DOI: 10.6977/IJoSI.202412_8(4).0007
M. A. Yusof, S. Saee, /Int. J. Systematic Innovation, 8(4), 103-118 (2024)

Code switching: exploring perplexity and coherence metrics for optimizing topic models of historical documents

Muhammad Abdullah Yusof[1], Suhaila Saee[2]
[1&2] Faculty of Computer Science and Information Technology, University Malaysia Sarawak, Sarawak, Malaysia
abdunimas459@gmail.com1, ssuhaila@unimas.my[2]
(Received 4 March 2024; Final version received 30 September 2024; Accepted 1 October 2024)

Abstract

The Latent Dirichlet Allocation (LDA) model has two important hyperparameters that control the document-topic distribution known as alpha (α), and topic-word distribution known as beta (β). It is important to find the suitable values for both hyperparameters to achieve an accurate topic cluster. Using a single evaluation method to determine the optimal hyperparameters values is insufficient due to the size and complexity of the dataset. Thus, an experiment was conducted to study the relationship between the hyperparameters with perplexity, coherence scores and to establish a baseline for further topic modelling studies. It is the first study that focuses on multiple languages in Sarawak Gazette data for topic modelling. The study was conducted on LDA using Gensim package. The result shows that while perplexity scores were good indicator of the model's ability to predict new or hidden data, the word cluster within topic does not always reflect the similarity or relationships between words which compromised topic interpretation. The lowest perplexity score was observed when α was set to 5 and β to 0.4. The coherence evaluation indicated the optimal number of topics for each set of hyperparameter values although the relationship with hidden words remains unclear. The coherence score is highest when the number of topics was 5 and 4. In conclusion, the perplexity scores are effective indicators of word prediction accuracy for each hyperparameter setting. While coherence captures the optimal number of topics needed to produce high-coherence word cluster within a topic. Combining both evaluation methods ensures optimal results, producing topics that are both accurate and interpretable.

Keywords: Hyperparameter, Latent Dirichlet Allocation, perplexity, topic coherence, topic modelling

1.Introduction

Topic modelling is a method to find a group of words or topics from a collection of words. According to Mohr & Bogdanov (2013), topic modelling is considered one of the text analysis methods and is used widely across disciplines in academia. The method provides an automatic way to code the content of a corpus into meaningful categories or also known as 'topic'.

According to Steyvers & Griffiths (2009), topic modelling was first described and detailed by Papadimitrou etc. (1998) paper. Seymore & Rosenfeld (1997) paper was focused on topic detection and building the topic clusters based on the similarities of data from a dataset of 5,000 elemental topics.

Hyperparameters in LDA model are used to conFig the model and to specify the algorithm that minimizes the loss function (Yang & Shami, 2020). Hyperparameter cannot be estimated from the learning data. Instead, it is set before the data training as the hyperparameters defined the model of topic modelling. Several recent studies, such as those by Muhajir etc. (2022), Xue etc. (2020), and Dieng etc. (2023), have explored the impact of hyperparameter tuning on topic model performance, demonstrating that varying the values of α and β significantly affects coherence and perplexity scores. Hyperparameters adjustment and tuning play a critical role in clustering. Two of the most important hyperparameters for LDA clustering task are α which controls document-topic distribution and β which controls topic-word distribution. Odden et. al. (2020) stating the need to find a stable hyperparameter of α to achieve accurate topics from an LDA model.

One of the topic modelling methods is Latent Dirichlet Allocation (LDA). Latent Dirichlet Allocation (LDA) was started back in 2003 from a study conducted by Blei et. al. (2003) that applied the method for the machine learning (Péladeau & Davoodi, 2018). It is one of the Bayesian Network methods and one of

the well-established methods for natural language processing (Teh et. al., 2006).

The problem that arises from the optimization of hyperparameters is the evaluation method to validate and qualify the output. A single evaluation method is not able to give an adequate explanation of the quality of the results. Perplexity and coherence are two common evaluation metrics used in topic modeling. Perplexity measures the model's ability to predict unseen or hidden data, while coherence assesses how interpretable the generated topics are to human readers. Previous research, including studies by Hassan etc. (2021), Pinto etc. (2021), and Wallach etc. (2009), has shown that relying solely on one evaluation method, such as perplexity, can negatively affect model interpretability, highlighting the importance of combining both perplexity and coherence scores for a more accurate assessment of topic model quality. Furthermore, Ding et. al. (2018) emphasized that relying on a single evaluation method will negatively affect the result. The paper explained the score of perplexity alone does not necessarily reflect the interpretability of the results which indicates the model scored higher in perplexity meanwhile the coherence of the model is low. Wallach et. al. (2009) also stressing the importance of hyperparameters to achieve performance gains.

Hence, this study will investigate the relation between hyperparameters values (α [Dirichlet prior parameter of per document-topic distribution], β [Dirichlet prior parameter of per topic-word distribution]), with the perplexity and coherence scores. The study is focusing on the process of topic modelling by using LDA model on a data that was extracted from historical documents. While previous studies have focused on applying LDA to social media and scientific datasets, there is a significant gap in exploring historical documents, especially those containing multiple languages. This study addresses that gap by applying LDA to the Sarawak Gazette, a code-switched historical dataset, to evaluate the impact of hyperparameter tuning on topic model performance.

2.Related Works

One of the key outputs of topic modeling is the identification of topic clusters, which group similar words based on their relationships. In the early years of topic modeling, approaches like n-grams and k-values were crucial for ensuring accurate topic clusters (Seymore & Rosenfeld, 1997). In LDA, the α and β hyperparameters are important as these values determine the LDA model algorithm and processes.

Muhajir etc. (2022) experimented with tuned-LDA by adjusting hyperparameters to achieve optimal clustering. The study compared these tuned hyperparameters against several other algorithms and found that tuned-LDA outperformed the rest, although no neural network algorithms were tested. The tuned-LDA outscored the rest of algorithms although no neural network algorithm is used in the experiment.

Several studies have investigated how the hyperparameters α and β influence both the coherence (which measures topic interpretability) and perplexity (which measures model accuracy in predicting unseen data) of LDA models. For example, Xue et. al. (2020) found that different values of α and β can significantly affect the coherence and perplexity scores of LDA models applied to Twitter data. Similarly, Dieng et. al. (2023) proposed a topic modeling approach that incorporates word embeddings and showed that the choice of hyperparameters can impact the coherence of the resulting topics.

Panichella (2021) also confirms the correlation between the hyperparameters α and β and the accuracy of the output of LDA. The paper detailed a study that manipulates hyperparameters by tuning it and tested on Gibbs iteration of the process. The study found that the result improved as both hyperparameters were set to 0 against the default value of 0.1 for both hyperparameters.

Gertis (2021) found out that the hyperparameters values are important to achieve a desirable and interpretable output. The paper gives two values for each of the α and β output. Wallach et. al. (2009) has a conclusion of the importance and effect of the hyperparameters. The study found that the asymmetric Direchlet prior over document-topic distribution is better than the symmetric prior while asymmetric Direchlet prior over topic-word distribution is insignificant compared to the symmetric prior.

Hassan etc. (2021) introduced a new method for determining the optimal number of topics in LDA models. They proposed using Normalized Absolute Coherence (NAC) and Normalized Absolute Perplexity (NAP) to balance coherence and perplexity, resulting in improved model interpretability and accuracy.

Pinto et. al. (2021) utilized both coherence and perplexity score to obtain optimal data and number of topics. The study found out that perplexity mathematical calculation is simpler than coherence although the quality of the result is inferior compared to the

coherence. Perplexity result indicates the best scores but the result that graded as having a good perplexity value does not reflect the quality of the output in form of word clusters. Newman et. al. (2011) used perplexity and Pointwise Mutual Information (PMI) score to manipulate the number of topics. The study utilized both measurements to achieve an optimal coherence based on the number of topics which dictates by perplexity and PMI.

Watanabe & Baturo (2023) utilized the hyperparameter to smooth out the topic clusters. The topic modelling process was repeated with the increment of hyperparameters values for each process to investigate the reaction between the likelihood of the next sample or output with the hyperparameters value. The paper found out that repetition also improves the inference of the unknown variable.

