Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-04 | Issue-04 | December-2023

Word-Graph Construction Techniques for Context Analysis

Rafique Yasir[1], Wu Jue[2], Mushtaq Muhammad Umer[3], Atif Nazma[4]
[1]PhD Scholar, School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China.
[2]Professor, School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China.
[3]PhD Scholar, School of Information Engineering, Southwest University of Science and Technology, Mianyang, China.
[4]MPhil Scholar, School of Information Engineering, Southwest University of Science and Technology, Mianyang, China.
yasirrafiquebscs@gmail.com, wujue@aliyun.com, umer449@gmail.com, nazmaatif30@gmail.com

DOI: 10.5281/zenodo.10594263

ABSTRACT

A Nomo-Word Graph Construction Analysis Method (NWGC-AM) is used to graph let the corresponding construction phrases into essential and non-essential citation groups. NMCS-NR, or Nomo Maximum Common Sub-graph edge resemblance, Maximum Common Subgraph Directed Edge resemblance (MCS-DER), and Maximum Common Subgraph Resemblance. The graph resemblance metrics used in this work are called Undirected Edges Resemblance (MCS-UER). The tests included five distinct classifiers: Random Forest, Naive Bayes, K-Nearest Neighbors (KNN), Decision Trees, and Support Vector Machines (SVM).Four sixty one (361) citations made up the annotated dataset used for the studies. The Decision Tree classifier exhibits superior performance, attaining an accuracy rate of 0.98.

Keywords: Lexical Network, Graph Based Language Representation, Node Link Structure, Citation Index, Resemblance.

Cite as: Rafique Yasir, Wu Jue, Mushtaq Muhammad Umer, & Atif Nazma. (2024). Word-Graph Construction Techniques for Context Analysis. *LC International Journal of STEM, 4(*4), 25–35. **https://doi.org/10.5281/zenodo.10594263**

INTRODUCTION

Why are certain papers cited by researchers? For many years, scholars in the fields of information sciences, discourse analysis, and sociology of science have been intrigued by this subject. The process of quoting data, findings, and conclusions from books, papers, or websites into a research project is known as citation. Citations support research questions and hypotheses, provide additional background information to the user, acknowledge previous work in the field, and contextualize the study in relation to other works in the field. Citation pattern analysis of scientific publications has been widely used to identify scientific collaboration, chart the domain boundaries of academic fields, evaluate the influence of research outputs, and track cross-domain knowledge transfer.

Metrics such as citation count are employed to gauge the significance and appeal of Scientific Research Papers (SRPs). Although some citations are significant from a semantic standpoint, some are not, the

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-04 | Issue-04 | December-2023

traditional paradigm of referring treats all citations equally [1] . One can ascertain the significance of a certain citation by looking at the context of citations. Understanding the type of citation is crucial because it enables us to evaluate the relative significance of each mentioned publication. Through citation analysis of their work, scientific publications provide a unique stream of data that may be traced back to individual researchers. Amount of bibliographic data published on the web is growing aggressive. Thus, it is crucial to analyze the mentioned work.

This study addresses the challenge of classifying citations into two distinct categories: significant and insignificant. Traditionally, the number of references a paper accumulates has been utilized as a metric to gauge the influence of the published research. Nevertheless, citations can serve different purposes, and not all references hold equal weight. References that are employed to build upon or expand existing research have a more substantial impact compared to those used for mere comparison or background information. Hence, this investigation also explores the potential of utilizing the surrounding text of a citation to determine its importance or insignificance.

However, we introduce a new method to classify citation sentences into essential and non-essential classes: the Nomo-Word Graphical Citation Analysis Method (NWGCAM). NWGCAM matches quoted phrases and places them into appropriate classifications using graphical similarity metrics. The Maximum common subgraph node Resemblance (MCS-NR), maximum common subgraph directed edge resemblance (MCS-DER), and maximum common subgraph undirected edge resemblance (MCS-UER) are the similarity metrics used in this work. The maximum number of common subgraph (MCS) nodes is determined by MCS-NS, and counter directed and directionless border at the M-C-S is determined by MCS-DER and MCS-UER, respectively.

