e-ISSN: 2614-8404
p-ISSN : 2776-3234

Automating the Extraction of Words and Topics in Indonesian Using the Term Frequency-Inverse Document Frequency Algorithm and Latent Dirichlet Allocation

Lalu Mutawalli[1], Mohammad Taufan Asri Zaen[2*)], Muhammad Fauzi Zulkarnaen
[1,2,3]Program Studi Sistem Informasi, STMIK Lombok
Email: [1]laluallistilo@gmail.com, [2]opanzain@gmail.com, [3]fauzi_tuan@gmail.com

Abstract – Keyword extraction and topic modeling in the analysis of Gojek user reviews in Indonesian are very important. By understanding user preferences and needs through keyword extraction, as well as grouping user reviews into different topics through topic modeling, stakeholders can use the information to further improve services. This research uses TF-IDF and LDA approaches to analyze text data from Gojek user reviews and feedback. The data spans from Nov 5, 2021, to Jan 2, 2024, totaling 225,002 rows. Each row includes username, content, time, and app version. The focus is on content reviews. The average length is 8 words, with a maximum of 104 and a minimum of a few words. The variability indicates a non-normal distribution. Preprocessing is conducted to maintain topic analysis accuracy. The TF-IDF method is used to extract relevant keywords, while the LDA approach is used to model the topics in user reviews. The topic analysis reveals patterns in Gojek user reviews. The first topic discusses experience, services, and affordable pricing. The second emphasizes app usability and benefits. The third relates to promos, discounts, and vouchers. The fourth reflects positive evaluations of service quality. However, the fifth topic highlights high costs and app issues. The sixth underscores overall user satisfaction and service convenience. Testing on the topic model yielded a coherence level of 0.509, indicating that the model's topics demonstrate a good level of consistency in finding relevant topics from Gojek user review data. The use of a combination of TF-IDF and LDA in Indonesian text analysis, particularly in the context of Gojek user reviews, is an important step in enhancing understanding and utilization of text data to improve overall user experience.

Keywords – word extraction; topic modeling; preferences; TF-IDF and LDA.

I. INTRODUCTION

The growth of digital data in Indonesia is very rapid, in 2022 experiencing a very high year-over-year (YOY) growth rate of 64.4% [1]. Driven by a variety of factors, including the popularity of digital platforms such as social media, online shopping trends (e-commerce), as well as online transportation service platforms [2]. One example is Gojek, recording the highest user growth in Indonesia, with an average application download reaching 957,000 every month [3]. Currently, Gojek is one of the largest online transportation service platforms in Indonesia, analyzing the usage patterns, preferences, and habits of Gojek users is crucial because it can provide valuable understanding for companies to improve user experience, improve services offered, and find new opportunities. A deep understanding of user behavior can be used to better respond to market needs, and enable the creation of more innovative solutions.

Keyword extraction and topic modeling have an important role, keyword extraction from Gojek user feedback reviews can help understand user preferences and needs better, and topic modeling allows to grouping of user reviews into various topics can provide knowledge about aspects that users want, such as service quality, convenience so that it can be used by stakeholders for service improvement. Various methods are used to extract words from a set of documents, including frequency-focused approaches. This research shows the success of TF-IDF in extracting stopwords in Indonesian [4]. Comparing InSet Lexicon and TF-IDF methods on Indonesian text emotion recognition datasets, the evaluation results of the two methods show that InSet Lexicon has an accuracy of 30% while TF-IDF reaches 62% [5]. Extracting sentence structure in Indonesian with a deep learning approach, resulted in an F1 score of 0.7 in testing [6]. Extracting mono lexical terms in Indonesian with a corpus method using AntConc for semi-automatic extraction, the results have the potential to create a broader bilingual terminology dictionary [7]. Collaboration between TF-IDF and linguistic knowledge is effective in extracting Uzbek stopwords with good accuracy [8]. In the application of RNN and transformers to extract attributions and statements of public figures in Indonesian, the test results on the RNN model were 81.34%, while the transformer was 81.01% [9]. Performing extraction in Indonesian to classify hate speech text data, testing the TF-IDF-ICSpF method, and improving KNN showed an average accuracy of 88.11%, 17.81% higher than KNN and TF-IDF [10]. The results show that TF-IDF is effective in extracting Indonesian words.

