Siamese Long Short-Term Memory for Detecting Conflict of Interest on Scientific Papers

Akhmad Bakhrul Ilmi, Diana Purwitasari, and Chastine Fatichah

***Abstract*—Scientific articles cited by other researchers have an impact on increasing author credibility. However, the citation process may be misused to unnaturally raise a bibliometric indicator value such as researcher's h-index. Researchers may overly cites their own works, referred as self-citation, even though the topic of the references are not related to the current article. Further misconduct is excessive citations on the works of peoples related to the researcher which can be coercive or not, referred as conflict of interest (CoI). The proposed method uses a deep learning approach, Siamese Long Short- Term Memory (LSTM), to recognize subject similarities between a scientific article and its references. Standard text similarity fails to do so because contextual relatedness of sentences in the articles need some learning process. Siamese-LSTM learns contextual relatedness of sentences in the article using two identical LSTM. Steps of the proposed method are (i) word-embedding to get weight values of terms but still considers their semantic relations, (ii) k-means clustering to generate training data for reducing time complexity in Siamese-LSTM learning of scientific articles, (iii) learns Siamese-LSTM weight from training data to identify contextual relatedness of sentences, (iv) calculate similarity of a scientific article with its references based on Siamese-LSTM. The empirical experiments are used to analyze similarity values and the possibility for conflict of interest in an article.**

Keywords—**Citation, Conflict of Interest, Scientific Text, Deep Learning, Similarity, Text Processing.**

I. Introduction[1]

Every scientific articles use citations to cite other researcher's work. Scientific articles cited by other researchers have an impact on increasing author credibility. However, there are researchers who misuse the citation process by making self-citation, negative citations, multi-authorship biased citations, etc. [1] that would raise bibliometric indicator value unnaturally, i.e. researcher's h-index. Researchers may also overly cites their own works even though the topic of the reference is not related to the current article. This case is referred as coercive self-citation. Coercive self-citation does not support the normal rules of research progress [2] and harm the integrity of the researcher.

Further misconduct of self-citation is excessive citations on the works of people related to the researcher which can be coercive or not. That coercive misconduct is called as conflict of interest (CoI). Conflict of Interest is a set of circumstances that create a risk that professional judgement or actions regarding a primary interest will be unduly influenced by a secondary interest [3]. In this case the primary interest is the purpose of professional activity which is integrity of research. Secondary interest is other interests to make profit of themselves such as increasing their h-index. That's why, conflict of interest needs to be detected before calculating researcher's h-index or another indicator value. This paper proposed a deep learning approach to recognize similarities between scientific articles and their references.

In this paper, we proposed a deep learning approach using Siamese Long Short-Term Memory (Siamese-LSTM) and Similarity distance to define relation between paper and each citation. Siamese-LSTM learns contextual relatedness of sentences in the article using two identical LSTM [4]. Steps of the proposed method are (i) word-embedding to know weight values of terms but still considers their semantic relations, (ii) k-means clustering to generate train data for reducing time complexity in Siamese-LSTM learning of scientific articles, (iii) learns Siamese-LSTM weight from train data to identify contextual relatedness of sentences, (iv) calculate similarity of a scientific article with its references based on Siamese-LSTM.

Akhmad Bakhrul Ilmi, Diana Purwitasari, and Chastine Fatichahare are with Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia. E-mail: abakhrul.ilmi@gmail.com.

II. Methodology

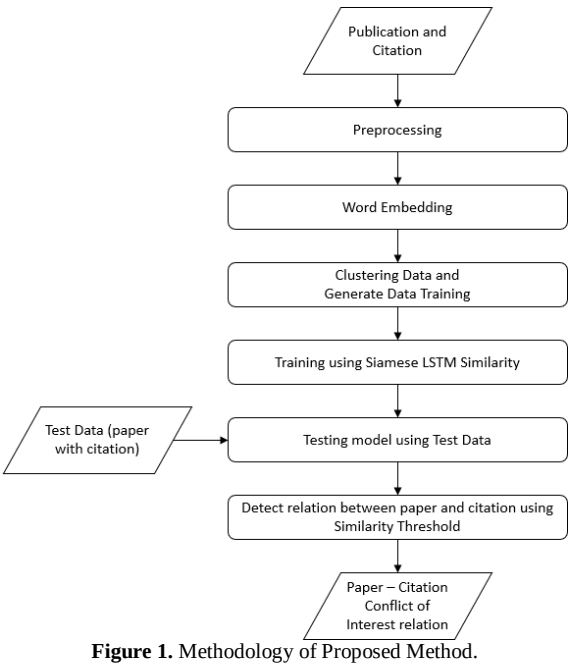

Figure 1. Methodology of Proposed Method.

