Revealing East Java Community Sentiments Towards Poverty: A Comparative Study Using LDA and BERT

Aries Dwi Indriyanti[1*], *Rahmat Gernowo*[2] and *Eko Sediyono*[3]

[1]Doctoral Program of Information System, School of Postgraduate Studies Diponegoro University Semarang, Indonesia iponegoro University, Information System, 50275 Semarang, Indonesia
[2]Doctoral Program of Information System, School of Postgraduate Studies Diponegoro University Semarang, Indonesia iponegoro University, Information System, 50275 Semarang, Indonesia
[3]Doctoral Program of Information System, School of Postgraduate Studies Diponegoro University Semarang, Indonesia iponegoro University, Information System, 50275 Semarang, Indonesia

Abstract. Through sentiment analysis on the social media platform Twitter, this study explores public opinion on poverty issues in East Java, Indonesia. By understanding public perception, policymakers can develop more effective poverty alleviation strategies. By applying the BERT and LDA models, two dominant themes were identified: public concern about poverty conditions and social comparison between the rich and the poor. The BERT model achieved an accuracy of 75.6%, demonstrating the potential of social media analysis to understand public perception and inform effective poverty alleviation strategies. Despite achieving fairly good accuracy values, it should be noted that data limitations and sample representation may affect the generalizability of the study results. The results of this study indicate that sentiment analysis has significant potential in informing public policy, especially in the context of poverty alleviation. However, limitations such as data quality and representation need to be considered for future research.

1 Introduction

Poverty is a multidimensional social phenomenon, especially in developing countries such as Indonesia. Although various efforts have been made to overcome it, the problem of poverty remains a significant challenge. Operationally, poverty is often defined as the inability to meet basic needs, both food and non-food[1]. The Indonesian Central Bureau of Statistics (BPS) adopts a basic needs approach in measuring poverty, using the poverty line as a monetary threshold. The poverty line reference in Indonesia consists of the food poverty line (GKM) and the non-food poverty line (GKNM) [2].

Based on data from the Central Statistics Agency (BPS), the percentage of poor people in Indonesia decreased in March 2024 to 9.03% or 25.22 million[3]. However, regional disparities are still a significant challenge. East Java's contribution to the total poor population in Indonesia is quite significant, reaching 9.79% or 3,982,690 people. Although there has been a decrease in the poverty rate nationally[4], East Java still faces challenges in reducing poverty rates, especially in rural areas reaching the highest figure in Indonesia, namely 13.3% or 2,399,990 people. This indicates regional disparities that need to be considered in poverty alleviation efforts.

To formulate an effective poverty alleviation strategy in East Java, a deep understanding of public perception is needed. A quantitative approach through sentiment analysis on text data from social media and news can provide significant contributions to a deeper understanding of public perception of poverty in East Java[5]. In this way, we can identify the causes of

* Corresponding author: author@email.org

poverty that are most often voiced by the community, thus enabling the formulation of more targeted policies[6].
Twitter or X is one of the most widely used platforms by the public with around 5.17 billion users to write opinions, criticisms and suggestions regarding government policies[7]. This study uses Twitter as a data source to analyze public perceptions of poverty in East Java. Through sentiment analysis techniques, we aim to identify and classify factors that are considered by the public as the main causes of poverty. Thus, this study is expected to provide empirical contributions to the development of more effective poverty alleviation policies [8]. Recent research utilizes a combination of data science, natural language processing (NLP), data visualization, and social network analysis to extract meaningful information from social media data. By combining topic modeling and sentiment analysis, this study is able to identify trending topics and analyze the polarity of public sentiment towards these topics[9]. Applying this method to Twitter data, for example, allows researchers to understand public perceptions regarding complex issues such as climate change, so that they can inform more effective public policies[5].
In this study, we combine two powerful deep learning models, namely Latent Dirichlet Allocation (LDA) to perform topic analysis[10] and Bidirectional Encoder Representations from Transformers (BERT) for text classification[11]. LDA is used to identify latent themes in the dataset, while BERT, trained on a large-scale text corpus, is able to deeply capture the context of words in sentences. By leveraging BERT's ability to understand natural language, we aim to improve the accuracy of text classification.

