Latent Semantic Indexing (LSI) and Hierarchical Dirichlet Process (HDP) Models on News Data

Arif Ridho Lubis, Santi Prayudani, Yulia Fatmi
Department of Computer Engineering and Informatics
Politeknik Negeri Medan
Medan, Indonesia
arifridho@polmed.ac.id, santi.prayudani@gmail.com,
yuliafatmi89@yahoo.co.id

Okvi Nugroho
SMA Muhammadiyah 2
Medan, Indonesia
okvinugroho3@gmail.com

2022 5th International Conference of Computer and Informatics Engineering (IC2IE) | 978-1-6654-5338-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/IC2IE56416.2022.9970067

***Abstract*—News has become a very important need in modern society. Almost every level of society needs information such as news. Online news gets the attention of writers because there are a lot of characteristics of the data available and can be used for analysis so as to get insights and current trends, such as on 4 news portals that are most frequently accessed by the public such as detik.com, okezone.com, kompas.com, and tribunnews.com. Related research in this field about Trend Analysis and Topic modeling on news media data in Indonesia then discusses news data collection methods and data collection techniques news pre-processing text, and 2 Topic Modeling algorithms applied in this research, namely Latent Semantic Indexing (LSI) and Hierarchical Dirichlet Process (HDP). In this study, the author has succeeded in extracting topics from news data from media in Indonesia. Various topics show the effectiveness of a learning model producing topics from the LSI and HDP models. This research has implemented natural language preprocessing (NLP) to process news data so that it can be processed and used as data from modeling a topic. This study finds topics based on the model of coherence and LSI and HDP contained in the news data with a total of 29,437. The accuracy obtained is 87% compared to the model from LDA-KNN which can determine the topic with an accuracy of 72%. Hence, the Latent Semantic Indexing (LSI) model and Hierarchical Dirichlet Process (HDP) have better accuracy than the LDA-KNN model.**

Keywords*—*NLP, topic, HDP, LSI, extraction

I. INTRODUCTION

News has become a very important need in modern society, almost every level of society needs information such as news [1]. It is the main information in the mass media physically and then develops into media that can be accessed online [2]. Online news gets the attention of writers because of the many characteristics of the data that are available and can be used for analysis so as to gain insight and current trends [3][4]. Analyzing trending issues such as event detection, topic detection, sentiment analysis, forecasting and some important research is related to the collected news data [5][6]. Currently, there are 4 news portals that are most frequently accessed by the public, such as detik.com, okezone.com, kompas.com, and tribunnews.com. All of them have characteristic data that can be utilized and processed into news trend analysis related to topic discovery news, predict news trends, and understand social behavior patterns, build recommendation systems, and document classification tasks[7]-[9]. The news data of each media finds this trending topic based on hashtags on news, news content, subject and hashtag analysis [10][11].

Other researchers have also studied this news trend analysis by using an API to retrieve news data and perform analysis on the news data [12]-[14]. So far, trend analysis was being carried out with traditional NLP techniques such as TF-IDF analysis, word2vector and machine learning based classification. LSI-based Topic Modeling was being applied to automatically find topics on trending news data[15][16]. There were also many researchers who have combined methods or algorithms that would be used to identify topics [17]-[19]. The data on the news could be seen as a collection of documents and text then Topic Modeling could be used to get the topic of the news by processing the data as a document and doing preprocessing.

Topic Modeling is a statistical method, mostly used to extract representative hidden data from the corpus [20][21]. The topic model is a probabilistic model representing the topic as a multinomial distribution of words, with the premise that each document of the corpus is described as a corresponding mixture of topics [22]-[24]. As far as the author's knowledge in doing Topic Modeling can use several algorithms such as Latent semantic indexing, Hierarchical Dirichlet Process, Latent Dirichlet Allocation and KNN. The contribution of this paper is about finding a topic in an online media news in Indonesia by applying Topic Modeling from news data resulted from the crawling technique collected from 2018 to 2021. This research was arranged as follows: First, this research is related in this field about Trend Analysis and Topic modeling on media news data in Indonesia then discussed news data collection methods and news data collection techniques pre-processing text, and 2 Topic Modeling algorithms applied in this study, namely Latent Semantic Indexing (LSI) and Hierarchical Dirichlet Process (HDP) models.

