International Journal of Research Studies in Science, Engineering and Technology
Volume 7, Issue 7, 2020, PP 01-06
ISSN 2349-476X

Afaan Oromo Sentence Based Plagiarism Detection: A Semantic Similarity Approach

[1]Gizaw Tadele Bekele, [2]Million Meshesha (PhD)

[1]*School of computing, Computer science department, Dilla University, Dilla, Ethiopia*

[2]*Information science department, Addis Ababa University, Addis Ababa, Ethiopia*

***Corresponding Author:** Gizaw Tadele Bekele, School of computing, Computer science department, Dilla University, Dilla, Ethiopia*

ABSTRACT

In current day the availability of digital technology enables world community to communicate and exchange information easily. As a result of which, we are in the era of information overloading where various types of information is collected from different sources. As the amount of available digital information increases it is difficult to access information efficiently from different sources. To address this problem, machine leaning based NLP has a great contribution. In this work we focused on semantic based similarity measure for plagiarism detection from Afaan Oromo documents. The study used LSI approach to decompose sentences into terms matrix for similarity calculation. We have collected 3 documents with 15 sentences, 14 sentences and 11 sentences. The documents are collected from different sources like two documents from Afaan Oromo published fiction and one document of personal bibliography from Afaan Oromo FBC. Preprocessing of text has been applied to the dataset. Java programming has been used to develop a prototype of the proposed model and SQL has been used to build sample dictionary.

The performance of the study work was tested on 10 sentences of suspicious query and 3 source documents of 275 key terms. The accuracy achieved in detecting plagiarism from suspicious query was 53.02 %.

Keywords: *Afaan Oromo, Semantic similarity, Machine learning, LSI, Plagiarism detection*

INTRODUCTION

Natural Language Processing (NLP) is a sub-field of Artificial Intelligence that is focused on enabling computers to understand and process human languages to get computers closer to a human-level understanding of language (Dan Jurafsky, 2012) (Resnik, 1999). Computers do not yet have the same intuitive understanding of natural language that humans do.

As a result, more research are expected to be done in NLP to enable computers communicate with human being. NLP embodies several techniques that can change the way people think, learn, and communicate with machine. One of these techniques is plagiarism detection. According to Merriam Webster's Dictionary (merriam, 2016), plagiarism is 'the act of using another person's words or ideas without giving credit to that person.'

Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions (Naghibzadeh, 2012) (Foltz, 1998). Plagiarism detection is a technique to find out the theft of single message, thesis, article, scientific paper, literary works, source code and others (Shams, April 2010).

RELATED WORK

Man Yan Miranda (Chong, 2013)have discussed the way and direction to detect plagiarism by textual similarity measure and contribute as machine learning bring benefits to the plagiarism detection framework.

LSI was proposed algorithm to detect plagiarism with similarity measure for Indonesia textsLucia D. Krisnawati (Krisnawati, August 29, 2016). There also Semantic concept included in the work by using WordNet Bahasa for meaning retrieving from WordNet Dictionary.

There was also research conducted by using supervised sentence embedding to identify semantic for advanced plagiarism detection in Russian language (A, 30, June 2, 2018).

King Abdulaziz (King Abdulaziz University, 2010) and Khalid Shams (Shams, April 2010)

has introduced as latent semantic indexing (LSI) or latent semantic analysis (LSA) is a technique in natural language processing to detect plagiarism.Khalid Shams have not test their program on a lot of data and they cannot mention the accuracy of their work.

METHODOLOGY

The aim of this study was to design plagiarism detection model for Afaan Oromo documents. To this end, we come up with architecture of the system, based on which different techniques and methods are identified that are followed for developing semantic sentences similarity for plagiarism detection with LSI algorism and customize the system.

For this study design science research methodology has been selected because it is an outcome based designing solution, which offers specific guidelines for evaluation and iteration within research procedures. The overall procedure and techniques occur in step by step from problem identification and motivation of semantic sentence similarity to communication stage are followed in this work.

