ICIC Express Letters
Part B: Applications
Volume **15**, Number **3**, March **2024**

ICIC International ©2024 ISSN 2185-2766
pp. **245–252**

DIMENSIONALITY REDUCTION AND VISUALIZATION OF WORD FORMATTING INFORMATION AS AUTHOR'S WRITING FEATURE IN CLASS ASSIGNMENT REPORTS

Asako Ohno* and Yoshihiro Ohata

Faculty of Engineering
Osaka Sangyo University
Nakagaito, Daito, Osaka 574-8530, Japan
e007725@ge.osaka-sandai.ac.jp
*Corresponding author: ohno@eic.osaka-sandai.ac.jp

Received June 2023; accepted August 2023

Abstract. *In this study, we parse Word .docx documents to obtain and quantify superficial features called word formatting information that are independent of the content of the document as author-specific features, aiming to support the process of visually identifying the author of a report by teachers. In this paper, we report on our attempt to reduce the dimension of the features using PCA (Principal Component Analysis) and t-SNE (t-distributed Stochastic Neighbor Embedding) and output them to visualize the similarity based on the author features of the target set of reports as feedback to teachers considering the application of the proposed method in actual classes.*
Keywords: Word formatting information, PCA, t-SNE, Class assignment report, Plagiarism detection, Dimensionality reduction, Visualization

1. **Introduction.** With the COVID-19 pandemic, university classes shifted online, and an increasing number of class assignment report documents are being created and submitted electronically. Since digital documents are easier to copy and paste than handwritten documents, report plagiarism is a growing concern. According to McCabe's 2003-5 academic integrity survey of more than 60,000 college students in the U.S. and Canada, 1 in 5 (21%) students reported that they had been involved in some form of cheating or plagiarism of assignments during the past year [1].

This trend has led to the use of plagiarism detection tools in many research and educational institutions, both domestic and international. However, according to Anson and Kruse, faculty members have been reluctant to use plagiarism detection tools for fear of false positives and student anxiety because many of the mainstream plagiarism detection methods in recent years have been based on content-based similarity [2].

In particular, unlike general academic papers, class assignment reports are written by multiple students on the same topic based on the knowledge acquired in class, and thus the content, such as terminology, expressions, and document structure within the report, is likely to be similar. For this reason, there is a risk of false positives due to the coincidence of existing content-based methods.

To deal with this problem, we proposed a method that quantifies superficial features independent of the document content as author-specific features. The aim is to assist class instructors in visually confirming the identity of the author and submitter of a report before directly confirming the authenticity of a "suspicion" of plagiarism detected by the plagiarism detection tool with the students.

In this study, we have analyzed *.docx* documents in Word, a document writing application widely used around the world, using an XML (eXtensible Markup Language)

DOI: 10.24507/icicelb.15.03.245

parser, and extracted 29 types of formatting information from the Word *.docx* documents as the author's features. Our method discriminated reports on different topics created by the same author with 90% accuracy using random forest, and visualized author-specific descriptive features using decision trees [3].

As feedback to the teachers who visually compare reports based on their appearance in actual classes, in addition to explaining and identifying the features of individual authors, providing a rough grouping of authors on a two-dimensional map is assumed to be helpful for teachers to identify groups of authors who are similar in terms of the appearance of their reports in the entire class. To realize the visualized feedback, in this paper as an initial attempt, we use PCA (Principal Component Analysis) [4] and t-SNE (t-distributed Stochastic Neighbor Embedding) [5], which are representative methods widely used in data analysis, to reduce the dimensionality of the features, and then create a map of the target reports to visualize and output the degree of similarity based on the descriptive features of the authors of the target set of reports.

The paper is organized as follows. Section 2 describes the trend of report plagiarism detection and basic knowledge of Word formatting information that forms the background of this study, and Section 3 explains the proposed method. Section 4 reports the results of dimensionality reduction and visualization attempts using PCA and t-SNE, and Section 5 provides a summary and future issues.

