrc\_fi}|}{\max(|\text{PDsusp}|, |\text{PDsrc}|)} \quad (1)$$

i ϵ {N,V,ADV,ADJ,NP>=1,VP>=1,

ADJP>=1,ADVP>=1,NP>=2,VP>=2,ADJP>=2,ADVP>=2}

Dsusp is the pre-processed suspicious document and PDsrc notate the pre-processed source document. In POS Tagging |PDsusp| and |PDsrc| are the cardinality of suspicious and source documents respectively and in Chunking it is the lengths of chunks. PDsusp_fi represents features extracted from pre-processed suspicious document and PDsrc_fi is the features extracted from the source document.

C. Two-Phase Feature Selection

The best features are selected from extracted features and as a result, the dimensionality of feature set can be reduced considerably. The filters and wrappers are two important algorithms for feature selection [27]. In the filter algorithm, the statistical tests and ranking are used to measure the feature relevance while the Wrapper Algorithm will search the best features in the feature space. In this work, a combination of Pearson correlation ranking and CFS subset evaluator is used for selecting the relevant features.

The Pearson correlation of the feature class set is computed for all feature class pairs and the features are ranked according to it.

$$rfc_i = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{f_i-\bar{f}}{s_f}\right)\left(\frac{c_i-\bar{c}}{s_c}\right) \quad (2)$$

rfc_i represents the Pearson's Correlation between the i[th] feature and the class. The average of feature and class values represented by $\bar{f}$and $\bar{c}$.s_f and s_c are the standard deviations of feature and class values. Each document pair contain 14 feature values and these values are represented by f_i .c_i which is the class value representation of n instances. The feature values greater than a threshold value $\alpha$is taken as the initial reduced feature set.

In the second stage of feature refinement, the CFS subset evaluator-measure of the relevance or the merit of the feature set- is used as a filtering technique.

$$MS = \frac{m''\,\overline{rfc}}{\sqrt{m''+m''(m''-1)\overline{rff}}} \quad (3)$$

MS is the Merit of Feature Subset S. It contains m'' features, $m'' \leq m'$. m is the features extracted from the suspicious – source document pairs. m' is the initial refined feature set which are selected with person's correlation. m'' is the subset of features set m'.$\overline{rff}$is the mean of feature to feature correlation and the denominator indicates the redundancy of the feature. The feature set S is obtained through the Forward Optimized Best First Search also called Beam Search [27].

D. **Machine Learning Based Classification**

Deep Learning [28, 29], one of the best and established machine learning approach, is used in this proposal for classifying the suspicious document. The capability of Deep Learning strategy to study and evolve the prominent features from a large set of data is already proven in the recent works [30]. In this work, the deep leaning method is used to classify the given suspicious-source document according to the degree of plagiarism.

E. Evaluation Measures

The proposed work is analyzed using the classical measures -Accuracy and F-measure. P and N represent the Plagiarised and non-plagiarised classes. The True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are obtained from this confusion matrix. True Positive (TP) is the number of plagiarised documents correctly classified as plagiarised and True Negative (TN) is the number of non-plagiarised document correctly classified as Non-Plagiarised. False Positive (FP) is the number of non-plagiarised document misclassified as plagiarised and False Negative (FN) is the number plagiarised documents misclassified as non-plagiarised.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6)$$

$$\text{F} - \text{Measure} = 2 * \frac{recall*precision}{recall+precision} \quad (7)$$

The harmonic mean of precision and recall represents the F-Measure and it is computed using Eq. (7).

IV. EXPERIMENT AND RESULTS

The experiments are conducted using the Plagiarized Short Answers (PSA) dataset [2]. The PSA contains plagiarized and non-plagiarized documents. Five computer science questions are used to create PSA corpus. The Wikipedia answers used as a source and the answers of students are treated as suspicious documents. Three levels of Plagiarisms namely Near Copy, Light Version and Heavy version are in the dataset. In near copy, the documents represent simple verbatim or literal plagiarism. Small rewritings and paraphrasing are employed in Light version. The heavy version contains structural changes, deep paraphrasing and rearrangements. The non-plagiarized documents are original and written by the students.

Table 1. The preliminary statistics of PSA dataset.