We aim to build a framework for finding conflict of interest between papers and their citations. The framework for detecting conflict of interest is given in the following steps and showed in Figure 1.

1. Retrieving article and each citation article from database.

2. Preprocessing text data (title and abstract) on each article
3. Word Embedding training to obtain the weight of each word in all articles
4. Clustering all data text used and generating data training from each cluster.
5. Training similarity score using Siamese LSTM.
6. By using threshold, detecting relation between papers and citations based on similarity scores.

A. Preprocessing

The text preprocessing stage aims to extract interesting and non-trivial data, so the text contains only the words needed for further process. In this study, the preprocessing consists of these following steps.

1. Case Folding
   Converting all alphabets in the documents to lowercase.
2. Punctuation marks removal
   Removing punctuation marks and special characters in the documents
3. Stop word Removal
   Removing words that occur too many times in the corpus so that they can't represent the content of documents.
4. Lemmatization
   Converting all terms to their base form by removing inflectional ending using vocabulary and morphological analysis.

B. Word Embedding

Word embedding is a collection of names from language modeling and feature extraction techniques in natural language processing (NLP) where each word or phrase of a vocabulary will be mapped to a vector of real numbers [5]. Word embedding is often used in neural networks, dimensionality reduction in vector space model, probabilistic models, etc. This word embedding method is also used as input to improve performance in human language processing such as syntactic parsing and sentiment analysis.

C. Clustering Data

This stage aims to divide a collection of scientific article documents into several clusters. These clusters will be used as input data on the calculation of similarity scores. The similarity scores between clusters that meet the given threshold are then transformed into distance matrix. This is done so that the training data used on the Siamese Neural Network will be able to map words with the same topic. The clustering algorithm used is K-Means ++ with Cosine Similarity [6].

Training data are obtained by filtering data from each cluster. The data used are those which are close to the cluster center and far away from the cluster center. Thresholding process is used to determine the closeness between data and the cluster center. The similarity matrix of the selected data will be used for the training process on Siamese LSTM. These steps intend to limit the number of data involved in training process so that it wouldn't take a considerable amount of time.

D. Siamese LSTM

Siamese network is a network that has multiple identical sub-networks. This network can be used for supervised similarity search. The combination of Siamese network and LSTM can create similarity search methods in text, which the training is done by LSTM. Siamese LSTM has several parts namely word embedding, 2 sub-network LSTM, and prediction function [4]. The prediction function used in this research is cosine similarity, where the desired result is the value of similarity between documents. An example of a Siamese LSTM process diagram is displayed in Figure 2.

Figure 2. shows a Siamese LSTM Network process diagram of words that has been weighted using word embedding. The input of LSTM is the probability value from document 1 and document 2, which is then carried out by training. Existing training weights on LSTM are used for both sub- networks. The output of each LSTM in the form of a hidden vector LSTM is then used as input to calculate similarity value using cosine similarity. MSE is then used to evaluate the reliability of resulting model.

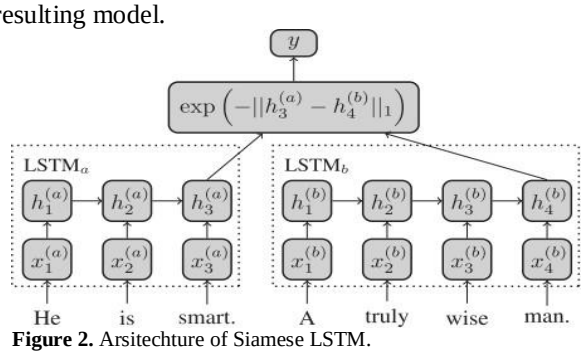

Figure 2. Arsitechture of Siamese LSTM.

III. RESULTS AND DISCUSSION

There are 2 scenarios in this paper. First, we conduct experiments to define value of k for the clustering process. In this experiment, we compare sum squared error and silhouette index to define the best value of k. Second, we perform experiments to compare the result of similarity scores between paper and citation with other similarity methods such as cosine similarity and jaccard similarity. Since there is no dependable ground truth, in this case we use other researches' definition of conflict of interest to build ground truth. After that we analyze the accuracy our proposed method based on the ground truth.

A. Dataset

Dataset used in this experiment is obtained from Arnetminer dataset [7]. This dataset contains from many scientific articles such as proceedings, journal paper, and dissertation related to computer science. We used citation analysis dataset that contain paper's, citation's, and author's data. Scientific article used in this paper only contain of title and abstract. Arnetminer dataset have around 2 million of scientific article. In this experiment we use data consist of 50% conflict of interest relation and the other 50% is from non-conflict relation.