In solving the research problem, the TF-IDF and LDA steps will be used sequentially to overcome the challenges in Indonesian text analysis. Firstly the TF-IDF method will be used to extract the most relevant keywords from each user review, this will help identify words that have a high weight. Next, the LDA approach is used to perform topic modeling of the user feedback reviews, clustering certain topics to identify naturally occurring patterns and trends in the text data.

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN : 2776-3234

II. RESEARCH METHODOLOGY

To extract, explore, and analyze Gojek user feedback reviews in Indonesian this research will involve a series of systematic stages, the stages are shown in Figure. 1 including data collection, and data preprocessing to prepare the data. Analyzing the distribution to understand emerging trends, word frequency analysis with TF-IDF, topic modeling using the LDA algorithm for clustering words into topics, and finally evaluating the performance of the model that has been developed to ensure the accuracy of the model in the context of extraction and topic modeling on Indonesian text case studies of Gojek application user reviews.

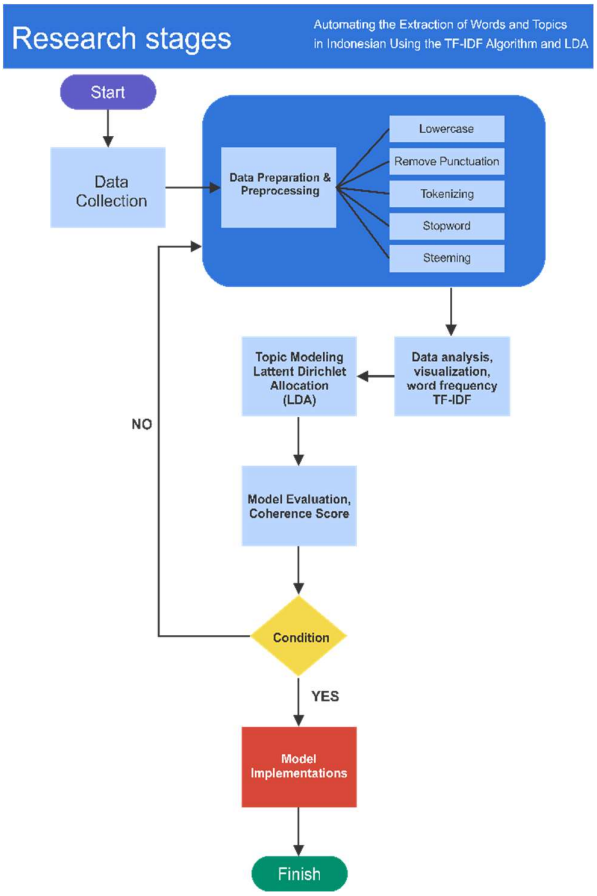

Fig. 1 Research stages

Figure. 1 shows the detailed steps in the research stages. The following is a further explanation for each stage:

1. Data Collection

This stage involves the process of collecting data from various relevant sources, the datasets used in this study are sourced from reviews of the Gojek application versions 4.0.0 to 4.9.3 from 2021 to 2024 [18]. The data spans from Nov 5, 2021, to Jan 2, 2024, totaling 225,002 rows. Each row includes username, content, time, and app version. The focus is on content reviews. The average length is 8 words, with a maximum of 104 and a minimum of a few words. The variability indicates a non-normal distribution. Preprocessing is conducted to maintain topic analysis accuracy.

2. Data preparation and preprocessing

After the data is collected, the next step is to clean and prepare the data for analysis, this stage involves five stages, namely lowercasing, removing punctuation, tokenizing, stopword removal, and stemming.

3. Descriptive Analysis and Data Visualization

This stage includes descriptive statistical analysis to understand the basic characteristics of the data, such as mean, median, and data distribution. In addition, data visualizations such as histograms, bar charts, or scatter plots are also used to visualize the distribution of the data.