Zhou et. al. (2023) also focusing on the relation between perplexity and coherence score with the hyperparameters. The study evaluated the LDA model and the hyperparameter values by analyzing the unigram and bigram topic results. However, the study only focused on the hyperparameter that controls the number of topic while the other hyperparameters including the α and β were ignored.

Furthermore, the evaluation of LDA models using perplexity and coherence scores has been applied in various domains. Agarwal et. al. (2020) used LDA to mine issues on Twitter during the COVID-19 pandemic and evaluated the quality of the generated topics using coherence scores. Griffiths & Steyvers (2004) introduced LDA and presented a Markov chain Monte Carlo algorithm for inference in the model, demonstrating its operation on a small dataset.

Based on Fig 1 by Lee etc. (2018), LDA vector space, M represents the total number of documents in the corpus, and N represents the number of words per document. Each word in a document (W) is assigned to a latent topic (Z), forming a topic-word distribution (φ) and a document-topic distribution (θ). The hyperparameters α control the distribution of topics per document, while β controls the distribution of words per topic. Within LDA architecture, every word (W) that exists in a document corresponds or is related to a latent topic (Z), which gives a topic-word distribution in the corpus (θ and φ).

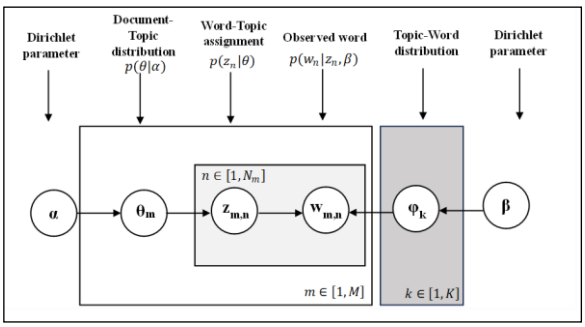

Fig 1. The vector space of LDA.

In the experiment, the dataset and the LDA model were treated as constant variables, while the hyperparameter values were manipulated. The resulting variables included the topics, coherence scores, and perplexity scores.

In conclusion, these studies emphasize the importance of tuning hyperparameters such as α and β to optimize topic clusters and balance coherence and perplexity. However, most of these studies focus on datasets like social media or scientific texts, leaving a gap in applying LDA to code-switched historical datasets, which this study aims to address. Manipulating the value of hyperparameters α and β were important in these papers as the values are determining the process of clustering in the LDA model and crucial step to obtain the appropriate number of topics except Zhou et. al. (2023) which focused on K hyperparameter that determines the number of topics.

The choice of hyperparameters α and β can impact the coherence and perplexity of the resulting topic models. These evaluation metrics have been applied in various domains, including social media analysis during the COVID-19 pandemic and scientific topic discovery. Researchers have explored the correlation between hyperparameters and model performance, highlighting the importance of selecting appropriate values for α and β to achieve coherent and interpretable topics. However, there is lack of utilization of historical data as a dataset in these studies and the process of detailing such dataset is not detailed.

3. Methodology

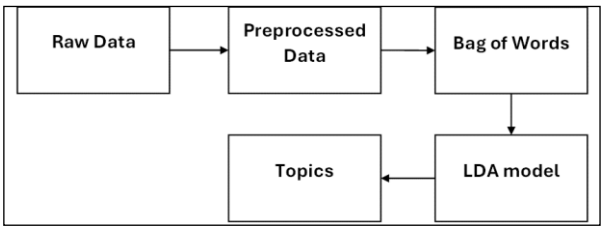

Fig 2. The workflow of LDA for baseline study.

Fig 2 shows the workflow of LDA for the study. The dataset for the experiment consists of 193 Sarawak Gazette documents, spanning the years 1907 to 1935, with a total of 2,696,635 words. The process begins by acquiring pdf documents from the Sarawak Gazette portal. These documents were scanned documents of the historical Sarawak Gazette which each page is treated as an image.

After acquisition, the documents were processed using an optical character recognition (OCR) tool. Due to the aging and damage of the Sarawak Gazette pages, the OCR results required extensive manual correction to handle errors.

All processed Sarawak Gazette documents were combined into a single corpus. The corpus was preprocessed by removing special characters, stop words, single-character words, punctuation, and numbers, and the text was converted to lowercase.

The study adapted the workflow of Lee etc. (2018) on vector space. The input data was fed into the model. The α and β hyperparameters were then distribute the topics for documents and words for topics. The manipulation of these hyperparameters was done to observe the influences and impacts of these manipulations towards the quality of word clusters for each topic and the coherence of the word clusters.

Removing single and double-letter words improved the model's accuracy. This step was necessary because the aging and deterioration of the Sarawak Gazette pages caused OCR errors, misinterpreting some defects as letters or symbols. By removing a single or double letters words and special characters, the number of errored words is able to be filtered out from the dataset. Then, the data was stored inside a bag of words and further divided into two sets of data for training and testing.

The experiment was run by using LDA model from the GENSIM package. The hyperparameters value was set during this stage. The hyperparameters alpha (α) and beta (β) were incrementally adjusted for each process, while the K value (number of topics) remained constant at its default value of 5 throughout the experiment. Each combination of hyperparameters resulted in different word clusters, perplexity scores, and coherence scores, as the distribution of document-topic and topic-word relations changed.

Then, the model was trained with the training data. After the training phase, the model was once again run with the test data. The K hyperparameter determines the number of topics produced by the model. The model iterated over the documents and randomly assigned words to topics. Next, it updated the proportion of documents assigned to each topic based on the words. This process was repeated for each word, allowing the clusters to be rearranged and reassigned with each iteration. The process is repeated, each time the hyperparameters were changed. The output of the model was recorded and saved. There are four outputs that were crucial for further analysis which includes a list of topics and its word clusters, word clouds, perplexity, and coherence score.

4. Results and Analysis

This section is split into perplexity score, word clouds, coherence score, and manual evaluation. Perplexity score measures the ability of a model to handle the hidden data. Word cloud demonstrates the word clusters per topic and coherence score measures the coherence between the words within the cluster. Manual evaluation is done by human validating the result of the experiments.

4.1 Perplexity Score

The experiment was observing several expected outputs which are the perplexity score, coherence score, and word cloud. The perplexity scores are beneficial to predict the ability of the model to deal with the hidden data while coherence scores reflect the coherent result of the model. Word clouds are important as a reference that reflects the interpretability of the result with the perplexity and coherence scores.

Table 1 shows the increase of hyperparameters' values causing the perplexity scores progressively larger. The results showed that the perplexity is at its best when $\alpha = 5$, $\beta = 0.4$ with a value of -9.99. The coherence score showed the highest score when the number of topics is set at 3.

Table 1. The perplexity score for hyperparameters.

Hyperparameters values	Perplexity scores
$\alpha = 5, \beta = 0.4$	-9.99
$\alpha = 6, \beta = 0.5$	-9.74
$\alpha = 7, \beta = 0.6$	-9.57
$\alpha = 8, \beta = 0.7$	-9.43
$\alpha = 9, \beta = 0.8$	-9.31
$\alpha = 10, \beta = 0.9$	-9.23

Based on Table 1, the word cloud is adequate and interpretable compared to other results. The higher

perplexity score recorded for α = 10, β = 0.9. The mean value for perplexity score is -9.545.

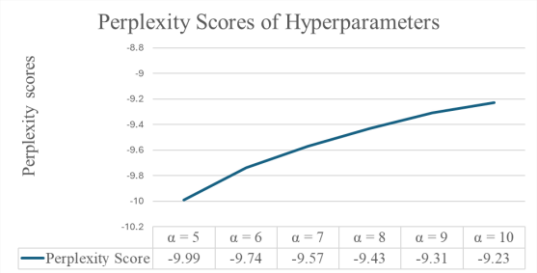

Fig 3. The perplexity graph.

The result in Fig3 shows that, as the values of hyperparameters increased, the perplexity score grew progressively. The perplexity for α = 5 is -9.99 while the perplexity scores for α = 10 is -9.23. The result signifies the findings as the higher the values of hyperparameters, the higher the values of perplexity which is the higher the inaccuracies.