2. **Backgrounds.**

2.1. **Plagiarism and its detection in academic report.**

Plagiarism in digital reports and research papers has become an ever-growing issue in recent years, and plagiarism detection tools have been introduced in educational and research institutions around the world [6]. Mainstream plagiarism detection tools, such as Turnitin [7], perform plagiarism detection by calculating the similarity between documents through text-based matching. Since simple text-based matching alone cannot cope with paraphrase deception, various methods have been proposed and implemented, including methods based on latent semantics and vectorization at the word or phrase level [6]. For example, Li et al. [8] proposed a recurrent neural network architecture for semantic similarity. The model uses conditional Bi-LSTM (Long Short-Term Memory) encoding and soft alignment attention mechanisms to identify semantic equivalence or inconsistency between pairs of words, phrases, and sentences. Experiments confirm that the model is effective in identifying paraphrases and semantic associations.

Plagiarism in class assignment reports can be classified into three categories: 1) In-class plagiarism, in which students copy the reports of other students in the same class, 2) Out-class plagiarism, in which students copy texts from external sources such as the Internet, and 3) Ghost-writing, in which a third person writes on behalf of the student [9]. ChatGPT [10], a generative AI application released by OpenAI in November 2022, instantly outputs sentences that look as if they were written by a human. With this, the fourth category of plagiarism, plagiarism by generative AI, is emerging. In contrast, Turnitin, a plagiarism detection tool, has already started providing a function to detect sentences generated by AI [7]. On the other hand, WebGPT [11], which is being developed by OpenAI, the developer of ChatGPT, can attach references to output sentences using a text-based web browsing environment. This may provide a new form of citation via generative AI that will become commonplace in the future. At present, however, many educational institutions have posted notes on their official Web sites warning students about the use of generative AI and urging careful use, and it is rare for students to be allowed to use generative AI without restrictions when writing reports and papers.

Thus, amidst this ever-evolving technological innovation, the work of double-checking, in which teachers use tools based on contents-based similarity of report documents to

detect "suspicious" plagiarism and then conduct their own visual checks to determine the identity of the author and submitter of the report by relying on visual features, is expected to become more important in the future. However, the process of visually comparing multiple reports is labor-intensive. In addition, there are no standards for judging the authors of the reports, and the process is left to the knowledge and experience of individual teachers, making it difficult to ensure fairness. In addition, class assignment reports are written by multiple students on the same topic based on the knowledge acquired in class, and thus the keywords and document structure are tending to be similar. Therefore, there is a risk of false positives in plagiarism detection methods based on content-based similarity. Therefore, this study proposes a method to assist teachers in their visual work by quantifying and representing superficial features that are independent of the content of the report. We have previously proposed a method for author authentication by training a set of hidden Markov models on the writing styles of multiple report documents created by the same author in the past [9]. However, this method is difficult to apply to actual classes because it requires multiple report documents created by the same author in the past to be collected in advance and trained into the model.

Therefore, we proposed a new method that focuses on formatting information, which is an appearance feature specific to Word documents, and uses it as a feature to identify the author [3].

2.2. **Word *.docx* file and office open XML.** Microsoft Word (Word) has long been the world's most popular digital writing application since the 1990s [12]. Users can format documents through intuitive operations on a graphical interface. Microsoft Office 2007 and later versions of Word have adopted OOXML (Office Open XML) as ECMA-376 as the file format, which stores text and layout information as XML files [13]. A *.docx* file consists of a set of XML files as shown in Figure 1.

