Examination of Document Clustering Based on Independent Topic Analysis and Word Embeddings

Riku Yasutomi[1][a], Seiji Yamada[2][b] and Takashi Onoda[1][c]
[1]Aoyama Gakuin University School of Science and Engineering, 5-10-1 Huchinobe, Chuo, Sagamihara, Kanagawa, Japan
[2]National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan

Keywords: Topic Model, NLP, Independent Topic Analysis, ITA, Word Embedding, Word2vec, Machine Learning, Document Clustering, News Paper.

Abstract: In recent years, research on text mining, which aims to extract useful information from textual data, has been actively conducted. This paper focuses on document classification methods that extract topics from textual data and assign documents to the extracted topics. Among these methods, the most representative is Latent Dirichlet Allocation (LDA). However, it has been pointed out that LDA often extracts similar topics due to the high amount of shared information between topics. Therefore, this paper proposes a document classification method based on Independent Topic Analysis (ITA), which extracts topics based on the independence of topics, and on Word Embedding, which learn word co-occurrence. This approach aims to avoid extracting similar topics and to achieve information grouping that is closer to human intuition. As a comparative metric, we used the agreement rate between the results of manually classifying documents into topics and those classified by each method. The results of the comparative experiment showed that the agreement rate for document classification based on ITA and Word Embedding was the highest. From these results, it was suggested that the proposed method could achieve document classification closer to human perception.

1 INTRODUCTION

In recent years, the widespread use of mobile devices and laptops, as well as rapid advancements in information technology, have been remarkable. As a result, anyone can now easily disseminate large volumes of textual data. These data are being accumulated daily, and the amount of available textual data continues to increase. In fact, the document information management service market is on a steady path of growth. Ministry of Internal Affairs and Communications of Japan has projected an increase in the volume of digital data. The amount of digital data worldwide is expected to expand fourfold over the five-year period since 2011. (MIC, 2012). However, on the other hand, there is a limit to human information processing capacity. Therefore, research on text mining, which aims to extract useful information from such data, has been actively conducted (Shinobu, 2015).

In particular, various studies have been conducted on methods for extracting topics from document data and classifying documents, including hierarchical clustering techniques applied to word embeddings (Grootendorst, 2022) and methods utilizing word distributions. Grouping a large amount of information into topics that align with human intuition is highly sought after (Vivek, 2024).

Among these document classification methods, the most representative approach is Latent Dirichlet Allocation (LDA), proposed by Blei et al. (Blei, 2003). LDA is a distribution-based method, but many studies have pointed out the problem of extracting highly similar topics (Section 2.2).

Therefore, this paper focuses on Independent Topic Analysis (ITA), proposed by Shinohara (Shinohara, 1999), which performs topic extraction with high independence. However, since ITA is a topic extraction method primarily used for topic tracking, it must be adapted to document classification problems by leveraging word embeddings (Takahiro, 2020).

[a] https://orcid.org/0009-0008-2675-4982
[b] https://orcid.org/0000-0002-5907-7382
[c] https://orcid.org/0000-0002-5432-0646

Yasutomi, R., Yamada, S. and Onoda, T.
Examination of Document Clustering Based on Independent Topic Analysis and Word Embeddings.
DOI: 10.5220/0013104100003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 185-192
ISBN: 978-989-758-737-5; ISSN: 2184-433X

As a method for obtaining word embeddings, this paper employs Word2Vec, which is the simplest in structure and facilitates analysis. The objective of this study is to achieve document classification that aligns closely with human intuition by focusing on topic independence while utilizing word embeddings.

The remainder of this paper is organized as follows: Section 2 explains LDA as the baseline method and reviews related works. Section 3 presents the proposed method for document classification. Section 4 describes the evaluation experiments of the proposed method. Finally, Section 5 discusses comparisons with methods other than LDA, highlights the limitations of the proposed method, and outlines future challenges, concluding the paper.

2 RELATED RESEARCH

This section discusses the most common document classification method, LDA, its issues, and the objective of this study.

