SINERGI Vol. 28, No. 1, February 2024: 153-162
http://publikasi.mercubuana.ac.id/index.php/sinergi
http://doi.org/10.22441/sinergi.2024.1.015

Enhancing Indonesian customer complaint analysis: LDA topic modelling with BERT embeddings

Mutiara Auliya Khadija[1*], Wahyu Nurharjadmo[2]
[1]Faculty of Vocational School, Universitas Sebelas Maret, Indonesia
[2]Department of Public Administration, Faculty of Social and Political Sciences, Universitas Sebelas Maret, Indonesia

Abstract
Social media data can be mining for recommended systems to know the best trends or patterns. The customers have the freedom to ask questions about the product, tell their demands, and convey their complaints through social media. By mining social media data, companies can gain valuable insights into customer preferences, opinions, and sentiments. This information can be utilized to improve products and services, tailor marketing strategies, and enhance overall customer satisfaction. Topic modelling is a text mining technique that extracts the content from the raw and unlabelled data. Latent Dirichlet Allocation is popular for topic modelling research cause flexible and adaptive. But that method has issues with sparsity, performs poorly when documented in the short text and there is no correlation between topics that are actually important in text data. BERT is Bidirectional Encoder Representations from Transformer is designed to pre-train deep bidirectional representations from unlabelled text. The result of this research proves that Latent Dirichlet Allocation and BERT can be arranged on the topic of Indonesian customer complaints. BERT-Base Multilingual Cased and LDA have the highest coherence score. The combination of BERT-Base Multilingual Uncased and LDA has the highest silhouette score. BERT Multilingual are potential for improving the LDA method for Indonesian customer complaints topic modelling.

Keywords:
BERT embeddings;
Enhancing analysis;
Indonesian customer complaints;
Latent Dirichlet Allocation;
Topic modelling;

Article History:
Received: August 11, 2023
Revised: October 10, 2023
Accepted: October 18, 2023
Published: February 2, 2024

Corresponding Author:
Mutiara Auliya Khadija,
Faculty of Vocational School,
Universitas Sebelas Maret,
Indonesia
Email:
mutiaraauliya@staff.uns.ac.id

INTRODUCTION

Social media is a suitable place to get an information from customers. Social media, such as Twitter, Instagram or Facebook can be used as a means of interaction with customer, business partners and also the competitors. Social media also can be used to determined, manage and determined the customer complaints [1]. The customers have freedom for asking the question about product, for telling their demands, convey their opinions or complaints. That considered important for the business or company for reach the customer satisfaction [2]. Considering that benefit of social media, further research and investigation is needed to understand how people will engage the product and also known the problem of customer. That can be used for handle customer comments and customer complaints [3]. This social media data can be mining for recommended system to know the best trends or pattern. The example of research that using social media data are sentiment analysis, opinion mining through sentiment analysis and topic modelling [4], and also topic modelling for getting customer opinion in ride hailing service [5].

Furthermore, social media data can also provide valuable insights for market research and competitor analysis. By monitoring discussions and interactions on social media platforms, businesses can gain a better understanding of customer preferences, identify emerging trends,

and keep track of their competitors' activities. This information can help businesses make informed decisions, improve their products or services, and stay ahead in the competitive market. In addition to benefiting businesses, social media also plays a significant role in empowering consumers. The ease of access and communication on social media platforms allows customers to voice their opinions and experiences publicly, influencing the reputation of a brand or company [6]. Companies that actively listen and respond to customer feedback demonstrate their commitment to customer satisfaction, which can positively impact brand loyalty and attract new customers. However, it is essential to recognize the challenges and ethical considerations associated with mining social media data [7]. Privacy concerns, data security, and ensuring compliance with regulations are critical aspects that businesses must address when using social media data for research or customer engagement.

Data mining has been applied to predict impact of social media and also gain feedback from customer. But social media data consists of text data that include unstructured data [5]. Hence, to negotiate with unstructured data, there are several technologies have been applied such as topic modelling and text mining especially in marketing and business related. Topic modelling is techniques for generate document in the form of keywords that can be indexing according user needs. Topic modelling is an approach used in the field of natural language processing to explore the associations among documents by assuming that words with comparable meanings often occur in similar contexts [6].

In the context of social media data mining, topic modelling plays a crucial role in identifying the most relevant and trending topics discussed by users. By analyzing large volumes of text data, topic modelling algorithms can extract key themes and patterns, allowing businesses to understand the interests and concerns of their target audience better. This information can be invaluable for shaping marketing strategies, content creation, and product development to align with customer preferences and demands [10]. Text mining techniques complement topic modelling by delving deeper into the textual content of social media posts and comments. Text mining involves extracting valuable insights from unstructured data, including sentiment analysis, entity recognition, and opinion mining. Sentiment analysis, for instance, helps businesses gauge the overall sentiment of customer opinions towards their products or services, enabling them to respond promptly to negative feedback or capitalize on positive sentiment [8].