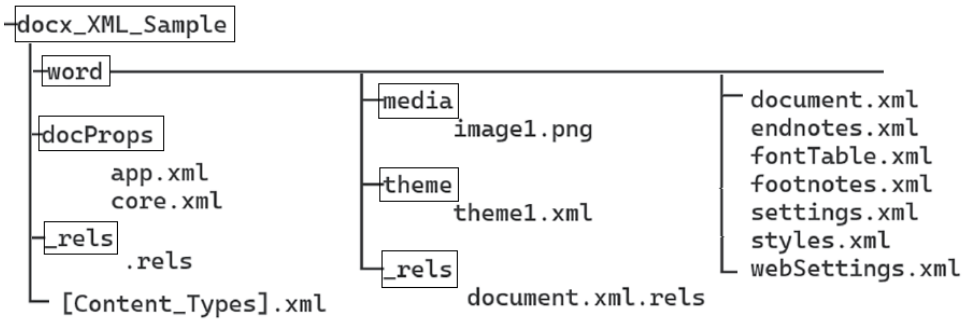

FIGURE 1. An XML file structure comprising a *.docx* document

The body of a Word *.docx* file is stored in an XML file document.xml. For example, a paragraph is marked up with the XML tag <w:p>. This allows the body of text and document structure and decoration information to be retained as an XML file. Such XML files are automatically generated and updated when users make edits on the Word application.

3. **Proposed Method.** In the proposed method, we parse these XML files and represent the author's description features using features called Word formatting information obtained from the parsed XML files. Specifically, 1) elements in the document structure such as paragraphs and tables described as the document object model, 2) counts of decorative information such as underlines and bold text for these elements, and 3) information such as number of characters and revisions provided as metadata are obtained, for a total of 29 explanatory variables.

The following is a list of 29 explanatory variables. We count the following 26 types of tags obtained by XML parsing of document.xml: 1) paragraph (p element), 2) section property (sectPr), 3) table (tbl), 4) hyperlink (hyperlink), 5) paragraph property (pPr), 6) run (r), 7) paragraph indent (ind), 8) paragraph justification (jc), 9) paragraph number property (numPr), 10) paragraph style (pStyle), 11) tab (tab), 12) line break (br), 13) inline shape (drawing), 14) graphic object (pict), 15) run property (rPr), 16) bold tags (b), 17) font color (color), 18) italic (i), 19) font type (rFonts), 20) character pitch (spacing), 21) font size (sz), 22) underline (u), 23) superScript/subscript (vertAlign), 24) table grid (tblGrid), 25) table property (tblPr), 26) table row (tr). In addition, the number of characters in a document 27) from app.xml, the number of footer XML files, 28) from footer.xml, and the number of revisions, 29) from core.xml are obtained.

In this method, the following procedure is used to quantify an author's feature:
[Step 1] Convert the *.docx* file to a set of XML files.
[Step 2] As 29 types of Word formatting information, values such as the counted number of XML tags mentioned above on document.xml and revision number obtained as metadata are used as 29-dimensional explanatory variables representing author characteristics.

In the evaluation experiment in [3], we used experimental reports on six themes written by 13 students in an actual university class. As hyperparameters, we used 100 decision trees, and a maximum depth of a tree was 5 and performed 5-fold cross-validation. Overall, we were able to identify the six different experimental reports written by the same student from those written by the other 12 students with approx. 90% average accuracy.

In this way, it is expected to provide useful quantitative information to assist teachers in visually checking the authors of reports by using the Word format information as an author's writing feature that is independent of the content of the report. However, while the previously mentioned information is useful for "identifying specific students among others", it is inconvenient for "overviewing many-to-many similarities, such as who's writing characteristics are similar to whom in the entire class".

To achieve this, we attempt to visualize the similarity of report writing features in the entire class in the next section. Specifically, the 29-dimensional features are reduced to two dimensions and plotted on two-dimensional coordinates.

4. **Dimensionality Reduction and Visualization.**

4.1. **Visualization using PCA (Principal Component Analysis).**

PCA (Principal Component Analysis) is a classical and widely used method of dimensionality reduction of multidimensional data. In principal component analysis, the coordinate axes expressing characteristics are rearranged to represent the original information with fewer dimensions.

The variables with the largest variance are called the first and second principal components, respectively. The original data is standardized to set the average of each variable to 0 and the variance to 1. Eigenvalues express how many variables each principal component has information on. The eigenvalue of the first principal component indicates whether one dimension of the first principal component can represent the information of n original variables. The contribution ratio of the n-th principal component is the eigenvalue of the n-th principal component divided by the sum of the eigenvalues of all principal components. By calculating the cumulative contribution ratio up to the n-th component, we can check how well the first through the n-th principal components explain the original data.