2.1 LDA (Latent Dirichlet Allocation)

In LDA, documents are represented as random mixtures of latent topics. The graphical model is shown in Fig 1. In this model, M represents the number of documents. N denotes the number of words. K stands for the number of topics. Each plate represents repeated structures. Furthermore, α is the parameter of the Dirichlet distribution, which serves as the prior distribution of the topic distribution θ_m for each document. And β is the parameter of the Dirichlet distribution, which serves as the prior distribution of the word distribution ϕ_k for each topic. $z_{m,n}$ denotes the topic assignment corresponding to the n-th word in document m, indicating from which topic each word was generated. $w_{m,n}$ represents the actual observed word in the n-th position of document m. Fig 1 illustrates the LDA process. The topic distribution per document and the word distribution per topic are controlled by the values of the Dirichlet parameters α and β. First, the latent topic z of a word is determined from the document-specific topic distribution θ through a multinomial distribution. Then, using the latent topic of the word and the word distribution ϕ for each topic, the observed word w (bag of words) is generated. In other words, LDA performs topic extraction and classification. The probabilistic generative model of words and documents forms the basis of LDA, allowing it to extract topics and classify documents effectively.

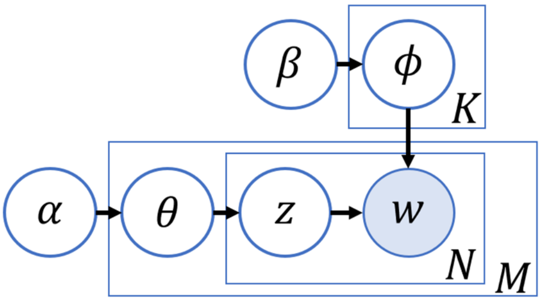


Figure 1: Graphical Model of LDA.

2.2 LDA Challenges

Many studies have pointed out the problem of extracting highly similar topics with LDA. Mu et al. (2024) discussed the limitations of traditional topic modeling methods like LDA, highlighting issues such as the extraction of similar topics and the difficulty in interpreting topics. As a result, it can become challenging to interpret the topics (Aletras and Stevenson, 2014), often leading to inaccurate topic labels during label assignment (Gillings and Hardie, 2023). To make the results interpretable by humans, post-processing of the model's output may sometimes be required (Vayansky and Kumar, 2020).

Due to these issues, there is a possibility that the classification results deviate from human intuition. Section 3 will provide an explanation of the proposed method, which focuses on ensuring independence between topics.

3 PROPOSED METHOD

In this section, we propose a document classification method that addresses the objective stated in Section 2.1 by utilizing ITA and Word2Vec. Section 3.1 explains ITA, Section 3.2 describes Word2Vec, and Section 3.3 provides an explanation of the document classification method that combines these approaches.

3.1 ITA

We introduce Independent Topic Analysis (ITA), proposed by Shinohara (Shinohara, 1999). As illustrated in Fig 2, ITA is a method for extracting topics with high independence. Highly independent topics are those that share a small amount of mutual information with each other. In other words, this method aims to avoid the extraction of similar topics and analyse the overall topic structure of documents. Let $t \in \{1, \dots, k\}$ denote the topic, $d \in \{1, \dots, n\}$ the document, and $w \in \{1, \dots, m\}$ the word, which are the common variables used in this method. ITA performs Singular Value Decomposition (SVD). On the word frequency matrix A of the documents to determine the significance of word w in topic t and the significance of document d in topic t. To evaluate the independence between topics, we use the kurtosis of higher-order statistics as a metric. Kurtosis is calculated as the difference between the fourth moment and a normal distribution with the same mean and variance. From this value, we calculate the concentration of words for the topics being used.

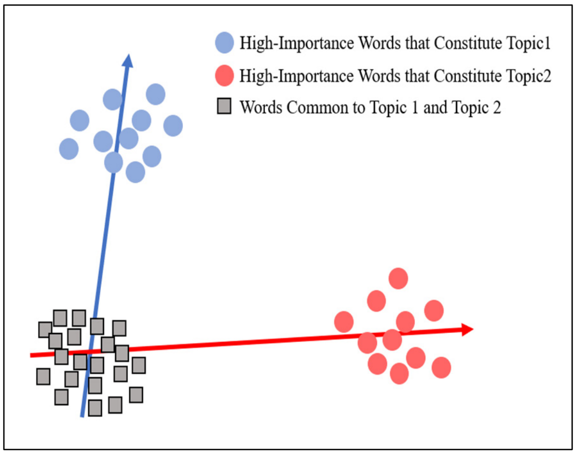

● High-Importance Words that Constitute Topic1
● High-Importance Words that Constitute Topic2
■ Words Common to Topic 1 and Topic 2

Figure 2: Image Diagram of Independent Topic Analysis (ITA).

