Aspect-Based Sentiment Analysis for Enhanced Understanding of 'Kemenkeu' Tweets

Priska Trisna Sejati [1], Farrikh Al Zami [2]*, Aris Marjuni [3], Heni Indrayani [4], Ika Dewi Puspitarini [5]
[1234]Faculty of Computer Science, Universitas Dian Nuswantoro
[5]Kementerian Keuangan Republik Indonesia
112202106676@mhs.dinus.ac.id [1], alzami@dsn.dinus.ac.id [2*], aris.marjuni@dsn.dinus.ac.id [3], heni.indrayani@dsn.dinus.ac.id [4],
ika.dewi@kemenkeu.go.id [5]

Article Info

Article history:

Received 2024-09-24
Revised 2024-10-06
Accepted 2024-10-08

Keyword:

Aspect Based Sentiment Analysis (ABSA),
Latent Dirichlet Allocation (LDA),
IndoBERT,
Kementerian Keuangan (Kemenkeu),
X (Twitter).

ABSTRACT

The perceptions and expressions shared by the public on social media play a crucial role in shaping the reputation of government institutions, such as the Ministry of Finance MOF (Kemenkeu) in Indonesia which also has faced increased scrutiny, particularly on Twitter. This study analyzes public sentiment towards the Indonesian Ministry of Finance (MoF) through Aspect-Based Sentiment Analysis (ABSA) on Twitter data. Using a dataset of 10,099 tweets from January to July 2024, this study combines IndoBERT for sentiment classification and Latent Dirichlet Allocation (LDA) for topic modeling. Here, LDA was tested across four scenarios that considered various combinations of stopwords removal and stemming techniques, resulting in coherence scores of 0.314256, 0.369636, 0.350285, and 0.541752. The most optimal results were achieved in the scenario of stopwords removal without stemming (with 0.314256 coherence score). The main results show: 1) Identification of four main topics related to MoF: Economy, Budget, Employees, and Tax; 2) The dominance of negative sentiment (6,837 tweets) compared to positive sentiment (198 tweets) across all topics; 3) The effectiveness of IndoBERT in handling the complexity of the Indonesian language, especially in interpreting context and language nuances; 4) The importance of proper preprocessing, with a scenario of removing stopwords without stemming resulting in the most relevant topics. This study provides valuable insights for MoF to understand public perception and identify areas that require special attention in public communication and policy.

I. Introduction

Social media platforms, including Twitter, have become essential to modern communication in the digital age, which has significantly transformed the way people exchange information[1]. Twitter has become one of the most powerful tools for voicing individual opinions, thoughts, and discussion globally upon various topics[2], including about the reputations of government institutions. The perceptions and opinions expressed by the public on social media play an important role in building the reputation of government institutions[3]. Recently, one of the Indonesian government institutions, The Ministry of Finance (Ministry of Finance) charted a lot of exposure and attention on social media especially at this point, Twitter.

One effective way to understand the public's perception of governmental institutions is by utilizing sentiment analysis, a technique that leverages natural language processing (NLP) and text mining to analyze opinions expressed in textual content, such as social media posts and reviews[4], by extracting and interpreting subjective information to classify sentiments as positive, negative, or neutral. By employing sentiment analysis, particularly on Twitter where most public opinions are shared through text, it is possible to uncover public attitudes toward governmental institutions[5] and gain detailed insights into opinions on government actions[6]. This is particularly useful during public emergencies, as it allows for the identification of critical shifts in sentiment, enabling governments to address public concerns more effectively[7].

While sentiment analysis with its advantage is a powerful tool for understanding public perceptions[8], its effectiveness can be hindered by the complexity and diversity of opinions expressed where multiple aspects are discussed on a single topic. This complexity can lead to misunderstandings and complicate the overall analysis[5]. Additionally, the use of cross-lingual terms and informal language commonly found on social media further complicates sentiment analysis, as traditional models may struggle to interpret these complexities[9]. Because of these problems, advanced text mining techniques are needed to address these challenges. One such method is Aspect-Based Sentiment Analysis (ABSA), which provides more detailed information by identifying specific aspects and their sentiments[10].

ABSA offers significant advantages, especially in analysing public sentiment toward governmental institutions by providing nuanced insights into specific elements of public opinion. Unlike traditional sentiment analysis, which often provides an overall sentiment, ABSA allows for the identification and evaluation of sentiments tied to individual aspects within complex sentences[11]. This method enhances the accuracy of sentiment predictions by focusing on specific aspects rather than broad categories, offering a deeper understanding of user opinions[12]. ABSA has proven to be a valuable tool for gaining detailed insights into how the public perceives various facets of government actions and policies[13].

However, current challenges in utilizing ABSA for government services include the scarcity of labelled training data, the complexity of natural language, and the difficulty in detecting implicit aspects, which collectively limit the effectiveness of deep learning model performance[14]. Furthermore, the complexity of language complicates the accurate extraction of aspects, as it involves differentiating sentiments tied to particular attributes rather than the overall sentiment[15], which in turn further complicates sentiment analysis in a linguistically diverse country like Indonesia. These constraints underscore the necessity for better techniques to enhance ABSA’s effectiveness in interpreting and understanding public sentiment towards government[16], especially in the context of Indonesian government discourse.

To address the challenges, this study integrates IndoBERT with ABSA. One important component of ABSA is topic modelling, which is a technique used to uncover hidden meanings and patterns in text, particularly in complex datasets like conversations[17]. Therefore, this study also combines Latent Dirichlet Allocation (LDA) for effective topic modelling and identifies key aspects of public sentiment and provides a detailed aspect-based sentiment analysis in Twitter data related to Indonesia's Ministry of Finance (Kemenkeu). IndoBERT is preferred for its outstanding performance in handling Indonesian NLP tasks, including ABSA, due to its proficiency in understanding the language's subtleties and context, which is crucial for accurate sentiment analysis[18]. Besides, LDA is chosen for its superior performance compared to other topic modelling techniques and its capability to uncover hidden patterns in large text corpora[19]. This approach offers valuable insights into public sentiment towards Indonesia's Ministry of Finance (Kemenkeu) through detailed aspect-based analysis of Twitter data. It provides actionable insights for policymakers while enhancing academic understanding by addressing gaps in aspect-based sentiment analysis within Indonesian and social media contexts.

To justify our research, in this study, we use a two-stage approach that combines Latent Dirichlet Allocation (LDA) and Aspect-Based Sentiment Analysis (ABSA) to provide a more comprehensive analysis. First, LDA is used to identify key topics in the tweet dataset related to the Ministry of Finance. These topics are then interpreted as ‘aspects’ in the context of ABSA. For example, topics dominated by words such as ‘tax’, ‘increase’, and ‘policy’ are identified as the ‘Tax Policy’ aspect. Second, we apply ABSA using IndoBERT to analyze specific sentiments related to each aspect identified by LDA. This approach allows us to link sentiments to specific aspects of the Ministry of Finance without the need for extensive manual labeling. Although LDA and ABSA do serve different purposes, the integration of both in this study allows for a more nuanced and contextual analysis. LDA helps in automatically identifying relevant aspects from a large corpus, while ABSA provides a sentiment analysis that focuses on those aspects. This approach allows us to analyze public sentiment towards various aspects of the Ministry of Finance's operations and policies more efficiently and comprehensively.