As technology continues to advance, social media data mining and text analytics will continue to evolve, offering businesses even more sophisticated tools to harness the power of social media data. The integration of these techniques into marketing and business strategies will facilitate better decision-making, enhance customer engagement, and ultimately lead to more successful and customer-centric businesses in the digital age [9].

There are several studies about topic modelling. Using Latent Dirichlet Allocation (LDA) for getting insight using Vkontakte social network in Russian language. Using LDA method for modelled the topic and produce clustering in each topic. They use Indonesian Twitter data with 4 different topics about economy, military, sports and technology [10].

LDA is basic approach for topic modelling. This method can understand unstructured data without labelled. But this method has issues with sparsity when encountering large vocabulary size in the sentences. LDA has certain assumptions in its computations that may not be conducive to producing precise or realistic topic models when applied to real-world datasets. Another issue is about the correlation between topics. LDA will produce some keyword that independent each other. LDA inherit to the Dirichlet probability and the independence of words [12].

To address the issues of sparsity and improve the performance of LDA [13], researchers have proposed various enhancements and extensions. One common approach is to incorporate word embeddings or pre-trained language models into the LDA framework. By leveraging the semantic relationships between words captured by these embeddings, the sparsity problem can be alleviated, leading to more coherent and informative topic models. Additionally, incorporating external knowledge sources or domain-specific constraints into the LDA model can help guide the topic modeling process and produce more accurate and meaningful results [14].

Another avenue for improving LDA is through the use of Bayesian non-parametric methods, such as Hierarchical Dirichlet Processes (HDP) or Indian Buffet Process (IBP). These methods allow for the automatic determination of the number of topics, addressing the need for a predefined number of topics, which is a limitation of traditional LDA. Moreover, non-parametric models can capture the hierarchical

structure of topics and account for topic correlations more effectively [9].

Despite the challenges and limitations, LDA remains a fundamental and widely used technique in topic modeling. Its simplicity, interpretability, and ability to uncover latent structures in unstructured data make it a popular choice for many applications [15]. However, as the field of natural language processing continues to evolve, researchers are continually exploring new methods and approaches to enhance topic modeling techniques, providing more robust and accurate insights into the underlying patterns and themes present in large text corpora [16].

For improve the topic modelling method, there are several research [14], conducted a study where they integrated a BERT-based multiclassification algorithm, utilizing supervised learning, with a Probabilistic and Semantic Hybrid Topic Inference (PSHTI) model, employing unsupervised learning, to identify topics in feedback and support data. Conducted a study where they integrated monolingual and multilingual topic analysis by employing LDA and BERT embeddings on both English and Chinese datasets [15].

In this paper, main contribution is to proposed a method topic modelling for Indonesian customer complaint using *Latent Dirichlet Allocation* and BERT embeddings. BERT is pretrained Bidirectional Encoder Representations from Transformers from Google Research. The researcher will use BERT-Base Multilingual Cased and Uncased for handling Indonesian customer complaints data in topic modelling based. The dataset are Domain and Hosting message that represent the complaint of the product. The goal of this research is proof the state-of-the-art technique that combined *Latent Dirichlet Allocation* and BERT to arrange customer complaint.

Topic Modeling and Latent Dirichlet Allocation (LDA)

Topic modelling is the technique that uses text clustering to organize some topic related with documents in to one place. Topic modelling is for getting the latent semantics in the corpus, that beneficial to identify topics better than text clustering. Topic modelling suitable for handling unstructured text documents. Topic modelling is an unsupervised learning that extracts the content from the raw and unlabelled data. The topics are generated with most frequently and likely occurring words [16].

One of popular topic modelling technique is LDA. LDA is handling topic modelling using probabilistic approaches discover latent topic. In LDA method assumes that documents are generated based on latent topics, with words having their own probabilities associated with these topics. LDA is a statistical model used for topic modeling in natural language processing. It is a generative probabilistic model that represents a corpus as a collection of random mixtures over latent topics. LDA is popular for topic modelling research cause flexible and adaptive if compare with text clustering. LDA considers the documents to be produced from random combinations of underlying topics, which are viewed as probability distributions over words [19].

Topic modelling, specifically LDA, has proven to be a valuable tool in various fields such as information retrieval, text mining, and natural language processing. It offers a robust approach to uncovering hidden structures and patterns within large and unstructured text datasets. By representing documents as mixtures of latent topics, LDA allows researchers and analysts to gain a deeper understanding of the underlying themes and concepts present in the corpus [20].