ITA seeks to extract topics from document data such that the word concentration of topics is maximized and the independence between topics is maximized. Moreover, since ITA is a topic extraction method, it requires a document classification method to utilize these extracted topics. In Section 3.2, we explain Word2Vec, which is used for document classification.

3.2 Word2Vec

We classify documents based on the topics extracted using ITA by employing Word2Vec, proposed by Tomas et al. (Tomas, 2013). Word2Vec is a model used to learn vector space representations of vocabulary in natural language processing. In this paper, we utilize the Continuous Bag of Words (CBOW) model, as illustrated in Fig 3. CBOW is a neural network comprising an input layer, hidden layer, and output layer. For each of the C words in document k, a one-hot vector is formed and used as the input. The error between the predicted output and the correct labels (one-hot vectors) for the actual target words are computed using the cross-entropy loss function. The weights $\boldsymbol{w'}$ of the output layer are obtained through weight updates via backpropagation. Each column of these weights represents the distributed representation of each word. In other words, CBOW is a model that predicts the central word from its surrounding context. Consequently, this model captures the semantic similarity between words through co-occurrences, representing it as a low-dimensional continuous vector.

3.3 ITA+Word2Vec

As mentioned in Section 3.1, ITA uses the frequency of word occurrences as input. From this, ITA seeks to identify topics that maximize the concentration of words and the independence of each topic from the document data. Due to this characteristic, there is a possibility that rare words, which appear only in specific articles, are preferentially extracted. Additionally, the co-occurrence of words is not taken into account during the topic extraction process. Therefore, we want to classify rare words that are assigned high importance within each topic while considering co-occurring words. To achieve this, we will introduce word embeddings. This classification method will henceforth be referred to as the "Vector Classification." By adopting this approach, it becomes possible

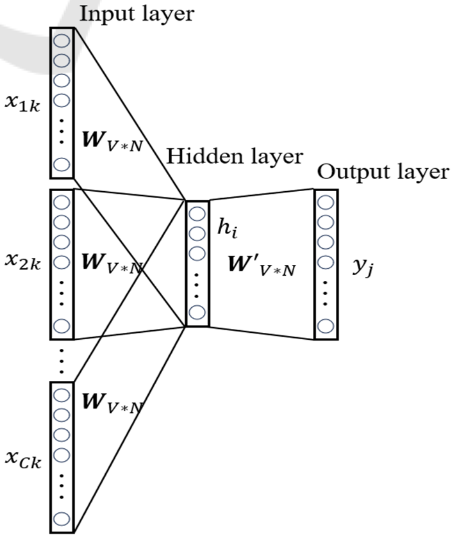


Figure 3: Image Diagram of ITA (Independent Topic Analysis).

to discover documents that include semantically similar words. Even if they don't contain rare, highly important words within the topics. The process of document classification using Vector Classification is illustrated below (Figure 4).

I. Perform the following process for topic $t \in \{1, \dots, k\}$.

(a) Convert all constituent words $w \in \{1, \dots, m\}$ of each topic into word embeddings.
(b) Convert the magnitude of each word embedding into the importance of word w in topic t.
(c) Add up all word embeddings for each topic to generate the word embedding $V_{topic,t}$ for topic t.

II. Perform the following process for document $d \in \{1, \dots, n\}$.

(a) Apply weighting to each word using TF-IDF.
(b) Similarly to step I. convert the constituent words into word embeddings, adjust their magnitudes, and sum them to generate the word embedding $V_{text,d}$ for document d.

III. Calculate the cosine similarity $\cos(V_{text,t}, V_{topic,t})$ between the $V_{text,t}$ of each document and the $V_{topic,t}$ of all topics.

IV. Assign the document to the topic with the highest similarity.

Through the aforementioned process, we classify documents into topics using the word embeddings obtained by Word2Vec. Word embeddings are learned based on word co-occurrences. ITA is a method that utilizes word occurrence frequencies and leverages relationships among topics. But it does not take word co-occurrences into consideration. Therefore, we incorporate word embeddings in the process of classifying articles into topics.

4 EVALUATION EXPERIMENT

In the evaluation experiments, we will compare the classification results obtained using LDA with the results from the proposed method, Vector Classification. We will also assess these results against human classification results. Based on these results, we will examine methods that are closer to human intuition.