The cumulative contribution ratio is shown in Figure 2. The contribution ratio of the first principal component is 0.379, that of the second principal component is 0.148, and that of the third principal component is 0.088. The cumulative contribution ratio up to the third principal component is 0.615 (61.5%). To represent approximately 80% of the original data, we need up to the seventh principal component. In this attempt, the first

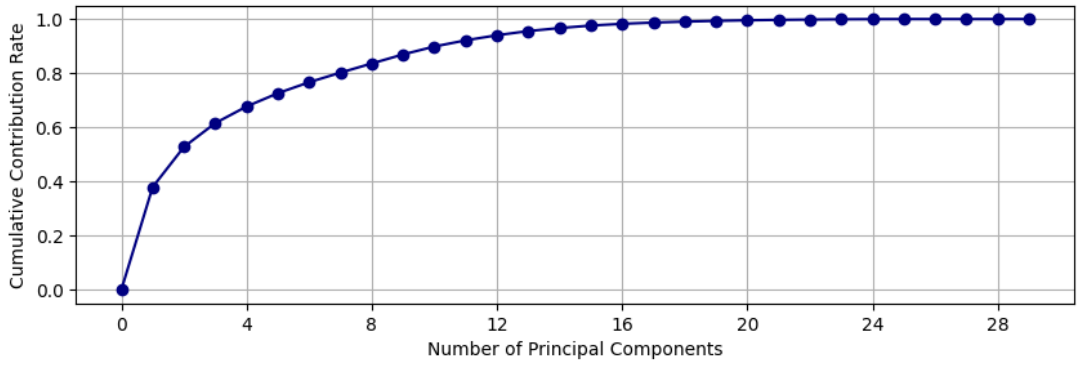

FIGURE 2. Cumulative contribution ratio

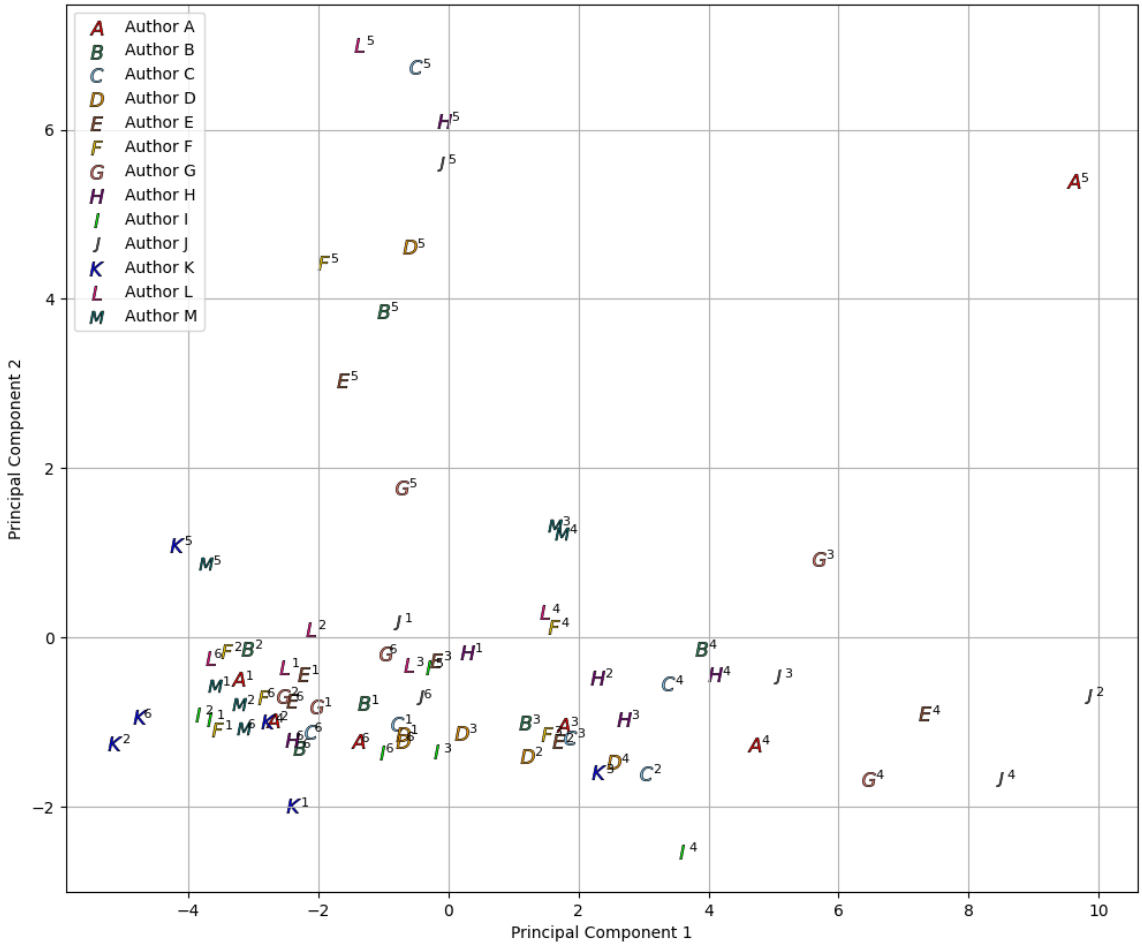

FIGURE 3. Visualization using PCA

and second principal components are used to plot the data in two-dimensional coordinates for visualization. The cumulative contribution ratio is 0.527, which means that almost half of the information is lost from the original data.

The value calculated by transforming the original data into each principal component is called the principal component score. By calculating the first and second principal component scores from each data variable, it is possible to convert the characteristics of each data into a two-dimensional graph. The characters A to M in Figure 3 represent the 13 students and the numbers 1 to 6 represent the themes of the reports, i.e., "B5" represents Author B's report written for theme 5. Each student's report on each of the six themes is plotted on two-dimensional coordinates using the two-dimensional features

obtained through principal component analysis. There are examples of reports from the same author plotted in close proximity to each other such as Author D and K, but overall the data is mixed.