Our study is in close agreement with that of [2] and [3], who similarly concentrate on categorizing referenced material in scholarly publications. Citations are divided into four categories by them: non-essential citations for using and expanding the work, and essential citations for related work and comparison. Nevertheless, our work goes above other methods by using the novel Word graph citation analysis method (WGCAM) for subgraph matching. NWGCAM uses various classifiers such as Support Vector Machine (SVM), Random Forest, Naive Bayes, K-Nearest Neighbors (KNN), and Decision Trees, to reach an outstanding accuracy of 0.98.

LITERATURE REVIEW

In the body of current literature, several citation classification techniques have been proposed. Using cue words, [2] present six novel traits that divide referenced chores into essential and insignificant citation categories. [3] Use a classification system with four unique types to distinguish between essential and non-essential citations in scientific journals. Use of chores and extension of chores are regarded as helpful citations in these classes, although related work and comparison are excluded. To categorise citations, [1] offer a strategy that takes into account elements like author overlap and per-section citations. [4] Separate referenced text into two groups—positive and negative—by using text mining techniques to extract the text from research papers.

[5] Compare their extra review citation categorizing with the most advanced methods currently used in the industry. Citation organization methodology quadruple unique Basic knowledge, basic ideas, comparisons, and technical basis of the course is presented by [6] . To identify the citation classes, they make use of text-based attributes. Ten categories of citations are proposed by [7] Positives, negatives, trials, hypotheses, developments, methodologies, further work, contrasts, book net warnings & concepts.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-04 | Issue-04 | December-2023

[8] Sorts construction into different categories such as Review Reference, Development Reference, and Methodology credentials, Academic Knowledge, Preliminary Credentials, Furthermore Research, etc. based on the citation's position within article and phrase used at the construction content. A citation classification system that divides citations into positive, negative, and neutral classifications is presented by [9]. [10] Provide a method that uses citation contexts to produce research article summaries. The referenced work is categorized into groups like prior state-of-the-art work, extensions, advantages, disadvantages, and synopsis. Citations are divided into three kinds by [11] aspects, polarity, and aim. Additionally, they describe a number of functions related to classification, such as hedging, supply, usefulness, acknowledgement, discussion, confirmation, comparison, and weakness. Diverse machine learning strategies and classification algorithms have been used in earlier research to divide citations into essential and non-important classes. On the other hand, the word-graph method uses three distinct graph similarity measures—MCSUER, MCSDER, and MCSNR—to provide a fresh way to classify citations.

METHODOLOGY

In this work, we present the Nomo-Word Graph Citation Analysis Method (NWGCAM). Figure 1 is used to illustrate how quotation phrases in research papers are classified into important and unimportant citation classes. Word graphs, which are directed, unweighted graphs created from the terms or words contained in the sentence, are used to represent each citation sentence

NWGCAM creates nomograph for one and two citation steps important and unimportant using annotated citation paragraph into data file [3] key in in order to carry out the classification. The method's output is the citations' classification. Each word graph is built using citation sentences from write to class. To evaluate upto a minute nature in respect to the citation classes, it is additionally converted in nomograph, which is called the target sentence. The target sentence's word graph and nomograph of citation step are then compared, and similarity values are determined by using maximum common subgraph graph similarity measures like MCS-NR, MCS-DER, and MCS-UER. The calculation procedure is demonstrated in equations 1, 2, and 3. The suggested methodology is illustrated in Figure 2.

The process of matching nomograph graph of the up to date sentence to the nomograph of the citation step is shown in Figure 2. Using the similarity measures MCS-NR, MCS-DER, and MCS-UER, this matching yields same values.