4. TF-IDF word frequency

The application of the TF-IDF method in the context of Indonesian word extraction in the case of Gojek application user reviews, is an important step to decipher text complexity and extract information and knowledge. TF-IDF consists of TF measures how often a word appears in a document. The following TF formula (1) is [19]:

$$tf(t,d) = \frac{f_{t,d}}{\sum t^l \in d f_{f^l,d}} \quad (1)$$

Meanwhile, IDF aims to measure how unique or important a word is in the document corpus. The following IDF formula (2) is:

$$idf(t,D) = log\frac{N}{|\{d \in D: t \in d\}|} \quad (2)$$

5. LDA Modeling

In the topic modeling stage for the extraction of Indonesian from Gojek application user reviews, forming groups of interrelated words into meaningful topics using LDA. The LDA formula as described in (3) provides the mathematical foundation underlying the topic formation process.

$$P(z_i = j|z_{-i}, w_i, d_i) \propto \frac{C^{WT}_{w_{ij}} + \eta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\eta} \frac{C^{DT}_{d_{ij}} + \alpha}{\sum_{t=1}^{T} C^{WT}_{d_{it}} + T\alpha} \quad (3)$$

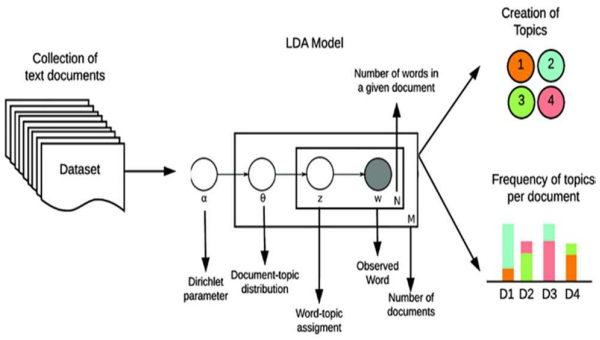

Fig. 2 Lattent Dirichlet Allocation Model [20]

The LDA stage, as described in Figure. 2, includes several important steps in the process. First, there is text to be analyzed, then processed by the LDA model. The Dirichlet parameter is used to define the topic distribution in the document known as document topic distribution, and is also the focus in this stage, it shows the proportion of each topic in the document. Each word in the document is assigned to a specific topic through the word topic assignment process. The words observed in the document are known as observed words. The number of words in a particular document, and the total number of documents in

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN : 2776-3234

the dataset are also important considerations. Form topics from the distribution of topics in the documents and calculate the frequency of occurrence of each topic in the document set, known as the frequency of topics.

6. Topic model evaluation

Topic model evaluation using cohesion score involves measuring how cohesive or related the topics generated by the model are [21]. Once the topic model has generated topics, a cohesion score is calculated for each topic. This involves measuring how often words within each topic appear together in Gojek customer reviews.

III. RESULTS AND DISCUSSION

Descriptive analysis was conducted to explore the data to gain a deep understanding of the data shape, structure, and characteristics contained in the datasets. The amount of data contained in the dataset is 224,913 reviews starting from 2021 to 2024, 2021 is 28,174, 2022 is 124,420, 2023 is 65584, and 2024 is 6,824 reviews. Figure.3 shows the number of datasets by year.

Fig.3 Number starting from 2021-2024

In the app rating analysis, the data shows a varied distribution of ratings given by users. Rating 5 dominates with a percentage of 65%, followed by rating 1 at 20%, rating 2 at 6%, and rating 3 and 4 at 4% each, so in general the application provides high satisfaction to users. Figure.3 is the result of visualizing the level of satisfaction based on users towards the application.

Based on the results of the distribution analysis on the dataset described in Table.1, it is found that the dataset consists of 224,913 entries covering several attributes, including year of publication and score. The average publication year of the data is around 2022 with a standard deviation of 0.70, indicating limited variation in publication year. The year of publication distribution shows that most of the data was published between 2021 and 2024, with the second quartile (median) and third quartile (75%) in 2022 and 2023. The rating scores have an average of 3.93, with most of the data falling in the range of 3 to 5. The second quartile (median) and third quartile (75%) of the scores are 5, indicating a tendency for the majority of users to give high scores. The dataset reflects a varied distribution in its attributes, indicating the complexity of the observed cases.