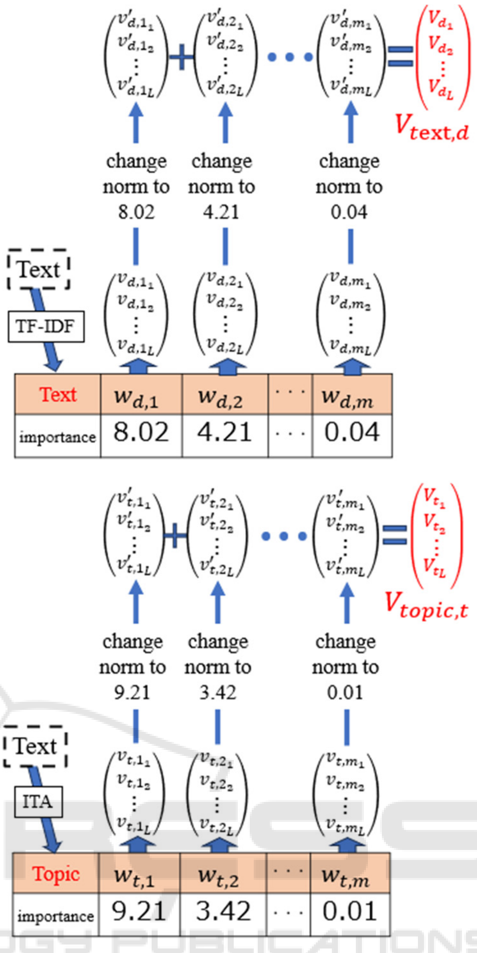


Figure 4: The method for converting documents and topics into word embeddings in vector classification.

4.1 Used Data and Evaluation Method

In the experiments, we utilize weekly article data from the Mainichi Shimbun in Japan for the year 2017 (The Mainichi, 2017). The data comprises articles from weeks 1, 2, 3, 4, 5, 10, 20, 30, 40, and 50. Initially, we extract and classify 13 topics for each week using LDA. Subsequently, we extract the same number of topics from the same data using Independent Topic Analysis. For these topics, we perform classifications using both a topic-article classification based on word importance (hereafter referred to as TF Classification) and the proposed method, Vector Classification.

TF Classification calculates the cosine similarity between the word importance derived from the topics extracted through ITA. And the word importance of each article, weighted using TF-IDF. The cosine similarity between vectors $\overrightarrow{x_1}$ and $\overrightarrow{x_2}$ is computed using the following formula.

$$\mathit{cos}(\overrightarrow{\boldsymbol{x}_1},\overrightarrow{\boldsymbol{x}_2}) = \frac{\overrightarrow{\boldsymbol{x}_1}\cdot\overrightarrow{\boldsymbol{x}_2}}{|\overrightarrow{\boldsymbol{x}_1}|\ |\overrightarrow{\boldsymbol{x}_2}|}$$

This represents the cosine of the angle between the two vectors, where a larger value indicates a greater similarity between the topic and the article. We calculate the cosine similarity for all topics and articles, assigning each article to the topic with the highest similarity. The number of extracted topics is set to 13. It is consistent with the number of article categories in the Mainichi Shimbun. Additionally, the parameters for LDA are set to $\alpha = 50/k$ and $\beta = 0.01$, following the research by Steyvers et al. (Steyvers, 2007) They are implemented using the Gensim library in Python. Parameter exploration is conducted based on coherence values to confirm the differences in results due to parameter settings. The parameters for Word2Vec were set to size=600 and window=10, based on the study by Oren et al. (Oren, 2016).

By comparing the classification results of ITA, LDA and human classification, we aim to investigate which method aligns more closely with human perception. The human classification involves the author manually assigning articles to topics extracted by

Table 1: The overlap of the top five words by probability for the 13 topics extracted using LDA.

topic t	probability				
	$w = 1$	$w = 2$	$w = 3$	$w = 4$	$w = 5$
1	scorers	Tokyo	Hyogo	Osaka	Chiba
2	cooperation	Saito	funeral	art museum	newspaper
3	cooperation	voluntary evacuation	Saito	funeral	newspaper
4	Japan	America	picture	world	Tokyo
5	cooperation	Saito	funeral	mother	newspaper
6	defense	Tanaka	opponent	champion	judgement
7	party chairman	constituency	incumbent	case	party
8	Kawasaki	Kashima	match	Urawa	Yokohama ma
9	cooperation	Saito	funeral	newspaper	art museum
10	joint research	suspect	bribe	arrest	university
11	Higashifukuoka	Toukaidai	final	opponent	victory
12	fertilized egg	transplant	consent	IVF	man
13	Aogakudai	Soudai	stage prize	victory	Jundai