Data Collection

We have collected three documents as training data with total of 40 sentences and 10 sentences for testing as testing data. Two documents are collected from manually published Afaan Oromo fiction of different title. One document was from Afaan Oromo FBC which concerned about bibliographic history of one women.

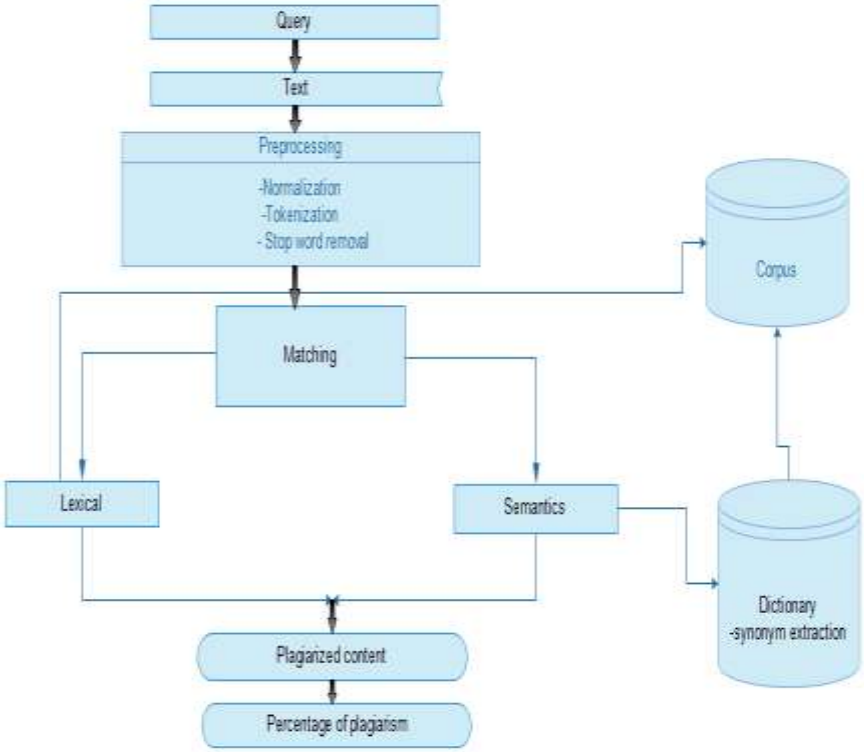

Figure1. *Architecture of the proposed work*

Text Preprocessing and Normalization

We have applied basic necessary preprocessing steps to our work. Normalization, tokenization and stop word removal has been applied on the work for text representation. So number, symbols, any punctuation markhas been not considered in our work to measure similarity for plagiarism detection in Afaan Oromo text. Stop words those are identified as stop word list by Debela (Debela, 2010), Fiseha (Tesema, 2013)and Eyob (Alemu, 2013) and other stop words identified by corresponding language are considered in this work.

Table1. *Sample stop word list*

Number	Words
1	Sun
2	isaan
3	Ol
4	Yoo
5	Fi
6	kee
7	kun
8	koo
9	As
10	garuu

The above table shows stop word must be removed from query that going to extract the stored documents and documents stored and going to extracted.

Semantic Similarity Measure

In this study to detect plagiarism sentence-level similarity measure approach is used including semantic similarity measure. To accomplish this work Latent Semantic Indexing (LSI) is applied as algorithm of semantic similarity measure for plagiarism detection. LSI is to find and fit a useful model of the relationships between terms and sentences. LSI examines the words used in a sentences and looks for their relationships with other words.

Table2. *Sample synonym for Afaan Oromo language*

Word	Synonyms
gaarii/good	dansaa, baroo, hosee, mishaa
baay'ee/ a lot	hedduu, danuu, bacaa
Waraabessa/ hyena	hobolaa, gulloo
ariitiin/ quickly	saffisaan, daddaffiin, hatattamaan
soba/ false	kijiba, dhara, waanyoo
rooba/rain	bokkaa
hiyyeesa/ poor	dhabaa
citaa/ grass	gaalala, marga
kabaja/ respect	ulfina, tabaroo

The above table shows sample synonym terms identified for Afaan Oromo those are collected and stored on database as dictionary. Since our approach was machine learning this synset representation system support our work as the algorithm learn from the sample dictionary whether the retrieved texts are from the same synset or not to decide whether they are similar or not semantically as result of similarity measure. Because all synonym and related terms or words are represented with their respective group.