II. METHOD

The research flow diagram below illustrates the step involved, followed by detailed explanation of each stage.

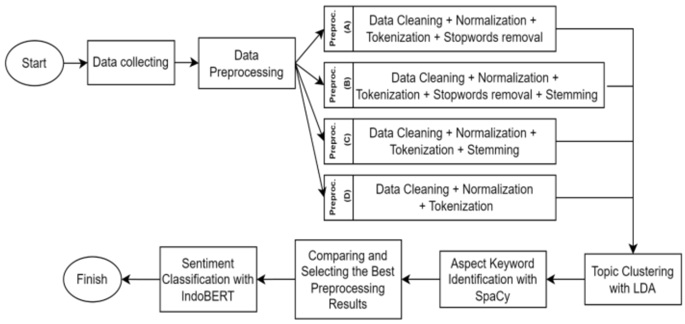

Start
Data collecting
Data Preprocessing
Preproc. (A)
Data Cleaning + Normalization + Tokenization + Stopwords removal
Preproc. (B)
Data Cleaning + Normalization + Tokenization + Stopwords removal + Stemming
Preproc. (C)
Data Cleaning + Normalization + Tokenization + Stemming
Preproc. (D)
Data Cleaning + Normalization + Tokenization
Topic Clustering with LDA
Aspect Keyword Identification with SpaCy
Comparing and Selecting the Best Preprocessing Results
Sentiment Classification with IndoBERT
Finish

Figure 1. Research Flow Diagram

Detail explanation from Figure 1 can be seen as follow:

A. Data Collecting

The first stage of this research involves data collection, which is a step to ensure the study is based on relevant and representative information. The data used in this study is sourced from X (Twitter), focusing on tweets containing the keyword “Kemenkeu” posted from January 1, 2024 to July 10, 2024. The data was collected using the Tweet Harvest tool, supported by Google Colab and the Python programming language, resulting in a dataset of 10,099 tweets with 13 columns of information (conversation_id_str, created_at, favorite_count, full_text, id_str, in_reply_to_screen_name, location, quote_count, reply_count, retweet_count, tweet_url, user_id_str, username, F1). This data serves as the foundation for subsequent analysis, including preprocessing, topic modeling, and sentiment analysis. About the dataset, the selection period covers several important events that affect the image of the Ministry of Finance, including tax increase policies and customs policies. This dataset focuses on Indonesian-language tweets to ensure the relevance of local context. While this dataset represents the opinions of Indonesian-speaking Twitter users about the Ministry of Finance during the period, it is important to acknowledge that it may not reflect the overall public sentiment, given the limited number of Twitter users and potential language bias. Several limitations need to be considered in interpreting the results. First, this dataset may not represent the views of people who do not use Twitter or communicate in languages other than Indonesian. Second, the sentiments revealed may have been influenced by specific events that occurred during the data collection period, such as customs and new tax policies. Thus, the period of January to July 2024 was chosen because it encompasses a series of significant events that have the potential to influence public perception of the Ministry of Finance, providing a rich context for sentiment analysis and topic identification.

B. Data Preprocessing

Pre-processing data is an essential activity in sentiment analysis and topic modelling to mitigate noise and irrelevant content in tweets, which often contain hashtags, emojis, and special symbols[20], by applying techniques as follow:

1) *Data Cleaning:* This process focuses on removing irrelevant elements (non-ASCII characters, extra spaces, emojis, tabs and escape characters, the word "RT", mentions, incomplete URLs, single characters, hyphens, punctuation (except underscores), and the word "user") and duplicates from the dataset to prevent redundancy and ensure each entry’s uniqueness, which is a crucial step for maintaining data integrity and reducing potential biases in the analysis[21].

2) Tokenization: This process involves splitting or breaking down sentences into units, typically individual words or tokens, allowing for more detailed analysis[20].

3) Normalization: converts non-standard words like slang and abbreviations into their standard forms. In this study, a manually compiled correction dictionary was used to accurately translate these variations into their correct standard equivalents.

4) Stopwords removal: In the stopwords removal stage, words that do not contribute significant meaning are eliminated[22]. Standard stopword lists are often based on general language sources, which might not be appropriate for specialized fields. This highlights the need for customized stopword lists in technical areas, which can be created by identifying non-informative terms. Moreover, removing stopwords can improve text mining tasks by enhancing feature extraction, as customized stopword lists have been shown to reduce information loss[23].

5) Stemming: reduces words to their root or base forms, consolidating different forms of the same word into a common term, such as the words ‘programs’, ‘programming’, and ‘programmer’ stem to ‘program’. While stemming can be helpful, it may also remove important language details and reduce accuracy by mixing different terms, especially in languages with complex word forms where subtle meanings can be lost[24]. Sastrawi module is used in this approach.

Considering the advantages of disadvantages determined from stopwords removal and stemming, this study experimented with several data preprocessing scenarios: (a) Stopwords removal without stemming, (b) Stopwords removal with stemming, (c) No stopwords removal with stemming, and (d) No stopwords removal and no stemming to identify and determine the most effective approach.

C. Topic Clustering with LDA

After successfully preprocessing the data, the next step is topic clustering, which involves grouping text documents into clusters based on the underlying themes or topics they represent. This process helps in identifying and organizing the main subjects discussed across the dataset. For topic clustering, Latent Dirichlet Allocation (LDA) implemented in Python using the Gensim library (version 4.3.3) was used. In this research, LDA is employed for topic clustering, which organizes preprocessed documents into clusters based on their contextual similarities[25].

As an unsupervised machine learning approach, LDA offers a robust and efficient method for summarizing document content. It is particularly effective for short texts and extensive review datasets, making it a suitable choice for topic extraction in this study[26].

A notable limitation of the LDA model is that it does not provide labels for the topics[22]. To address this issue, this research manually assigned labels to the resulting clusters of words to provide meaningful interpretations.

D. Aspect Keyword Identification with SpaCy

In this stage, aspect keywords are identified using SpaCy's part-of-speech (version 3.7.6) tagging to enhance the accuracy of aspect and opinion term extraction. By focusing on POS tagging, we can better distinguish significant aspect-related terms from other words in the text. This approach refines the identification of relevant aspects and opinion words, thus improving the overall effectiveness of feature extraction and subsequent analysis[27].

E. *Comparing and Selecting the Best Preprocessing Results*

In this section, the best preprocessing scenario from the four approaches, (a) Stopwords removal without stemming, (b) Stopwords removal with stemming, (c) No stopwords removal with stemming, and (d) No stopwords removal and no stemming, are compared and selected. The evaluation considers the effectiveness of each method across the entire pipeline, including preprocessing, topic modeling, and aspect keyword identification. Based on the results from each stage, the most robust approach identified and utilized for the subsequent labeling phase. This process ensures that the chosen method provides the most accurate and meaningful insights for further analysis.