II. RELATED WORK

Topic modeling is part of an approach at the text mining stage being used to extract hidden topics in data or documents or conceptually the process of grouping text data based on a particular topic [25][26]. This is a statistical concept that will be used for text mining and then identify hidden patterns in a collection of having unstructured data [27][28]. The concept of Topic Modeling was first introduced by using Latent Semantic Indexing (LSI). This non-probabilistic model performs contextual identification between words in a collection of documents [29][30]. Hierarchical Dirichlet Process (HDP) for modeling documents on each "topic" as the distribution of words in the data or corpus against known vocabulary in the document or corpus. The topic combination is taken from the Dirichlet distribution then word for word in the corpus is independently taken [31][32].

Many researchers have applied Topic Modeling to data in documents or corpus. The research of D. Zhao, J. He, and J.

Liu [33] had conducted research on the latent topic of a corpus with large amounts of data, the LDA model assumed that a document closest to a topic Latent Dirichlet Allocation is a classic topic model that could extract latent topics from big data corpus. This model assumed that if a document was relevant to a topic, this study discussed the LDA algorithm for topic classification by adding a topic tagger distribution parameter criterion to LDA which could produce relevant categories. The results obtained would then use the Gibbs side to determine the estimated infrensi and the results of the study so as to show the effectiveness of the LDA method.

Research by W. Chen and X. Zhang [34] has worked on text classification by discussing the similarity between texts that need to be calculated, taking into account that the classification that utilizes the method using only words from features and categories does not incorporate semantic similarities between the available word features. This study uses the classification model of LDA and K-Nearest Neighbor. LDA serves to solve the problem of semantic similarity in the text then the K-Nearest Neighbor algorithm is used to classify text data from Corpus China. The result obtained from the classification was 0.933. The results showed that the LDA-KNN model had superior classification performance in text categorization

Research by J. Bian, Z. Jiang, and Q. Chen [35] had been carried out work on LDA (Latent Dirichlet Allocation). This model could determine the hidden topics in a corpus so that it was useful to form a good summary. In this study, based on the LDA model, the distribution of topics to documents and terms or topic categories were obtained. In this study proposed a sentence ranking method that could show sentences having the highest frequency. This experiment showed by using summarization to get more performance and by getting a good rouge value.

Research by Al-Anzi and D. AbuZeina [36] had been conducted by utilizing a vector space model performing document representation in the data, but this technique had drawbacks such as high dimensional space and semantic losses, therefore LSI or latent semantic indexing was used to reduce feature dimensions and semantic losses in data, besides LSI has been successfully implemented in search engines and texts for classifying. This study took an approach so that it could improve the quality of the data taken in the search. This research proposed the development of the LSI technique based on the use of words that form a matrix in data on a corpus.

III. Methodology

This study discussed the use of topic modeling techniques that had been regulated by the Latent Semantic Indexing (LSI) and Hierarchical Dirichlet Process (HDP) techniques to find topics from news data. The methodology of this research consisted of five different stages such as data collection using crawling techniques, preprocessing, TF IDF then Topic Modeling using LSI together with HDP. The following was the research methodology as shown in Fig. 1.

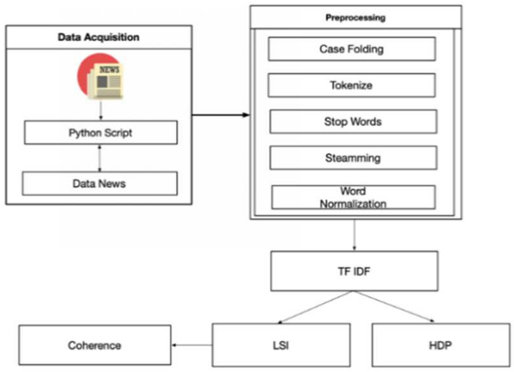

Fig. 1. The process in topic modelling

A. Data Collection

A crawling technique was applied to get news data on online media in Indonesia. This study collected news data created between 07 November 2017 to 20 May 2020 from 4 media news sites in Indonesia, i.e., detik.com, okezone.com, tribun.com, and kompas.com. This study succeeded in collecting data with a total of 29.437. The data from the research can be seen in Table I.