Table 2: The overlap of the top five words by importance for the 13 topics extracted using ITA.

topic t	importance				
	$w = 1$	$w = 2$	$w = 3$	$w = 4$	$w = 5$
1	opponent	Toyko	Tanaka	champion	defense
2	man	woman	fertilized egg	transplant	IVF
3	China	America	Japan	Taiwan	Takahara
4	dissolution	prime minister	reporter	Suzuki	this year
5	Koike	laugh	decision	Tomin-first	political
6	Goyoukai	Nihon-buyo	everyone	recital	musical
7	Higashifukuoka	Toukaidai	final	Toin Gakuin	semifinal
8	abdication	emperor	citizens	coverage	constitution
9	Aogakudai	stage prize	victory	Soudai	outward way
10	America	company	influence	UK	production
11	party chairman	case	constituency	incumbent	party
12	Kashima	Kawasaki	Yokohama ma	Urawa	Iwata
13	Saitama	Hyogo	Kyoto	Chiba	Okayama

Table 3: Comparison of the matching rates between article classifications by each method and human classifications.

	average number of articles	ITA			LDA	
		The number of articles classified manually	Agreement rate with vector classification (%)	Agreement rate with TF classification (%)	The number of articles classified manually	Agreement rate with LDA (%)
1-5 weeks	1423.8	423.0	64.8	81.0	439.4	40.6
10-50 weeks※1	1680.0	505.8	69.9	83.8	518.2	43.8

Table 4: Comparison of the matching rates between the top 10 articles and topics for each method.

	number of articles	number of matched articles	matching rate (%)
Vector classification	129	98	76.0
LDA (coherence)	88	52	59.1

both LDA and ITA. During this process, only assign articles that are clearly determined to belong to each topic. Articles that are judged not to belong to any topic or those that are determined to belong to multiple topics are excluded. We had other experiment participants perform a similar assignment task beforehand. As a result, the agreement rate with the author's classification was 85.6%. Therefore, we confirmed in advance that there was no significant bias in classification accuracy.

From these results, we will compare LDA, a topic extraction and classification method that utilizes a probabilistic generative model, with ITA, which leverages the relationships between the extracted topics. Furthermore, we will compare the TF classification method that uses word importance with the Vector Classification method that employs distributed representations obtained from word co-occurrences.

4.2 Results

Table 1 presents the 13 topics extracted by LDA from the article data of the first week of 2017. The parameters used are those that yielded the best coherence values. Table 2 displays the topics extracted by ITA from the same data. As shown in Table 2, ITA assigns importance to all constituent words of each topic. Allowing the user of the method to infer meaning from the most important words of each topic, similar to LDA. LDA exhibits a high degree of commonality in words across topics, suggesting a substantial amount of mutual information. In contrast, ITA extracts topics with high independence, resulting in fewer common words among topics compared to LDA. Similar results were obtained with other datasets.

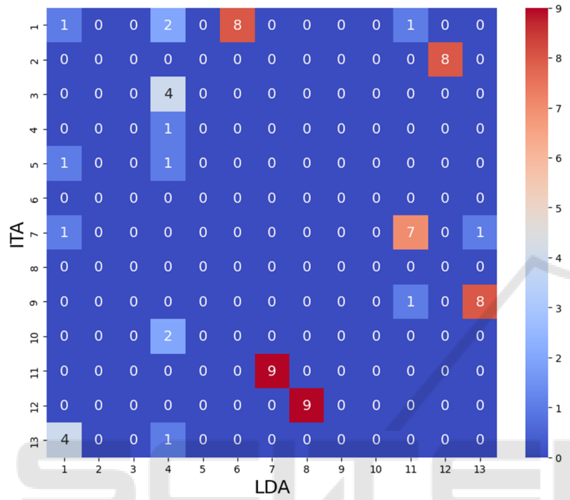
Next, Table 3 shows the average number of articles and the number of articles assigned to the 13 topics by human classification. And the average agreement rates between classifications by each method and human classification for the article data over one year. Specifically for weeks 1-5 (one week each) and weeks 10-50 (every ten weeks). The average number of articles assigned to each topic by human classifiers was 423.0 and 505.8 for ITA. While for LDA, the numbers were 439.4 and 518.2, indicating no significant difference. Furthermore, the agreement rates between articles assigned by each method and those assigned by human classifiers were 64.8% and 69.9% for TF Classification. 81.0% and 83.8% for Vector Classification, while LDA yielded 40.6% and 43.8%. Thus, Vector Classification achieved the highest agreement rates with human classification, whereas LDA yielded the lowest.

