ISSN 2085-4552

Topic Modelling Using VSM-LDA For Document Summarization

Luthfi Atikah[1], Novrindah Alvi Hasanah[2], Agus Zainal Arifin[2]
[1] Management of Informatic: Politeknik Astra, Kabupaten Bekasi, Indonesia
[2] Departement of Informatic: Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia
[3] Departement of Informatic: Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[1]luthfiatikah@polythecnic.astra.ac.id , [2]novrindah@uin-malang.ac.id, [3]agusza@cs.its.ac.id

Accepted 01 November 2022
Approved 24 December 2022

***Abstract*— Summarization is a process to simplify the contents of a document by eliminating elements that are considered unimportant but do not reduce the core meaning the document wants to convey. However, as is known, a document will contain more than one topic. So it is necessary to identify the topic so that the summarization process is more effective. Latent Dirichlet Allocation (LDA) is a commonly used method of identifying topics. However, when running a program on a different dataset, LDA experiences "order effects", that is, the resulting topic will be different if the train data sequence is changed. In the same document input, LDA will provide inconsistent topics resulting in low coherence values. Therefore, this paper proposes a topic modelling method using a combination of LDA and VSM (Vector Space Model) for automatic summarization. The proposed method can overcome order effects and identify document topics that are calculated based on the TF-IDF weight on VSM generated by LDA. The results of the proposed topic modeling method on the 1300 Twitter data resulted in the highest coherence value reaching 0.72. The summary results obtained Rouge 1 is 0.78, Rouge 2 is 0.67 dan Rouge L is 0.80.**

***Index Terms*— LDA; Order Effects; Summarization; Topic Modelling; VSM-LDA**.

I. INTRODUCTION

The rapid development of the internet has made the information circulating on the internet also increasing. This makes it difficult for information seekers to conclude what news happened. So, needs a way to find useful information efficiently. Summarization can help readers quickly understand the themes and concepts of the entire document, and effectively save time reading.

Summarization is a process to simplify the contents of a document or text by eliminating elements that are deemed unimportant but do not reduce the core meaning wants to convey. However, as it is known, a document will contain more than one topic. So it is necessary to search for topics so that the summarization process is more effective. To find out the topics contained in a document, there have been many methods used to formulate a document making it easier for readers to find out important information contained in a document based on that topic. One of the most commonly used methods is LDA. LDA is a model that calculates the probability of each word on a random topic [1]. LDA has two stages, namely modelling each word into the topic and calculating the probability of each topic repeatedly.

Some studies like [2][3] use the LDA method to determine the topic as the basis for automatic summarization. However, in the process of running a program on a different dataset, LDA experiences an "order effect", that is, the resulting topic will be different if the training data sequence is changed [4]. Such sequencing effects lead to systematic errors for any study in text mining. Then the resulting topic becomes inaccurate. To solve this problem, this paper proposes a Vector Space Model (VSM). This is based on several previous studies that used VSM as a method of retrieval of information on documents such as research conducted by [5][6]. In the VSM method, several online documents will be indexed and sorted based on the weight of search words contained in online documents using the TF-IDF algorithm. In the process, if there is other data that is executed in the LDA program it will still be processed properly because the data train is not randomized, but the sequence has been determined by word weight using the TF-IDF algorithm on VSM [5]. It aims to produce relevant topics based on modelling topics.

However, before doing topic modelling with VSM-LDA, in this study data clustering was carried out which aims to retrieve data in certain clusters so that it can be processed in the next process. Based on [7] using sentence clustering before doing topic modelling for automatic summarization of documents. This study determines the threshold of a particular cluster so that not all clusters are taken. The threshold used is a cluster with a minimum number of members of 110 data. Many clustering methods have been developed so far. This study uses the k-means clustering method. This is based on previous research which is used as a reference such as research conducted by [8] [9][10] which shows that effective clustering is used for data clustering techniques.

Furthermore, the summarization method used in this study is TextRank. The TextRank method is a graphically based method that sorts words in commonly

used text processing [11]. TextRank is used in several studies including [12] performing tweet text summarization to analyze user interest. This method generates a rank for each word which is then mapped to a data set of user interests. Several other studies previously used TextRank as a summarization method, such as in the study [11][13] which improved the TextRank method as a summary method.