4.2 Word Clouds

The word cluster are visualized by using word cloud. The larger size words signify high occurrence within the text and its relevancy to the topic and the smaller size words signify lower occurrence. Based on the interpretation of Fig4, Topic 0 focusing on exhibition about Borneo; Topic 1 focusing on report of district affairs from second, third, fourth, and fifth divisions (encompassing every division in Sarawak during the Brooke's administration except the first division that encompassing an area of today's Kuching, Samarahan, and Serian divisions). Below are the word clouds for the outputs of α = 5, β = 0.4:

Fig 4. The word clusters for each topic when α = 5, β = 0.4

In **Fig4**, Topic 2 focusing on shipping and mining industry in Brooketon (now known as Muara Coal Mine in Brunei); Topic 3 focusing on regatta that was held annually in Sarawak; and Topic 4 focusing on central government policy and activity of central government in Kuching.

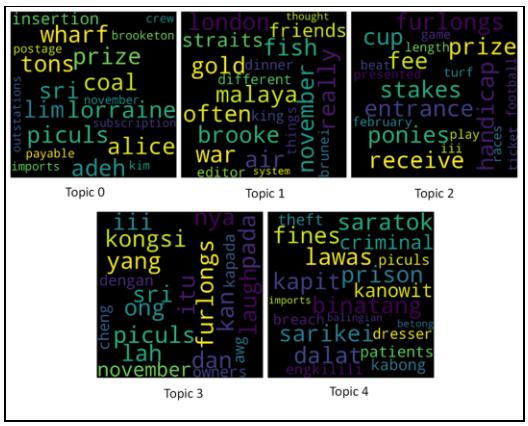

Fig 5. The word clusters for each topic when α = 6, β = 0.5

Fig 5. shows the word clouds for α = 6, β = 0.5. Topic 0 focusing on economic activities such as shipping and mining; Topic 1 focusing on the political relation of Sarawak with the outside world as well as Brooke's relation with the natives; Topic 2 focusing on the horse racing sports in Sarawak; Topic 3 contains random words and phrases; and Topic 4 focusing on reports from other divisions except the first division.

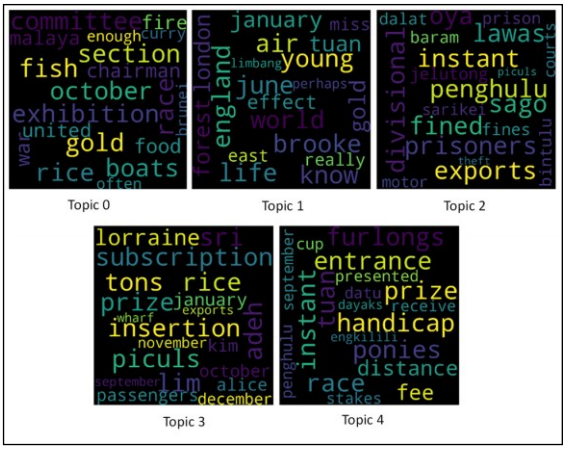

Fig 6. The word clusters for each topic when α = 7, β = 0.6

Fig 6. shows the word clouds for α = 7, β = 0.6. Topic 0 focusing on events happen in the October; Topic 1 focusing on report of events occur in the January and June; Topic 2 focusing on criminal and other problem from second, fourth, and fifth divisions; Topic 3 focusing on economic activities; and Topic 4 focusing on horse racing sports.

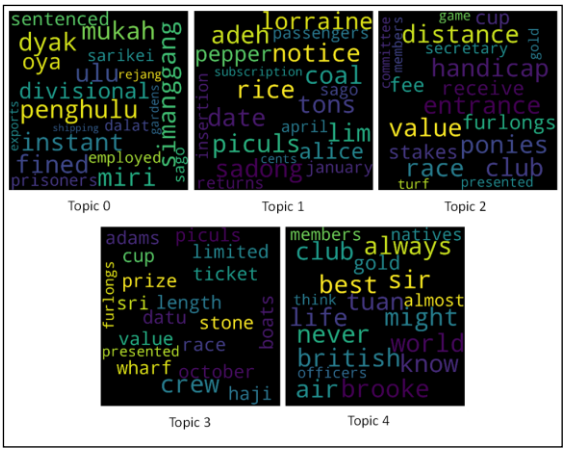

Fig 7. The word clusters for each topic when α = 8, β = 0.7

Fig 7. above are the word clouds for α = 8, β = 0.7. Topic 0 focusing on reports from other divisions except the first division; Topic 1 focusing on economic activities of the state; Topic 2 focusing on horse racing sport activities; Topic 3 focusing on regatta events; and Topic 4 focusing on the colonial life during the Brooke's era.

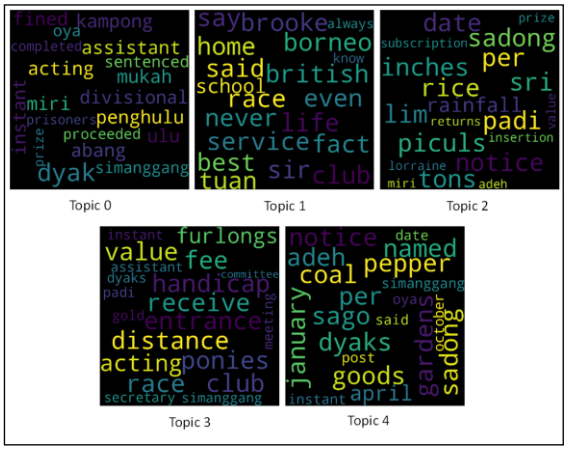

Fig 8. The word clusters for each topic when α = 9, β = 0.8

Fig 8. above are the word clouds for α = 9, β = 0.8. Topic 0 focusing on reports from other divisions except the first division; Topic 1 focusing on a life during colonial era; Topic 2 focusing on rice production; Topic 3 focusing on horse racing; and Topic 4 focusing on the economic activities.

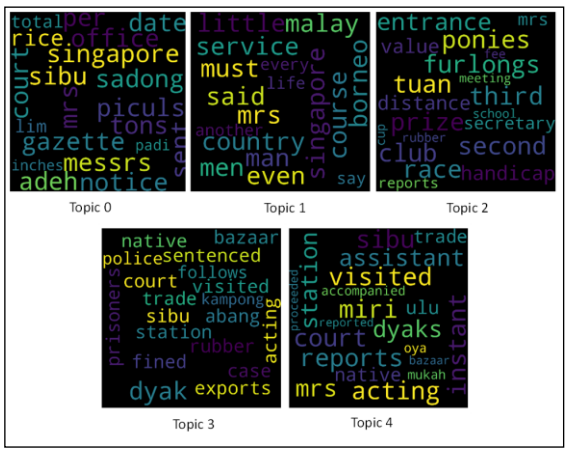

Fig 9. The word clusters for each topic when α = 10, β = 0.9

Fig 9. above are the word clouds for α = 10, β = 0.9. Topic 0 focusing on economic activities; Topic 1 several unrelated words and phrases; Topic 2 focusing on horse racing sport activities; Topic 3 focusing on report about Sibu; and Topic 4 focusing on the report of third and fourth division.