Table 4 presents the results of the agreement rates for the top 10 articles assigned to each topic when LDA parameters were determined based on coherence values. It also includes the results when using Vector Classification. In this case, the agreement rate for Vector Classification was again the highest. This indicates that the results did not significantly vary with different LDA parameters.

4.3 Discussion

From the results in Tables 1 and 2, it is evident that the amount of common information between topics is lower in ITA compared to LDA. However, it is also apparent that there are topics with overlapping content between the two methods. Figure 5 illustrates the number of matching top ten words between the 13 topics extracted by LDA and ITA. The top ten words of topics 6, 7, 8, 11, 12, and 13 extracted by LDA match over 70% with those of topics 1, 2, 7, 9, 11, and 12 in ITA. This suggests that the granularity of topics extracted by ITA is not significantly different from that of LDA. Simultaneously, it can be inferred that the issue of redundant topics with a high number of common words in LDA is mitigated. This allows for a more multifaceted understanding of the overall topic structure of the documents.

In the comparative experiment with human classification using newspaper articles, it was found that the results of ITA and the Vector Classification based on word2vec exhibited the highest agreement rates with human classification. This indicates the potential for achieving document classification that aligns more closely with human perception through ITA and document classification based on word embeddings. Moreover, the lack of significant differences in the average number of articles assigned to each topic by humans between LDA and Independent Topic Analysis implies that there is no substantial difference in the granularity of the extracted topics. This finding can also be understood from the comparison in Figure 5. Specifically, it appears that discrepancies in precision arise during the classification process of articles into topics beyond the common content identified by LDA and ITA. The number of articles classified by humans for each topic accounted for approximately 30% of the total articles. This characteristic of newspaper articles, which comprehensively documents events occurring each day, likely contributes to the difficulty in extracting large topics with a high number of assigned articles.



ITA \ LDA	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	0	0	2	0	8	0	0	0	0	1	0	0
2	0	0	0	0	0	0	0	0	0	0	0	8	0
3	0	0	0	4	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0	0
5	1	0	0	1	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	0	0	0	0	0	0	0	0	0	7	0	1
8	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	0	8
10	0	0	0	2	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	9	0	0	0	0	0	0
12	0	0	0	0	0	0	0	9	0	0	0	0	0
13	4	0	0	1	0	0	0	0	0	0	0	0	0

Figure 5: The number of matching top ten words between topics extracted by LDA and ITA.

Additionally, the agreement rate of TF Classification was lower compared to Vector Classification. ITA aims to minimize the common information among topics in its extraction method. Consequently, rare words that appear in a limited number of documents may be assigned a higher level of importance. However, when calculating the similarity between topics and documents using the TF Classification method, it may only be high for a few documents containing these important yet rare words. This can lead to classification results that diverge from human perception. Simultaneously, the similarity to articles containing important words within the topics consistently produces high values. This characteristic could explain why the results for agreement rates were higher than those of LDA.

While LDA estimates the agreement rates with each topic based on word distribution, the proposed method calculates the agreement rates between topics and the constituent words of articles. This is achieved using word embeddings derived from co-occurrence learning. Methods utilizing probabilistic generative models are known to achieve high precision in grouping documents. However, due to the abundance of common words among the extracted topics and high mutual information, a single document may correspond to multiple topics. Therefore, when evaluated based on proximity to human perception, it is plausible that precision may be lower.

Regarding the number of articles classified by humans, ITA showed little variation across weeks, while LDA exhibited variability in its values. A common characteristic attributed to LDA is the difficulty in determining optimal parameters. Since the optimal solution varies depending on the dataset and purpose, adjustments are required for each application. Consequently, experience and trial-and-error may be necessary. In this study, the parameters for LDA were determined based on prior research; however, the output of LDA is highly sensitive to its parameters. Therefore, changing the parameters could potentially alter the experimental results. The same applies to Word2Vec. In contrast, the results of ITA do not fluctuate significantly based on user input. It is believed that the minimal variability in the number of articles classified by humans is a result of this characteristic.