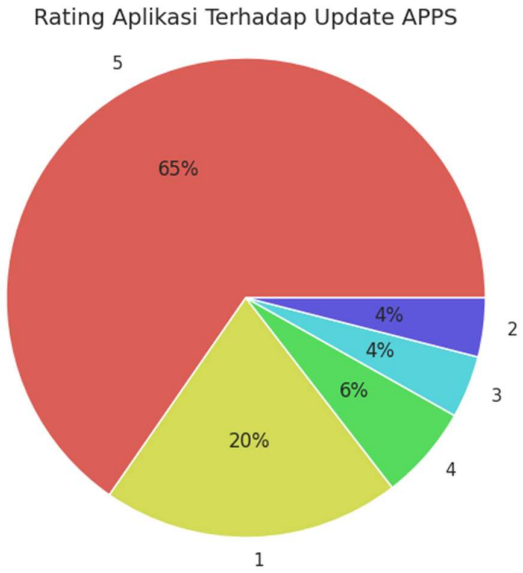

Fig.4 Application user satisfaction rating

Table 1. Data distribution

	Sore	***Publish Year***
Count	224913.000000	224913.000000
Mean	3.929061	2022.226959
Standar Deviasi	1.630582	0.697499
Minimum	1.000000	2021.000000
25%	3.000000	2022.000000
50%	5.000000	2022.000000
75%	5.000000	2023.000000
Maximum	5.000000	2024.000000

After going through the data preprocessing process, the text data is ready for further analysis. At this stage, each step aims to prepare the text optimally so that it can be processed efficiently and accurately. The next stage is making wordcloud for representative visualization of the dominant words in the text, as well as calculating the frequency of word occurrence using the Term Frequency-Inverse Document Frequency (TF-IDF) method to evaluate the importance of words in the text. Based on the results of wordcloud visualization in Figure. 5, it can be concluded that the keywords are very dominant in the analyzed text. Words that appear more often give the most prominent topics or themes in the text.

In the results of wordcloud analysis, the dominant words are Gojek, driver, already, thank you, promo, steady, disappointed, really, good, like, discount, please, difficult. In the wordcloud visualization, the analysis highlights the words that dominate in the text. However, there is no objective assessment because there is no weight on each word. To overcome this, it is necessary to analyze using the TF-IDF method to assess the importance of a word based on the frequency of occurrence of the word in the document as well as the uniqueness of the word among all documents.

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN : 2776-3234

Fig. 5 Wordcloud visualization results

The results of analysis using the TF-IDF method provide a deep understanding of the importance of each word in the text being analyzed. Figure. 6 explains 20 types of words that have a high weight in analysis using TF-IDF. The TF-IDF visualization results provide a clear understanding of the order of keywords based on the highest weight, which includes words such as "gojek, application, helpful, driver, good, great, gopay, good, really, already, please, use, fast, accurate, expensive and difficult." By looking at these words, you can quickly identify the most dominant topics or themes in the document, thus providing valuable insight into understanding the essence of the text and allowing readers (stakeholders) to capture the essence of the message that application users want to convey.