This study focuses on the performance of LDA which has an order effect. In the same document input, LDA will provide inconsistent topics resulting in low coherence values. Therefore, this paper proposes a topic modelling method using a combination of LDA and VSM for automatic summarization. Our proposed method can overcome "order effects" so that it can identifies the document topics that are calculated based on the TF-IDF weights on the VSM generated by the LDA. Our topic modeling method can find relevant topics in documents that can produce a high rouge score on the summary results.

II. Related Work

Several previous works have referred us to using this method. In this study we used k-Means as a clustering method before determining topics using topic modelling with LDA. A study, [8] conducted k-means grouping for Al-Qur'an Verses Grouping using the K-Means Algorithm. Another study from [9] used k-Means grouping to apply to the results of modelling topics using LDA to summarize document grouping.

This study combines VSM and LDA. The LDA reference used is based on previous research [3]. Computation of text similarity based on the LDA topic model and coincidence Word is carried out to analyze the semantic correlation of text themes, the results of this study, indicate that LDA will cover text topics better if processed together with coincidence Word. Another study [2] used LDA to perform document summarization, they proposed punitive-based LDA to be summarized. In that study, LDA showed good results. Another study, the LDA method is also used in analyzing public sentiment in various industrial fields, including economics in terms of product marketing [14]. Meanwhile, VSM has shown good performance in word weighting using the TF-IDF algorithm based on previous research conducted by [5]. Then, [6] did study using VSM to query similarity indexing and the results show that each word in the document is weighted so that it can be sequential according to weight.

To summarize the document, this study uses the TextRank for summarization method. This is based on research [11] which applies Text Rank for Automatic Summarization, and the summarization results show that the TextRank method is better than BM25. Meanwhile, another study [13] compared the summary results using TF-IDF and TextRank which showed that the summary results with TextRank were better than TF-IDF for the precision, recall and f-measure values.

III. Proposed Method

To achieve the desired results, this research will go through several stages, namely data preparation, data clustering, topic modelling, and the last id summerizing the document as shown in Figure 1. The following subsections present the explanation of the stages in detail.

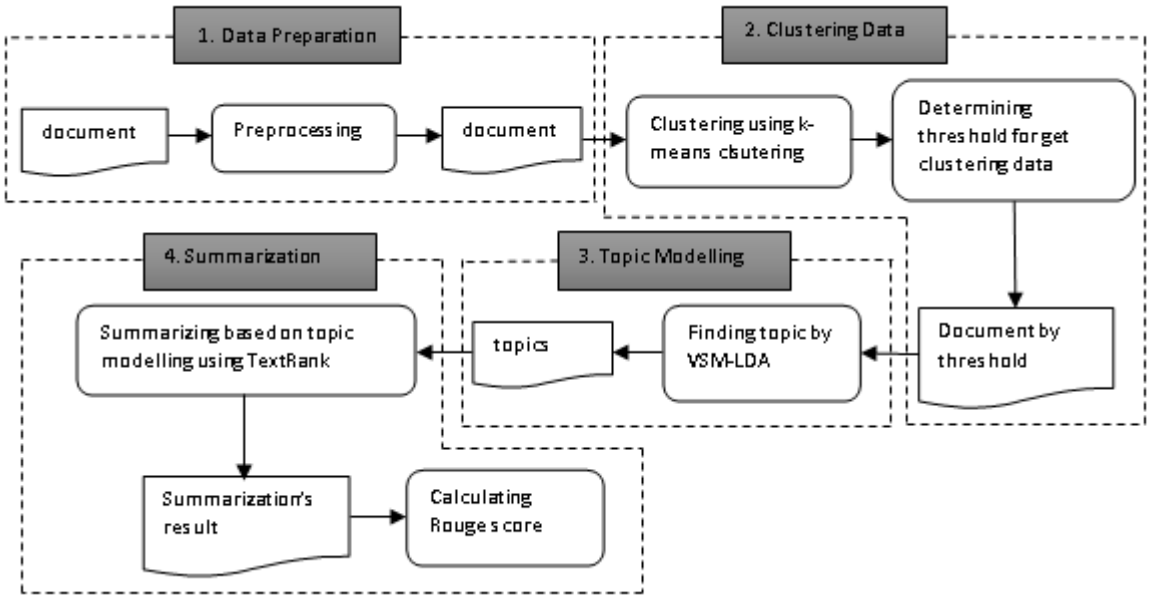


Fig. 1. Research Method

TABLE I. SAMPLE DATA

No	Original data	Clean data
1.	@MeltingIce Assuming max acceleration of 2 to 3 g's, but in a comfortable direction. Will feel like a mild to moder? https://t.co/fpjmEgrHfC	assuming max acceleration of to gs but in a comfortable direction will feel like a mild to moder
2.	Tesla Semi truck unveil & test ride tentatively scheduled for Oct 26th in Hawthorne. Worth seeing this beast in person. Unreal.	tesla semi-truck unveil test ride tentatively scheduled for oct th in hawthorne worth seeing this beast in person unreal
3.	@Nickg_uk @Model3Owners Feature coming soon	feature coming soon