TABLE I. Pattern of News Data

Date	News Type	Headline
07/11/2017	Megapolitan	Pembunuhan Wanita di Hotel Menteng Berawal dari Open BO, Selengkapnya di iNews Sore Minggu Pukul 16.00 WIB
08/11/2017	National	Terkuak! Identitas Pelaku Pembunuhan Wanita di Kawasan Menteng Jakarta Pusat, Selengkapnya di iNews Sore Jumat Pukul 16.00WIB
09/11/2017	Celebrity	Drakor Mouse, Perbuatan Licik Seorang Pembunuh
10/11/2017	News	Polisi Tangkap Penyebar Hoaks Genosida Warga Papua di Facebook
11/11/2017	National	Selegram Tewas di Tangan Kekasih, Selengkapnya di iNews Siang Sabtu Pukul 11.00 WIB
14/11/2017	National	TNI AD dan Marinir Diterjunkan Buru Kelompok MIT
15/11/2017	News	Bentrok Pendukung Trump dan BLM di Portland, Satu Orang Tewas Ditembak
…	…	…
…	…	…
…	…	…
03/01/2018	Celebrity	Polisi Temukan Ganja Milik Anji Selain di Rumah
04/01/2018	National	Cerita Menkumham Yasonna Minta Maaf ke Warga Tanjung Priok
05/01/2018	Megapolitan	Polisi Tangkap Pelaku Spesialis Begal Handphone yang Viral di Tebet
…	…	…
…	…	…
…	…	…
01/01/2020	National	Penembakan saat Shalat Idul Adha, Meleset Karena Penembak Lihat Bung Karno Ada Dua
18/05/2020	National	PP Muhammadiyah: Pembunuhan Laskar FPI Harusnya Masuk Pelanggaran HAM Berat
19/05/2020	National	"Realita" di iNews TV Pukul 15.00 Ini: Miris! Jasad Misterius Terbungkus Kardus Mie Instan
20/05/2020	News	Dipicu Pembunuhan Warga Suku Arab, 83 Orang Tewas dalam Serangan Milisi di Sudan

B. Preprocessing

Pre-processing is required to make the input data usable for further analysis. Because news is a sentence that only has a few punctuation marks, special characters, it is also necessary to have a preprocessing stage to improve the word structure of the news data. The preprocessing stages that this study carried out were as shown in Fig. 1. This study removed punctuation and special characters. The study then performed the removal of all English letters and numbers. Table II is the result of preprocessing from the news data in Table I.

TABLE II. PREPROCESSING RESULTS

Headline
pembunuhan wanita di hotel menteng berawal dari selengkapnya di inews sore
terkuak identitas pelaku pembunuhan wanita di kawasan menteng jakarta pusat, selengkapnya di inews sore
perbuatan licik seorang pembunuh
polisi tangkap penyebar genosida warga papua di facebook
selebgram tewas di tangan kekasih, selengkapnya di inews siang
tni ad dan marinir diterjunkan buru kelompok
bentrok pendukung trump dan satu orang tewas ditembak
...
...
...
polisi temukan ganja milik anji selain di rumah
cerita menkumham yasonna minta maaf ke warga tanjung priok
polisi tangkap pelaku spesialis begal handphone yang viral di tebet
...
...
...
penembakan saat shalat idul adha, meleset karena penembak lihat bung karno ada dua
Muhammadiyah pembunuhan laskar harusnya masuk pelanggaran ham berat
realita di inews tv pukul ini miris jasad misterius terbungkus kardus mie instan
dipicu pembunuhan warga suku arab 83 orang tewas dalam serangan milisi di sudan

After the preprocessing results obtained in Table II, then the modeling process will be carried out from Latent Semantic Indexing, Hierarchical Dirichlet Process, Latent Dirichlet Allocation and KNN to obtain a model with each accuracy.