Document Class	Plagiarism type	No of suspicious-source document pairs	Total number of suspicious – source document pairs
Plagiarised	Near Copy	19	57
	Light Version	19	
	Heavy Version	19	38
Non—Plagiarised		38	

The match value is computed from the suspicious -source document using the Eq. (1). The match values are used to select the prominent features which have done in two phases as explained in the previous section. The ranking of the features after the first filter processing is shown in table 2. It is noted in the experiment that the initial 14 features are reduced to 9 .

Table 2. Feature ranking Based on Correlation Score.

Feature	Correlation Score
FNP>=1	0.798
FN	0.790
FADJ	0.785
FNP>=2	0.781
FADV>=1	0.752
FVP>=1	0.752
FV	0.737
FVB>=2	0.726
FADV	0.722

In the next level of refinement, the CFS subset evaluator is used as per Eq.(3). After the reduction process we have obtained the Noun(FN), adjective(FADJ), verb phrase >=1(FVP>=1), adverb phrase >=1(FADV>=1) and Noun phrase(FNP>=2) as the best features.

Table 3. Classification accuracy of the experiments with different conditions

Strategy	Attribute evaluator and search method	Number of features selected	Accuracy			
			NB (%)	SVM (%)	DT (%)	DL (%)
Without feature reduction	-	14	89.19	90.08	92.63	92.78
With feature reduction	Phase 1-Correlation evaluator with rank search	9	90.01	92.23	94.74	95.2
	Phase 2-CFS with beam search (Best features)	5	95.79	95.79	97.89	99.378

The refined feature set obtained in this phase is fed into the Deep Learning based classification module for training. The dataset is divided into 80:20 ratio for learning and testing purposes. The results obtained in the different experiments are abstracted in table 3. The result of the classification experiment without the refinement of features is also noted in the table. It is clear in the results that the proposed Deep Learning (DL) based approach is better than the other Naïve bayes(NB), Support Vector Machine (SVM) and Decision Tree(DT) methods. It is interesting that the results obtained in the experiment with five features performed well comparing to the other experiments with a greater number of features.

V. CONCLUSION

In this work, we have explored the influence of context-based features in the classification of given suspicious-source document according to the degree of plagiarism. The results show that the proposed deep learning based method performed well compared to the other strategies. The results obtained for the experiments using a smaller number of features is comparably better with other experiments with more number feature set. That results show the reliability of feature refinement and deep learning method in plagiarism detection and classification.

REFERENCE

[1] Potthast, M. , Stein, B. , Barrón-Cedeño, A. , & Rosso, P. (2010). An evaluation frame- work for plagiarism detection. In Proceedings of 23rd international conference on computational linguistics, COLING, Beijing, China.

[2] Clough, P.,& Stevenson, M. (2010). Developing a corpus of plagiarised short answers. In Language Resources and Evaluation: 45 (pp. 5–24). Springer.

[3] Lancaster,&Culwin (2005). Classifications of plagiarism detection engines. ITALICS, 4 (2).

[4] Alzahrani, S. M., Salim, N.,& Abraham, A. (2012). Understanding plagiarism linguis- tic patterns, textual features, & detection methods. IEEE Transactions on Systems, Man, & Cybernetics, Part C: Applications & Reviews, 42 (2), 133–149.

[5] Alzahrani, S. M.,& Salim, N. (2010). Fuzzy semantic-based string similarity for ex- trinsic plagiarism detection- lab report for PAN at CLEF 2010. Proceedings of the 2nd international workshop PAN-10, Padua, Italy.

[6] Barrón-Cedeño, A., Basile, C., Esposti, M. D.,& Rosso, P. (2010). Word length n-grams for text re-use detection. In 11th international conference on intelligent text pro- cessing and computational linguistics (pp. 687–699). LNCS.

[7] Barrón-Cedeño, A., Rosso, P.,&Bened, J. M. (2009). Reducing the plagiarism de- tection search space on the basis of the Kullback-Leibler distance. In Inter- national conference on intelligent text processing and computational linguistics (pp. 523–534).

[8] Ravi, N. R.,& Gupta, D. (2015). Efficient paragraph-based chunking and download filtering for plagiarism source retrieval -Notebook for PAN at CLEF 2015. CLEF 2015 e valuation l abs and w workshop –Working notes papers, Toulouse, France.

[9] Rafiei, J.,Mohtaj, S.,Zarrabi, V.,&Asghari, H. (2015). Source Retrieval Plagiarism Detection based on Noun Phrase and Keyword Phrase Extraction—Notebook for PAN at CLEF 2015. CLEF 2015 evaluation labs and workshop –Working notes papers, Toulouse, France.