A. *Data Preparation*

This study uses an experimental object in the form of a dataset available online from the Kaggle website. The dataset is a collection data from the Elonmusk Twitter timeline of around 1.300 data which contains automotive issues. The column 'Text' in the dataset becomes the main object that is summarized. The first step in experimenting is cleaning the dataset in the preprocessing process to remove elements that are not needed in the next process to get a summary. Noise in data such as punctuation marks, numbers, and upper letters are replaced with lowercase letters so that the data obtained is cleaned. Data that has been cleaned becomes data that is ready for use in the next process. Table 1 shows an example of original data from Elon Musk that has been preprocessed to obtain clean data.

B. *Clustering Data*

Data that had been cleaned from preprocessing were grouped using k-means clustering. In k-means clusters, the resulting word classes are grouped in semantic similarity under the Euclidean metric boundary. In this study, the threshold is used to take the results of the cluster clustering to be processed at the topic modelling and summarization stage. The k-means algorithm procedure is presented as follows:

1. Step 1. Determine the cluster *K* value. This study uses the value of $K = 7$.
2. Step 2. Allocate data into clusters randomly.
3. Step 3. Calculate the average centroid in each cluster.
4. Step 4. Allocate each data to the nearest centroid/average.
5. Step 5. Return to Step 3 if there are still changes to the centroid of the data cluster has been moved. And stop when nothing changes.
6. Step 6. Count the amount of data in each cluster.
7. Step 7. Determine the threshold value.
8. Step 8. Get the data according to the threshold value.

C. *Topic Modelling*

The vector space model (VSM) is a method used to represent documents as spatial vectors and calculate the similarity between vectors to measure the similarity between documents [3]. VSM is generally run with the TF-IDF (Term frequency-inverse document frequency) algorithm to weight words. This study uses VSM which is processed using the TF-IDF algorithm. We process VSM with Latent Dirichlet Allocation (LDA) as a method of modelling topics so that the topics contained in the document can be found with high coherence values. Meanwhile, LDA is a topic modelling method. Topic modelling aims to find topics automatically in the data or corpus. LDA is based on a Bayesian probabilistic model in which each topic has a separate word probability distribution, and each document consists of a mixture of topics [3]. The basic idea of LDA is that a document is represented as a random mix of latent (invisible) topics [15]. The VSM-LDA we use is shown in Figure 2 below. Cospus document are processed using VSM run with the TF-IDF algorithm. Each word in the document is weighted with TF-IDF weight in VSM before being processed to identify its topic.

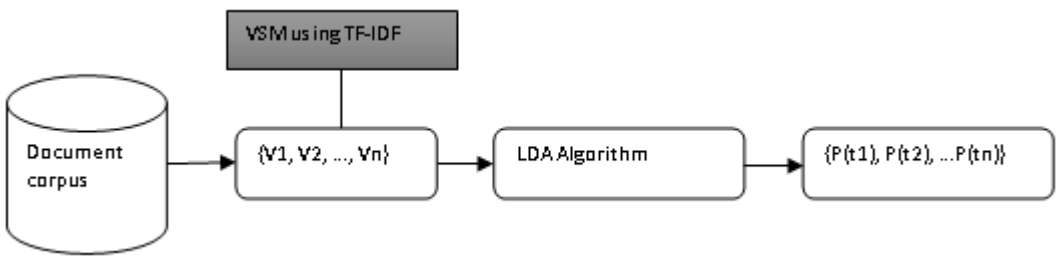


Fig. 2. VSM+LDA Algorithm

D. Summarization

The results of the modelling topic were summarized using the TextRank algorithm as the method used to summarize documents. Each sentence is ranked via the TextRank algorithm. Based on the similarity index, important sentences are selected. Term Frequency - Inverse Document Frequency is used to calculate the importance of terms in each sentence in the input document. According to [10] research, the TextRank algorithm will work in summarizing a document, that is, each sentence in the document represents a node and vertex that represents the similarity. After that, to get the sentence value calculation from a vertex, all the similarity values of the edges connected to each node are added. Sentence values of a primary node will be compared to all existing sentence nodes, sentence nodes that have the same sentence values as a primary node rank them highest because they have similarities.