F. *Sentiment Classification with IndoBERT*

Sentiment Classification is a critical step in the Aspect-Based Sentiment Analysis (ABSA). Sentiment analysis, a form of text classification, involves analyzing text to determine the underlying sentiment. This process is closely tied to text classification, where words and sentences are grouped based on their emotional tone or discussed topics, with Natural Language Processing (NLP) techniques classifying the text according to its context and content[28].

In this research, IndoBERT, a pre-trained model specifically developed for the Indonesian language and based on the BERT architecture, is employed for sentiment classification. The model (indobenchmark/indobert-base-p1) was loaded from the Hugging Face model hub using version 4.44.2 of the Transformers library. IndoBERT excels in capturing word relationships and understanding the context within sentences, making it highly effective for sentiment classification in Indonesian text[28]. By leveraging IndoBERT, this study ensures accurate sentiment classification, which is crucial for understanding the emotional tone associated with different aspects within the dataset.

III. Result And Discussion

This research covers the entire process previously outlined in the Method section, including data collection, preprocessing, topic clustering, aspect keyword identification, and sentiment classification. The results of each stage are presented as the outcomes of the research.

A. *Data Collecting*

As described in the Method chapter, a total of 10,099 tweet records containing the keyword "Kemenkeu" were collected. Table I below provides an example of the raw data before any cleaning or preprocessing was applied.

TABLE I
Raw Data Crawling Results

indeks	full_text
0	Kementerian Keuangan (Kemenkeu) secara resmi telah menaikkan tarif cukai hasil tembakau (CHT) rata-rata 10% pada awal 2024. Harga rokok jadi makin mahal. https://t.co/014NtplWgL
1	Kemenkeu Pastikan Gaji PNS Naik 8 Persen per 1 Januari Namun Dirapel https://t.co/JgeCQd9b3c #aktualcom #Aktualofficial
2	@I***A**y Food Estatenya Gagal dan Lahan Terbengkalai (Data Walhi dan Greenpeace).!! Kalau Hilirisasi.. kenapa tidak ada Pemasukan yg Signifikan untuk RI (Data di Kemenkeu).Jilat boleh.. Tolol jangan..!!

B. *Data Preprocessing*

In this study, the main phases of data preprocessing include data cleaning, tokenization, and normalization. Stemming and stopwords removal were carefully considered through various preprocessing scenarios, as trials revealed that applying either of these processes, or even combining them, could alter the meaning of sentences.

1) *Data Cleaning*

After data collection, the cleaning process involves several steps to enhance text clarity. This process is necessary because the raw data often contains symbols or words that are less meaningful and non-standard, which makes the text untidy and difficult to read as illustrated in Figure 2.

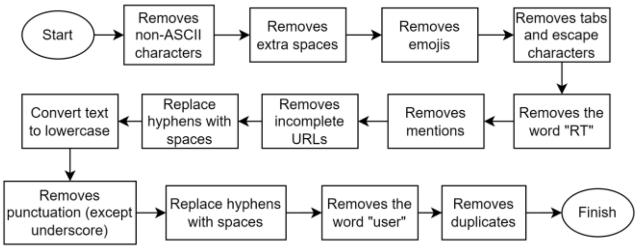

Start
Removes non-ASCII characters
Removes extra spaces
Removes emojis
Removes tabs and escape characters
Removes the word "RT"
Removes mentions
Removes incomplete URLs
Replace hyphens with spaces
Convert text to lowercase
Removes punctuation (except underscore)
Replace hyphens with spaces
Removes the word "user"
Removes duplicates
Finish

Figure 2. Data Cleaning Flow Diagram

These cleaning steps ensure that the data is more uniform and relevant. The data cleaning result can be seen in Table II.

TABLE II
Comparison of Sample Data Before and After Cleaning

full_text	clean_text
Kementerian Keuangan (Kemenkeu) secara resmi telah menaikkan tarif cukai hasil tembakau (CHT) rata-rata 10% pada awal 2024. Harga rokok jadi makin mahal. https://t.co/014NtplWgL	kementerian keuangan kemenkeu secara resmi telah menaikkan tarif cukai hasil tembakau cht rata rata percent pada awal number harga rokok jadi makin mahal
Kemenkeu Pastikan Gaji PNS Naik 8 Persen per 1 Januari Namun Dirapel https://t.co/JgeCQd9b3c #aktualcom #Aktualofficial	kemenkeu pastikan gaji pns naik number persen per number januari namun dirapel aktual com aktualofficial
@I***A**y Food Estatenya Gagal dan Lahan Terbengkalai	food estatenya gagal dan lahan terbengkalai data walhi dan

full_text	clean_text
(Data Walhi dan Greenpeace).!! Kalau Hilirisasi.. kenapa tidak ada Pemasukan yg Signifikan untuk RI (Data di Kemenkeu).Jilat boleh.. Tolol jangan..!!	greenpeace kalau hilirisasi kenapa tidak ada pemasukan yg signifikan untuk ri data di kemenkeu jilat boleh tolol jangan

2) *Tokenization*

Following the data cleaning process, tokenization is performed to further refine the text data. The results of tokenization can be seen in Table III.

TABLE III
SAMPLE DATA AFTER TOKENIZATION

clean_text	tokens
kementerian keuangan kemenkeu secara resmi telah menaikkan tarif cukai hasil tembakau cht rata rata percent pada awal number harga rokok jadi makin mahal	kementerian,keuangan,kemenkeu,secara,resmi,telah,menaikkan,tarif,cukai,hasil,tembakau,cht,rata,rata,percent,pada,awal,number,harga,rokok,jadi,makin,mahal
kemenkeu pastikan gaji pns naik number persen per number januari namun dirapel aktual com aktualofficial	kemenkeu,pastikan,gaji,pns,naik,number,persen,per,number,januari,namun,dirapel,aktual,com,aktualofficial
food estatenya gagal dan lahan terbengkalai data walhi dan greenpeace kalau hilirisasi kenapa tidak ada pemasukan yg signifikan untuk ri data di kemenkeu jilat boleh tolol jangan	food,estatenya,gagal,dan,lahan,terbengkalai,data,walhi,dan,greenpeace,kalau,hilirisasi,kenapa,tidak,ada,pemasukan,yg,signifikan,untuk,ri,data,di,kemenkeu,jilat,boleh,tolol,jangan

3) *Normalization*

After tokenization, the next step is normalization, which is essential for standardizing text data. As mentioned in the Method section, informal language, slang, and abbreviations are converted into their formal counterparts using a manually compiled correction dictionary. The outcomes of the normalization process are illustrated in Table IV.