Each similarity number indicates degree to similarity between the new sentence and its relevant cite step. A very similar number suggests that the cite paragraph related to the nearly all-applicable cite step.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-04 | Issue-04 | December-2023

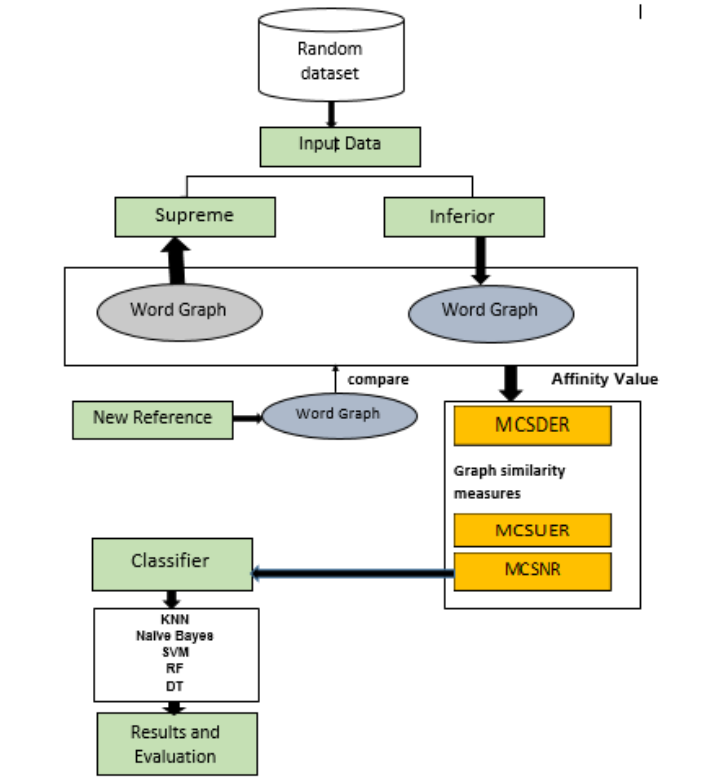

Figure1: proposed approach

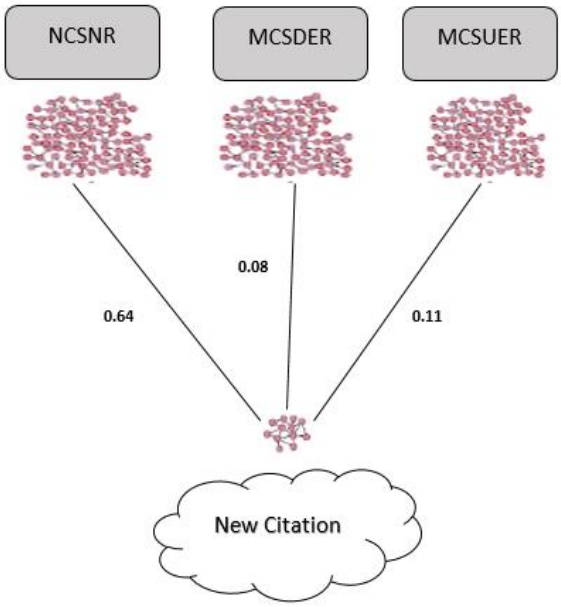

Figure 2 graph matching between the words graphs of the main citation with the Nomo-word-graph of the construction category.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-04 | Issue-04 | December-2023

The degree of similarity between the new sentence and its corresponding citation class is indicated by each same number. The cite paragraph may belong to very suitable cite step if the similarity value is higher.

Data

The dataset that Valenzuela introduced in the study is very familiar [3]. The citations in academic papers are classified using this benchmark dataset as a guide. With 19,638 research articles included in the dataset, it offers a comprehensive resource for our analysis.

Model Development

Nomograph is built form on the word to word that are contained in a cite paragraph and how close together they are. A vertex or node represents each word or term in the citation sentence. An edge forms between the corresponding vertices of two words or terms that surround within the cite paragraph and connected in a directed manner. This procedure maintains the original word order. The border that connect the apex captures the proximity of the words. To illustrate this, body cage moves over text, identifying the nodes, border of nomogram, as depicted in Fig 3. The area of the frame ranges from one to 15 words. Figure 4 provides an example of a word graph for the cite paragraph: "undersign are word vectors (Lin and Hovy, 2000)". Of these papers, 106,509 have general citations; 298 research articles have been randomly annotated from them. The citation classes supreme and inferior are displayed in Table 1. Class labeled 0 denotes inferior labor, whereas labeled 1 denotes comparison. Label 1 designates the important class whereby the work is used in steps two and three, respectively. In the dataset, 85.5% of the citations are classified as inferior and 14.4% as supreme.