1.1 Related Work

The use of the Latent Dirichlet Allocation (LDA) algorithm to identify potential topics related to fake news, as well as more accurate classification and detection of information that spreads or is wrong[12]. By analyzing bold news articles from three leading platforms during the early period of the Covid-19 pandemic. And utilizing the BERT machine learning model, it is able to extract news and investigate the relationship between the pandemic and the dynamics of the money market[13]. Showing the results that there is a significant positive correlation between news and market performance and differences in the influence of sentiment and news categories between media platforms.
In a study conducted by B. Wan, et al. combined BERT with the OCC emotion model and used an adaptive fusion algorithm to integrate emotion knowledge[14]. The experimental results showed that the approach or the ECR-BERT model significantly improved the model's performance in classifying sentiment and providing better explanations. S. Ulthirapathy and D. Sandanam compared the performance of LDA in topic modeling with the BERT model in classifying Twitter user sentiment related to climate change[9]. With the sentiment labeled as pro news, support, neutral, and anti. The uncased BERT model has shown the best results such as precision of 91.35%, recall of 89.65%, and accuracy of 93.50% compared to other methods.

1.1 Literatur Reviews

1.1.1 BERT Model

Introduced in 2018, the BERT (Bidirectional Encoder Representations from Transformers) revolutionized the field of Natural Language Processing (NLP) with its ability to read text bidirectionally, greatly enhancing its understanding of language context[15]. BERT stands for "Bidirectional Encoder Representations from Transformers" and is used to obtain pre-

trained bidirectional representations from text input by combining left and right context conditioning16. Because it is pre-trained, BERT is able to generate context-rich word representations. By simultaneously combining information from previous and subsequent words, BERT is able to capture deeper nuances of meaning in text.

1.1.2 LDA Topic Modelling

The topic modeling process aims to obtain the distribution of words that form a topic and documents with a particular topic. LDA is a probabilistic model used to identify latent topics in a collection of documents. This model assumes that each document is a mixture of several topics, and each topic is a probability distribution over words. The LDA topic model is unsupervised machine learning[17]. The model is useful in identifying hidden information in large collections of documents. The implementation of LDA in Python, by utilizing the Gensim library and the LdaModel module, allows researchers to explore dominant topics in their document collections and visualize the results in graphical form.

2 Research Methodology

The visualization of this research flow can be seen in Figure

Fig. 1. Research Procedure Flowchart

This study adopts a Natural Language Processing (NLP) approach to classify public sentiment regarding the issue of poverty in East Java as reflected in tweets on the social media platform Twitter or X. The methods used include: (1) Sentiment Analysis with the Bidirectional Encoder Representations from Transformers (BERT) model to determine the sentiment polarity (positive, negative, or neutral) of each tweet, and (2) Topic Modeling with the Latent Dirichlet Allocation (LDA) algorithm to identify the main themes that dominate conversations about poverty in East Java. The stages in the research method include:

2.1 Preparation

In the initial stage includes determining the research topic to be implemented, the method to be used, collecting data on community sentiment regarding poverty cases in East Java through the Twitter or X platform. This study intends to explore public perceptions regarding poverty problems in East Java through sentiment analysis on the Twitter social media platform. Raw data is obtained using web scraping techniques using the Twitter API, using certain parameters. Data retrieval is carried out using the Twitter API which allows easier data collection with search parameters such as language, tweet type, and date range [18]. By using these parameters, the data obtained becomes more accurate and in accordance with the objectives of the study.

2.2 Planning

In the second stage, the classification system is designed by determining the model architecture. This stage includes feature extraction, data division into training and test data, and hyperparameter optimization to achieve maximum model performance.

2.3 Programming

The third stage will focus more on the process of compiling program code using the Python language and the help of Google Collaboratory to write and run program code.

2.3.1 Data Preprocessing

The data preprocessing stage is carried out to clean and filter the dataset, so that it produces data that is ready for the model training process. Some commonly used preprocessing techniques include text cleaning, tokenizing, filtering, stopword removal, stemming and Slang Word Removal [19].