IV. Result and Discussion

Automatic summarization is done based on the results of clustering using K-means clustering. In this experiment, data from the Elonmusk Twitter timeline, which is available online on the Kaggle website, becomes the object for system testing. The value of *K* used is *K* = 7. The value of *K* is determined based on fine tuning the best *k* on the dataset. This study uses a threshold value to retrieve data at a certain value for processing at the topic modelling stage. Cluster data to be processed on the modelling topic is a cluster with data more than 110 data. The clustering results are shown in Table 2. In the cluster results, cluster data from *n* = 1 and *n* = 3 are taken and used in the next process because the two clusters have data more than 110 data.

TABLE II. The Result Of Clustering Data Using K-Means Clustering

Cluster –n	Count
0	108
1	894
2	45
3	117
4	71
5	66
6	21

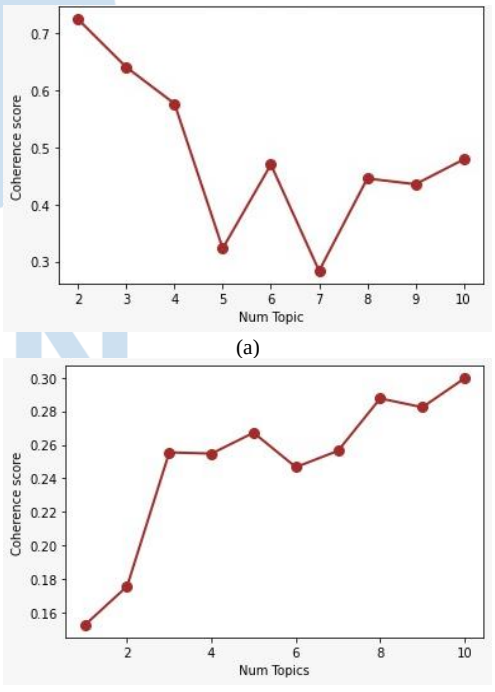
The runtime measurement results in Table I show After getting the cluster data that has been determined using the threshold value, the data is processed to search for topics using topic modelling. At this stage, the research conducted two experiments, namely by using VSM-LDA and LDA to compare how the results of the coherent value of the resulting topic. The coherent topics obtained by the VSM-LDA and LDA shown in Figure 2.

Figure 3 part (a) shows that the topics obtained from VSM-LDA and LDA both get 10 topics for the dataset used in the experiment but with different coherent values. The coherent value obtained by the VSM-LDA method is higher than the coherent value obtained by the LDA method. After getting the topic of the document, the next step is to summarize the document. The summarization trial was conducted twice. The first experiment was carried out using the proposed method, namely k-Means + VSM-LDA + textRank. The second experiment uses the k-means sampling method + LDA + textRank. The results of the two summaries are presented in Table 3 which shows the rouge of each summarized result.

TABLE III. Rouge Score

No	Method	Rouge-1	Rouge-2	Rouge-L
1	K-means+VSM-LDA+ textRank	0.78	**0.67**	**0.80**
2	K-means+LDA+ textRank	0.61	0.48	0.66

Table 3 shows that the scores for Rouge 1, Rouge 2, and Rouge 3 show that the proposed summarization method using k-means clustering + VSM-LDA + TextRank shows better results than the comparison method using k-means clustering + LDA + TextRank.



(a)

(b)

Fig. 3. Topic Coherence by (a) VSM-LDA (b) LDA

Topic modelling using LDA has become the most frequently used method of identifying topics. But in reality, as revealed in research from [5], when running a program on a different dataset, LDA will experience an "order effect", that is, the resulting topic will be different if the training data sequence is changed [4]. Such a sequence effect causes systematic errors for any study, for example text classification and grouping. So that the resulting topics are inaccurate and reduce the

effectiveness of text mining being carried out. This affects the coherent topic score. From the analysis, LDA is suitable for large-scale data because LDA can reduce dimensions. In addition, with LDA data is mapped into topics so that the relationship of each topic can be found.

Meanwhile, VSM is a method that is run based on the TF-IDF algorithm where each word is sorted based on the weight of search words contained in the document [5]. VSM can cover the shortcomings of LDA which often experience order effects so that the resulting topic has a low coherence value. In the process, if there is other data that will be executed in the LDA program, it will still be processed properly because the train data order will not be random, but the sequence has been determined by word weight which has previously been processed by VSM using the TF-IDF algorithm.