Table 1 dataset statistics

Label	Citation	No of Citation(361)
0(0-1)	Relevant(related work)	298
1(2-3)	Irrelevant	63

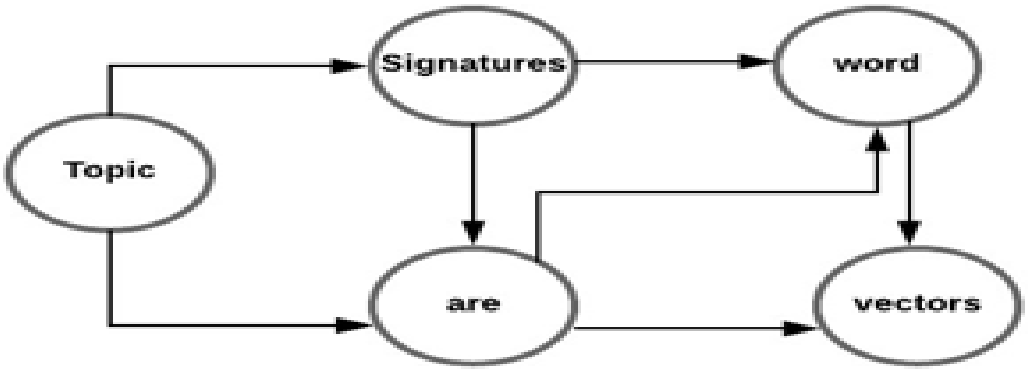

Figure 3: Nomo-word graph construction of a citation sentence.

Maximum Common Sub-Graph (Mcs) Parameters

This here we utilized similarity matric based on the maximum common sub-graph (MCS) size in one two Nomo-word graphs. There are three distinct variations on these similarity matric are denoted as equations 1, 2, and 3 [12] . The M-C-S-based similarity measures are as follows:

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-04 | Issue-04 | December-2023

$$\text{MCS-NR} = \frac{MCSNR(|G_T|,|G_S|)}{min(|G_T|,|G_S|)} \quad (1)$$

$$\text{MCS-UER} = \frac{MCSUER(|G_T|,|G_S|)}{min(|G_T|,|G_S|)} \quad (2)$$

$$\text{MCS-UER} = \frac{MCSDER(|G_T|,|G_S|)}{min(|G_T|,|G_S|)} \quad (3)$$

The number of nodes that included in the MCS of graphs. Nomo-word graph targeted by latest cite paragraph is the Nomo-word graph source of important and un-important cite.
On the other hand, represent the area of the origin and earmark monograph. The area of a Nomo graph can be determined by the integer of interchange and border surround within the M-C-S. Refers to the count of undirected border that are encompassed within the MCS. Denotes area of directed area accommodate in the M-C-S.

The proposed approach involves the categorization of citations into two distinct parts. In the initial step, Nomo-word graphs are constructed for every cite step, and those cite are utilized to instruct various angle matric including SVM, KNN, Bayes, Random Forest, and Decision Tree. The other step involves representing instruct dataset of every citation as a Nomo-word graph. Eventually, Nomo-word graphs of the citations are compared to the word graphs of the supreme and inferior citation classes in which graph-based resemblance scale, namely MCSNR, MCSDER, or MCSUER. The results of the subgraph matching generate three similarity vectors based on the graph similarity measures, which represent the cite steps. The angle from the instructing are then employed to instruct various steps of classes.
An updated cite is transformed into Nomo-graph let and compared to the more than one existing graph let. This process yields same numbers that characterize each cite paragraph. The analysis as adorn in Figure 4 and utilizing the set of feature outlined in Table 4.1, assigns the citation vectors to their respective classes. The classification is performed using five distinct matric: Random Forest, Naive Bayes, SVM, KNN, and Decision Tree.