5 CONCLUSIONS

In this paper, we conducted a classification of newspaper articles into various topics using ITA, a method known for its high independence in topic extraction, along with word embeddings. We compared the classification results of our proposed method with those obtained through LDA, a widely used topic extraction method, as well as with results manually classified by humans. The findings revealed that our proposed method demonstrated a higher accuracy in terms of the agreement rate between the extracted topics and the assigned articles compared to existing methods. This suggests that our approach effectively avoids the extraction of topics characterized by a high number of common words, which is a noted issue with LDA, allowing for a more multifaceted representation of document topic structures.

The number of words in the topics extracted by ITA corresponds to the number of constituent words in the input documents, resulting in a high computational cost for calculating the similarity between topics and documents. Consequently, there are current limitations in applying ITA to larger-scale document datasets.

In this study, LDA, the most representative method for document classification, was chosen as a comparison benchmark. However, recent models such as BERTopic, which employs BERT and hierarchical clustering (Grootendorst, 2022), and LDA2vec, which integrates word embeddings prior to LDA application (Akihiro, 2019), have also been extensively studied. Future work must include comparisons with these state-of-the-art methods. Additionally, it is necessary to explore the use of word embeddings before applying ITA.

In the present experiment, Japanese newspaper articles were used as the dataset. Newspaper articles were chosen because they cover a wide variety of topics, making it challenging to categorize them into a small number of groups. As future work, we plan to explore the application of comparison methods to other datasets, such as review sites or documents maintained by companies, as well as to English documents.

REFERENCES

Ministry of Internal Affairs and Communications (MIC). (2012). Information and Communications White Paper, 2024 edition (in Japanese)

Shinobu, O. (2015). Technologies and Trends in Text Mining. Japanese Journal of Statistics and Data Science (JJSD). 28-1, pp.31–40 (in Japanese)

Grootendorst Maarten. (2022). BERTopic : Neural topic modeling with a class-based TF-IDF procdure. arXiv : 2203.05794

Vivek M., Mohit A., Rohit K. K. (2024). A comprehensive and analytical review of text clustering techniques. International Journal of Data Science and Analytics. 18:239–258

David M. Blei. Andrew Y. Ng. Michael I. Jordan. (2003). Latent Dirichlet Allocation. The Journal of Machine Learning Research. Vol.3, pp993-1022

Yasushi Shinohara. (1999). Independent Topic Analysis - Extraction of Distinctive Topics through Maximization of Independence. Communications Techniques. OFS99-14 (in Japanese)

Takahiro Nishigaki. Kenta Yamamoto. Takashi Onoda. (2020). Topic Tracking and Visualization Method using Independent Topic Analysis. AIRCC Publishing Corporation. Vol.10, No.3, pp. 1-18

Yida Mu. Chun Dong. Kalina Bontcheva. Xingyi Song. (2024). Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling

Nikolaos Aletras. Mark Stevenson. (2014). Labelling topics using unsupervised graph-based methods. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Vol.2, pp.631–636

Mathew Gillings. Andrew Hardie. (2023). The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice. Digital Scholarship in the Humanities, 38(2):530–543

Ike Vayansky. Sathish AP Kumar. (2020). A review of topic modeling methods. Information Systems, 94: 101582

Tomas Mikolov. Kai Chen. Greg Corrado. Jeffrey Dean. (2013). Efficient Estimation of Word Representations in Vector Space

Provider: The Mainichi - Japan Daily News. Released by: Nikkai Associates. (2017). CD - Mainichi Shimbun Data Collection

Mark Steyvers. Tom Griffiths. (2007). Probabilistic Topic Models. Lawrence Erbaum Associates

Oren Melamud. David McClosky. Siddharth Patwardhan. (2016). Mohit Bansal:The Role of Context Types and Dimensionality in Learning Word Embeddings. NAACL-HLT 2016. pp.1030–1040

Akihiro Ito. Rina Shirai. Hiroshi Uehara. (2019). Topic Extraction Using Similar Vocabulary with Word2Vec. IPSJ SIG Technical Report. Vol.2019-ITS-77, No.21 (in Japanese).