One of the significant advantages of LDA is its ability to handle large and diverse datasets effectively. It can process vast amounts of text data from different sources and identify meaningful topics without the need for manual annotation or supervision. This unsupervised learning method is particularly useful in scenarios where labeled training data is scarce or expensive to obtain [18].

Moreover, LDA's probabilistic nature makes it a flexible and adaptive technique for topic modelling. It can easily accommodate new data and dynamically adjust topic distributions as the corpus evolves over time. This adaptability allows researchers to continuously refine and update their topic models, ensuring that they remain relevant and accurate in capturing emerging trends and patterns in the data [19].

Topic modelling, particularly through LDA, has become an indispensable tool for extracting valuable insights and hidden structures from unstructured text data. Its ability to handle large datasets, adapt to changing information, and uncover latent topics make it a powerful technique for various applications in natural language processing, information retrieval, and beyond. As the volume of textual data continues to grow exponentially, topic modelling techniques like LDA will continue to play a crucial role in making sense of vast amounts of information and

aiding decision-making processes across diverse domains.

LDA also faces challenges with sparsity when dealing with a large vocabulary size in corpora. LDA heavily relies on specific assumptions in its calculations, which might not be appropriate for producing precise outcomes when applied to real-world datasets. In addition to issues with instability and parameter inference, the topic modelling approach of LDA tends to yield poor performance when dealing with documents that are insufficiently long or composed of short text [21]. Other issues in LDA is about the independency of the topic. There is no correlation between topics. Actually, the correlation of the topic is important in text data cause here is context based [20].

LDA has limitations in capturing the temporal dynamics of the data, as it treats all documents as independent and static snapshots. In dynamic environments, such as social media, news articles, or online forums, where topics can evolve rapidly over time, this lack of temporal modeling can lead to suboptimal results [21]. Researchers have explored extensions of LDA to address temporal aspects, such as *Dynamic Topic Models* (DTM) and Temporal-LDA (tLDA), which attempt to capture topic evolution over time. However, incorporating temporal dynamics often introduces additional complexity and computational overhead, making it a trade-off between accuracy and efficiency.

The interpretability of LDA-generated topics can be challenging, especially when the model includes a large number of topics. As the number of topics increases, it becomes more difficult for human analysts to make sense of the topics and their associations with specific themes. Techniques such as topic summarization and visualization are employed to improve interpretability, but striking the right balance between the number of topics and interpretability remains an ongoing research challenge.

While LDA is a widely used and powerful topic modeling technique, it is not without its challenges and limitations. Addressing issues related to sparsity, parameter inference, short text documents, and incorporating temporal dynamics will continue to be areas of active research. As the field of natural language processing and topic modeling advances, researchers aim to develop more robust and flexible approaches that can handle the complexities of real-world textual data and improve the overall performance and interpretability of topic modeling techniques [23].

LDA assumes the Dirichlet distribution to obtain the distribution of every document topic. Afterwards, the results of Dirichlet are used to allocate words in the different topics. *Latent Dirichlet Allocation* representation shown in Figure 1. In LDA, the documents are represented by the variable M, while α represents the parameter controlling the distribution of topics in the documents. θ represents the distribution of topics in each document, z represents the assignment of topics for the nth word in the document, and w represents the words. A higher value of α indicates a greater combination of topics in a document, whereas a lower value of α indicates a lesser combination of topics in a document. Additionally, the words found in the documents are denoted by the variable N. The parameter β is used to control the distribution of words in topics, and ϕ represents the distribution of topics. The variable w represents the number of topics and w represents the number of words. A higher value of β signifies a higher mixture of words within a topic, whereas a lower value of β signifies a lower mixture of words within a topic. In that formula there is K which is randomly assign each word in the document to one of k topics.

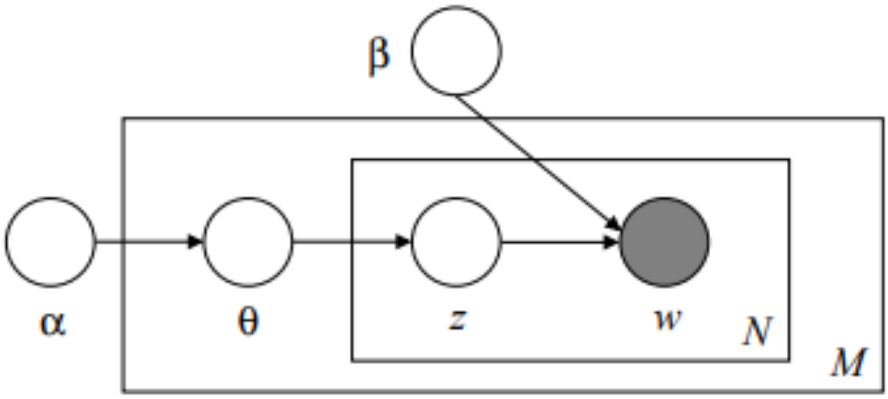

Figure 1. Latent Dirichlet Allocation Representation

Bidirectional Encoder Representations from Transformer (BERT)