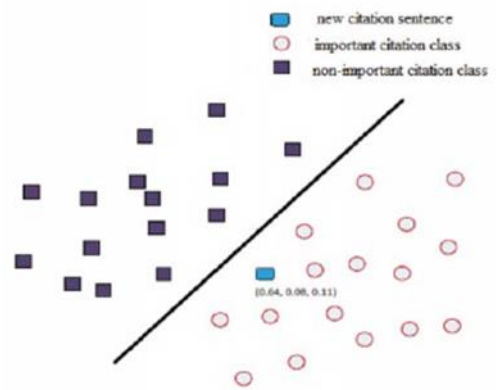

Figure 4: Classification of a new citation construction sentence using its vector representation.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-04 | Issue-04 | December-2023

Set No	Feature Set	Description
1	Imp_node_sim	Important citation mode similarity
2	Not_imp_mode_sim	Not Important citation mode similarity
3	Imp_directed_sim	Important directed edges similarity
4	Not_imp_directed_sim	not-important citation directed edges similarity
5	Imp_undirected_sim	important citation undirected edges similarity
6	Not_imp_undirected_sim	not-important citation undirected edges similarity

4.1 Table 2 feature set

Using the 10-fold cross validation method, we looked at our suggested WGCAM technique [13] . In each fold, 10% are utilized for testing and 90% are used for training. The experiments, classification, and implementation for creating a graph, comparing a graph, and creating a vector representation of citations are all done using the Python programming language.

We employ Nomo Receiver Operating Characteristic (NROC) curve analysis to assess our suggested methodology. Citation subgraph matching to essential and not-so-important citation classes is done using NWGCAM. The positive class represents the significant class, while the negative class represents the unimportant class. NROC curves for five distinct classifiers are shown in Figure 5-9, which demonstrates that Decision Trees outperform with an AUC (mean) = 0.98.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-04 | Issue-04 | December-2023