B. Result

First Scenario, to define value of k in the clustering process, we used SSE and silhouette index to determine it. From Figure 3, it is found that the value of SSE has decreased as number of cluster increases. This is because the smaller the value of k used, the greater the area of cluster. To determine the number of clusters by using

SSE, it can be done using elbow's method. We obtain that k = 2 until k = 8 have a big different of error compared to next amount of cluster and after k = 12 it has similar error to each SSE. From that, we conclude that by using number of clusters from 8 to 12 will provide good cluster results. But, to confirm our opinion, we will use another cluster validity's method to determine the number of clusters.

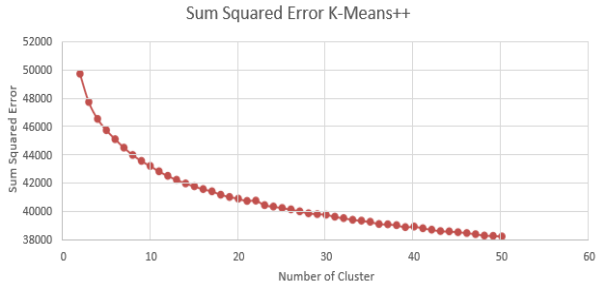

Figure 3. Sum Squared Error of K-Means++ from K = 2 to K = 50.

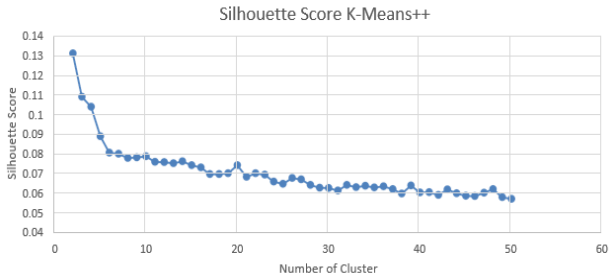

Figure 4. Silhouette of K-Means++ from K = 2 to K = 50.

To obtain the best number of clusters in K-Means, we analyze it using the silhouette score. From k = 8 to k = 12 obtained from the previous analysis, the silhouette score will be analyzed. From Figure 4, it is found that the highest score of silhouette score is k = 10. Therefore, in this research, the process of clustering the document used k = 10 in k-means ++.

Second scenario, we perform experiment to compare our proposed method with other methods. In this case, we will compare it with cosine similarity and jaccard similarity. We used accuracy to define the performance of each method. From Table 1, we obtain that Siamese LSTM outperform other similarity method by 62.50%. This is because Siamese LSTM is able the check the similarity based on semantics. On the other hand, cosine similarity only detects 50% of all test data. This is because, cosine similarity on use term frequency to determine the sameness of each article and fail to obtain the semantic of it. Although Jaccard similarity also have a good accuracy in Table 1, however it fails to determine conflict of interest. Its because the result of jaccard similarity is "Conflict" on all dataset which make its accuracy is high compared to cosine similarity.

TABLE 1.
RESULT OF PROPOSED METHOD

	Threshold Similarity		
	0.3	**0.35**	**0.4**
Siamese LSTM	62.50%	62.50%	50%
Cosine Similarity	37.50%	37.50%	50%
Jaccard Similarity	50%	50%	50%

IV. CONCLUSION

In this paper, we proposed a method to detect conflict of interest from scientific articles for each paper and each citation, by using deep learning approach, we have provided a new method to detecting conflict of interest. A-Miner dataset is used in this experiment for scientific article. From the experiment, we conclude that deep learning approach outperform other similarity measure in detecting conflict of interest with the accuracy of 62.50%. We have some issue faced in this research. One of them is the lack of ground-truth. Ground-truth used in this research is obtained from other paper but its not dependable. Our future work will be focused on the useful of Siamese LSTM to detecting the authors habit of citing by combine it with self-citation analysis.

REFERENCES

[1] K. Moustafa, "Aberration of the Citation," *Account. Res.*, vol. 23, no. 4, pp. 230–244, Jul. 2016.

[2] T. Yu, G. Yu, and M.-Y. Wang, "Classification method for detecting coercive self-citation in journals," *J. Informetr.*, vol. 8, no. 1, pp. 123–135, Jan. 2014.

[3] Institute of Medicine, *Conflict of Interest in Medical Research, Education, and Practice*. Washington, D.C.: National Academies Press, 2009.

[4] J. Mueller, "Siamese Recurrent Architectures for Learning Sentence Similarity," in *Proc. 30th Conf. Artif. Intell. (AAAI 2016)*, 2016, pp. 2786–2792.

[5] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining Word Embeddings Using Intensity Scores for Sentiment Analysis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 671–681, Mar. 2018.

[6] R. Srivastava and H. Gupta, "K-means Based Document Clustering with Automatic ' K ' Selection and Cluster Refinement," *Int. J. Comput. Sci. Mob. Appl*, 2014.

[7] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: extraction and mining of academic social networks," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, 2008, p. 990.