2.3.2 Data Labelling

The sentiment classification process for the dataset is carried out by assigning three label classes: negative (-1), positive (1), and neutral (0). Given the limited availability of Indonesian language labeling libraries, the dataset that has been pre-processed is first translated into English. For the efficiency of the translation process, the dataset is divided into several partitions. After the translation is complete, the partitions are combined back into a single dataset. The final stage is the implementation of sentiment labeling using the TextBlob library on the English language dataset that has been formed.

2.3.3 Sentiment Analysis Using BERT Model

The BERT model is adopted to analyze the dataset that has gone through the labeling process. This study aims to train a sentiment analysis model and simultaneously measure the quality of data labeling. The following flowchart illustrates the classification process using the BERT model.

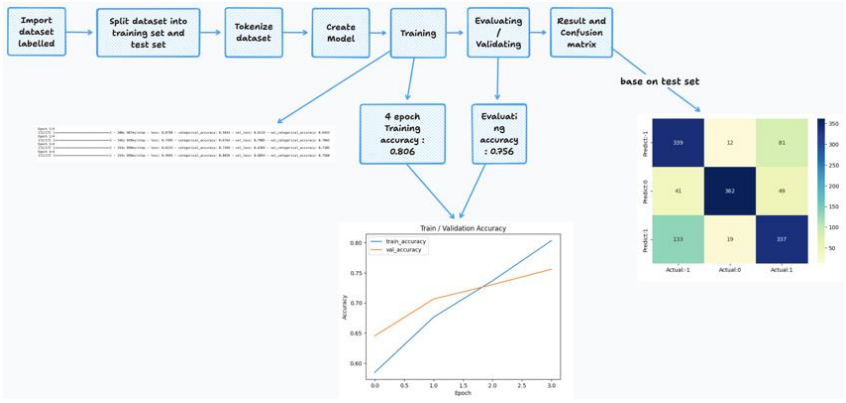

Fig. 2. Flowchart Of Sentiment Analysis Process Using BERT Model

2.3.4 Topic Modelling Using LDA

After data accuracy validation is carried out, the next stage is to group the data (clustering) based on the dominant topics that appear in the dataset. This clustering process will produce 2 to 10 topic groups, from which several groups with the highest coherence scores will be selected. The topic model will then be built based on the two selected topic groups.

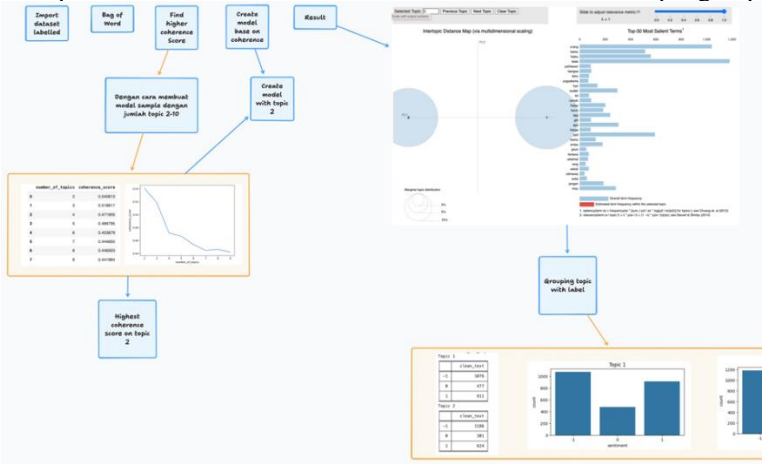

Fig. 3. Flowchart Of Topic Modelling Using LDA

2.4 Testing

After the topic grouping stage (topic modeling) is complete, the next step is to evaluate the sentiment classification model that has been developed. This evaluation will use a test dataset that is independent of the training dataset to measure the model's performance in classifying East Java community sentiments related to poverty issues.

2.5 Evaluation

The final stage of the research process involves an in-depth evaluation of the analyzed data. The findings are then integrated into a broader research framework, opening up opportunities for publication in international journals and intellectual property protection (HKI).