DATA ANALYSIS AND RESULTS

Results

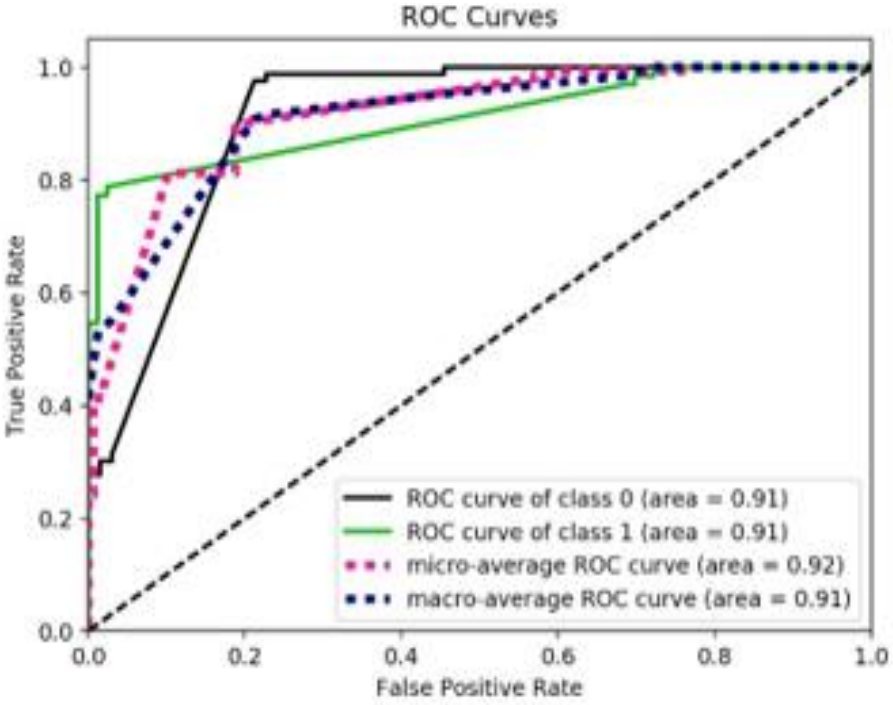

Figure 5: NROC curve on SVM

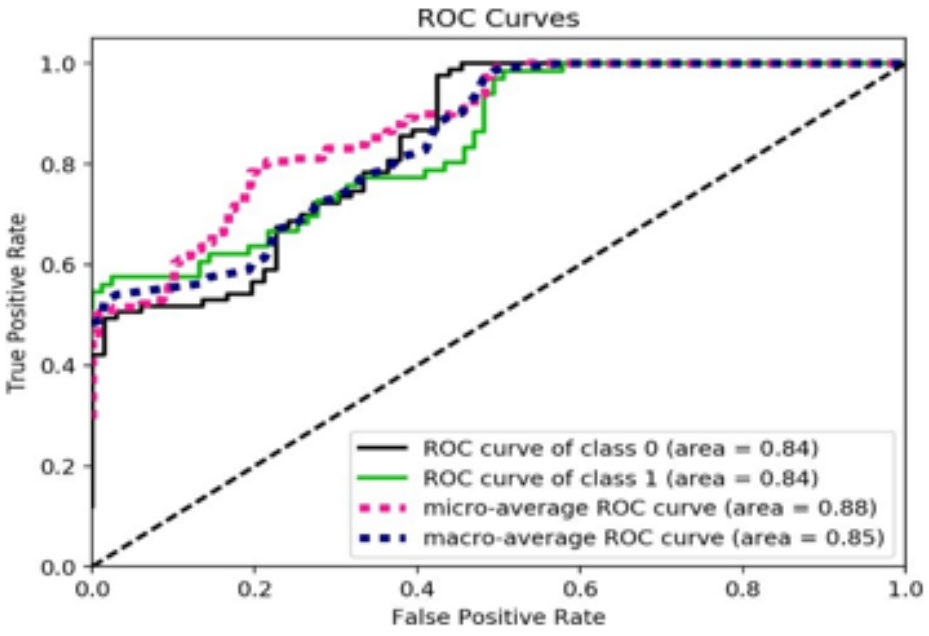

Figure 6: NROC curve on Naïve Bayes

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-04 | Issue-04 | December-2023