C. LSI Process for Topic Extraction

We applied Latent Semantic Indexing (LSI), a technique in the field of NLP (Natural Language Processing), which is often used to analyze the relationship between a dataset of terms associated with those data. It applies a mathematical technique called Singular Value Decomposition (SVD) to identify patterns in the relationships between words and sentences contained in a corpus of documents (unstructured collections of documents). The following is the structure of the latent semantic index.

Fig. 2 explained that latent semantic indexing performed the process of word distribution stating that words closing in meaning would appear in similar pieces of text. The starting point was the matrix of the distribution of words in the document set. It is an m times n matrix where m is the number of unique words and n is the cardinality of the corpus.

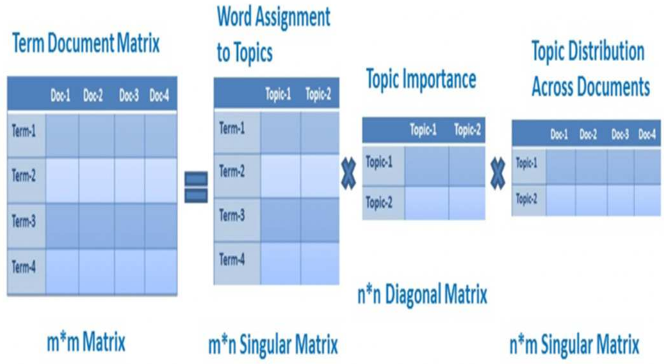

Fig. 2. Structure of laten semantic indexing

IV. DISCUSSION

This stage will explain the analysis of the data in this study based on the methodology described in section 3 of the research methodology:

A. Coherence Model

After the document data had been formed and before the topic model stage was created, it was necessary to determine how many topics would be formed with the reason to avoid the weakness of topics that were too narrow and broad. To determine the topic, the process of measuring topic coherence was carried out. For this reason, the UCI, UMass, and CV methods would be applied. The calculation process was carried out using the Gensim library. The graph of the resulting coherence score was up and down, in this study it would display a graph of the starting topic. Based on Fig. 3 on the resulting coherence graph, the topics displayed were topics 0 – 5.

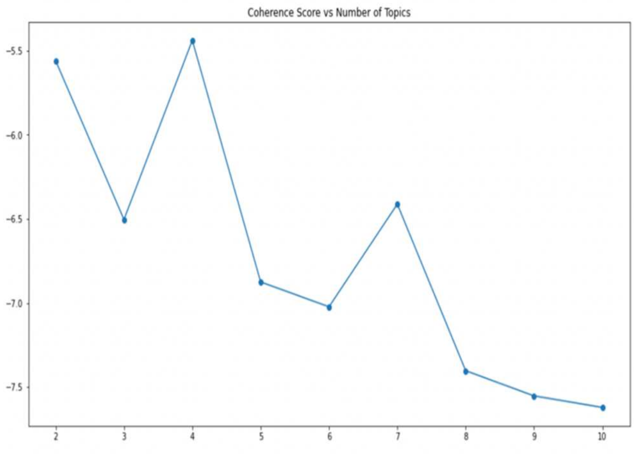

Fig. 3. Coherence Model

B. Topic Extraction from News Data

Topic extraction from news model data from Latent Semantic Indexing and Hierarchical Dirichlet Process (HDP) applied in modeling a topic on news using TF-IDF forming a vector then was processed with Latent semantic analysis and Hierarchical Dirichlet Process (HDP), the results of LSI and HDP can be seen in Table III and Table IV. In this study, the number of topics to be extracted was obtained into four topics. Results and Visualizations are shown in the next section.