3 Result and Discussion

3.1 Hasil Collection Data

Through web scraping techniques, qualitative data was collected from the social media platform Twitter (now X) to measure public perceptions of poverty issues in East Java. This study used the keywords "poverty" and "poor" in the period from December 30, 2023 to August 17, 2024, resulting in 4,771 tweets. After data preprocessing to eliminate duplication, further analysis was carried out on 4,703 unique tweets.

3.2 Preprocessing Data

As an initial step in sentiment analysis of unstructured text data, a text pre-processing process is carried out. The initial stage involves extracting text features in the form of tweet columns.

Next, a series of text cleaning techniques, tokenization, stopword removal, stemming, and slang word removal are carried out. This process aims to obtain a clean and consistent text representation, so that it is ready for further analysis stages. Table 1 shows an example of the results of data pre-processing.

Table 2. Tweet Results That Have Gone Through Data Preprocessing

Tweet Before	Tweet After
@AsahPolaPikir Lho katanya pengangguran dan kemiskinan makin turun. Kok Bapak berani bilang sebaliknya?	lho kata anggur miskin makin turun kok bapak berani bilang balik
Kini sudah 79 Tahun Indonesia berdiri masih banyak masyarakat di bawah garis kemiskinan Mari kita wujudkan Indonesia Emas dengan memerdekakan diri dari kemiskinan dan berkontribusi terhadap kemajuan Negara Selamat Hari Kemerdekaan Indonesia!	kini tahun indonesia diri banyak masyarakat bawah garis miskin mari wujud indonesia emas merdeka diri miskin kontribusi maju negara selamat hari merdeka indonesia
hingga 5%. Angka kemiskinan diturunkan dlm rentang 7 hingga 8%. Rasio gini dalam kisaran 0 379 hingga 0 382. Indeks Modal Manusia pd level 0 56. Nilai Tukar Petani ditingkatkan dikisaran 115 hingga 120. Nilai Tukar Nelayan dijaga dikisaran 105 hingga 108.	hingga angka miskin turun dalam rentang hingga rasio gin kisar hingga indeks modal manusia pada level nilai tukar tani tingkat kisar hingga nilai tukar nelayan jaga kisar hingga
@5teV3n_Pe9eL Iya kemiskinan makin turun semakin tidak sanggup beli kebutuhan pokok	peel iya miskin makin turun makin sanggup beli butuh pokok

3.3 Labelling Data

The Twitter dataset (now X) consisting of 4,703 tweets about poverty in East Java has been classified into three sentiment classes: negative (2,287 tweets), positive (1,605 tweets), and neutral (811 tweets). The negative class reflects negative public sentiment towards poverty conditions in East Java, while the positive class shows support for the poverty reduction narrative. The neutral class represents tweets that do not explicitly state support or rejection of the poverty issue.

3.4 Analysis Sentiment Using BERT Model

In the initial stage, the dataset will be partitioned into two subsets. As much as 80% of the data will be used to train the model, while the rest (20%) will be used to test the performance of the trained model. This data division process is very important to ensure that the model can generalize well to new data that has never been encountered before. After the data division process, the distribution of sentiment classes (negative, positive, and neutral) will be visualized using the Matplotlib library. This visualization will provide a clearer picture of the proportion of each class in the dataset, thus helping to understand the characteristics of the data that will be used for model training.

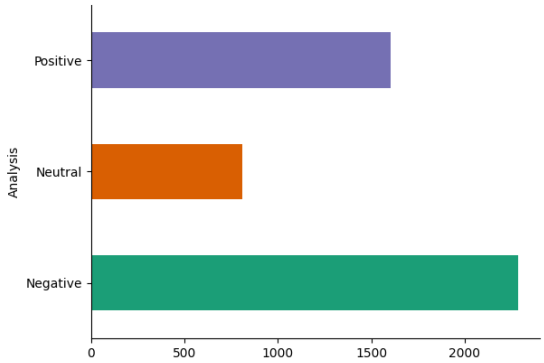

Fig. 4. Visualization of the Number of Negative, Positive and Neutral Data Distribution.