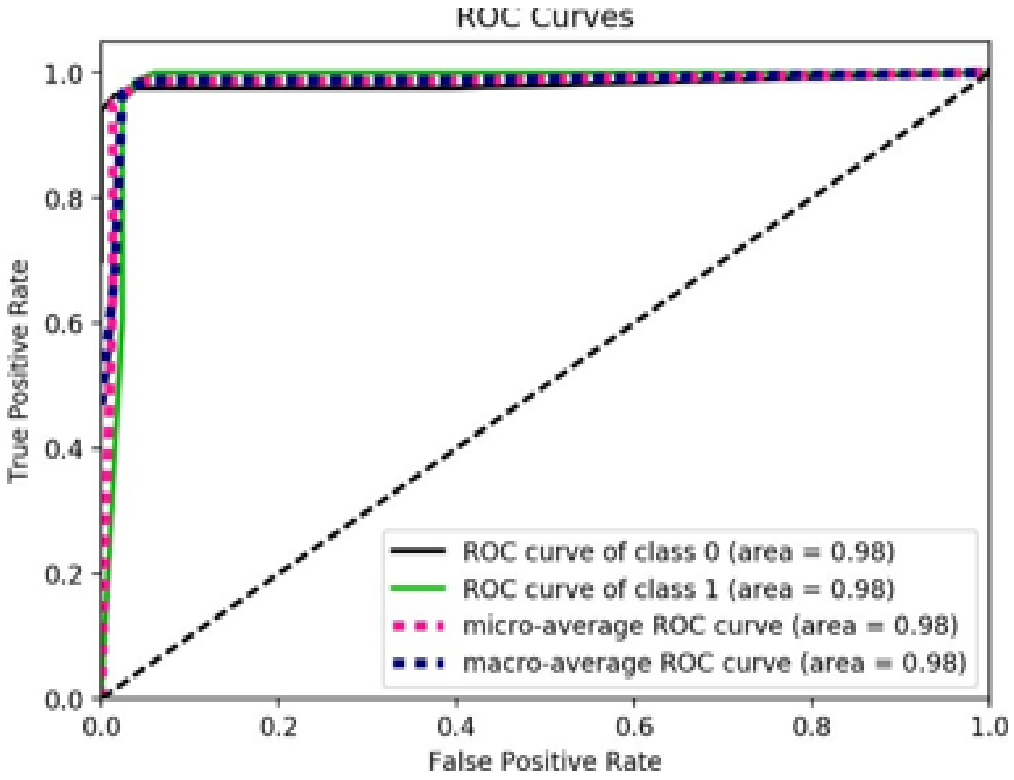

Figure 7: NROC curve On Random Forest

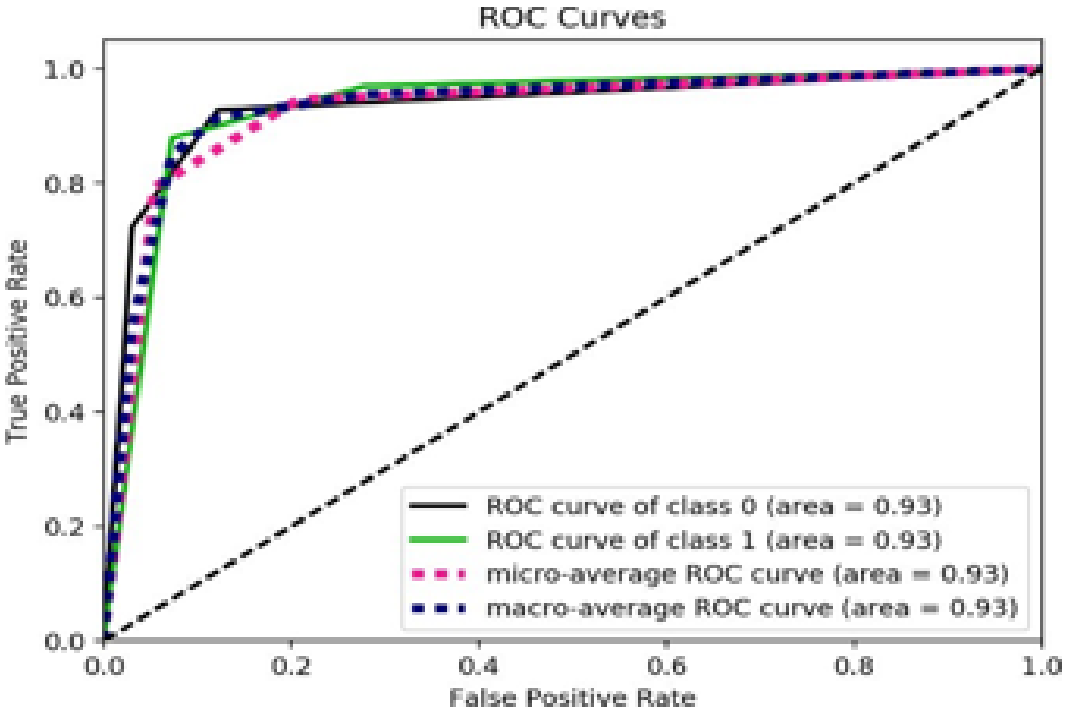

Figure 8: NROC curve On KNN

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-04 | Issue-04 | December-2023