BERT stands for *Bidirectional Encoder Representations from Transformer* is designed to pre train deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in layers. The best concept of BERT is the namely the idea of *Masked Language Modelling* (MLM) [24]. BERT extends semi supervised learning by analysing whether two sentences are related or not. Multi-task learning has a positive impact because it can act as a form of control in word prediction and sentence classification. BERT is the development of transfer learning in Natural Language Processing. Some feature extraction that builds transfer learning such as *Term Frequency Inverse Document Frequency* (TF-IDF) and *Bag of Words* (BOW).

Then there is an embedding method that can decide how to convert words into numeric vector forms [13]. Transfer learning has been successfully implemented in computer vision with a pre-training model with ImageNet.

Before BERT, there are several transfer learning models like ULMFiT and ELMo. Embedding from Language Models (ELMo), one of the transfer learning applied to Natural Language Processing, serves as feature extraction. ULMFiT, ELMo and Google BERT focuses on storing knowledge from training and applying I to the different but related to the problem [19]. BERT models consist of several transformer encoders stacked together and designed from unlabelled text to pretrain deep representations in all layers. BERT will trained a large model in 12 layer to 24 layer transformers on large corpus BookCorpus with 800 M words and Wikipedia with 2,500M words [25].

Until 2020, BERT already releases several models:

a. *BERT-Large, Uncased* (Whole Word Masking): 24-layer, 1024-hidden, 16-heads, 340M parameters
b. *BERT-Large, Cased* (Whole Word Masking): 24-layer, 1024-hidden, 16-heads, 340M parameters
c. *BERT-Base, Uncased*: 12-layer, 768-hidden, 12-heads, 110M parameters
d. *BERT-Large, Uncased*: 24-layer, 1024-hidden, 16-heads, 340M parameters
e. *BERT-Base, Cased*: 12-layer, 768-hidden, 12-heads , 110M parameters
f. *BERT-Large, Cased*: 24-layer, 1024-hidden, 16-heads, 340M parameters
g. *BERT-Base, Multilingual Cased*: 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
h. *BERT-Base, Multilingual Uncased*: 102 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
i. *BERT-Base, Chinese*: Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters

BERT-Base consist of 12 layers and BERT-Large consist of 24 layers. *BERT-Base Uncased* means that the text has been lowercased before tokenization processed. And the *BERT-Base Cased* means that true case and accent markers are preserved. *BERT-Base Uncased* is better for *Named Entity Recognition* and *Part of Speech Tagging* [27].

Mining Customer Complaint

Customer review in social media is interesting for analysis. Several customers prefer to post their comment, suggestion also complaint about the product through social media. This costumer review has insights that can be mining for business intelligence and decision-making purpose. Customer complaint gained from review sentences in the social media using linguistic feature-based sentences analysis. Using the customer review data, the stakeholder can know the perception of their customers. Mining customer complaint also can be handling using sentiment analysis.

Sentiment analysis can be done using Machine Learning method like Multinomial Naïve Bayes Classifier that have results positive and negative opinions. Mining customer complaint can be done using clustering model [14].

METHOD

Dataset

The study utilized social media data from DomaiNesia, an Indonesian Domain and Hosting Provider. The data collection process involved gathering customer complaints from direct messages on Instagram and Twitter in August 2020. The researcher employed *Tweepy* for extracting data from Twitter and Selenium for acquiring Instagram direct message data.

Table 1 presents an example of the Indonesian customer complaints dataset used in the research. Number 1 contain the question about an error when uploading website. Number 2, the customer complaints about the difficulties when adding the domain addons even though the domain is active have added it to cPanel. Number 3 is about the example of customer complaints related with the condition of server.

Table 1. Example of Customer Complaints Data

No.	Customer Complaints
1.	Error ketika upload web, gimana ya min?
2.	*Ini saya juga kesulitan untuk menambah addons domain,caranya gimana ya? Padahal domain sudah aktif dan sudah saya tambahkan ke cpanel*
3.	*Min, server domainesia ada masalah ga min? drtd on off (putus nyambung) terus*

Proposed Method

In this paper, the researcher proposed topic modelling for Indonesian customer complaint using Latent Dirichlet Allocation and BERT embeddings. The research framework is depicted in Figure 2, outlining the overall process. Initially, the data of Indonesian customer complaints related to Domain and Hosting will undergo pre-processing. Next, the application of LDA will be employed for topic modelling. The researcher will then utilize BERT-Base Multilingual Cased and BERT-Base Multilingual Uncased to embed the topic modelling. LDA will be used for probabilistic topic assignment vector, while BERT will be used for sentence embedding. Subsequently, the results will be collected for further evaluation by the researcher.