TABLE IV
SAMPLE DATA AFTER NORMALIZATION

tokens	normalize
kementerian,keuangan,kemenkeu,secara,resmi,telah,menaikkan,tarif,cukai,hasil,tembakau,cht,rata,rata,percent,pada,awal,number,harga,rokok,jadi,makin,mahal	kementerian,keuangan,kemenkeu,secara,resmi,telah,menaikkan,tarif,cukai,hasil,tembakau,cht,rata,rata,percent,pada,awal,nomor,harga,rokok,jadi,makin,mahal
kemenkeu,pastikan,gaji,pns,naik,number,persen,per,number,januari,namun,dirapel,aktual,com,aktualofficial	kemenkeu,pastikan,gaji,pns,naik,nomor,persen,per,nomor,januari,namun,dirapel,aktual,com,aktualofficial
food,estatenya,gagal,dan,lahan,terbengkalai,data,walhi,dan,greenpeace,kalau,hilirisasi,kenapa,tidak,ada,pemasukan,yg,signifikan,untuk,ri,data,di,kemenkeu,jilat,boleh,tolol,jangan	food,estatenya,gagal,dan,lahan,terbengkalai,data,walhi,dan,greenpeace,kalau,hilirisasi,kenapa,tidak,ada,pemasukan,yang,signifikan,untuk,ri,data,di,kemenkeu,jilat,boleh,tolol,jangan

4) *Stopwords removal*

In this study, a custom stopwords list was used to avoid issues found with standard lists from libraries like NLTK and Sastrawi. These standard lists often remove important negations, such as 'tidak' (meaning 'not'), commonly used in negative sentiment expressions, might be removed, thereby altering the entire meaning of a sentence. For example, the sentence 'saya tidak suka nasi goreng' (I don't like nasi goreng) would become 'saya suka nasi goreng' (I like nasi goreng) after applying standard stopwords, which conveys the opposite meaning. The use of a custom stopwords list aimed to preserve the integrity of sentence meaning in sentiment analysis. Table V below provides an example of the result of stopwords removal.

TABLE V
SAMPLE DATA AFTER STOPWORDS REMOVAL

normalize	stopwords
kementerian,keuangan,kemenkeu,secara,resmi,telah,menaikkan,tarif,cukai,hasil,tembakau,cht,rata,rata,percent,pada,awal,nomor,harga,rokok,jadi,makin,mahal	kementerian,keuangan,kemenkeu,resmi,menaikkan,tarif,cukai,hasil,tembakau,cht,percent,nomor,harga,rokok,mahal
kemenkeu,pastikan,gaji,pns,naik,nomor,persen,per,nomor,januari,namun,dirapel,aktual,com,aktualofficial	kemenkeu,pastikan,gaji,pns,nomor,persen,nomor,januari,namun,dirapel,aktual,com,aktualofficial
food,estatenya,gagal,dan,lahan,terbengkalai,data,walhi,dan,greenpeace,kalau,hilirisasi,kenapa,tidak,ada,pemasukan,yang,signifikan,untuk,ri,data,di,kemenkeu,jilat,boleh,tolol,jangan	food,estatenya,gagal,lahan,terbengkalai,data,walhi,greenpeace,hilirisasi,tidak,pemasukan,signifikan,ri,data,kemenkeu,jilat,tolol

5) *Stemming*

Similar to stopwords removal, stemming was also carefully avoided to preserve the accuracy of sentiment analysis. Although stemming is a common technique used to reduce words to their root forms, it can sometimes alter the meaning of sentences. Table VI illustrates examples comparing the effects of stemming and no stemming.

TABLE VI
SAMPLE DATA AFTER STEMMING APPLICATION

stopwords	stemming
kementerian,keuangan,kemenkeu,resmi,menaikkan,tarif,cukai,hasil,tembakau,cht,percent,nomor,harga,rokok,mahal	menteri, uang, kemenkeu, resmi, naik, tarif, cukai, hasil, tembakau, cht, percent, nomor, harga, rokok, mahal

stopwords	stemming
kemenkeu,pastikan,gaji,pns,nomor,persen,nomor,januari,namun,dirapel,aktual,com,aktualofficial	kemenkeu, pasti, gaji, pns, nomor, persen, nomor, januari, rapel, aktual, com, aktualofficial
food,estatenya,gagal,lahan,terbengkalai,data,walhi,greenpeace,hilirisasi,tidak,pemasukan,signifikan,ri,data,kemenkeu,jilat,tolol	food, estatenya, gagal, lahan, bengkalai, data, walhi, greenpeace, hilir, tidak, pasu, signifikan, ri, data, kemenkeu, jilat, tolol

C. Topic Clustering with LDA

In this section, the process begins by preparing the data that has undergone previous preprocessing steps. The preparation involves converting the preprocessed results into a list of words that are easier to analyze. Bigrams and trigrams are then created to capture word associations that frequently co-occur in the text. Additionally, lemmatization is performed on the remaining words, with only nouns, adjectives, verbs, and adverbs being retained.

Next, a dictionary and corpus are created from the cleaned data, which are then used to identify word distributions within the text. However, before this step, coherence scores are calculated to assess how cohesively words form a topic. After determining the optimal number of topics, the LDA model is created and analyzed to obtain the topic distribution across the documents.

1) Coherence Score Analysis: Before proceeding with topic clustering using LDA, coherence scores are calculated for each scenario. This step is essential as higher coherence scores indicate how well the generated topics are defined.

TABLE VII
COHERENCE SCORES FOR EACH PREPROCESSING SCENARIO AND ITS OPTIMAL NUMBER OF TOPICS

Preprocessing Scenario	Optimal Number of Topics	Coherence Score
A (Stopwords removal + No Stemming)	4	0.314256
B (Stopwords removal + Stemming)	4	0.369636
C (No stopwords removal + Stemming)	2	0.350285
D (No stopwords removal + No Stemming)	3	0.541752

According to Table VII, it's clear that the optimal number of topics varies across scenarios, with some scenarios having an optimal number of 4 topics, while others have 3 or 2.

TABLE VIII
SUMMARY OF COHERENCE SCORES

Average Coherence Score	0.393982
Standar Deviation	0.103291
Minimum Value	0.314256 (Scenario A)
Maximum Value	0.541752 (Scenario D)
Range	0.227496

Based on Tables VII and VIII, the comparison of the scenarios reveals several findings. Scenario D (No stopwords removal + No Stemming) achieved the highest coherence score of 0.541752, which is 72.39% higher than the lowest score. Besides, Scenario A (Stopwords removal + No Stemming) had the lowest coherence score of 0.314256. Scenarios B (Stopwords removal + Stemming) and C (No stopwords removal + Stemming) had relatively close scores of 0.369636 and 0.350285, with only a 5.52% difference.

The results of each scenario showed various impact of using stopwords removal and stemming. The effects of stopwords removal are shown. With stemming, it slightly improves the score (C vs. B), while without stemming, it decreased the score (D vs. A). Stemming also showed varying effects. without stopwords removal, it pulled the score down (D vs. C), but with stopwords removal it increased the score (A vs. B).

2) Topic Modelling: Latent Dirichlet Allocation (LDA) is utilized to form a topic model, grouping words into specific topics based on their co-occurrence within the same document. This process yields several topics, each composed of key terms that have a high probability of representing the topic. Here are the results.