The next step is data modeling, which involves the process of encoding and tokenization. The BERT training model aims to feed the dataset into the model so that it can be trained and tested, resulting in a robust language model that is able to process text naturally. In other words, this model is trained to understand and produce text that is similar to human text.

```
Model: "model"
_________________________________________________________________________________________________
 Layer (type)                Output Shape              Param #   Connected to
=================================================================================================
 input_1 (InputLayer)        [(None, 128)]             0         []

 input_2 (InputLayer)        [(None, 128)]             0         []

 tf_bert_model (TFBertModel  TFBaseModelOutputWithPooli  109482240  ['input_1[0][0]',
 )                           ngAndCrossAttentions(last_            'input_2[0][0]']
                             hidden_state=(None, 128, 7
                             68),
                              pooler_output=(None, 768)
                             , past_key_values=None, hi
                             dden_states=None, attentio
                             ns=None, cross_attentions=
                             None)

 dense (Dense)               (None, 3)                 2307      ['tf_bert_model[0][1]']

=================================================================================================
Total params: 109484547 (417.65 MB)
Trainable params: 109484547 (417.65 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________________________________________
```

Fig. 5. BERT Training Model setup results

Then, after the BERT model is installed correctly, the data computation is continued into the model that has been created with epochs = 4 and batch_size = 32 which refers to [20].

```
Epoch 1/4
172/172 [==============================] - 200s 907ms/step - loss: 0.8799 - categorical_accuracy: 0.5844 - val_loss: 0.8119 - val_categorical_accuracy: 0.6453
Epoch 2/4
172/172 [==============================] - 146s 849ms/step - loss: 0.7495 - categorical_accuracy: 0.6764 - val_loss: 0.7005 - val_categorical_accuracy: 0.7065
Epoch 3/4
172/172 [==============================] - 154s 898ms/step - loss: 0.6233 - categorical_accuracy: 0.7369 - val_loss: 0.6384 - val_categorical_accuracy: 0.7305
Epoch 4/4
172/172 [==============================] - 154s 898ms/step - loss: 0.4845 - categorical_accuracy: 0.8036 - val_loss: 0.6054 - val_categorical_accuracy: 0.7560
```

Fig. 5. Evaluation Results On Bert Training Model For 4 Iterations

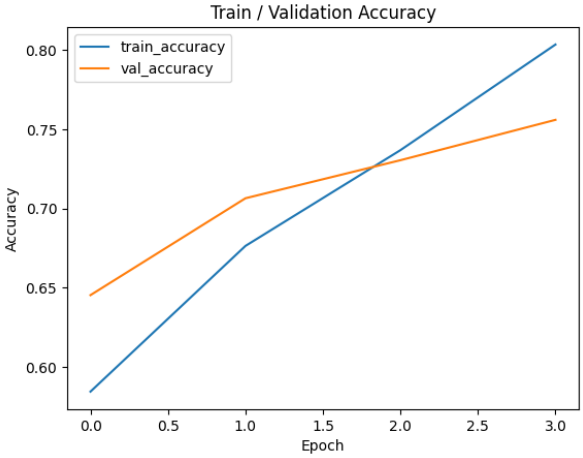

Fig. 6. Score Accuracy Graph

Graph analysis shows that the model has not reached optimal performance. Although the training accuracy reached 0.8 in the 4th epoch, the validation accuracy only reached 0.75. This indicates an overfitting problem, where the model memorizes the training data too much and is less able to generalize to new data. The possible cause is poor data quality, such as inconsistent sentiment labels or noisy data. To overcome this problem, it is recommended to manually relabel the data by humans to make it more accurate and consistent.

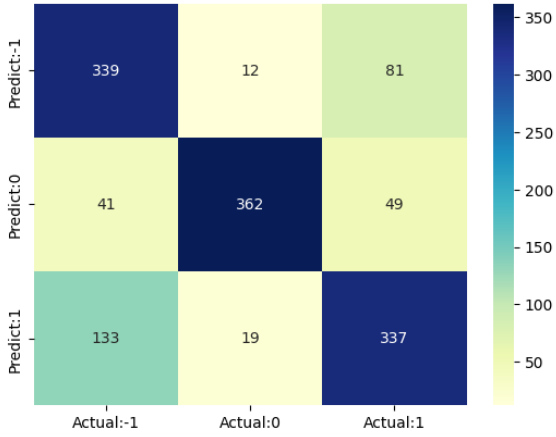

Fig. 7. Result and Confusion Matrix

The evaluation in Figure 7 shows that the TextBlob module successfully classifies data sentiment with an accuracy of 75.6%. This level of accuracy indicates that the model is reliable for automatically tagging text sentiment according to predetermined categories.