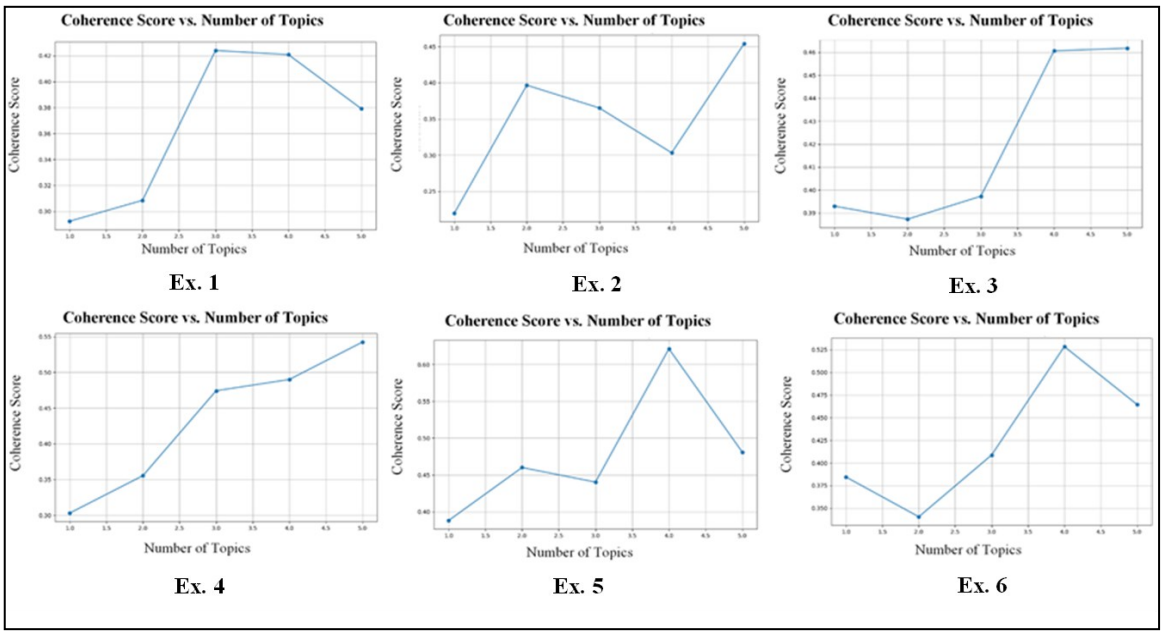

Fig 10. Coherence score for each of the experiment

4.3 Coherence Score

Table 2 shows the coherence result of the experiment. The highest coherence value is 0.621 when α value is set to 9 and β to 0.8. The coherence values across number of topics when α value is set to 9 and β to 0.8 are consistently high with one fall below 0.300

Table 2. The coherence values for every hyperparameter values

Experiment	α	β	Topics				
			1	2	3	4	5
Ex. 1	5	0.4	0.292	0.308	0.424	0.421	0.379
Ex. 2	6	0.5	**0.220**	0.397	0.365	0.303	0.454
Ex. 3	7	0.6	0.393	0.387	0.397	0.461	0.462
Ex. 4	8	0.7	0.303	0.355	0.474	0.490	0.542
Ex. 5	9	0.8	0.388	0.460	0.440	**0.621**	0.481
Ex. 6	10	0.9	0.384	0.340	0.409	0.529	0.464

Table 2 shows lowest coherence score is 0.220 α value is set to 6 and β to 0.5 the coherence values across number of topics when α value is set to 6 and β to 0.5 is consistently low with the highest value within the range is 0.454 which fall when the number of topics is 5, the second lowest after α = 5 and β to 0.4.

The best number of topics when the hyperparameters were set to α = 5 and β = 0.4 is 3 topics. The graph shows improvement except for the slight decrease when the topic is 5.

The best number of topics when the hyperparameters were set to α = 6 and β = 0.5 is 5 topics. The graph shows improvement with a decrease in value at topic 3 and 4.

The best number of topics when the hyperparameters were set to α = 7 and β = 0.6 is 5 topics. The graph shows improvement although the values are relatively consistent at topic 4 and 5. The best number of topics when the hyperparameters were set to α = 8 and β = 0.7 is 5 topics. There is consistent improvement in values across topics. The best number of topics when the hyperparameters were set to α = 9 and β = 0.8 is 4 topics. There is a sudden drop of value at topic 5. The best number of topics when the hyperparameters were set to α = 10 and β = 0.9 is 4 topics. The value drops at the beginning before increasing and finally a slight decrease at the end.

4.4 Manual Evaluation

Manual evaluation is conducted involving human validator. Human validators are asked to score the coherence of the word cluster for each topic based on a scale of 0 to 5 with 0 signify a very low coherence between words in the cluster to 5 as the most agreeable cluster. Human validators are also asked for the appropriate label for each topic.

Table 3. Topic coherence rate by experiment

Experiment	Hyperparameters	Average Coherence Rate
Ex. 1	$\alpha = 5, \beta = 0.4$	3.36
Ex. 2	$\alpha = 6, \beta = 0.5$	3.16
Ex. 3	$\alpha = 7, \beta = 0.6$	2.92
Ex. 4	$\alpha = 8, \beta = 0.7$	3.72
Ex. 5	$\alpha = 9, \beta = 0.8$	3.60
Ex. 6	$\alpha = 10, \beta = 0.9$	3.08

Table 3 shows the result of manual evaluation. Manual evaluation by expert has been done to validify and interpret the results of word clusters. The validators were asked to rate each of the topics for each experiment from the scale of one to five with one signifying a severe lack of coherence within the word clusters of a topic and five as highly agreeable to the word clusters. The validators are also asked to state their interpretation of a topic based on the word clusters

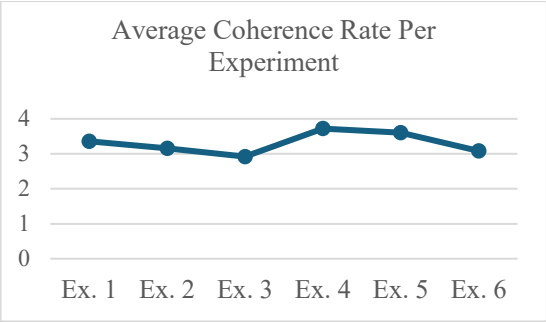

Fig 12. Average coherence rate per experiment

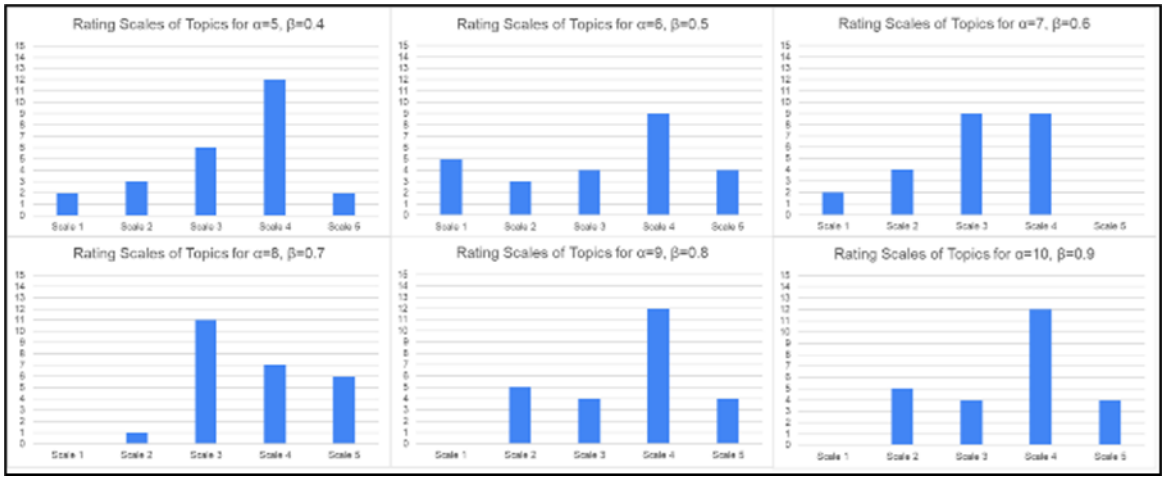

Fig 11. Rating scales for each of the experiments

Fig 11 shows that no validators have used Scale 5 to rate the results of the experiment for $\alpha = 7$, and $\beta = 0.6$ while **Fig 12** shows the average coherence score for every experiment. The highest average coherence rate for the experiments is 3.72 when $\alpha = 8$, and $\beta = 0.7$. Based on the following Fig of the rating scales for each of the experiment, the highest average coherence rate is caused by the higher occurrences of Scale 5 being used to rate the results of that experiment. The lowest coherence rate is when $\alpha = 7$, and $\beta = 0.6$ at 2.92.

Table 4. Topic coherence rate per topic for each experiment

Experi-ments	α	β	Average Coherence Rate of Topics				
			0	1	2	3	4
Ex. 1	5	0.4	2.80	4.00	2.60	3.00	**4.40**
Ex. 2	6	0.5	2.80	3.60	3.60	**1.80**	4.00
Ex. 3	7	0.6	3.00	3.20	3.40	2.80	2.20
Ex. 4	8	0.7	3.60	3.80	4.00	3.20	4.00
Ex. 5	9	0.8	3.40	3.40	4.20	3.20	3.80
Ex. 6	10	0.9	3.00	2.80	3.20	3.80	2.60

Table 4 shows the average coherence rate per topic by each of the experiment. Topic 4 of experiment 1 has the highest average of coherence rate at 4.40. Further investigation found that the validators have chosen only Scale 4 and scale 5 to rate the result of Topic 4 in Experiment 1, hence explaining the highest average. The lowest average is Topic 3 in Experiment 2 at 1.80. The validators have rated the result of this experiment with the lower scales arranging from Scale 1 to Scale 3. The lower scales given has caused the average to fall below 2.00.