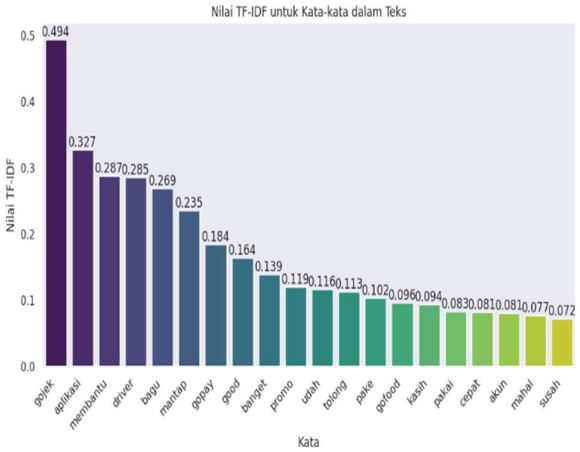

Fig. 6 Results of word weight analysis in the document

The choice is to proceed to the topic modeling stage using Latent Dirichlet Allocation (LDA) after analysis using TF-IDF to gain a deeper understanding of the underlying topic structure in the text. TF-IDF analysis provides insight into the importance of words in a text based on frequency of occurrence but does not provide an understanding of how these words are tied to each other in the context of a broader topic or theme. By involving LDA, hidden patterns in documents can be identified that group words into related topics. Further analysis by applying LDA makes it possible to get a more holistic picture of the text content and identify the main topics so as to gain deeper knowledge about the meaning of the text being analyzed. Table 2 Results of topic modeling into 7 topics with 10 words in each topic, showing variations in user sentiment and experience with Gojek services. While there are some topics that reflect positive sentiments, such as convenience, speed, and service benefits, there are also topics that describe negative experiences, such as high costs or difficulties with app features. Figure 4.5 is the result of the topic modeling visualization. This conclusion illustrates the importance of continually paying attention to user feedback and making necessary improvements to improve the overall user experience.

Table 2. Topic Modeling

Topic number	Word	Topic interpretation
Topic 0	0.059*"pua" + 0.038*"upgrad" + 0.038*"murah" + 0.036*"mantul" + 0.032*"layanan" + 0.030*"pelayanannya" + 0.027*"terbantu" + 0.027*"skrg" + 0.023*"jasa" + 0.023*"enak"	This topic relates to user experiences related to Gojek service improvements or upgrades, such as an increase in service quality at an affordable price. Words such as "pelayanan" and "terbantu" indicate that users feel helped or satisfied with the services they receive.
Topic 1	0.038*"bermanfaat" + 0.035*"senang" + 0.034*"nice" + 0.032*"gojek" + 0.030*"best" + 0.023*"mempermudah" + 0.020*"error" + 0.020*"menyenangkan" + 0.017*"menu" + 0.016*"muncul"	Keywords in this topic include "bermanfaat", "senang", "nice", and "mempermudah". This topic shows users' positive sentiments towards Gojek services, focusing on the ease of use and benefits provided by the application.
Topic 2	0.065*"promo" + 0.050*"gojek" + 0.047*"akun" + 0.031*"voucher" + 0.024*"pakai" + 0.022*"bintang" + 0.018*"promonya" + 0.018*"diskon" + 0.018*"kasih" + 0.017*"udah"	This topic includes key words such as "promo", "akun", "voucher", and "diskon". This topic relates to the use of promos or discounts in Gojek services, such as the use of promo codes or vouchers to get discounts.
Topic 3	0.070*"banget" + 0.045*"mahal" + 0.038*"biaya" + 0.035*"aplikasi" + 0.026*"pake" + 0.024*"harga" + 0.021*"ongkir" + 0.019*"beli" + 0.018*"gofood" + 0.018*"coba"	Key words in this topic include "banet", "mahal", "biaya", and "aplikasi". This topic describes negative user sentiments related to the high price or cost of using Gojek services, as well as problems or inconveniences in using the application.
Topic 4	0.353*"bagu" + 0.059*"pelayanan" + 0.036*"aplikasinya" + 0.032*"aplikasi" + 0.027*"banget" + 0.023*"jelek" + 0.017*"buruk" + 0.015*"payah" + 0.012*"parah" + 0.009*"siip"	This topic is dominated by the keyword "bagus", Other keywords include "pelayanan" and "aplikasi". This topic may relate to a positive evaluation or praise of Gojek's service quality or application.

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN : 2776-3234

Topic 5	0.073*"gopay" + 0.048*"updat" + 0.038*"aplikasi" + 0.033*"gojek" + 0.029*"saldo" + 0.028*"masuk" + 0.021*"udah" + 0.020*"suka" + 0.017*"tolong" + 0.016*"buka"	Keywords in this topic include "gopay", "update", "aplikasi", and "suka". This topic relates to the use of the GoPay payment feature in the Gojek application, as well as updates related to the feature.
Topic 6	0.382*"mantap" + 0.079*"mudah" + 0.057*"memuaskan" + 0.039*"mantab" + 0.021*"jalan" + 0.021*"tingkatkan" + 0.013*"penumpang" + 0.010*"mantappp" + 0.010*"prose" + 0.009*"ngak"	This topic is dominated by the keywords "matap", "mudah", and "memusakan". Other keywords include "road" and "improve". This topic shows the positive sentiment of users regarding the easy and satisfying experience of using Gojek services.