3.5 LDA Topic Modelling

The results of the word cloud visualization in Figure 3 confirm that poverty is a central issue in public opinion in East Java. The frequency of occurrence of words such as 'poor' and 'rich' indicates that people feel the direct impact of this problem and consider it an urgent social problem. And the dominance of words related to poverty and wealth shows that government

policies in the economic and social fields need to focus more on poverty alleviation and reducing inequality.

Fig. 8. Tweets Word Cloud View

Referring to the findings of Michael R and Both A et al., the topic coherence score was chosen as a quantitative index of the optimal number of topics in the LDA model[21]. Analysis in Figure 4 shows that the highest topic coherence score of 0.540813 was achieved when the coherence value in the LDA model was 2. However, to ensure that the resulting topics were specific and informative enough, experiments were conducted with various numbers of topics. The experimental results consistently showed that the model with two topics provided the best balance between coherence and information detail. So it can be concluded that choosing 2 topics will produce a model with the best quality in terms of coherence

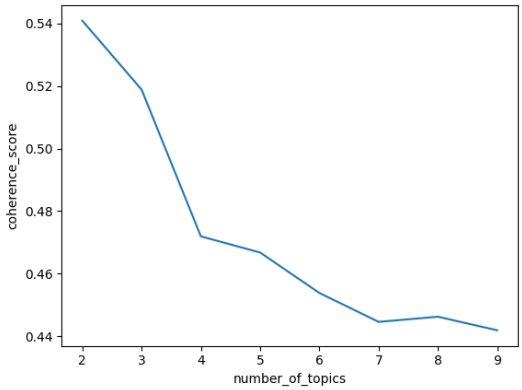

Fig. 9. The coherence score of different number topics

Using the Gensim module to build a bag-of-words model from pre-processed data. Each document in the corpus is represented as a word vector, where each vector dimension represents the frequency of occurrence of a word in the document. This process begins with the formation of a dictionary that maps each unique word to a numeric index.
The LDA model in this study involves determining the Number of topics as much as 2 and the word of topic as much as 10. The modeling results produce a probabilistic distribution of words on each topic, with the highest probability words as topic representations. Table 2 presents the top keywords for each topic. The selection of these parameters, namely the number of topics and the length of words per topic, has significant implications for the quality and interpretability of the modeling results, so it requires careful consideration.

Table 3. Topic Modeling Results Output

Topic	Probabilitas * Kata
0	0.060*"miskin" + 0.019*"yang" + 0.012*"tidak" + 0.011*"jadi" + 0.010*"orang" + 0.006*"kaya" + 0.006*"apa" + 0.006*"rakyat" + 0.005*"banyak" + 0.005*"buat"

1	0.059*"miskin" + 0.020*"tidak" + 0.019*"orang" + 0.012*"yang" + 0.011*"kamu" + 0.011*"kalau" + 0.007*"kaya" + 0.006*"banyak" + 0.006*"sudah" + 0.006*"saya"

Based on the results of topic modeling, researchers conducted further analysis by measuring the probability of word occurrence in each topic. The results of the analysis are presented in Table 4, it was found that both topics were closely related to the problems of poverty and social inequality.

Table 4. Analysis of Topic Modeling Results

Topic	Hasil Analisis Topik berdasarkan Probabilitas Kata
0	Permasalahan kemiskinan yang dialami oleh sebagian besar masyarakat, serta keresahan mereka terhadap kondisi tersebut dan harapan akan adanya perubahan.
1	Membandingkan kondisi hidup antara orang miskin dan kaya, serta mengeksplorasi perasaan pribadi dan harapan individu dalam menghadapi tantangan kemiskinan.