TABLE IX
TOPIC MODELLING RESULTS

Preprocessing Scenario	Topic Modelling Results
A (Stopwords removal + No Stemming)	Topic: 1 Words: 0.107*"nomor" + 0.065*"kemenkeu" + 0.016*"uang" + 0.009*"pemerintah" + 0.009*"apbn" + 0.008*"gaji" + 0.008*"ekonomi" + 0.008*"percent" + 0.007*"semester" + 0.007*"terjaga" Topic: 2 Words: 0.098*"kemenkeu" + 0.043*"tidak" + 0.026*"ya" + 0.011*"negara" + 0.008*"nya" + 0.008*"data" + 0.007*"anggaran" + 0.006*"nih" + 0.005*"kementerian" + 0.005*"biaya" Topic: 3 Words: 0.046*"kemenkeu" + 0.010*"indonesia" + 0.009*"negeri" + 0.007*"keuangan" + 0.007*"ri" + 0.007*"pegawai" + 0.007*"lpdp" + 0.007*"program" + 0.006*"kerja" + 0.006*"anak" Topic: 4 Words: 0.021*"pajak" + 0.015*"orang" + 0.014*"pakai" + 0.011*"bumn" + 0.010*"informasi" + 0.009*"kegiatan" + 0.008*"bidang" + 0.007*"selengkapnya" + 0.006*"masyarakat" + 0.006*"laporan"
B (Stopwords removal + Stemming)	Topic: 1 Word: 0.062*"kemenkeu" + 0.023*"a" + 0.016*"ja" + 0.014*"bank" + 0.014*"negara" + 0.011*"k" + 0.010*"anggar" + 0.009*"gaji" + 0.009*"menteri" + 0.007*"program"

Preprocessing Scenario	Topic Modelling Results
	Topic: 2 Word: 0.106*"kemenkeu" + 0.047*"tidak" + 0.009*"data" + 0.008*"tan" + 0.008*"ri" + 0.007*"tingkat_miskin" + 0.007*"bansos_ekonomi_domestik_turun" + 0.007*"nih" + 0.006*"jabat" + 0.005*"tri" Topic: 3 Word: 0.101*"nomor" + 0.061*"kemenkeu" + 0.019*"n" + 0.010*"kerja" + 0.010*"menteri_uang" + 0.010*"apbn" + 0.009*"negeri" + 0.008*"jaga" + 0.007*"menja" + 0.007*"percent" Topic: 4 Word: 0.029*"uang" + 0.017*"pajak" + 0.015*"ekonomi" + 0.014*"perintah" + 0.014*"indonesia" + 0.012*"orang" + 0.011*"pakai" + 0.010*"lapor" + 0.009*"laku" + 0.008*"s"
C (No stopwords removal + Stemming)	Topic: 1 Words: 0.056*"kemenkeu" + 0.027*"tidak" + 0.021*"yang" + 0.014*"kalau" + 0.014*"saja" + 0.011*"n" + 0.010*"ke" + 0.009*"sudah" + 0.008*"sama" + 0.008*"gaji" Topic: 2 Words: 0.043*"kemenkeu" + 0.017*"yang" + 0.015*"tidak" + 0.012*"bea" + 0.012*"cukai" + 0.011*"negara" + 0.009*"apa" + 0.007*"anggar" + 0.007*"n" + 0.006*"presiden"
D (No stopwords removal + No Stemming)	Topic: 1 Words: 0.048*"kemenkeu" + 0.032*"yang" + 0.017*"di" + 0.016*"tidak" + 0.014*"itu" + 0.012*"dan" + 0.012*"ada" + 0.012*"ini" + 0.011*"kalau" + 0.011*"ya" Topic: 2 Words: 0.051*"nomor" + 0.032*"kemenkeu" + 0.025*"dan" + 0.018*"untuk" + 0.011*"di" + 0.011*"dari" + 0.008*"yang" + 0.007*"dalam" + 0.006*"dengan" + 0.006*"oleh" Topic: 3 Words: 0.042*"kemenkeu" + 0.008*"gaji" + 0.008*"asn" + 0.007*"nih" + 0.007*"bumn" + 0.006*"ke" + 0.006*"pns" + 0.005*"aturan" + 0.005*"bi" + 0.005*"kemenperin"

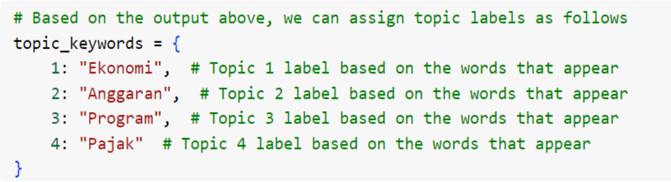
Based on the topic modeling results, a comparison between scenarios can be made:

- Scenario A (Stopwords removal + No Stemming) shows more specific topics focusing on terms like "kemenkeu," "uang" (money), "pemerintah" (government), "apbn" (state budget), and "gaji" (salary), which are highly relevant to the themes of Kemenkeu policies, financial issues, and the economy. Although it has a lower coherence score (0.314256) than other scenarios, the topics are rich in information and relevance.
- Scenario B (Stopwords removal + Stemming) includes terms like "negara" (country) and "anggar" (budget), indicating a good focus on relevant aspects such as fiscal and economic policies. However, the appearance of unclear terms like "a," "ja," and other abbreviations suggests that the combination of stopwords removal and stemming might remove too much context.
- Scenario C (No stopwords removal + Stemming) produces topics dominated by stopwords such as "tidak" (not), "yang" (which), "sama" (same), "apa" (what), and "saja" (only), indicating reduced topic relevance. Both topics show the appearance of "kemenkeu" as a dominant word. However, the presence of short words like "n" suggests that stemming may reduce topic quality by producing less informative words.
- Scenario D (No stopwords removal + No Stemming) has the highest coherence score (0.541752) but features many stopwords and common words such as "yang" (which), "di" (in), "dan" (and), "ada" (there is), and "ini" (this), which do not provide much information. This indicates that even though the coherence score is high, the quality of information is low due to the dominance of stopwords.

In our analysis, although scenario D (without stopword removal and without stemming) produced the highest coherence score (0.541752), we chose scenario A (stopword removal without stemming) with a coherence score of 0.314256 as the optimal one. This decision was based on a comprehensive evaluation that considered not only the coherence score, but also the quality and relevance of the generated topics. Qualitative analysis showed that the topics from scenario A were more meaningful and relevant to the context of the Ministry of Finance. For example, topics such as ‘kemenkeu’, ‘uang’, ‘pemerintah’, ‘apbn’, and ‘gaji’ that appeared in scenario A were more suitable for aspect-based sentiment analysis related to the policies and performance of the Ministry of Finance. In contrast, scenario D, despite having a higher coherence score, produced topics dominated by stopwords and general words that were less informative. We acknowledge the trade-off between coherence score and topic quality, but in the context of this study, topic relevance was considered more important to support accurate sentiment analysis. This decision is also supported by expert evaluations that confirm that the topics from scenario A are more appropriate for analyzing public sentiment towards the Ministry of Finance. Thus, our selection of the optimal scenario involves both quantitative and qualitative considerations to ensure more meaningful and contextual analysis results.