With a coherence score of 0.509, it can be concluded that the LDA model has a moderate level of coherence. This shows that the topics generated by the model tend to have a fairly good relationship between the words in each topic. Nevertheless, improvement is still needed in terms of increasing the coherence of the model, either by adjusting the model parameters or performing better data preprocessing. Higher coherence will ensure that the topics generated are easier to understand and more meaningful, making it easier to analyze and interpret the result. The topic analysis uncovers patterns in Gojek user reviews. The first topic discusses experience, services, and affordable pricing. The second highlights app usability and benefits. The third relates to promos, discounts, and vouchers. The fourth reflects positive evaluations of service quality. However, the fifth focuses on high costs and app issues. The sixth emphasizes overall user satisfaction and service convenience. It effectively outlines the six main topics identified in the analysis, each focusing on different aspects such as user experience, pricing, app usability, promotions, service quality, and overall satisfaction. This summary offers a clear understanding of the key themes discussed within the reviews, highlighting both positive and negative aspects of the service.

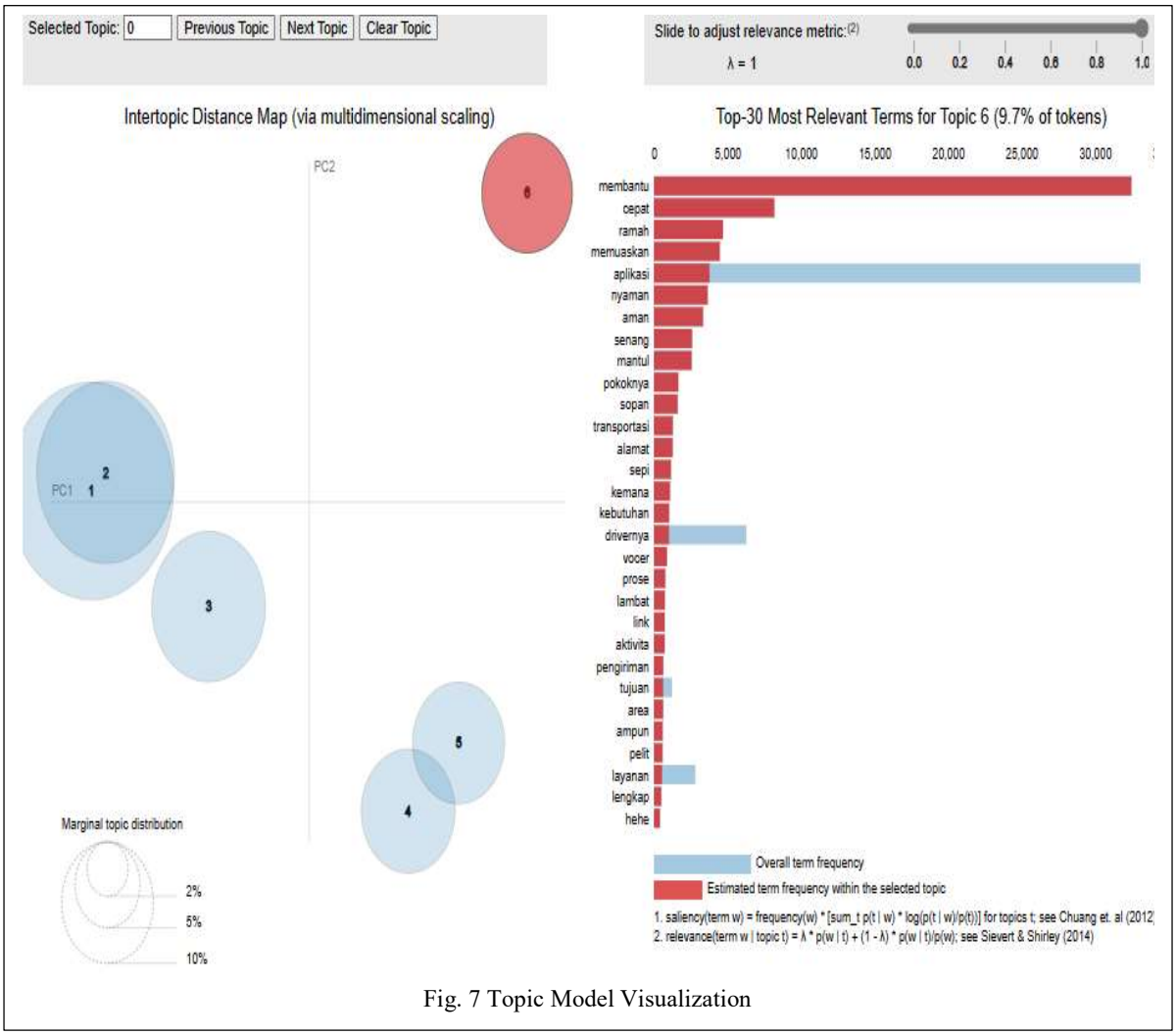

Fig. 7 Topic Model Visualization

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN : 2776-3234

IV. CONCLUSION

This research highlights the rapid growth of digital data in Indonesia, particularly in the use of platforms such as Gojek, with annual growth reaching 64.4% by 2022. Driving factors include the popularity of social media, e-commerce, and ride-hailing services. Gojek is a clear example of this growth with an average of nearly one million downloads each month, making it one of the largest platforms in Indonesia. Analysis of Gojek's usage patterns and preferences is important to improve user experience and services. Text analysis, especially from user reviews, is crucial in understanding user preferences and needs. Keyword extraction and topic modeling play