4.2. **Visualization using t-SNE (t-distributed Stochastic Neighbor Embedding).** t-SNE (t-distributed Stochastic Neighbor Embedding) is an improved version of SNE. SNE [14] is a dimensionality reduction method that can handle nonlinear data that linear dimensionality reduction methods such as PCA cannot handle and is superior to PCA in plotting similar multidimensional data in close locations on a two-dimensional plane.

SNE converts distances between high-dimensional data points x_i and x_j into conditional probability $p_{j|i}$, shown as Equation (1), assuming that they follow a Gaussian distribution.

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|x_i - x_k\|^2/2\sigma_i^2\right)}, \tag{1}$$

where σ_i^2 is the variance of the Gaussian distribution centered at data point x_i set as a parameter called *Perplexity* by analyzer. The distance between the same data is assumed to be $p_{i|i} = 0$.

The similarity between data points y_i and y_j after dimensionality reduction is also expressed as a conditional probability $q_{j|i}$ as in Equation (2). Note that the variance is fixed at $\frac{1}{\sqrt{2}}$.

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k\neq i}\exp(-\|y_i - y_k\|^2)} \tag{2}$$

The distance between the same data is also assumed to be $q_{i|i} = 0$.

The distance is obtained by minimizing the KL divergence as a loss function shown as Equation (3) so that the distance between the data points in the higher dimension $p_{j|i}$ and the distance in the lower dimension $q_{j|i}$ after dimensionality reduction are as close as possible.

$$C = \sum_i KL(P_i\|Q_i) = \sum_i\sum_j p_{j|i}\log\frac{p_{j|i}}{q_{j|i}} \tag{3}$$

Due to the difficulty of minimizing this loss function shown as Equation (3), t-SNE, which uses t-distribution instead of Gaussian distribution, is widely used. Here, we also use t-SNE.

In t-SNE, the distance between high-dimensional data points x_i and x_j is symmetric, defined as Equation (4).