TABLE III. RESULT OF TOPIC EXTRACTION BY LSI

Topic 0	N	Topic 1	N	Topic 2	N	Topic 3	N	Topic 4	N
Polisi	0.901	Narkoba	-0.736	Tewas	0.742	Sabu	0.691	Perampok	-0.827
Sabu	0.157	Sabu	-0.442	Perampok	0.284	Narkoba	0.595	Tewas	-0.595
Narkoba	0.132	Polisi	0.246	Sabu	-0.202	Kg	0.288	Rp	-0.186
Pelaku	0.132	Bnn	-0.224	Polisi	-0.198	Tewas	0.142	Juta	-0.16
Ditangkap	0.113	Kg	-0.18	Anak	0.166	Ditangkap	0.126	Ditangkap	-0.114
Perampok	0.103	Ditangkap	-0.129	Temukan	0.152	Polisi	-0.115	Uang	-0.1
Tewas	0.081	Bandar	-0.086	Diduga	0.14	Perampok	0.074	Emas	-0.097

TABLE IV. RESULT OF TOPIC EXTRACTION BY HDP

Percent Contribution	Topic Keyword
0.8758	Polisi. Narkoba. Perampok. Tewas. Sabu. Pelaku
0.8759	Polisi. Narkoba. Perampok. Ditangkap
0.8897	Polisi. Perampok. Pelaku. Ditangkap
0.8896	Polisi. Narkoba. Sabu. Perampok. Tewas
0.8581	Polisi. Narkoba. Perampok. Tewas. Sabu

C. *Latens Semantic Indexing*

Table III showed the five topics that had been determined by the latent semantic indexing process contained in the news data with 5 topics sorted by the highest weight on the top and the lowest on the bottom. Table III showed that topic 0 was about police and methamphetamine at the top. topic 1 was about drugs and shabu. topic 2 was about death and robbers. topic 3 was about methamphetamine and drugs. topic 4 was about robbers and being killed. Then in Table IV showed the results of the Hierarchical Dirichlet Process (HDP) model process which determined the topic with the highest presentation based on news data. Then in Fig. 3 showed that in the data there were words representative of the topic. Fig. 4 showed that persecution. narcotics. robbery. theft and murder representative words that had the highest frequency.

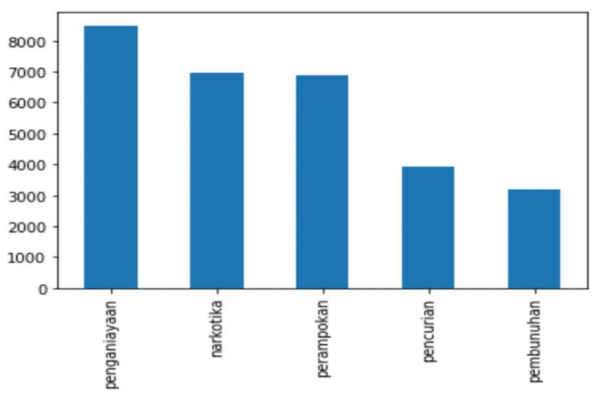

Fig. 4. Representative word on the topic

D. *Visualization*

This study uses sns.boxplot. a python library displaying visualization of the news data used. Fig. 5 showed that there was a header that had an average and length of sentence was 8 to 9 words some headlines had a length of 15 words and a maximum of 24 words. Fig. 6 shows headline length for each news type.