The initial Indonesian customer complaint data will undergo pre-processing, which involves cleaning up the Twitter data prior to entering it into the model also Instagram data. Pre-processing is carried out to remove special characters, filter out specific words, or standardize them. In the pre-processing there are several stages such as remove stop word, remove sentence, case folding, remove non-Ascii, remove URL, remove punctuations, remove digit and number, stemming, slang word conversion.

After pre-processing, the data will undergo feature extraction, where the researcher applies transfer learning from pre-trained transformers.

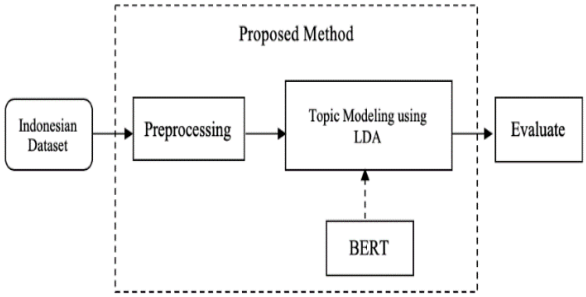

Figure 2. Proposed Method

BERT is utilized in this research, specifically BERT-Base Multilingual Uncased and BERT-Base Multilingual Cased. For handling the Indonesian language, the researcher utilizes BERT-Base Multilingual.

The two types of BERT-Base Multilingual used in this study are as follows:

a. BERT-Base, Multilingual Cased
   Consists of 104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters
b. BERT-Base, Multilingual Uncased
   Consists of 102 languages, 12-layer, 768-hidden, 12-heads, 110M parameters

The researcher used Sentence Transformers, TensorFlow for our base model. And also, Gensim for Latent Dirichlet Allocation, using Pandas, Numpy, Matplotlib for data allocation.

RESULTS AND DISCUSSION

The researcher used 979 sentences in Indonesian customer complaint data after pre-processing process. The researcher is making some experiment in 1000, 1500 and 2000 samples, when number of the topics is 3. Figure 3, shown the result of combination *Latent Dirichlet Allocation* and *BERT-Base Multilingual Cased* and *BERT-Base Multilingual Uncased* clustering. The researcher also compared it with combination of LDA and TF-IDF. The result as shown in Figure 4.

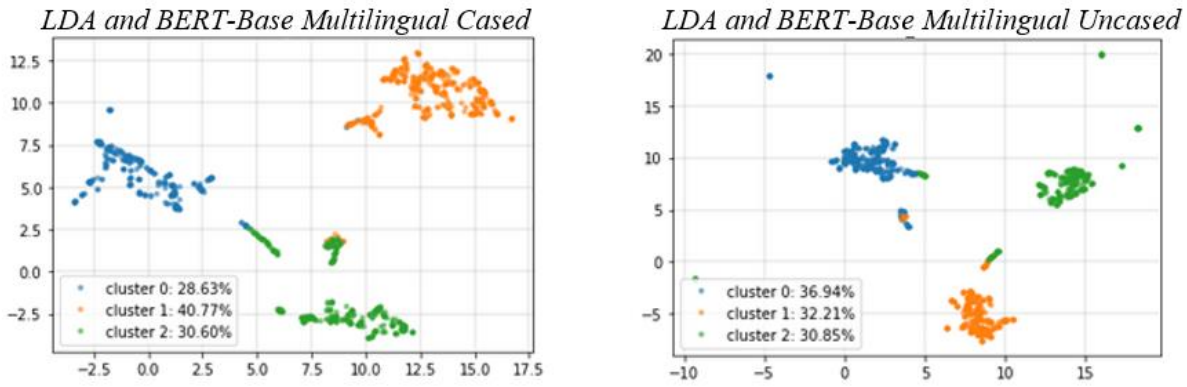

Figure 3. The Clustering Result of Combination Latent Dirichlet Allocation and BERT-Base Multilingual