an important role in this regard. The combination of TF-IDF and LDA is proposed to address the challenges of text analysis in Indonesian. The results show the effectiveness of this combination in improving the understanding of text analysis, especially on Gojek user review data. Topic modeling results show variations in user sentiment towards Gojek services. While there are positive sentiments such as convenience and speed of service, there are also issues such as high costs and difficulties with application features. Nonetheless, improving the coherence of the LDA model is necessary to ensure more meaningful and understandable analysis results for readers or interested stakeholders.

REFERENCES

[1] F. Tjandradinata, J. J. Q. Yap, and S. Putra, "Indonesia Big Data and Analytics Software Market Grew," *IDC Corporate*, 2023. https://www.idc.com/getdoc.jsp?containerId=prAP50219423 (accessed Mar. 25, 2024).

[2] N. N. Arief and A. Gustomo, "Analyzing the impact of big data and artificial intelligence on the communications profession: A case study on Public Relations (PR) Practitioners in Indonesia," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 10, no. 3, pp. 1066–1071, 2020, doi: 10.18517/ijaseit.10.3.11821.

[3] E. Santika, "Aplikasi Transportasi Online Terbanyak Diunduh di RI 2023," *Kata Data*, 2024. https://databoks.katadata.co.id/datapublish/2024/01/23/aplikasi-transportasi-online-terbanyak-diunduh-di-ri-2023-gojek-juaranya (accessed Mar. 27, 2024).

[4] H. T. Y. Achsan, H. Suhartanto, W. C. Wibowo, D. A. Dewi, and K. Ismed, "Automatic Extraction of Indonesian Stopwords," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, pp. 166–171, 2023, doi: 10.14569/IJACSA.2023.0140221.

[5] A. Nurkasanah and M. Hayaty, "Feature Extraction using Lexicon on the Emotion Recognition Dataset of Indonesian Text," *Ultim. J. Tek. Inform.*, vol. 14, no. 1, pp. 20–27, 2022, doi: 10.31937/ti.v14i1.2540.

[6] J. Petrus, Ermatita, Sukemi, and Erwin, "A Novel Approach: Tokenization Framework based on Sentence Structure in Indonesian Language," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, pp. 541–549, 2023, doi: 10.14569/IJACSA.2023.0140264.