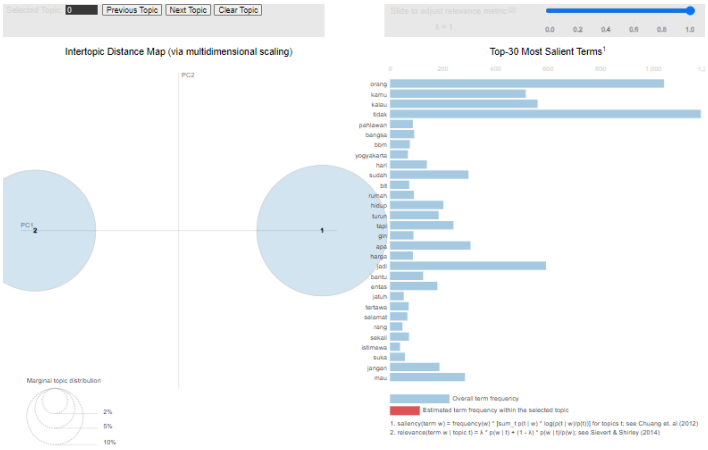

Fig. 10. Poverty Topic Visualization from December 2023 – August 2024

Documents or tweets that have gone through the pre-processing process are then grouped into several topics using the trained LDA model. Each topic group contains 3 documents or tweets. The results of this classification are presented in Table 5 and Table 6.

Table 5. Analysis of Topic 1 Modeling Results

Tweets	Class
@AsahPolaPikir Lho katanya pengangguran dan kemiskinan makin turun. Kok Bapak berani bilang sebaliknya?	-1
Kemiskinan Ekstrem Indonesia mendekati 0 Persen Tahun 2024. #Pidatopresiden2024 NusantaraBaru https://t.co/Ohx8nSD5xm	1
@ardisatriawan Senjata terkuat jokowi itu kemiskinan &amp	0

Table 5. Analysis of Topic 2 Modeling Results

Tweets	Class
yaiyalah tolol itu makanya ada istilah kemiskinan struktural monyettttt ya emang tugas pemerintah dgn duit pajak dan resources mereka bisa bikin aturan ina inu gini aja ga paham bener emang masyarakat kyk elu gini yg bego	-1
ya pemerintahnya sendiri yg memelihara kemiskinan struktural ini lol	1
Tau kemiskinan struktural ga?	0

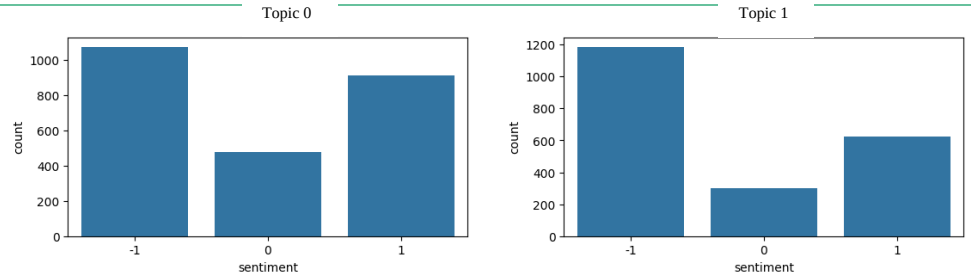

Fig. 11. Visualization of the Distribution of the Number of Data on Topics 1 and 2 in Tweets Data

The visualization in Figure 7 shows an unbalanced distribution of sentiment classes in each topic. Topic 0, whether in the negative, positive, or neutral classes, is dominated by a very high number of tweets from the negative class of Topic 1.

4 Conclusion

This study contributes to a better understanding of public perceptions of poverty in East Java. LDA was used to identify the main topics discussed, while BERT was used to classify the sentiment of each tweet into positive, negative, or neutral. Through sentiment analysis of Twitter data, we successfully identified the main topics and dominant sentiments. The classification model we developed, which combines LDA and BERT, proved effective in classifying sentiments with a high accuracy of 75.6% and found 2 topic findings with the highest coherence value of 0.540813. The results of this study can be a basis for the government and other stakeholders in designing more targeted policies to address poverty issues.