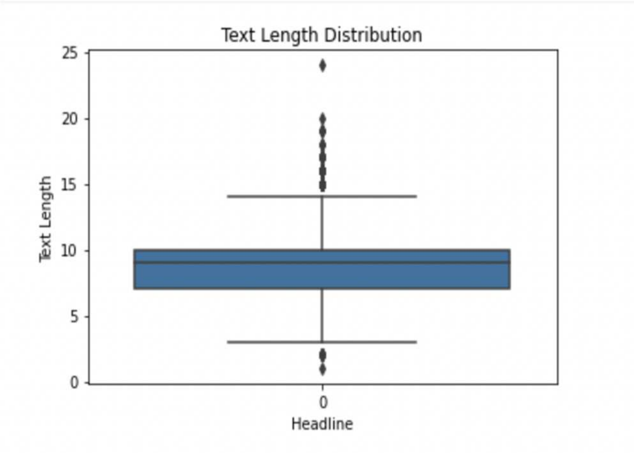

Fig. 5. Text length distribution

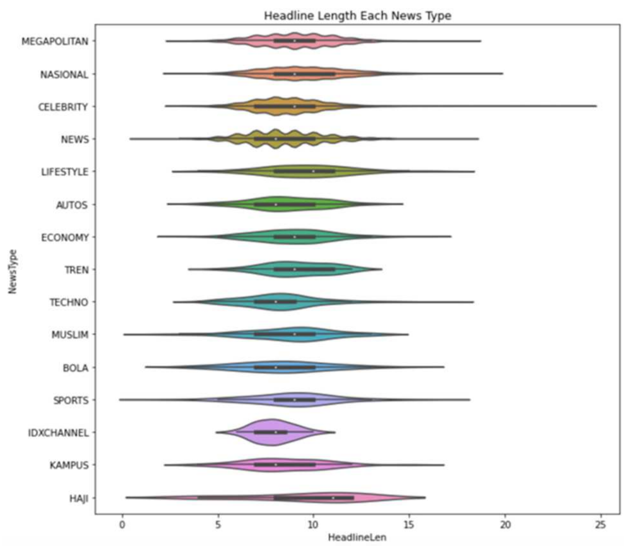

Fig. 6. Headline length each news type

E. *LDA-KNN Model*

In modeling this topic, it will compare with LDA-KNN to see which accuracy model is the best in determining the topic of news. The results of LDA-KNN can be seen in Fig. 7. Based on the modeling of LDA-KNN, the results of the confusion matrix show that the value of precision is 79%. recall is 88% and accuracy is 72%.

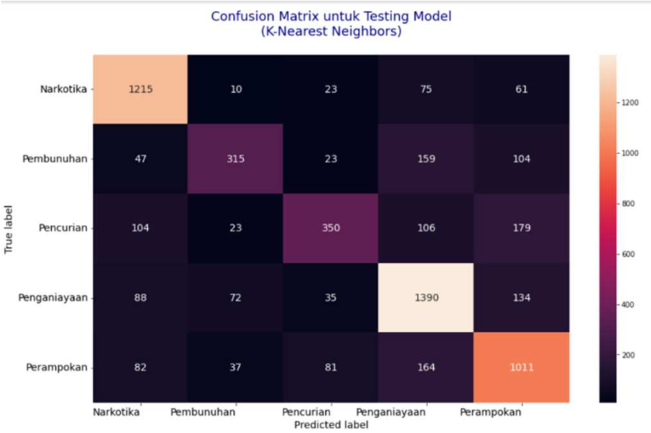

Fig. 7. Confusion matrix results from LDA-KNN

V. Conclusion

In this study, the author succeeded in extracting topics from news data from media in Indonesia. Various topics showed the effectiveness of a learning model producing topics from the LSI and HDP models. This research implemented natural language preprocessing (NLP) to process news data so that it could be processed and used as data from modeling a topic. This study found topics based on the model of coherence and LSI and HDP contained in the news data with a total of 29.437. Then the application of the LSI and HDP models produces an accuracy value of 87% which means that the modeling in determining the topic is good. Then when compared to the LDA-KNN model, the accuracy in determining the topic is 72% so it can be concluded that the modeling of LSI and HDP is better when compared to LDA-KNN. In the future, this research will conduct on a larger number of news datasets using techniques from natural language preprocessing, such as developing a model from LSI-HDP by comparing it to other models using techniques from Gibbs Sampling.

References

[1] K. Shu. S. Wang. and H. Liu. “Understanding user profiles on social media for fake news detection.” in *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. 2018. pp. 430–435.

[2] I. Shklovski. L. Palen. and J. Sutton. “Finding community through information and communication technology in disaster response.” in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 2008. pp. 127–136.