As worth mentioning, the selection of scenario A (stopwords removal without stemming) as the optimal one was based on several empirical considerations: 1) Context Preservation: Stemming tends to remove important nuances in Indonesian, especially for words related to policy and economy. For example, 'policy' and 'wise' have different connotations in the context of the Ministry of Finance. 2) Stopwords Removal Effectiveness: Stopwords removal increases the focus of the analysis on more meaningful words, reducing noise in the data without removing important context. 3) Topic Quality: Qualitative analysis shows that the topics generated from scenario A are more coherent and relevant to the Ministry of Finance domain compared to other scenarios. In addition, we also applied additional preprocessing steps: 1) Text Normalization: Converting text to lowercase and removing non-alphanumeric characters for consistency. 2) Noise Removal: Removing URLs, mentions (@user), and hashtags to focus on the text content. 3) Spelling Correction: Using a custom dictionary to correct common spelling errors and abbreviations in the Indonesian financial context.4) Tokenization: Breaking text into individual tokens using an Indonesian-specific tokenizer. This combination of steps has proven effective in improving the quality of input for both LDA and IndoBERT models, resulting in more meaningful topics and more accurate sentiment classification.

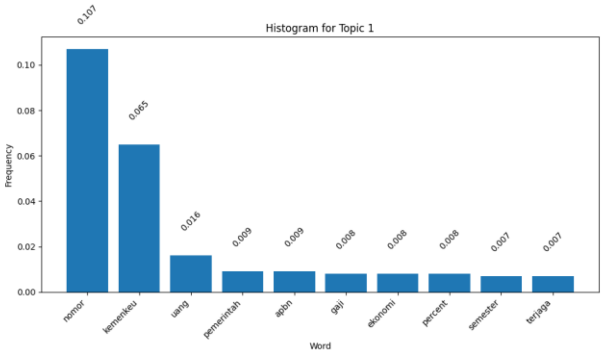
Manual labeling is also performed for each topic based on the identified keywords. For each topic, manual labeling is carried out by analyzing the prominent words that appear within the topic, and a single representative word is chosen to encapsulate the main theme of that topic. The results of the manual labeling for each topic can be seen in Figure 2 below.



```
# Based on the output above, we can assign topic labels as follows
topic_keywords = {
    1: "Ekonomi",  # Topic 1 label based on the words that appear
    2: "Anggaran",  # Topic 2 label based on the words that appear
    3: "Program",  # Topic 3 label based on the words that appear
    4: "Pajak"  # Topic 4 label based on the words that appear
}
```

Figure 2. Manual Labeling of the Topics

Furthermore, to facilitate understanding, visualizations such as histogram for each topic, word clouds, document distribution plots based on dominant topics, and PyLDAvis are also included in Figure 3.



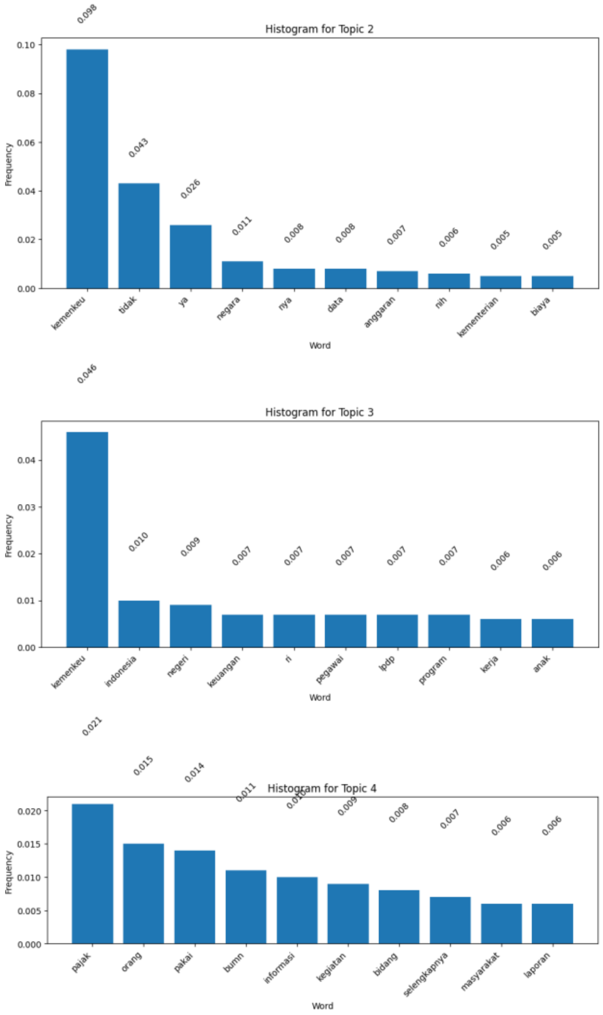


Figure 3. Histogram for Each Topics

Figure 3 illustrates the term distribution across topics from LDA topic modeling by its frequency, providing insights into the thematic focus of each topic. The analysis reveals that while there is significant overlap in terms like "kemenkeu" across topics, each topic uniquely addresses different facets of financial and governmental discourse, indicating a nuanced understanding of financial issues within the dataset.

Moreover, a word cloud is presented, offering a visual representation similar to the previous histogram. In this word cloud, the size of each word corresponds to its frequency in the text. Words that appear more frequently are shown in larger sizes, while less frequent words are shown in smaller sizes. Figure 4 showcases the Word Cloud for each topic.

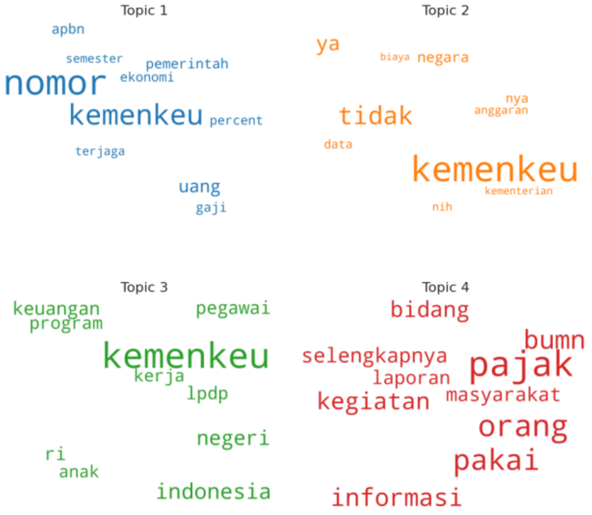


Figure 4. Word Cloud of Top 10 Words in Each Topic

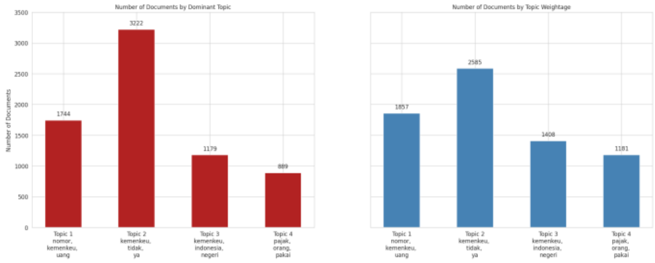


Figure 5. Document Distribution Plots

Figure 5 presents two histograms that offer a comprehensive overview of how documents are distributed across different topics. The first histogram reveals the primary topic for each document, indicating which topic is most prominent in the majority of the documents. Topic 2 emerges as the most dominant, with 3,222 documents. In contrast, the second histogram provides a nuanced view by showing the cumulative importance or "weightage" of each topic across all documents. Topic 2 still leads with 2,585 documents.