References

1. Badan Pusat Statistik. Penjelasan Data Kemiskinan. *Press Release BPS* 1–2 (2011).
2. Kurnianingsih, T. Dimensi Kemiskinan. *Biro Anal. Anggar. dan Pelaks. APBN DPR RI* 47–56 (2012).
3. BPS Provinsi Jawa Timur. Profil Kemiskinan Di Jawa Timur Maret 2024. 1–7 (2024).
4. BPS. Persentase Penduduk Miskin (P0) Menurut Provinsi dan Daerah (Persen), 2024. https://www.bps.go.id/id/statistics-table/2/MTkyIzI=/persentase-penduduk-miskin--p0--menurut-provinsi-dan-daerah.html.
5. Nurmawati, E. & Amanda, A. Analisis Sentimen Dan Pemodelan Topik Pada Tweet Terkait Data Badan Pusat Statistik. *J. Sist. Inf. dan Inform.* **6**, 165–176 (2023).
6. Depaula, N. & Harrison, T. The EPA under the Obama and Trump administrations : Using LDA topic modeling to discover themes , issues and policy agendas on Twitter. 1–29 (2018).
7. Reportal, D. Global Social Media Statistic. https://datareportal.com/social-media-users (2024).
8. Ayong, N., Azizah, N., Zakiyyah, A., Nur, M. & Pendahuluan, A. Faktor-Faktor yang Memengaruhi Kemiskinan di 38 Kabupaten / Kota Jawa Timur. 222–232 (2024).
9. Uthirapathy, S. E. & Sandanam, D. Topic Modelling and Opinion Analysis on Climate Change Twitter Data Using LDA and BERT Model. *Procedia Comput.*

Sci. **218**, 908–917 (2022).

10. Ali, T., Omar, B. & Soulaimane, K. Analyzing tourism reviews using an LDA topic-based sentiment analysis approach. *MethodsX* **9**, 101894 (2022).
11. Islam, M. J., Datta, R. & Iqbal, A. Actual rating calculation of the zoom cloud meetings app using user reviews on google play store with sentiment annotation of BERT and hybridization of RNN and LSTM. *Expert Syst. Appl.* **223**, 119919 (2023).
12. Nair, V., Pareek, D. J. & Bhatt, S. A Knowledge-Based Deep Learning Approach for Automatic Fake News Detection using BERT on Twitter. *Procedia Comput. Sci.* **235**, 1870–1882 (2024).
13. Costola, M., Hinz, O., Nofer, M. & Pelizzon, L. Machine learning sentiment analysis, COVID-19 news and stock market reactions. *Res. Int. Bus. Financ.* **64**, 101881 (2023).
14. Wan, B., Wu, P., Yeo, C. K. & Li, G. Emotion-cognitive reasoning integrated BERT for sentiment analysis of online public opinions on emergencies. *Inf. Process. Manag.* **61**, 103609 (2024).
15. Wu, D., Yang, J. & Wang, K. Exploring the reversal curse and other deductive logical reasoning in BERT and GPT-based large language models. *Patterns* 101030 (2024) doi:10.1016/j.patter.2024.101030.
16. Patel, A., Oza, P. & Agrawal, S. Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model. *Procedia Comput. Sci.* **218**, 2459–2467 (2022).
17. Zhao, N., Fan, G., Qi, Z. & Shi, J. Exploring the current situation of cultural tourism scenic spots Exploring the current situation of cultural tourism scenic spots based on LDA model based on LDA model Take Nanjing, Jiangsu Province, China as an example-Take Nanjing, Jiangsu Province, Chi. *Procedia Comput. Sci.* **00**, 826–832 (2023).
18. Qorib, M., Oladunni, T., Denis, M., Ososanya, E. & Cotae, P. Covid-19 vaccine hesitancy : Text mining , sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. **212**, (2023).
19. Fitri, V. A. *et al.* Sentiment Analysis of Social Media Twitter with Case of Anti-Sentiment Analysis of Social Media Twitter with Case of Anti- LGBT Campaign in Indonesia using Naïve Bayes , Decision Tree , LGBT Campaign in Indonesia. **00**, (2019).
20. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.* **1**, 4171–4186 (2019).
21. Jelodar, H. *et al.* Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. (2018).