[10] Ravi, N. R.,& Gupta, D. (2016). A plagiarized source retrieval system developed using efficient download filtering and POS tagged query formulation with effective paragraph based chunking. International Journal of Artificial Intelligence, 14 , 145–160 .

[11] Ravi, N. R., Vani, K., & Gupta, D. (2015). Exploration of fuzzy C means clustering algorithm in external plagiarism detection system. In Proceedings of the symposium on intelligent systems technologies and applications (ISTA-2015). (pp. 127–138) .

[12] Kong, L., Lu, Z, Han, Y., Qi, H., Han, Z.,& Wang, Q. (2015). Source retrieval and text alignment corpus construction for plagiarism detection—Notebook for PAN at CLEF 2015. CLEF 2015 e valuation l abs and workshop –Working notes papers, Toulouse, France .

[13] Prakash, A. ,&Saha, S. K. (2014). Experiments on document chunking and query formation for plagiarism source retrieval—Notebook for PAN at CLEF 2014. CLEF 2014 e valuation l abs and workshop –Working notes papers, Sheffield, UK .
[14] Williams, J. (2002). The plagiarism problem: Are students entirely to blame. InProceedings of the 19th annual conference of the Australasian society for computers in learning in tertiary education (ASCILITE) (pp. 721–730) .
[15] Suchomel, S. ,&Brandejs, M. (2014). Heterogeneous queries for synoptic and phrasal search—Notebook for PAN at CLEF 2014. CLEF 2014 e valuation l abs and workshop –Working notes papers, Sheffield, UK .
[16] Zubarev, D. ,&Sochenkov, I. (2014). Using sentence similarity measure for plagiarism source retrieval. CLEF , 1027–1034 (Working Notes) .
[17] Alzahrani, S. M. , Salim, N. , & Palade, V. (2015). Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. Journal of King Saud University –Computer and Information Sciences, 27 (3), 24 8–26 8 .
[18] Chong, M. (2013). A Study on plagiarism detection and plagiarism direction identification using natural language processing techniques . U.K.: University of Wolver- hampton .
[19] Gupta, D. , Vani, K. , & Singh, C. K. (2014). Using natural language processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection. In International conference on advances in computing, communications and informatics (pp. 2694–2699) .
[20] Kalleberg, R. B. (2015). Towards detecting textual plagiarism using machine learning methods Master thesis . University of Agder .
[21] Vani, K. ,& Gupta, D. (2015). Investigating the impact of combined similarity metrics and POS tagging in extrinsic text plagiarism detection system. In Proceedings of the international conference on advances in computing, communication and informatics (pp. 1578–1584) .
[22] Vani, K. ,& Gupta, D. (2017). Detection of idea plagiarism using syntax–semantic concept extractions with genetic algorithm. Expert Systems with Applications, 73 (1), 11–26 .
[23] Banea, C. , Chen, D. , Mihalcea, R. , Cardie, C. , & Wiebe, J. (2014). SimCompass: Us- ing deep learning word embeddings to assess cross-level similarity. In Proceed- ings of the 8th international workshop on semantic evaluation, Dublin, Ireland (pp. 560–565) .
[24] Ferrero, J. ,Besaceir, L. , Schwab, D. , & Agnes, F. (2017). Using word embedding for cross-language plagiarism detection. Conference of the European chapter of the association for computational linguistics, (EACL 2017) . Valencia, Spain: IEEE .
[25] Zhang, Qi , Kang, J. , Qian, J. , & Huang, X. (2014). Continuous word embeddings for detecting local text reuses at the semantic level. In Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval (SIGIR '14) ACM, New York, NY, USA (pp. 797–806) .
[26] Chong, M. ,Specia, L. , &Mitkov, R. (2010). Using natural language processing for automatic plagiarism detection. In Proceedings of the 4thinternational plagiarism conference .
[27] Hall, M. A. (1999). Correlation-based feature selection for machine learning Hamilton PhD thesis . University Of Waikato .
[28] Kohavi, R. (1994). Feature subset selection as search with probabilistic estimates. AAAI fall symposium on relevance .
[29] Caruana, R. ,& Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on ma- chine learning, Pittsburgh, PA .
[30] Jitendra Yasaswi, Suresh Purini, C. V. Jawahar. Plagiarism Detection in Programming Assignments Using Deep Features.