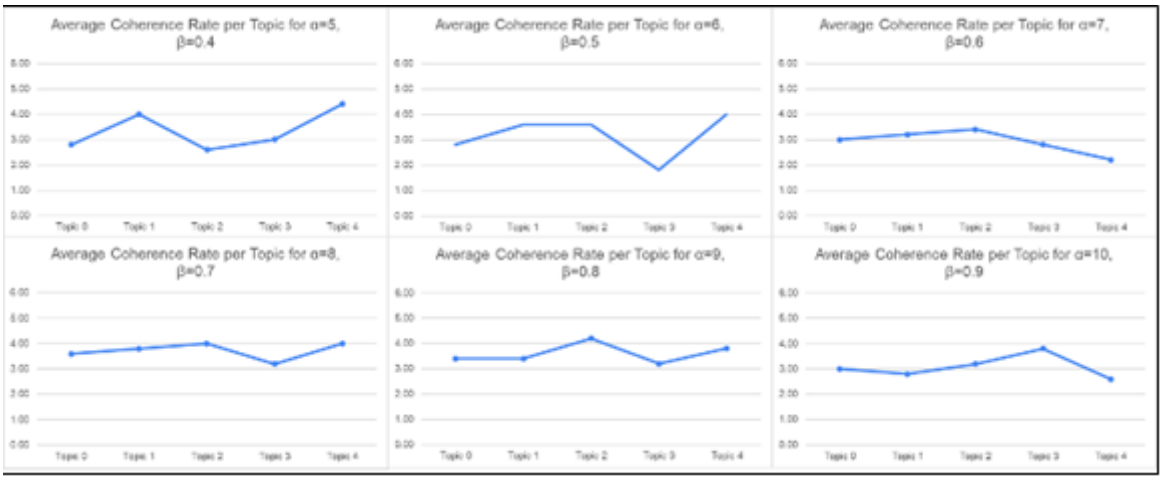

Fig 13. Average coherence rate per topic for every experiments

Fig 13 shows the average coherence rate per topic for every experiment. Experiment 3 has the lowest overall coherence rate and shows a slow progress before falling in average coherence rate at Topic 3 and Topic 4. The highest overall average coherence rate is Experiment 4. The average coherence rate for all topics is stable and only has a slight dip in Topic 3 before rising up again at Topic 4.

Table 5. Topic coherence rate per topic across all experiments

Topics	Average Coherence Rate
Topic 0	3.10
Topic 1	3.47
Topic 2	3.50
Topic 3	2.97
Topic 4	3.50

The following **Table 5** shows the average coherence rate per topic across all experiments. Topic with the highest coherence rate is Topic 2 and Topic 4 with an average of 3.50. **Fig 14** shows that Topic 4 has accumulated the highest number of Scale 4. Topic with the lowest average coherence rate is Topic 3 at 2.97. Topic 3 has a lower Scale 3 accumulation and has an almost accumulation size of Scale 2 and 3.

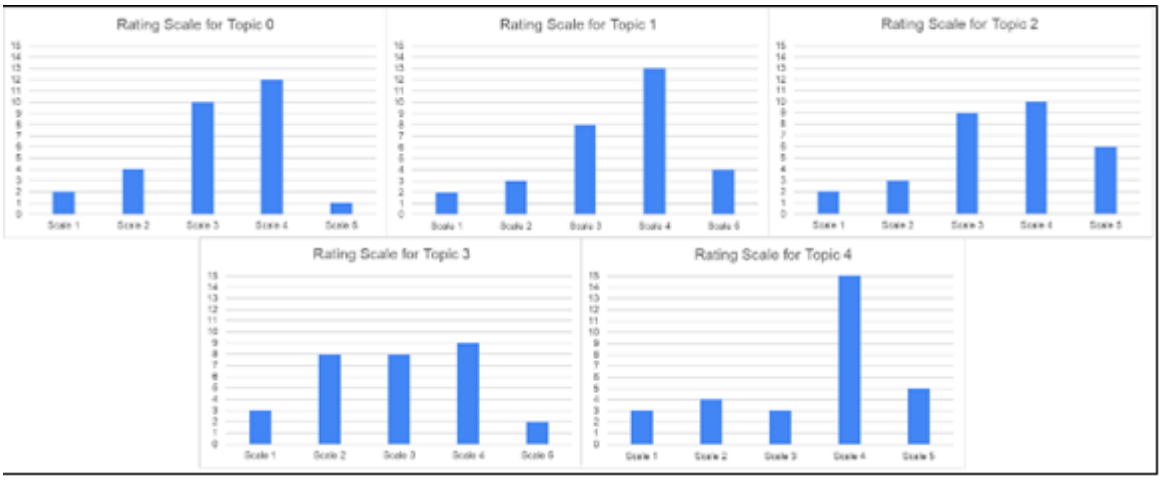

Fig 14. Rating scales for each of the topic

The graph shows the pattern of average coherence score for each topic across all experiments. The graph shows a rise from Topic 0 to Topic 2. A sudden dip happens in Topic 3 before a sudden rise at Topic 4. The average coherence rate for the entire topics is 3.31.

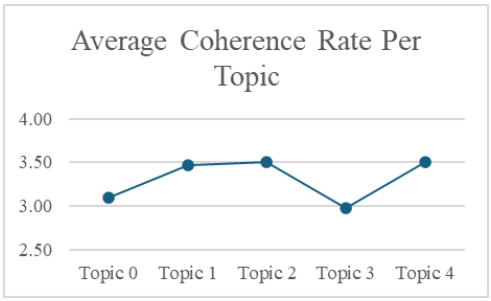

Fig 15. Average coherence rate per topic across all experiments

The accumulation of scale used by the validators shows that the highest Scale 5 was used when rating the results of Topic 2 and the lowest is Topic 0 as shown in **Fig 15** The application of Scale 5 shows a drastic influence on the average coherence rate as shown in Topic 2 and Topic 4, both of which accumulate a higher proportion of Scale 5 compared to other topics. The smaller scales such as Scale 1 and Scale 2 dragged the average coherence rate. Topic 3 suffered from a high accumulation of lower scales and lack of Scale 5. Topic 1 also has an almost similar scale accumulation except for the high accumulation of Scale 4 which slightly increase the average coherence rate compared to Topic 3.

Based on the average coherence rate, Topic 1 has the highest rate at 3.39. Topic 1 has a high application of Scale 3, Scale 4 and 2 applications of Scale 5. The second highest is Topic 4 with a rate of 3.28. Topic 4 has a high application of Scale 4 which influenced the average coherence rate to be higher than other topics except Topic 1. Topic 3 is the lowest in terms of average coherence rate, only at 2.56. Topic 3 also has a high application of Scale 2 which is at 8 applications.

The topic labels have been given by the validator after rating the clusters. Although the topics labelled differently from one validator to another, there are consistent pattern within these labels. For example, based on Topic 1 in the first experiment when α=5 and β=0.4, the validators gave several labels such as Sarikei-Mukah, places, and places in Sarawak. The labels given clearly state something that is related to locations or places. However, there are certain topics that have mixed unrelated labels such as Topic 0 in the experiment that involves α=8 and β=0.7, the validators gave labels such as exports, social justice, and unidentified. The answers for this demonstrate the varied interpretation of the word cluster which is caused by the lack of coherence between words in the cluster.