Afterward, the LDA model analysis reveals a clear distinction between the four identified topics. On the right side of the figure, 30 keywords with the highest frequency across all topics are listed. Meanwhile, on the left, the circles represent the occurrence frequency of each topic, as shown by the considerable distance between them, indicating low correlation and strong differentiation among the topics. Topic 1 emerges as the most dominant, with the largest circle, signifying its higher frequency and central importance compared to the other topics. The significant spacing between the circles suggests that each topic is distinct and does not overlap substantially with the others, highlighting the effectiveness of the model in categorizing the data into well-defined, separate themes. Figure 6 showcases this visual representation of LDA Model Analysis.

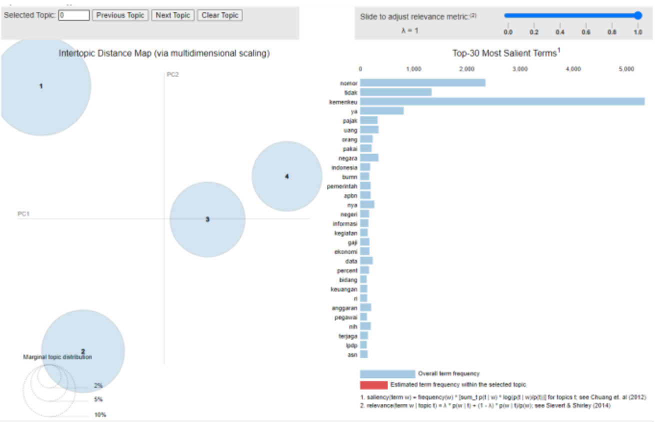

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms1
0
1,000
2,000
3,000
4,000
5,000
nomor
tidak
kemenkeu
ya
pajak
uang
orang
pakai
negara
indonesia
bumn
pemerintah
apbn
nya
negeri
informasi
kegiatan
gaji
ekonomi
data
percent
bidang
keuangan
ri
anggaran
pegawai
nih
terjaga
lpdp
asn
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 6. LDA Model Results

D. Aspect Keyword Identification with SpaCy

This section focuses on the identification of aspect keywords using SpaCy, following the LDA topic modeling process. Aspect keyword identification is performed to understand the specific elements discussed in each topic. This step is conducted concurrently with the LDA topic modeling, ensuring that the identified keywords align with the topics derived.

TABLE X
SAMPLE OF ASPECT KEYWORD IDENTIFICATION WITH SPACY RESULT

full_text	Kementerian Keuangan (Kemenkeu) secara resmi telah menaikkan tarif cukai hasil tembakau (CHT) rata-rata 10% pada awal 2024. Harga rokok jadi makin mahal. https://t.co/014NtplWgL
preprocessed	kementerian, keuangan, kemenkeu, resmi, menaikkan, tarif, cukai, hasil, tembakau, cht, percent, nomor, harga, rokok, mahal
Dominant_topic	4
Keywords	pajak, orang, pakai, bumn, informasi, kegiatan, bidang, selengkapnya, masyarakat, laporan
topic_keywords	Pajak

E. Sentiment Classification with IndoBERT

Following the topic modeling and clustering stages, the next step involved sentiment classification using IndoBERT. At this stage, each record was already associated with a specific topic, allowing us to label each one as positive or negative. This process was crucial for understanding the overall sentiment distribution across the identified topics. To visualize the results, a Sentiment Bar Graph was created to display the count of positive and negative sentiments.

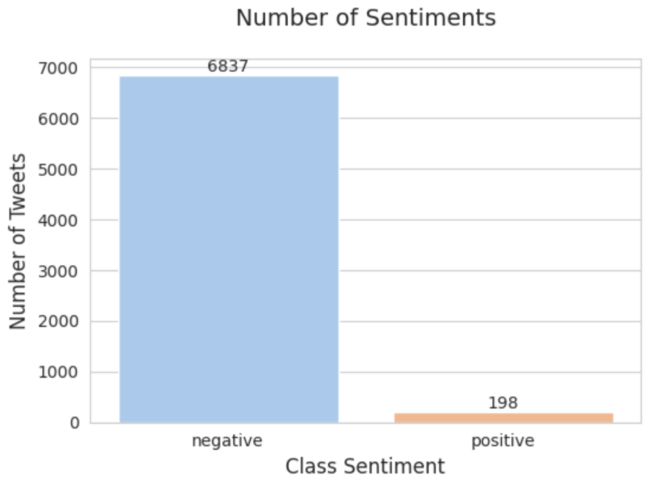


Figure 7. Sentiment Bar Graph

The Sentiment Bar Graph (Figure 7) provides a notable contrast between the volumes of positive and negative sentiments within the dataset. With 6,837 negative sentiments compared to only 198 positive sentiments, the graph highlights a predominance of negative sentiment throughout the documents.

In this study, we use a pre-trained IndoBERT model for sentiment analysis without performing any fine-tuning or additional evaluation. This decision is based on IndoBERT's proven reputation in natural language processing tasks for Indonesian, including sentiment analysis. IndoBERT was chosen for several advantages: 1) Contextual Understanding: IndoBERT is able to understand the meaning of words based on the context of the sentence, which is very important in sentiment analysis of tweets that often contain informal language and complex contexts; 2) Ambiguity Handling: The model is better at handling ambiguous words in Indonesian, which often appear in discussions about finance and government policy topics; 3) Pre-training on Local Data: IndoBERT has been trained on a large corpus of Indonesian, making it more adaptive to the nuances and variations of Indonesian used in tweets about the Ministry of Finance.

Given IndoBERT's inherent strength in understanding Indonesian, we apply this model directly for sentiment classification on our dataset. The results of this classification are then used as a basis for further analysis of public sentiment towards various aspects of the Ministry of Finance that have been identified through topic modeling. Thus, while we did not conduct a formal evaluation of IndoBERT's performance in the specific context of this study, the decision to use this model was based on its well-documented performance in the literature for similar tasks in Indonesian. This approach allowed us to focus on analyzing and interpreting sentiment results in the context of topics relevant to the Ministry of Finance, rather than on the technical evaluation of the model itself.

Additionally, a table titled Sample of Data Labelling Result was provided to showcase examples of the labeled data.