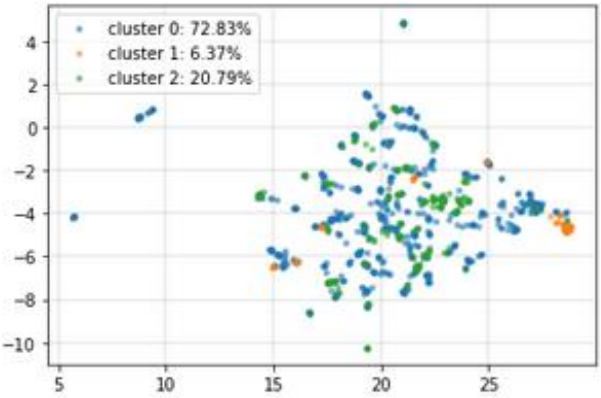

Figure 4. The Clustering Result of Combination Latent Dirichlet Allocation and TF-IDF

For evaluation, the researcher employs coherence score and silhouette score. As presented in Table 2, three different sample sizes, namely 1000, 1500, and 3000, are used for the models, and the results are obtained. The findings indicate that the highest Silhouette score is achieved by *Latent Dirichlet Allocation* and BERT-*Base Multilingual Uncased*. Specifically, in 1000 samples, the researcher obtained a score of 0.603, in 1500 samples a score of 0.616, and in 2000 samples a score of 0.608. While the highest coherence score is obtained by LDA and *BERT-Base Multilingual Uncased*. In 1000 samples is obtained 0.382, in 1500 samples is obtained 0.388, in 2000 samples is obtained 0.346. LDA and TF-IDF get the lowest silhouette score cause TF-IDF is also bag-of words-based word order. That will not separate the words and do not have balanced clusters. TF-IDF also loses the contextual information and the coherence of the words are not be considered as important aspects.

The wordcloud of the result, has shown in Table 3. There are 3 topics in this research. Topic 1, all of our models have concerns about domain complaint. The customer has concerns about how the company can help them for doing something related with domain. The researcher will investigate how to assist them and explore methods to pay off the invoice. In the application of LDA and BERT-*Base Multilingual Uncased* and *Cased*, the topics such as "email", "blog", "data", "bayer", and "basic" are revealed, which were not prominently captured in LDA and TF-IDF models.

That can be shown that embedding words has work for building the contextual topic. Topic 2, all of our models have concerns about term and service and the product error complaint. On LDA and *BERT-Base Multilingual Cased* can show the “password” “wordless” words. Topic 3, has talking about domain transfer, domain error and also domain promo complaint.

As shown in Table 3, the topic found were related to customer complaint are about domain complaint, product error complaint and promo complaint. The customer can complaint because they do not satisfy with the services like server down or website error. With this result, the company can improve their products and services better.

With this result, the topic modelling using LDA and pre-trained BERT has been proven for getting topic in customer complaint. But need some improvement here. As shown in Table 3, in the wordcloud result the researcher shown many words that not suitable and become outliers such as “sudan”, “moon”, “vision”, “mint”, and etc. That is comes from the BERT pre-trained. BERT train in a large model on large corpus Wikipedia and BookCorpus. Although there is *BERT-Base Multilingual* and compatible with Indonesia language, but BERT still need improvement.

Table 2. The Result of Coherence Score and Silhouette Score

Sample	LDA and TF-IDF		LDA and BERT-Base Multilingual Uncased		LDA and BERT-Base Multilingual Cased	
	Coherence Score	Silhouette Score	Coherence Score	Silhouette Score	Coherence Score	Silhouette Score
1000	0.377	0.022	0.292	**0.603**	**0.382**	0.578
1500	0.306	0.019	0.315	**0.616**	**0.388**	0.602
2000	0.307	0.022	0.298	**0.608**	**0.346**	0.542

Table 3. The Wordcloud Result Every Topic

	Topic 1	Topic 2	Topic 3
LDA and TF-IDF			
LDA and BERT-Base Multilingual Uncased			
LDA and BERT-Base Multilingual Cased			

When compared to previous research on customer complaints or customer comments, this research has a higher silhouette score and a higher coherence score. It is shown from previous research [6] that with only the LDA method, topics can be generated but have word limitations. Unlike this research which uses BERT which functions as a sentence embedding vector so that it can enrich each word in the resulting topic. There are previous studies that are also related to the level of customer satisfaction in restaurants using customer review data and the Latent Dirichlet Allocation method [14]. The result shows unstable and the words produced need "translators from experts". In contrast to this research where LDA is combined with BERT as a language embedding which is useful for generating topic models from embedding data [28, 29, 30].