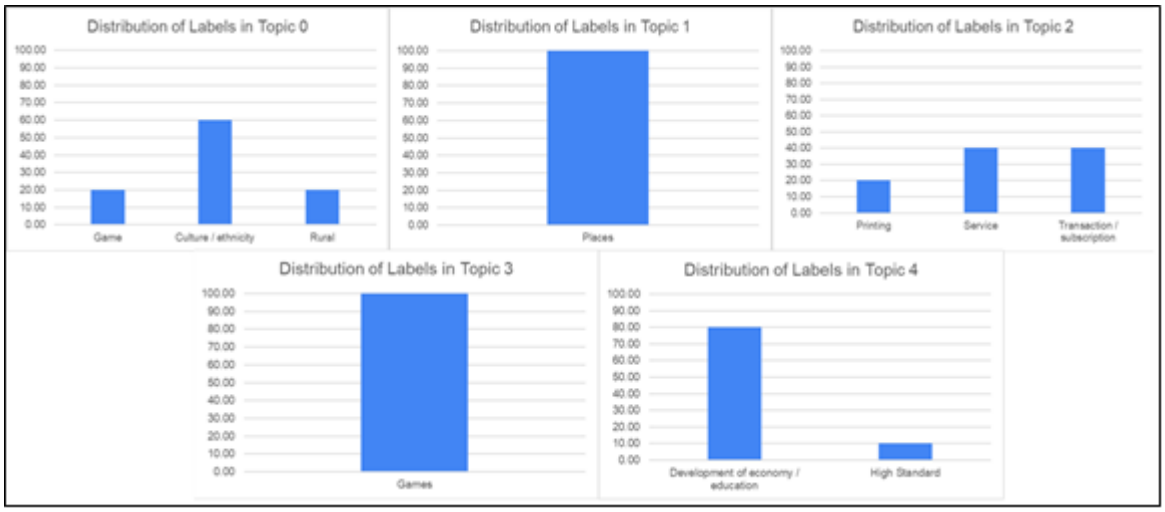

Fig 16. Distribution of labels in topics of Experiment 1.

Fig 16 shows the distribution of labels for Experiment 1 with 60% of the evaluators label Topic 0 as culture and ethnic related terms while 20% label the topic as game and another 20% as rural. For Topic 1, all evaluators associate it as places. Topic 2 of Experiment 1 is classified as printing by 20% of the evaluators and 40% each as service and transaction. Topic 3 is unanimously categorized as games or sports. The last topic of Experiment 1 is associated with development of economy or education by 80% of the evaluators and the remaining 20% label the topic as high standard.

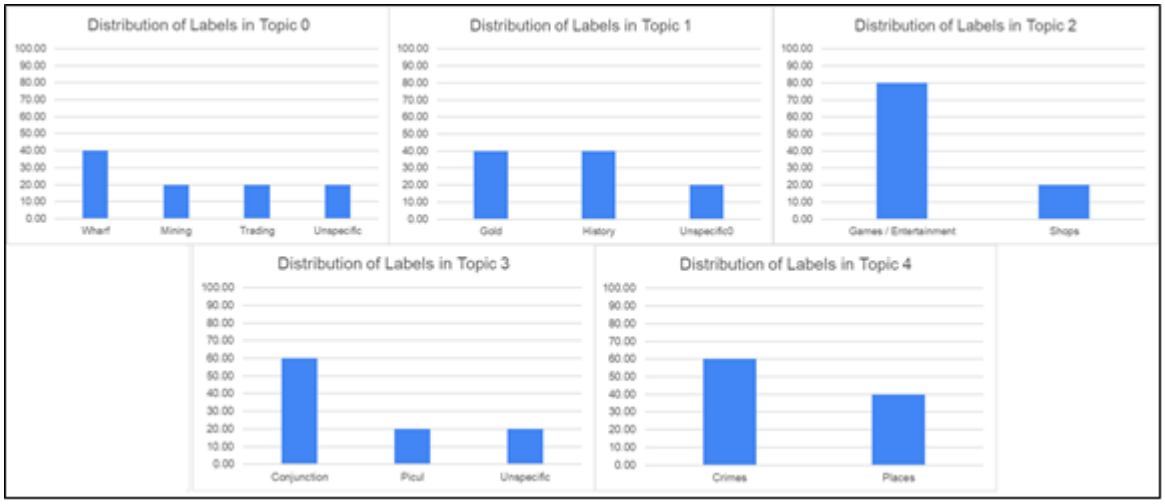

Fig 17. Distribution of labels in topics of Experiment 2.

As shown in **Fig 17**, Topic 0 of Experiment 2 shows 40% of the evaluators label the topic as wharf and 10% each labelling it as mining town and trading. Another 10% classify the topic as unspecific. For Topic 1, 40% of the evaluators classify the topic as gold and another 40% classify it as history or politic while 20% classify the topic as unspecific. Topic 2 is associated with games or entertainment by 80% of the evaluators and 20% as shops. Topic 3 of Experiment 2 is classified as unspecific by 60% of the evaluators and 20% each as conjunction or measurement unit picul, indicating the ambiguity of the word cluster. The last topic of Experiment 2 is classified as places by 40% of the evaluators and the remaining 60% classify it as crimes.

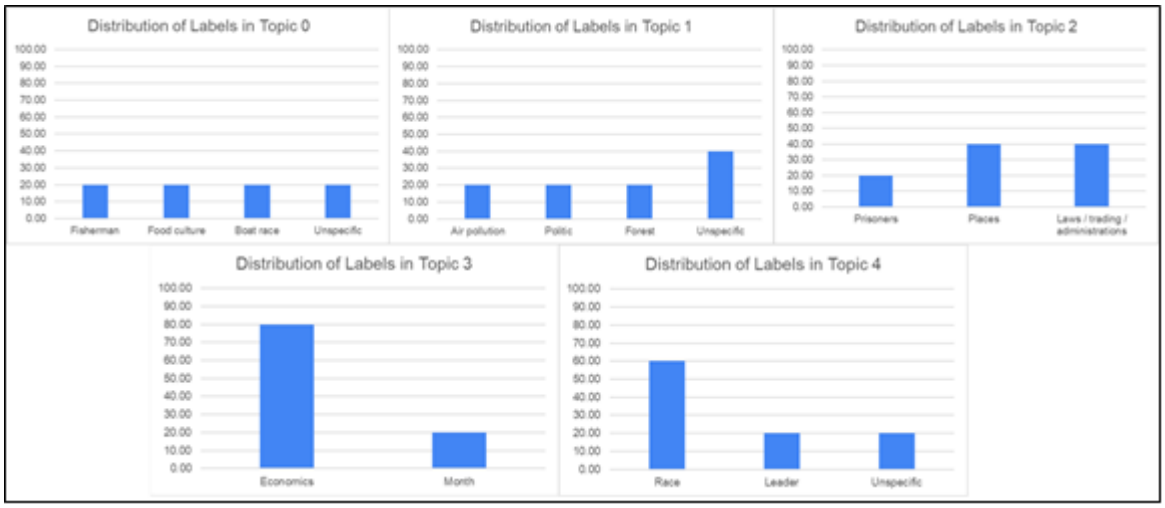

Fig 18. Distribution of labels in topics of Experiment 3.

Throughout Experiment 3 as shown in **Fig 18**, Topic 0 is associated with fisherman, culture, and boat race, with each label selected by 20% of the evaluators. Additionally, 20% of the evaluators label the topic as unspecific. For Topic 1, 20% of evaluators label the topic as air pollution, forestry and politics while the remaining 40% classify the topic as unspecific, reflecting the vagueness of the word cluster. As for Topic 2, 20% of evaluators label the topic as prisoner while 40% of the evaluators categorize it as either places or administrative and laws related. 80% of the evaluators classify Topic 3 as economic related and the remaining classify it as month. The last topic of the experiment is classified as race or competition by 60% of the evaluators while 20% classify it as leader. Another 20% classify the topic as unspecified.