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n} \tag{4}$$

Furthermore, they represent the distance q_{ij}, as shown in Equation (5), in low-dimensional space using a t-distribution with a single degree of freedom, which allows t-SNE to represent closer data in higher-dimensional space closer in low dimensional space and vice versa for more distant data.

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}} \tag{5}$$

Figure 4 shows the results plotted on a two-dimensional map by dimensionality reduction using t-SNE. Compared to the PCA results in Figure 3, clusters by author are formed to some extent. By checking the cases where some of the reports of the same author are placed far from the cluster of the author, we found scattered cases corresponding to the number of report groups that match the classification rules indicated by the nodes of the decision tree and the number of report groups that do not.

The value of perplexity, one of the hyperparameters, is considered appropriate by the originator to be between 5 and 50. This time, the value was set to 35 as a result of trial and error.

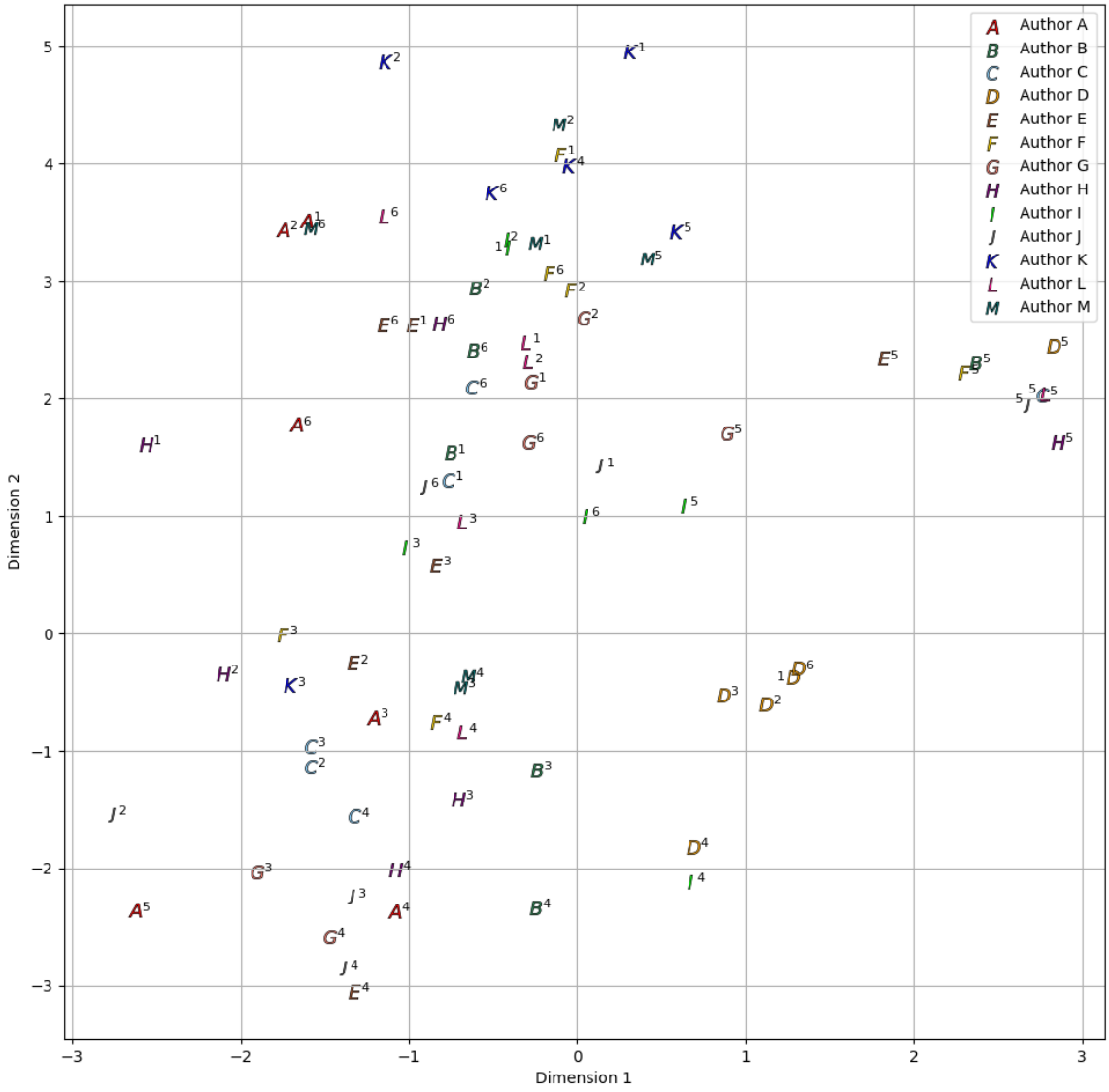

FIGURE 4. Visualization using t-SNE (perplexity = 35)

5. **Summary and Future Works.** The aim of this study is to extract and quantitatively represent authors' features independent of report content as Word format information, and to provide this information as a reference for teachers to identify report authors when conducting plagiarism detection in class reports. In this paper, we report an attempt to reduce 29 features defined as authors' writing features using PCA and t-SNE and plot them on two-dimensional maps to support teachers to visually check the similarity of many-to-many report authors' writing features. We confirmed that t-SNE more closely represents the similarity between multidata in a way that is similar to real data. Further examinations of the explanatory variables are needed. The visualization approach and content also need to be improved.

Acknowledgment. This work was partially supported by a JSPS KAKENHI Grant-in-Aid for Scientific Research (C), 23K02723.

REFERENCES