[3] A. R. Lubis. M. K. M. Nasution. O. S. Sitompul. and E. M. Zamzami. “The effect of the TF-IDF algorithm in times series in forecasting word on social media.” *Indones. J. Electr. Eng. Comput. Sci.*. vol. 22. no. 2. p. 976. 2021. doi: 10.11591/ijeecs.v22.i2.pp976-984.

[4] K. Kucher and A. Kerren. “Text visualization techniques: Taxonomy. visual survey. and community insights.” in *2015 IEEE Pacific visualization symposium (pacificVis)*. 2015. pp. 117–121.

[5] P. Chauhan. N. Sharma. and G. Sikka. “The emergence of social media data and sentiment analysis in election prediction.” *J. Ambient Intell. Humaniz. Comput.*. vol. 12. no. 2. pp. 2601–2627. 2021.

[6] K. H. Manguri. R. N. Ramadhan. and P. R. M. Amin. “Twitter sentiment analysis on worldwide COVID-19 outbreaks.” *Kurdistan J. Appl. Res.*. pp. 54–65. 2020.

[7] M. Muslikhin and M. Rizkiansyah. “Google Control Journalism Practices in Indonesian Cyber Media.” in *2021 International Conference on Information Management and Technology (ICIMTech)*. 2021. vol. 1. pp. 673–677.

[8] H.-C. Shin *et al.*. “Deep convolutional neural networks for computer-aided detection: CNN architectures. dataset characteristics and transfer learning.” *IEEE Trans. Med. Imaging*. vol. 35. no. 5. pp. 1285–1298. 2016.

[9] B. Jang and J. Yoon. “Characteristics analysis of data from news and social network services.” *IEEE Access*. vol. 6. pp. 18061–18073. 2018.

[10] M. Li and L. Wang. “A survey on personalized news recommendation technology.” *IEEE Access*. vol. 7. pp. 145861–145879. 2019.

[11] P. Qi. J. Cao. T. Yang. J. Guo. and J. Li. “Exploiting multi-domain visual information for fake news detection.” in *2019 IEEE International Conference on Data Mining (ICDM)*. 2019. pp. 518–527.

[12] A. R. Lubis *et al.*. “Obtaining Value From The Constraints in Finding User Habitual Words.” pp. 8–11. 2020.

[13] B. Batrinca and P. C. Treleaven. “Social media analytics: a survey of techniques. tools and platforms.” *AI Soc.*. vol. 30. no. 1. pp. 89–116. 2015. doi: 10.1007/s00146-014-0549-4.

[14] R. Batool. A. M. Khattak. J. Maqbool. and S. Lee. “Precise tweet classification and sentiment analysis.” in *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)*. 2013. pp. 461–466.

[15] R. A. Stein. P. A. Jaques. and J. F. Valiati. “An analysis of hierarchical text classification using word embeddings.” *Inf. Sci. (Ny).*. vol. 471. pp. 216–232. 2019.

[16] B. Altınel and M. C. Ganiz. “Semantic text classification: A survey of past and recent advances.” *Inf. Process. Manag.*. vol. 54. no. 6. pp. 1129–1153. 2018.

[17] A. R. Lubis. M. K. M. Nasution. O. S. Sitompul. and E. M. Zamzami. “A Framework of Utilizing Big Data of Social Media to Find Out the Habits of Users Using Keyword.” 2020. pp. 140–144.

[18] B. Liu. C. W. Chin. and H. T. Ng. “Mining topic-specific concepts and definitions on the web.” in *Proceedings of the 12th international conference on World Wide Web*. 2003. pp. 251–260.

[19] G. P. Zhang. “Neural networks for classification: a survey.” *IEEE Trans. Syst. Man. Cybern. Part C (Applications Rev.*. vol. 30. no. 4. pp. 451–462. 2000.

[20] H. Jelodar *et al.*. “Latent Dirichlet allocation (LDA) and topic modeling: models. applications. a survey.” *Multimed. Tools Appl.*. vol. 78. no. 11. pp. 15169–15211. 2019.