Result And Discussion

For implementation development of tool used in LSI model for this study was java NetBeans version 8.2 and SQL server 2012 as backend to represent synonym terms in relational database. Because java NetBeans has more sophisticated library and package like hashmap, array List and soon those can assist to implement similarity measure for plagiarism detection. In this work SQL server used for design of synonym terms representation on database to serve as WordNet during semantic similarity measure for plagiarism detection.

For semantic similarity handling we collect seventy eight (78) synonym words from different domain of Afaan Oromo Daily communication. For example wild animal in English "Hyena" is named in Afaan Oromo as "hobolaa" in Bale and Arsi zone, "hobolaa" in Jimma and all Shoa and Wallaga zone it named as "Waraabessa".

We have compared collected documents with query inserted for similarity measure. We have considered text similarity in two aspect (lexically and semantically). Lexically, physically similar words are measured and semantically the meaning (synonym) of word has been retrieved from sample WordNet designed by SQL server.

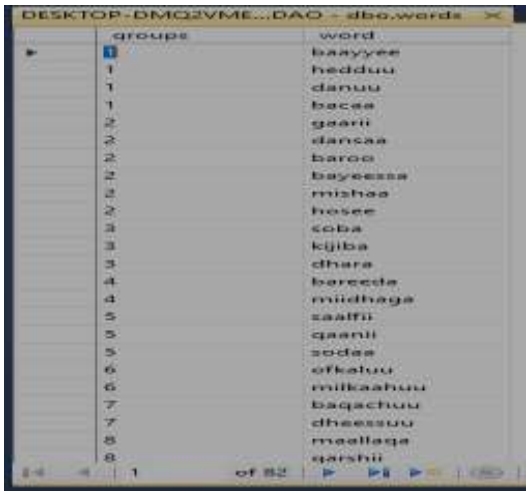

Figure 2. *Synonym representation in relational database*

We can understand from figure 2 how synonym terms are classified into the same group with their respective meaning semantically in relational database. For example the English term "false" by Afaan Oromo "**soba"** word has two synonyms "**dhara**" and "**kijiba"** as our sample relational database representation and the other words or terms has their own synonyms.

Red color highlighted and blue color highlighted are there to indicate lexically plagiarized part of documents and semantically plagiarized part of documents respectively for all three documents independently. But black color highlighted parts indicates stop words of the documents retrieved by query request. Black color highlighted or stop words has no effect on similarity calculation to get result of ratio plagiarism because the algorithm exclude stop words from any task of calculation. We have used only two documents with one sentence of query to show lexically and semantically plagiarized part of D0 and D1 in below figure. But we have used all three documents collected in the evaluation part of this study.