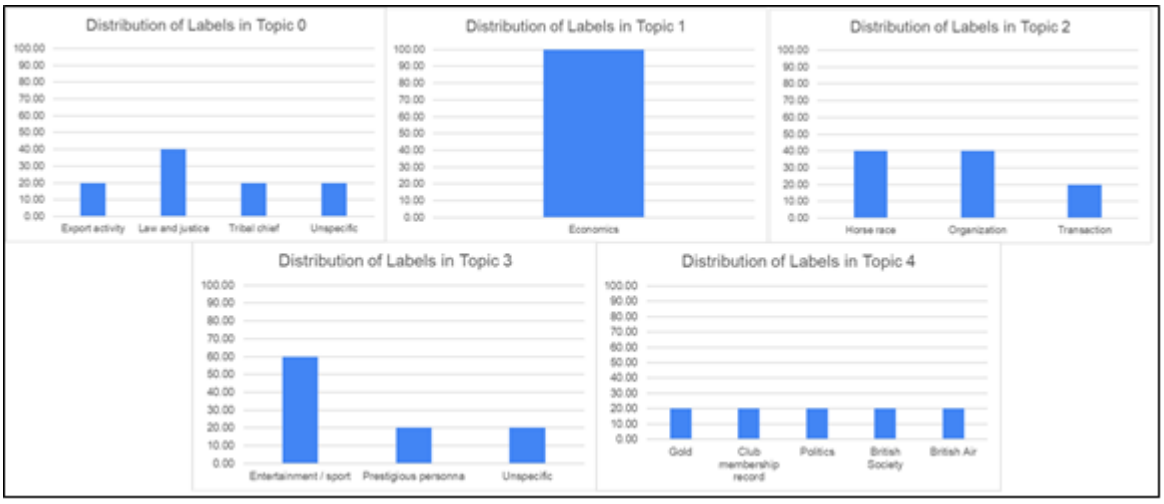

Fig 19. Distribution of labels in topics of Experiment 4.

In Experiment 4 as shown in **Fig 19**, 40% of the evaluators classify Topic 0 as justice or law and 20% each as export activities and tribal chief. Another 20% classify the topic as unspecific as shown in **Fig 19**. For Topic 1, there is 100% agreement among the evaluators, who classify the topic as related to economics. Topic 2 is classified as horse race and organizational terms by 40% each and the remaining classify the term as transaction. Topic 3 has a 30% label that is related to entertainment or sport and the 20% as prestigious persona, and another 20% as unspecific. Topic 4 has a heterogeneous label with each label representing 20% of the evaluators. The labels include gold, club membership record, politics, British society and British air.

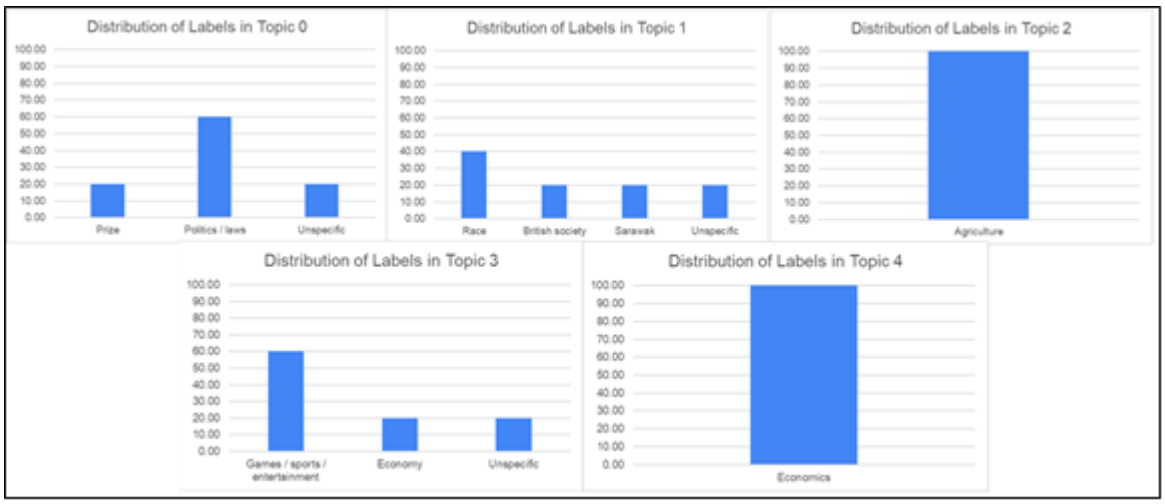

Fig 20. Distribution of labels in topics of Experiment 5

As observed in **Fig 20** for Experiment 5, the evaluators identify Topic 0 as prize, politics or law. 20% of the evaluators failed to interpret the result and classify the topic as unspecific. Next, as for Topic 1, 40% classify it as race, 20% as British society and another 20% as Sarawak. The remaining 20% label it as unspecific. The evaluators unanimously relate Topic 2 as agriculture related term. Topic 3 is labelled as game or sport or entertainment by 60% of the evaluators and 20% each to economy and unspecific. Lastly, all evaluators relate the last topic to agricultural and economic terms.

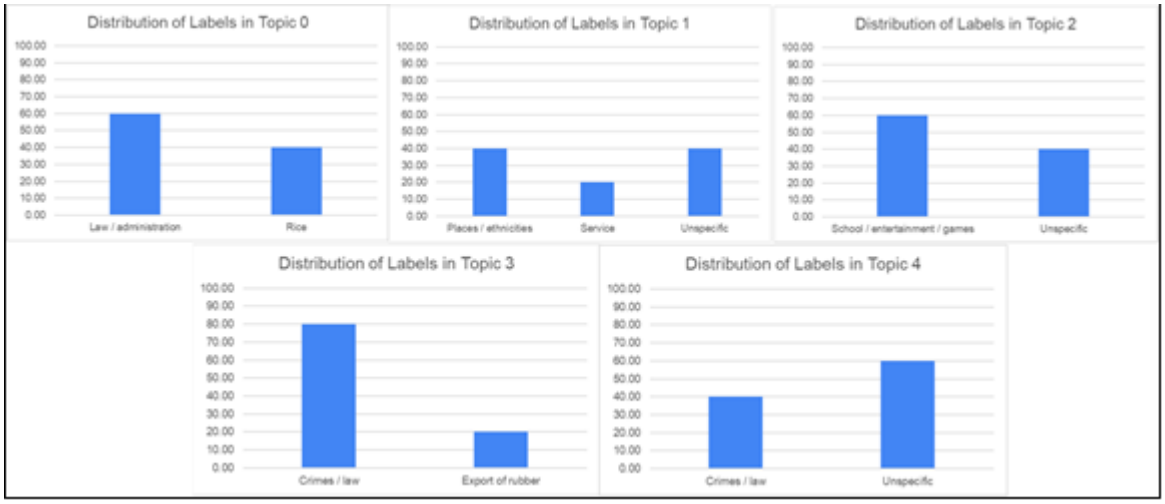

Fig 21. Distribution of labels in topics of Experiment 6.

In the last experiment, 60% of the evaluators classify Topic 0 as laws or administration and the remaining label it as rice as shown in Fig 21. In Topic 1, 20% label it as service, 40% as places or ethnicity, and the remaining 40% as unspecific. Topic 2 comprises 60% of the evaluators classify the topic as school or entertainment, and another 40% as unspecific. 20% of evaluators categorize Topic 3 as exportation of rubber and the rest as crimes. The last topic shows a high degree of ambiguity as 60% of the evaluators label it as unspecific. The remaining distribution shows the topic is classified as crime and law.

The study shows the most heterogeneous results are obtained from Topic 3 and 4 of the Experiment 4, with each of the labels constituting 20% of the evaluators. Five topics from Experiment 1, 4, and 5 receive

unanimous labelling by the evaluators. The label 'unspecific' label is assigned 21 times with the highest proportion belonging to Topic 3 of Experiment 2 and Topic 4 of Experiment 6, as 60% of the evaluators classify these topics as unspecific.

5. Discussion

The perplexity scores increased steadily and peaked at $\alpha = 10$. Pinto et. al. (2021) shows the perplexity score tends to be lower which is good when the number of topics and the values of hyperparameters are lower. Thus, the result shows the optimum result of perplexity score is when $\alpha = 5$.

The coherence scores for this hyperparameters values suggested that the most coherent result is obtained when the number of topics are 4. The word clouds have a sufficient number of word clusters although it does not necessarily reflect the interpretability of the result. Different hyperparameter produced different word clusters and coherence scores due to the varied distribution of document-topic and topic-word. The coherence score charts across hyperparameters show an increase with $\alpha = 5$, $\alpha = 9$, and $\alpha = 10$ shows a slight dip at topic 4.

The experiment also showed that the perplexity value does not necessarily reflect the interpretability of results. When $\alpha = 6$, $\beta = 0.5$, the perplexity indicates a high accuracy. The word cloud of the result showed slightly random words such as iii within the word cluster and words that occurred twice in each of the word clouds which makes the interpretation for each topic rather difficult. Coherence score for this hyperparameters values show that the highest coherent level is when the number of topics is 5. The fifth word cloud shows the words that are linked to criminal activities and reports. Pinto et. al. (2021) presenting the results of coherence score improving as the number of topics are increasing meanwhile showing the perplexity does not necessarily representing the coherence of the result.