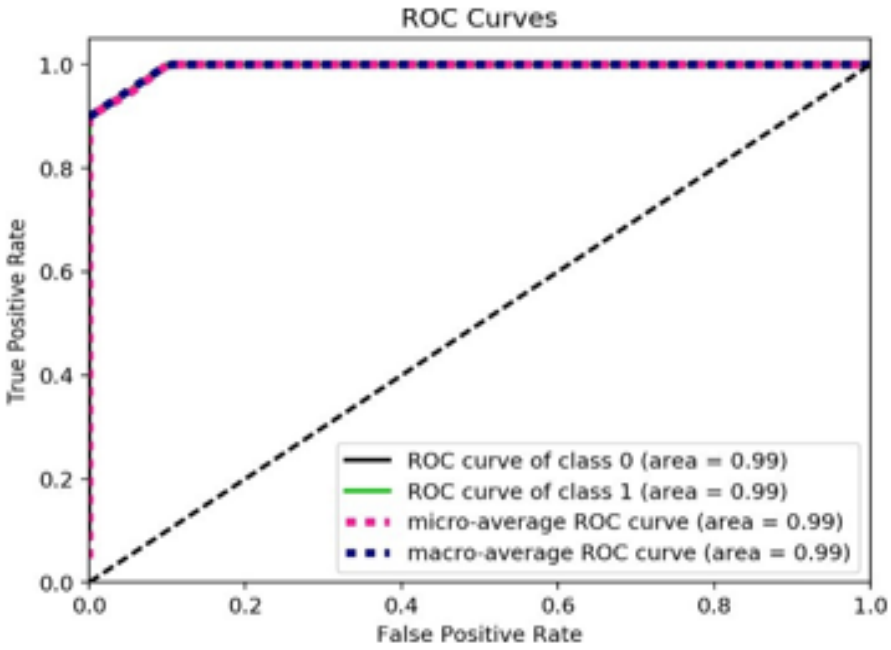

Figure 9: NROC Curve On Decision Three

CONCLUSION AND RECOMMENDATIONS

This paper presents the Nomo-Word Graph Citation Analysis Method (NWGCAM), a technique towards categorizing cite paragraph of research publications towards essential and inferior citation steps by matching subgraphs. Every citation sentence is represented by a directed, unweighted word graph that is created using the terms and/or words present in the sentence. Word graphs for the significant and non-important citation classes have been produced. Graph likeness metrics such as MCS-NS, MCS-DES, and MCS-UES are used by the WG-CAM approach to match citation sentences to classes. We used Bayes, Random Forest, SVM, KNN, and Decision Tree classifiers for classification. With an AUC of 0.98, Decision Tree performed the best out of all of them. Positive outcomes were also obtained by KNN and Random Forest, with AUC (mean) values of 0.98 and 0.92, respectively. These outcomes show how successful the suggested strategy is; with an AUC of 0.98, it performs better. We intend to improve the NWG-CAM in further studies by adding weighted word graphs. With this enhancement, we will be able to better refine the subgraph matching process by identifying the significance of specific words within citation sentences.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-04 | Issue-04 | December-2023

REFERENCES

Zhu, X., et al., *Measuring academic influence: Not all citations are equal.* Journal of the Association for Information Science and Technology, 2015. **66**(2): p. 408-427.

Hassan, S.-U., A. Akram, and P. Haddawy. *Identifying important citations using contextual information from full text*. in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 2017. IEEE.

Valenzuela, M., V. Ha, and O. Etzioni. *Identifying Meaningful Citations*. in *AAAI workshop: Scholarly big data*. 2015.

Sula, C.A. and M. Miller, *Citations, contexts, and humanistic discourse: Toward automatic extraction and classification.* Literary and Linguistic Computing, 2014. **29**(3): p. 452-464.

Jochim, C. and H. Schütze. *Towards a generic and flexible citation classifier based on a faceted classification scheme*. in *Proceedings of COLING 2012*. 2012.

Dong, C. and U. Schäfer. *Ensemble-style self-training on citation classification*. in *Proceedings of 5th international joint conference on natural language processing*. 2011.

Garzone, M. and R.E. Mercer. *Towards an automated citation classifier*. in *Advances in Artificial Intelligence: 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2000 Montéal, Quebec, Canada, May 14–17, 2000 Proceedings 13*. 2000. Springer.

Finney, B., *The reference characteristics of scientific texts*. 1979, City University (London, England).

Teufel, S., A. Siddharthan, and D. Tidhar. *Automatic classification of citation function*. in *Proceedings of the 2006 conference on empirical methods in natural language processing*. 2006.

Tandon, N. and A. Jain. *Citation context sentiment analysis for structured summarization of research papers*. in *35th German conference on artificial intelligence*. 2012. Citeseer.

Hernández Álvarez, M., J.M. Gómez, and P. Martínez-Barco, *Annotated corpus for citation context analysis.* 2016.

Violos, J., et al. *Sentiment analysis using word-graphs*. in *Proceedings of the 6th International Conference on Web Intelligence, mining and semantics*. 2016.

Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *Ijcai*. 1995. Montreal, Canada.