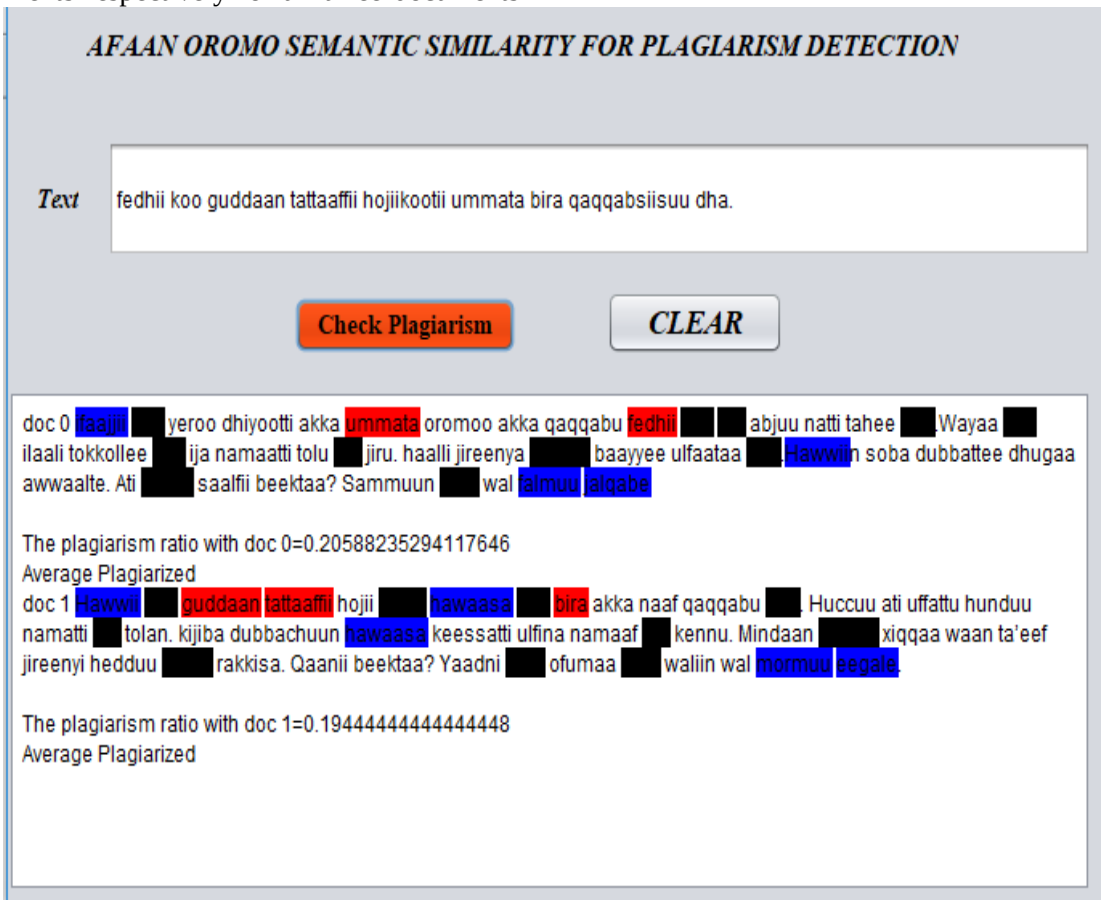

Figure3. *Highlighted part to indicate lexically plus semantically*

For example term "tattaaffii" which is provided in the text (query) has available directly as it is in D1 which copy and pasted in query provided from D1.

That is why it was highlighted with red color in figure 3 to show as it was plagiarized part of D1 and "tattaaffii" was not physically present in D0 where as its synonym term "ifaajjii" is there in D0 which highlighted blue color.

Evaluation Procedure

To evaluate the performance of the proposed work, the study has been proposed to use Afaan Oromo collected synonym terms without any domain specification for semantic extraction. The performance of test has been evaluated manually and by applying IR system performance evaluation techniques those are recall and precision measure techniques. Precision is fraction of retrieved sentences that are relevant whereas recall is fraction of relevant sentences that retrieved. We classified our dataset into training data and test data. The classification ratio of our dataset was 90 % training data and 10 % testing data. All terms of individual document for all three documents are counted as training data whereas all terms of query are considered as testing data for our work.

Table3. *Result of manual work*

NO	Query	Document 0	Document 1	Document 2
1	Q1	64.8	59.16	7.67
2	Q2	20.12	36.67	6.2
3	Q3	31.67	80	16.67
4	Q4	9.33	22.67	71.83
5	Q5	26.17	47.33	3.67
6	Q6	25.83	44.16	5.3
7	Q7	100	78.67	4.4
8	Q8	44.33	47.5	0
9	Q9	0	0	2
10	Q10	0	0	0

The above table show that result gained from human judgement (manual) of suspected queries and stored documents.

Table4. *Result of proposed model*

NO	Query	Document 0	Document 1	Document2
1	Query 1	12.1	10.6	1.1
2	Query 2	4.5	6.1	1.1
3	Query 3	13	11	2
4	Query 4	8.1	5.4	18.2
5	Query 5	16.67	10.5	1.3
6	Query 6	4.5	4.1	1.4
7	Query 7	7.6	7.5	0
8	Query 8	6	6	1.1
9	Query 9	0	1.4	2.4
10	Q1uery 10	0	0	1.1

The result achieved from proposed model of suspected queries and stored documents. Each terms of suspected queries are compared with each terms of stored documents. The comparison of terms for sentences similarity measures are based on the LSI algorism principle of matrix decomposition.