[21] D. Liu. Y. Zeng. Y. Luo. H. Pang. and X. H. Wu. “Window-Based Topic Model for HDP.” *2019 16th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. ICCWAMTIP 2019*. pp. 70–75. 2019. doi: 10.1109/ICCWAMTIP47768.2019.9067737.

[22] L. Shu. B. Long. and W. Meng. “A latent topic model for complete entity resolution.” in *2009 IEEE 25th International Conference on Data Engineering*. 2009. pp. 880–891.

[23] J. Huang. M. Peng. P. Li. Z. Hu. and C. Xu. “Improving biterm topic model with word embeddings.” *World Wide Web*. vol. 23. no. 6. pp. 3099–3124. 2020.

[24] J. Xuan. J. Lu. G. Zhang. and X. Luo. “Topic model for graph mining.” *IEEE Trans. Cybern.*. vol. 45. no. 12. pp. 2792–2803. 2015.

[25] L. L. Benites-Lazaro. L. Giatti. and A. Giarolla. “Topic modeling method for analyzing social actor discourses on climate change. energy and food security.” *Energy Res. Soc. Sci.*. vol. 45. pp. 318–330. 2018.

[26] M. Allahyari *et al.*. “A brief survey of text mining: Classification. clustering and extraction techniques.” *arXiv Prepr. arXiv1707.02919*. 2017.

[27] S. Vijayarani. M. J. Ilamathi. and M. Nithya. “Preprocessing techniques for text mining-an overview.” *Int. J. Comput. Sci. Commun. Networks*. vol. 5. no. 1. pp. 7–16. 2015.

[28] S. A. Salloum. M. Al-Emran. A. A. Monem. and K. Shaalan. “Using text mining techniques for extracting information from research articles.” in *Intelligent natural language processing: Trends and Applications*. Springer. 2018. pp. 373–397.

[29] A. R. Pathak. M. Pandey. and S. Rautaray. “Topic-level sentiment analysis of social media data using deep learning.” *Appl. Soft Comput.*. vol. 108. p. 107440. 2021.

[30] F. Viegas *et al.*. “Semantically-enhanced topic modeling.” in *Proceedings of the 27th ACM international conference on*

information and knowledge management. 2018. pp. 893–902.

[31] X. Wang. X. Ma. and W. E. L. Grimson. “Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models.” *IEEE Trans. Pattern Anal. Mach. Intell.*. vol. 31. no. 3. pp. 539–555. 2008.

[32] M. Shimosaka. T. Tsukiji. S. Tominaga. and K. Tsubouchi. “Coupled hierarchical Dirichlet process mixtures for simultaneous clustering and topic modeling.” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2016. pp. 230–246.

[33] D. Zhao. J. He. and J. Liu. “An improved LDA algorithm for text classification.” *Proc. - 2014 Int. Conf. Inf. Sci. Electron. Electr. Eng. ISEEE 2014*. vol. 1. pp. 217–221. 2014. doi: 10.1109/InfoSEEE.2014.6948100.

[34] W. Chen and X. Zhang. “Research on text categorization model based on LDA - KNN.” *Proc. 2017 IEEE 2nd Adv. Inf. Technol. Electron. Autom. Control Conf. IAEAC 2017*. pp. 2719–2726. 2017. doi: 10.1109/IAEAC.2017.8054520.

[35] J. Bian. Z. Jiang. and Q. Chen. “Research on multi-document summarization based on LDA topic model.” *Proc. - 2014 6th Int. Conf. Intell. Human-Machine Syst. Cybern. IHMSC 2014*. vol. 2. pp. 113–116. 2014. doi: 10.1109/IHMSC.2014.130.

[36] F. S. Al-Anzi and D. AbuZeina. “Enhanced search for Arabic language using latent semantic indexing (LSI).” *2018 Int. Conf. Intell. Innov. Comput. Appl. ICONIC 2018*. pp. 1–4. 2019. doi: 10.1109/ICONIC.2018.8601096.