CONCLUSION

The customers have freedom for asking the question about product, for telling their demands, convey their opinions or complaints through social media. And the researcher can mine that social media through topic modelling techniques. Topic modelling is an unsupervised learning that extracts the content from the raw and unlabelled data. The topics are generated with most frequently and likely occurring words. *Latent Dirichlet Allocation* is popular for topic modelling research cause flexible and adaptive. But, LDA has issues with sparsity when encountering large vocabulary size and performs poorly when documents in the short text. There is no correlation between topics that actually important in text data. BERT is *Bidirectional Encoder Representations from Transformer* is designed to pre train deep bidirectional representations from unlabelled text. There are *BERT-Base Multilingual Cased* and *BERT-Base Multilingual Uncased* for handling multi language such as Indonesian language. The result of this research shown that combination between *BERT-Base Multilingual Cased* and LDA have the highest coherence score. And the combination between *BERT-Base Multilingual Uncased* and LDA will have the highest silhouette score.

For future works, there are potential improvements on the models. Further research is required to develop pre-trained models specifically for the Indonesian language. Creating domain-specific pre-trained models could lead to more enhanced results.

ACKNOWLEDGMENT

This work is sponsored by DomaiNesia, that give opportunity to do research. Since 2009, DomaiNesia has been providing reliable, scalable, and secure web hosting solutions.

REFERENCES

[1] N. F. Naini, S. Santoso, T. S. Andriani, U. G. Claudia, and Nurfadillah, "The Effect of Product Quality, Service Quality, Customer Satisfaction on Customer Loyalty," *J. Consum. Sci.*, vol. 7, no. 1, Art. no. 1, Feb. 2022, doi: 10.29244/jcs.7.1.34-50.

[2] R. A. Wayasti, I. Surjandari, and Zulkamain, "Mining Customer Opinion for Topic Modeling Purpose: Case Study of Ride-Hailing Service Provider," in *2018 6th International Conference on Information and Communication Technology (ICoICT)*, May 2018, pp. 305–309. doi: 10.1109/ICoICT.2018.8528751.

[3] X. Duan, J. Li, and Y. Chen, "Analysis of Amazon Market Product Satisfaction Based on LDA Theme Model," in *2020 International*

Conference on Computer Vision, Image and Deep Learning (CVIDL), Chongqing, China: IEEE, Jul. 2020, pp. 693–696. doi: 10.1109/CVIDL51233.2020.00048.

[4] L. C. Cheng and L. R. Sharmayne, "Analysing Digital Banking Reviews Using Text Mining," in *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, The Hague, Netherlands: IEEE, Dec. 2020, pp. 914–918. doi: 10.1109/ASONAM49781.2020.9381429.

[5] P. M. Aji, V. Nadhila, and L. Sanny, "Effect of social media marketing on Instagram towards purchase intention: Evidence from Indonesia's ready-to-drink tea industry," *Int. J. Data Netw. Sci.*, pp. 91–104, 2020, doi: 10.5267/j.ijdns.2020.3.002.

[6] R. Batra and D. Pramod, "Exploring Customer comments using Latent Dirichlet Allocation," in *2021 IEEE Pune Section International Conference (PuneCon)*, Pune, India: IEEE, Dec. 2021, pp. 1–6. doi: 10.1109/PuneCon52575.2021.9686484.

[7] X. Ren, "Application of Apriori Association Rules Algorithm to Data Mining Technology to Mining E-commerce Potential Customers," in *2021 International Wireless Communications and Mobile Computing (IWCMC)*, Jun. 2021, pp. 1193–1196. doi: 10.1109/IWCMC51323.2021.9498773.

[8] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, and Rudy, "News Article Text Classification in Indonesian Language," *Procedia Comput. Sci.*, vol. 116, pp. 137–143, 2017, doi: 10.1016/j.procs.2017.10.039.

[9] E. S. Negara, D. Triadi, and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," in *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, Oct. 2019, pp. 386–390. doi: 10.1109/ICECOS47637.2019.8984523.

[10] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Inf. Syst.*, vol. 94, p. 101582, Dec. 2020, doi: 10.1016/j.is.2020.101582.

[11] M. Hagras, G. Hassan, and N. Farag, "Towards Natural Disasters Detection from Twitter Using Topic Modelling," in *2017 European Conference on Electrical Engineering and Computer Science (EECS)*, Nov. 2017, pp. 272–279. doi: 10.1109/EECS.2017.57.

[12] G. Harshvardhan, M. K. Gourisaria, A. Sahu, S. S. Rautaray and M. Pandey, "Topic Modelling Twitterati Sentiments using Latent Dirichlet Allocation during Demonetization," *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, 2021, pp. 811-815.

[13] Q. Xie, X. Zhang, Y. Ding, and M. Song, "Monolingual and multilingual topic analysis using LDA and BERT embeddings," *J. Informetr.*, vol. 14, no. 3, p. 101055, Aug. 2020, doi: 10.1016/j.joi.2020.101055.