The similarity measure of the sentences concerns lexical and semantical similarity terms. Summation of result achieved from both (lexical and semantical) similarity was our result as a finding of this work.

CONCLUSION

This study proposed and designed a system called AOSSS measure to solve Afaan Oromo plagiarism problem. The system was designed based on machine learning approach. We have implemented the machine learning features by using LSI algorithm concept to decompose the sentence into term for vector representation for query provided and documents stored. LSI algorism was used to index the term with its value in java hashmap and adopt the model for similarity measure.

For evaluation target we first considered human judgment manually with different respondents of language speakers and plagiarized ration calculated from the adopted LSI algorism to this study. The result obtained from both point can be counted as accuracy of the system. Hence we have got accuracy of calculated ratio of 53.02 %.

FUTURE WORK

Dataset size was the critical parameter for evaluation to achieve better results. In our case we have used small dataset with medium result. N-gram word matching parameter is also a best plagiarism detector technique we have planned for this study and recommend for other local languages.

Standardize the sample dictionary we prepared for synonym term as synset by relational database. Means this sample dictionary must be standardized as WordNet for English throughout further step for Afaan Oromo and other local language with respective rule if not. Enhance performance of the current work in the future steps by deeply focusing on stemming and POS tagging since we didn’t apply POS tagging in this work.

REFERENCES

[1] T. H. Dan Jurafsky, "Document Similarity in Information Retrieval," in *Document Similarity in Information Retrieval*, 2012, p. 81.

[2] P. Resnik, "Semantic similarity in taxonomy: An information-based measure and its application to

problems of ambiguity in natural language," *Journal ofArtificial Intelligence Research*, vol. 11, p. 95–130, 1999.

[3] merriam, "dictionary/plagiarism," merriam-webster, 4 september 2016. [Online]. Available: http://www.merriam-webster.com/dictionary/plagiarism. [Accessed 12 june 2019].

[4] M. Naghibzadeh, "Semantic similarity assessment of words using weighted WordNet," *Springer-Verlag Berlin Heidelberg*, 2012.

[5] P. W. Foltz, "An Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 25, pp. 259-284, 1998.

[6] K. Shams, "Plagiarism Detection Using Semantic Analysis," *Masters thesis*, April 2010.

[7] M. Y. M. Chong, "A Study on Plagiarism Detection and Plagiarism Direction Identification Using Natural Language Processing Techniques," degree of Doctor of Philosophy, University of Wolverhampton, 2013.

[8] L. D. Krisnawati, "Plagiarism Detection for Indonesian Texts," Masters thesis, August 29, 2016.

[9] B. A. V. A. D. M. A, "Framework For Russian Plagiarism Detection Using Sentence Embedding Similarity and Negative Sampling," C*omputational Linguistics and Intellectual Technologies, Proceedings of the International Conference* , 30, June 2, 2018.

[10] S. A. King Abdulaziz University, "A Framework For Plagiarism Detection In Arabic Documents," *DOI : 10.5121/csit.2015.50201*, 2010.

[11] Debela, "Designing a Stemmer for Afaan Oromo Text: A Hybrid Approach," *Masters thesis*, 2010.

[12] F. B. Tesema, "Afaan Oromo Automatic News Text Summarizer Based on Sentence Selection Function," *Masters thesis*, 2013.

[13] E. N. Alemu, "Afaan Oromo –Amharic Cross Lingual Information Retrieval: A corpus Based Approach," *Masters thesis*, 2013.

Citation: *Gizaw Tadele Bekele, Million Meshesha, "Afaan Oromo Sentence Based Plagiarism Detection: A Semantic Similarity Approach" International Journal of Research Studies in Science, Engineering and Technology, 7(7), 2020, pp. 01-06.*