[1] D. L. McCabe, Cheating among college and university students: A North American perspective, *International Journal for Educational Integrity*, vol.1, no.1, 2005.

[2] C. M. Anson and O. Kruse, Plagiarism detection and intertextuality software, in *Digital Writing Technologies in Higher Education*, O. Kruise et al. (eds.), Cham, Springer, 2023.

[3] A. Ohno, Could authors of academic reports be discerned using formatting information obtained by parsing XML of .docx documents?, *IEEJ Transactions on Electronics, Information and Systems*, vol.143, no.1, pp.91-100, 2023.
[4] I. T. Jolliffe and J. Cadima, Principal component analysis: A review and recent developments, *Philosophical Transactions A*, vol.374, no.2065, https://doi.org/10.1098/rsta.2015.0202, 2016.
[5] L. van der Maaten and G. Hinton, Visualizing data using t-SNE, *Journal of Machine Learning Research*, pp.2579-2605, 2008.
[6] J. Wang and Y. Dong, Measurement of text similarity: A survey, *Information*, vol.11, no.9, DOI: 10.3390/info11090421, 2020.
[7] *Turnitin*, https://www.turnitin.com/, Accessed on Dec. 19, 2023.
[8] X. Li, C. Yao, Q. Zhang and G. Zhang, Semantic similarity modeling based on multi-granularity interaction matching, *International Journal of Innovative Computing, Information and Control*, vol.15, no.5, pp.1685-1700, 2019.
[9] A. Ohno et al., Similarity measurement based on author's writing styles for academic report plagiarism detection, *IEEJ Transactions on Electronics, Information and Systems*, vol.140, no.2, pp.235-241, 2020.
[10] Open AI, *ChatGPT*, https://openai.com/product/chatgpt, Accessed on Dec. 19, 2023.
[11] R. Nakano, J. Hilton et al., *WebGPT: Browser-Assisted Question-Answering with Human Feedback*, https://arxiv.org/pdf/2112.09332v3.pdf, Accessed on Dec. 19, 2023.
[12] C. Rapp, T. Heilmann and O. Kruse, Beyond MS Word: Alternatives and developments, in *Digital Writing Technologies in Higher Education*, O. Kruise et al. (eds.), Cham, Springer, 2023.
[13] ECMA International, *ECMA-376 Office Open XML File Formats*, 5th Edition, https://ecma-international.org/publications-and-standards/standards/ecma-376/, Accessed on Dec. 19, 2023.
[14] G. E. Hinton and S. Roweis, Stochastic neighbor embedding, *Advances in Neural Information Processing Systems*, vol.15, pp.857-864, 2002.