[14] S. Karmakar, N. Sivakumar, and A. S. Pillai, "Exploring Satisfaction Level of Customers in Restaurants by Using Latent Dirichlet Allocation(LDA) Algorithm," in *2023 International Conference on Inventive Computation Technologies (ICICT)*, Lalitpur, Nepal: IEEE, Apr. 2023, pp. 880–887. doi: 10.1109/ICICT57646.2023.10134169.

[15] D. Hendry *et al.*, "Topic Modeling for Customer Service Chats," in *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Depok, Indonesia: IEEE, Oct. 2021, pp. 1–6. doi: 10.1109/ICACSIS53237.2021.9631322.

[16] E. Atagun, B. Hartoka, and A. Albayrak, "Topic Modeling Using LDA and BERT Techniques: Teknofest Example," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, Ankara, Turkey: IEEE, Sep. 2021, pp. 660–664. doi: 10.1109/UBMK52708.2021.9558988.

[17] X. Deng, R. Smith, and G. Quintin, "Semi-Supervised Learning Approach to Discover Enterprise User Insights from Feedback and Support," *ArXiv200709303 Cs Stat*, Jul. 2020, Accessed: Aug. 14, 2020. [Online]. Available: http://arxiv.org/abs/2007.09303

[18] S. Sendhilkumar, M. Srivani, and G. S. Mahalakshmi, "Generation of Word Clouds Using Document Topic Models," in *2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)*, Feb. 2017, pp. 306–308. doi: 10.1109/ICRTCCM.2017.60.

[19] V. Sharifian-Attar, S. De, S. Jabbari, J. Li, H. Moss, and J. Johnson, "Analysing Longitudinal Social Science Questionnaires: Topic modelling with BERT-based Embeddings," in *2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan: IEEE, Dec. 2022, pp. 5558–5567. doi: 10.1109/BigData55660.2022.10020678.

[20] Y. Huang, R. Wang, B. Huang, B. Wei, S. L. Zheng, and M. Chen, "Sentiment Classification of Crowdsourcing Participants' Reviews Text Based on LDA Topic Model,"

IEEE Access, vol. 9, pp. 108131–108143, 2021, doi: 10.1109/ACCESS.2021.3101565.

[21] M. Kretinin and G. Nguyen, "Topic Modeling on News Articles using Latent Dirichlet Allocation," in *2022 IEEE 26th International Conference on Intelligent Engineering Systems (INES)*, Georgioupolis Chania, Greece: IEEE, Aug. 2022, pp. 000249–000254. doi: 10.1109/INES56734.2022.9922609.

[22] D. M. Blei et al., "Latent Dirichlet Allocation," Journal of Machine Learning Research, *The Journal of Machine Learning Research*, vol. 30, pp. 993-1022. 2003.

[23] S. P. M C, B. R. Reddy, D. S. Tharun Reddy, and D. Gupta, "Comparative Analysis of Research Papers Categorization using LDA and NMF Approaches," in *2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon)*, Vijaypur, India: IEEE, Nov. 2022, pp. 1–7. doi: 10.1109/NKCon56289.2022.10127059.

[24] C. Zheng, Z. Wang, and J. He, "BERT-Based Mixed Question Answering Matching Model," in *2022 11th International Conference of Information and Communication Technology (ICTech))*, Wuhan, China: IEEE, Feb. 2022, pp. 355–358, doi: 10.1109/ICTech55460.2022.00077.

[25] 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, pp. 908–917, 2023, doi: 10.1016/j.procs.2023.01.071.

[26] A. Ali, Y. Xia, Q. Umer, and M. Osman, "BERT based severity prediction of bug reports for the maintenance of mobile applications," *J. Syst. Softw.*, vol. 208, p. 111898, Feb. 2024, doi: 10.1016/j.jss.2023.111898.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[28] J. Yang *et al.*, "BERT and hierarchical cross attention-based question answering over bridge inspection knowledge graph," *Expert Syst. Appl.*, vol. 233, p. 120896, Dec. 2023, doi: 10.1016/j.eswa.2023.120896.

[29] I. N. Joharee, N. N. W. Nik Hashim, N. S. Mohd Shah, " Sentiment Analysis and Text Classification for Depression Detection," *Journal of Integrated and Advanced Engineering (JIAE)*, vol. 3, no. 1, pp. 65-78, 2023, doi: 10. 51662/jiae.v3i1.

[30] A. Irwanto and L. Goeirmanto, "Sentiment Analysis from Twitter About Covid-19 Vaccination in Indonesia Using Naive Bayes and Xgboost Classifier Algorithm," *SINERGI*, vol. 27, no. 2, pp. 145-152, 2023, doi: 10.22441/sinergi.2023.2.001