[7] W. Maulana, E. Setia, and T. Lubis, "Corpus-Based Terms Extraction in Linguistics Domain for Indonesian Language," *J. Kata*, vol. 6, no. 2, pp. 257–270, 2022, doi: 10.22216/kata.v6i2.908.

[8] K. Madatov, S. Bekchanov, and J. Vičič, "Dataset of stopwords extracted from Uzbek texts," *Data Br.*, vol. 43, 2022, doi: 10.1016/j.dib.2022.108351.

[9] Y. S. Yohanes, Y. J. Kumar, N. Z. Zulkarnain, and B. Raza, "Extraction and attribution of public figures statements for journalism in Indonesia using deep learning," *Knowledge-Based Syst.*, vol. 289, no. October 2023, p. 111558, 2024, doi: 10.1016/j.knosys.2024.111558.

[10] N. A. Saputra, K. Aeni, and N. M. Saraswati, "Indonesian Hate Speech Text Classification Using Improved K-Nearest Neighbor with TF-IDF-ICSρF," vol. 11, no. 1, pp. 21–30, 2024, doi: 10.15294/sji.v11i1.48085.

[11] J. H. Lee and M. J. Ostwald, "Latent Dirichlet Allocation (LDA) topic models for Space Syntax studies on spatial experience," *City Territ. Archit.*, vol. 11, no. 1, p. 3, 2024, doi: 10.1186/s40410-023-00223-3.

[12] J. Akbar, T. A. M., Y. Tolla, A. E. Ahmad, A. Yaqin, and E. Utami, "Pemodelan Topik Menggunakan Latent Dirichlet Allocation pada Ulasan Aplikasi PeduliLindungi," *InComTech J. Telekomun. dan Komput.*, vol. 13, no. 1, p. 40, 2023, doi: 10.22441/incomtech.v13i1.15572.

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.

e-ISSN: 2614-8404
p-ISSN : 2776-3234

[13] M. A. Khadija and W. Nurharjadmo, "Enhancing Indonesian customer complaint analysis: LDA topic modelling with BERT embeddings," *Sinergi (Indonesia)*, vol. 28, no. 1, pp. 153–162, 2024, doi: 10.22441/sinergi.2024.1.015.

[14] D. Yu and B. Xiang, "Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling," *Expert Syst. Appl.*, vol. 225, no. September 2022, p. 120114, 2023, doi: 10.1016/j.eswa.2023.120114.

[15] S. E. Uthirapathy and D. Sandanam, "Topic Modelling and Opinion Analysis on Climate Change Twitter Data Using LDA and BERT Model.," *Procedia Comput. Sci.*, vol. 218, no. 2022, pp. 908–917, 2022, doi: 10.1016/j.procs.2023.01.071.

[16] G. S. Buana, R. Tyasnurita, N. C. Puspita, R. A. Vinarti, and F. Mahananto, "Text-Based Content Analysis on Social Media Using Topic Modelling to Support Digital Marketing," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 1, pp. 88–95, 2024, doi: 10.62527/joiv.8.1.1636.

[17] M. A. E. Ignaco and M. A. Ballera, "Optimize Searching Using Latent Dirichlet Allocation," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 3s, pp. 161–166, 2024.

[18] A. Zidane, "Gojek App Reviews Bahasa Indonesia," *Kaggle*, 2024. https://www.kaggle.com/datasets/ucupsedaya/gojek-app-reviews-bahasa-indonesia (accessed Mar. 25, 2024).

[19] M. Kumar and R. Vig, "Term-Frequency Inverse Document Frequency Definition Semantic (TIDS) Based Focused Web Crawler," in *Global Trends in Information Systems and Software Applications*, New Yoork: Springer, 2011, pp. 31–36.

[20] M. Bakrey, "All about Latent Dirichlet Allocation (LDA) in NLP," *Medium*, 2023. https://mohamedbakrey094.medium.com/all-about-latent-dirichlet-allocation-lda-in-nlp-6cfa7825034e (accessed Apr. 02, 2024).

[21] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

JISA (Jurnal Informatika dan Sains) (e-ISSN: 2614-8404) is published by Program Studi Teknik Informatika, Universitas Trilogi
under Creative Commons Attribution-ShareAlike 4.0 International License.