V. CONCLUSIONS

This is indicated by the coherence score obtained in the experiment which shows that the coherence score with VSM-LDA shows a good value when compared to the coherence score in the LDA method. The output of Latent Dirichlet Allocation (LDA) is a topic defined by K. Topic coherence is a measure used to evaluate modelling topics based on the top words, the higher the coherence score obtained, the higher the interpretation of the words in the modelling topic [16].

The results of modelling topics with good coherence values become the basis for summarization. This study uses the TextRank method. The results showed that the scores of Rouge 1, Rouge 2, and Rouge L which were calculated based on the summary results showed that the results of the summary based on the results of topic modelling using the proposed method, namely K-means clustering + VSM-LDA + TextRank had a higher value when compared to the summary results using the K-means clustering method + LDA + TextRank. This is influenced by the topics identified using modelling topics with good coherence. The better the coherence value, the better the summary results.

REFERENCES

[1] L. Hagen, "Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models" *Inf Process Manag*, vol. 54, no. 6, pp. 1292–1307, 2018, doi: 10.1016/j.ipm.2018.05.006.

[2] Y. L. Chang and J. T. Chien, "Latent Dirichlet learning for document summarization," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, no. April 2009, pp. 1689–1692, 2009, doi: 10.1109/ICASSP.2009.4959927.

[3] M. Shao and L. Qin, "Text Similarity Computing Based on LDA Topic Model and Word Co-occurrence," no. Sekeie, pp. 199–203, 2014, doi: 10.2991/sekeie-14.2014.47.

[4] A. Agrawal, W. Fu, and T. Menzies, "What is wrong with topic modeling? And how to fix it using search-based software engineering," *Inf Softw Technol*, vol. 98, pp. 74–88, 2018, doi: 10.1016/j.infsof.2018.02.005.

[5] Irmawati, "SISTEM TEMU KEMBALI INFORMASI PADA DOKUMEN DENGAN Irmawati," *Jurnal Ilmiah FIFO*, vol. IX, no. 1, pp. 74–80, 2017.

[6] P. K. Reshma, S. Rajagopal, and V. L. Lajish, "A Novel Document and Query Similarity Indexing using VSM for Unstructured Documents," *2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020*, pp. 676–681, 2020.

[7] I. Lukmana, D. Swanjaya, A. Kurniawardhani, A. Z. Arifin, and D. Purwitasari, "Sentence Clustering Improved Using Topic Words," *Juti*, pp. 1–8, 2014.

[8] C. Slamet, A. Rahman, M. A. Ramdhani, and W. Dharmalaksana, "Clustering the Verses of the Holy Qur'an using K-means Algorithm," *Asian Journal of Information Technology*, vol. 15, no. 24, pp. 5159–5162, 2016.

[9] E. Y. Hidayat, F. Firdausillah, K. Hastuti, I. N. Dewi, and Azhari, "Automatic text summarization using latent drichlet allocation (LDA) for document clustering," *International Journal of Advances in Intelligent Informatics*, vol. 1, no. 3, pp. 132–139, 2015, doi: 10.26555/ijain.v1i3.43.

[10] M. Cha, Y. Gwon, and H. T. Kung, "Language modeling by clustering with word embeddings for text readability assessment," *International Conference on Information and Knowledge Management, Proceedings*, vol. Part F1318, pp. 2003–2006, 2017, doi: 10.1145/3132847.3133104.

[11] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the Similarity Function of TextRank for Automated Summarization," 2016.

[12] R. Niu and B. Shen, "Microblog User Interest Mining Based on Improved TextRank Model," *J Comput (Taipei)*, vol. 30, no. 1, pp. 42–51, 2019.

[13] N. Kumari and P. Singh, "Automated Hindi Text Summarization Using Tf-Idf and Textrank Algorithm," vol. 7, no. 17, pp. 2547–2555, 2020.

[14] X. Liu, A. C. Burns, and Y. Hou, "An Investigation of Brand-Related User-Generated Content on Twitter," *J Advert*, vol. 46, no. 2, pp. 236–247, 2017.

[15] Z. Tong and H. Zhang, "A Text Mining Research Based on LDA Topic Modelling," *Computer Science & Information Technology (CS & IT)*, pp. 201–210, 2016.

[16] F. Rosner, A. Hinneburg, M. Röder, M. Nettling, and A. Both, "Evaluating topic coherence measures," no. December, pp. 0–4, 2014.