Several problems have been encountered during the experiment and evaluation phase of topic modelling by utilizing LDA. The first problem is the requirement of a large dataset to achieve high accuracy. It is problematic especially for the historical data that required familiar data due to the peculiarity of format and spelling although such data is rare and difficult to acquire. Certain words also occurred in two different topics such as the word 'acting' appeared in both Topic 0 and Topic 3 of $\alpha = 9$ due to the similar problem. The lack dimensional reduction caused some words to occur in more than one topic cluster. This lack of dimensional reduction compromises the interpretability of topics.

The perplexity score is useful to indicate the capability of the model to handle and predict the unknown and hidden data although this is not sufficient in determining the optimal number of topics. However, the experiment showed that there is no correlation between perplexity score with coherence. A high scored perplexity score does not reflect the interpretability of word cluster within a topic. The perplexity is not reflecting the semantic interpretability and as a result, the perplexity score is not correlated with the manual inspection and interpretation of topic quality (O'Callaghan et. al.,2015), (Yuan et. al., 2023).

Coherence score is observed to be efficient to identify the best number of topics that are able to ensure the interpretability of the result although the status of hidden or unknown data that may be important to the topic cluster is ambiguous. In the experiment, the coherence score is capable of establishing the correlation between the score and the interpretation of topics compared to the perplexity score (Bretsko et. al., 2023).

Based on the manual validation and the original coherence score, it is cleared there are several differences. In the coherence score generated by LDA, α=9 and β=0.8 shows the highest range of coherence score for its topics, the manual validation founds that α=8 and β=0.7 yield the highest coherence rate. However, the value is not far off as the manual validation for the former one is the second highest.

In conclusion, the LDA has several weaknesses. The repetition of certain words hindered a proper interpretation either from the model itself or based on human inspection. The study also found that the perplexity does not reflect the quality of word cluster.

6. Conclusions

The results showed that as the hyperparameter values increased, the model produced more inaccuracies. Specifically, the word clouds became more random, and the word clusters grew sparse. Additionally, the same words frequently appeared in multiple topics, leading to redundancy and making topic interpretation more challenging. The result also showed that as the hyperparameters increase, the result becomes harder to interpret, and the multiple occurrences of the same

words also caused confusion of assigning the name or summarizing the content of the topics.

The coherence values for each hyperparameter setting were also analyzed. The results indicate that the best coherence values typically occur when the number of topics exceeds 3. The highest coherence scores were observed when the number of topics was set to 5, the maximum used in this experiment.

Future studies should explore additional aspects of hyperparameter tuning and evaluation methods to further optimize the number of topics. A key focus could be comparing LDA with more recent algorithms, such as transformer-based topic modeling or other advanced LDA variants. Fine-tuning models for specific datasets may also yield improvements over traditional LDA techniques. It is also feasible to build a fine-tuned model that can be compared with the LDA results.

Acknowledgement

The authors of this paper would express an appreciation to University Malaysia Sarawak for the support of the publication. The authors acknowledged the financial support from the Ministry of Higher Education Malaysia through Fundamental Research Grant Scheme (FRGS) (F08/FRGS/2029/2020). The authors also extend their gratitude to the validators that contributed to the study by providing their valuable validation and feedback.

References

Agarwal, Ankita, Preetham Salehundam, Swati Padhee, William L Romine, and Tanvi Banerjee. (2020). “Leveraging Natural Language Processing to Mine Issues on Twitter during the COVID-19 Pandemic.” ArXiv (Cornell University), December. https://doi.org/10.1109/bigdata50022.2020.9378002.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Bretsko, D., Belyi, A., & Sobolevsky, S. (2023, June). Comparative Analysis of Community Detection and Transformer-Based Approaches for Topic Clustering of Scientific Papers. In International Conference on Computational Science and Its Applications (pp. 648-660). Cham: Springer Nature Switzerland.

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics, 8, 439-453.

Ding, R., Nallapati, R., & Xiang, B. (2018). Coherence-aware neural topic modeling. arXiv preprint arXiv:1809.02687.

Gertis, E.M. (2021) "Extracting Philosophical Topics from Reddit Posts via Topic Modeling."

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101(suppl_1), 5228-5235.

Hasan, M., Rahman, A., Karim, M. R., Khan, M. S. I., & Islam, M. J. (2021). Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA). In Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020 (pp. 341-354). Springer Singapore.

Lee, J., Kang, J. H., Jun, S., Lim, H., Jang, D., & Park, S. (2018). Ensemble modeling for sustainable technology transfer. Sustainability, 10(7), 2278.

Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. Poetics, 41(6), 545-569.

Muhajir, D., Akbar, M., Bagaskara, A., & Vinarti, R. (2022). Improving classification algorithm on education dataset using hyperparameter tuning. Procedia Computer Science, 197, 538-544.

Newman, D., Bonilla, E. V., & Buntine, W. (2011). Improving topic coherence with regularized topic models. Advances in neural information processing systems, 24.

O’Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications, 42(13), 5645-5657.

Odden, T. O. B., Marin, A., & Caballero, M. D. (2020). Thematic analysis of 18 years of physics education research conference proceedings using natural language processing. Physical Review Physics Education Research, 16(1), 010142.

Panichella, A. (2021). A Systematic Comparison of search-Based approaches for LDA hyperparameter tuning. Information and Software Technology, 130, 106411.

Papadimitriou, C. H., Tamaki, H., Raghavan, P., & Vempala, S. (1998, May). Latent semantic indexing: A probabilistic analysis. In Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (pp. 159-168).

Péladeau, N., & Davoodi, E. (2018). Comparison of latent Dirichlet modeling and factor analysis for topic extraction: A lesson of history.

Pinto Gurdiel, L., Morales Mediano, J., & Cifuentes Quintero, J. A. (2021). A comparison study between coherence and perplexity for determining the number of topics in practitioners interviews analysis.

Seymore, K., & Rosenfeld, R. (1997). Large-scale topic detection and language model adaptation. Carnegie-Mellon University. Department of Computer Science.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. Handbook of latent semantic analysis, 427(7), 424-440.

Teh, Y., Newman, D., & Welling, M. (2006). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. Advances in neural information processing systems, 19.

Wallach, H., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. Advances in neural information processing systems, 22.

Watanabe, K., & Baturo, A. (2023). Seeded Sequential LDA: A Semi-supervised Algorithm for Topic-specific Analysis of Sentences. Social Science Computer Review, 08944393231178605.

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the covid 19 pandemic: using latent dirichlet allocation for topic modeling on twitter. Plos One, 15(9), e0239441. https://doi.org/10.1371/journal.pone.0239441

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295-316.

Yuan, M., Lin, P., Rashidi, L., & Zobel, J. (2023). Assessment of the Quality of Topic Models for Information Retrieval Applications.

Zhou, Z., Liu, M., & Tao, Z. (2023). Quantitative Analysis of Citi's ESG Reporting: LDA and TF-IDF Approaches. Financial Engineering and Risk Management, 6(3), 53-63.

AUTHOR BIOGRAPHIES

Muhammad Abdullah Yusof is a master student at University Malaysia Sarawak in Malaysia. Before then, he had industrial experience in the Western Digital. He received his bachelor's degree from Universiti Malaysia Sarawak. His areas of interest include Natural Language Processing, Text Processing, Artificial Intelligence, and Computational Linguistics.

Suhaila Saee is a senior lecturer at the Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak. She holds a Bachelor's (2001) and Master's (2002) in Computer Science from Universiti Putra Malaysia. Suhaila worked as a researcher at MIMOS Berhad, specializing in knowledge engineering and machine translation. In 2007, she co-founded the Sarawak Language Tech-nology (SaLT) research group with Professor Dr. Alvin Yeo Wee, focusing on digitizing and preserving 63 Sarawak indigenous languages using computational linguistics and natural language processing techniques. Her research interests include computational morphology, NLP, and preserving under-resourced languages. The area of interest is preserving under-resourced languages through speech and text processing.