TABLE XI
SAMPLE OF DATA LABELLING RESULT

full_text	@I***A**y Food Estatenya Gagal dan Lahan Terbengkalai (Data Walhi dan Greenpeace).!! Kalau Hilirisasi.. kenapa tidak ada Pemasukan yg Signifikan untuk RI (Data di Kemenkeu).Jilat boleh.. Tolol jangan..!!
preprocessed	food, estatenya, gagal, lahan, terbengkalai, data, walhi, greenpeace, hilirisasi, tidak, pemasukan, signifikan, ri, data, kemenkeu, jilat, tolol
Dominant_topic	2
Keywords	kemenkeu, tidak, ya, negara, nya, data, anggaran, nih, kementerian, biaya
topic_keywords	Anggaran
label	negative

Furthermore, Sentiment Distribution Plots by Topic below were generated to provide a more detailed view of how sentiments are distributed within each topic.

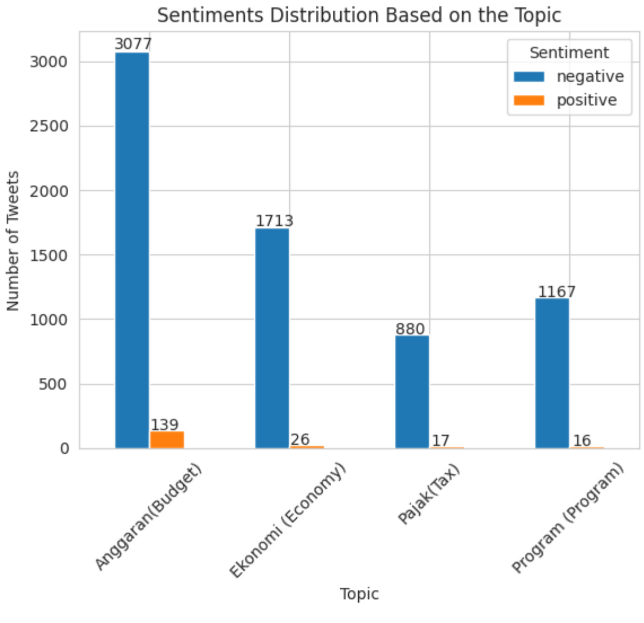


Figure 8. Sentiment Distribution Plots by Topic

The Sentiment Distribution Plots by Topic reveal significant variations in sentiment across different topics within the document. This distribution highlights a consistent pattern of negative sentiment across all topics, indicating a widespread dissatisfaction or criticism that pervades various aspects of the document.

In another words, our aspect-based sentiment analysis (ABSA) reveals a complex relationship between various aspects of the Ministry of Finance and public sentiment. Across the four main topics identified – ‘Economy’, ‘Budget’ (Anggaran), ‘Employees’ (Pegawai) and ‘Taxes’ (Pajak) – we found significant variation in sentiment. The ‘Economy’ aspect showed mixed sentiment, with positive views towards economic stabilization efforts, but also concerns about inflation and global economic uncertainty. ‘Budget’ received predominantly negative sentiment, mainly related to perceptions of inefficiency and lack of transparency in the allocation of public funds. The ‘Employees’ aspect showed positive sentiment towards the professionalism of the Ministry of Finance staff, but also criticism related to corruption cases involving several officials. ‘Taxes’ emerged as the aspect with the most negative sentiment, reflecting public dissatisfaction with tax hike policies and the complexity of the tax system. The analysis revealed that negative sentiment was often rooted in perceptions of unfairness or lack of transparency, while positive sentiment was related to policies that were perceived to support economic stability and public welfare. These findings highlight areas where the Ministry of Finance may need to improve public communication and policy adjustments to improve public perception and trust.

IV. Conclusion

In conclusion, the use of IndoBERT in this study demonstrated significant effectiveness in handling the complexity of the Indonesian language. In particular, the model successfully overcomes several typical linguistic challenges. First, in terms of complex morphology, IndoBERT is able to accurately identify various forms of affixation that are common in Indonesian. For example, the model successfully interprets the nuances of meaning between the words ‘menaikkan’ and 'dinaikkan' in the context of tax policy, where both have the same root but different implications. Second, IndoBERT demonstrates good ability in handling diverse sentence structures, including subject-predicate inversions that are often found in formal Indonesian. For example, in the sentence 'Diumumkan oleh Kemenkeu Kebijakan baru tentang pajak', the model successfully identifies 'Kemenkeu' as the subject even though its position is not at the beginning of the sentence. Third, IndoBERT is able to correctly interpret the use of particles and pronouns that are typical of Indonesian, such as '-lah', '-kah', and 'nya', which often affect the nuances of sentiment in sentences. This ability is clearly seen in the analysis of tweets such as ' Apakah Kebijakan Barunya efektif?' where the model successfully captures a skeptical tone. In addition, IndoBERT shows good performance in recognizing and interpreting loanwords and technical financial terms that are often used in the context of the Ministry of Finance, such as 'APBN' or 'defisit'. This success shows the model's adaptability to specific domains in the Indonesian language.

IV. Acknowledgement

This research has been conducted in collaboration with the Bureau of Communications and Information Services of the Ministry of Finance.

References

[1] A. Chaudhuri and C. F. Prendes, “Social Media and EJVES 2013–2023: From Inception to Evolution,” *Eur. J. Vasc. Endovasc. Surg.*, vol. 65, no. 6, pp. 769–771, Jun. 2023, doi: 10.1016/j.ejvs.2023.04.012.

[2] Aldinata, A. M. Soesanto, V. C. Chandra, and D. Suhartono, “Sentiments comparison on Twitter about LGBT,” *7th Int. Conf. Comput. Sci. Comput. Intell. 2022*, vol. 216, pp. 765–773, Jan. 2023, doi: 10.1016/j.procs.2022.12.194.

[3] M. Mansoor, “Citizens’ trust in government as a function of good governance and government agency’s provision of quality information on social media during COVID-19,” *Gov. Inf. Q.*, vol. 38, no. 4, p. 101597, Oct. 2021, doi: 10.1016/j.giq.2021.101597.

[4] M. Bordoloi and S. K. Biswas, “Sentiment analysis: A survey on design framework, applications and future scopes,” *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 12505–12560, Nov. 2023, doi: 10.1007/s10462-023-10442-2.

[5] S. Bengesi, T. Oladunni, R. Olusegun, and H. Audu, “A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets,” *IEEE Access*, vol. 11, pp. 11811–11826, 2023, doi: 10.1109/ACCESS.2023.3242290.

[6] N. Parveen, P. Chakrabarti, B. T. Hung, and A. Shaik, “Twitter sentiment analysis using hybrid gated attention recurrent network,” *J. Big Data*, vol. 10, no. 1, p. 50, Apr. 2023, doi: 10.1186/s40537-023-00726-3.

[7] Z. Pi and H. Feng, “The evolution of public sentiment toward government management of emergencies: Social media analytics,” *Front. Ecol. Evol.*, vol. 10, p. 1026175, Dec. 2022, doi: 10.3389/fevo.2022.1026175.

[8] A. Al-Adaileh, M. Al-Kfairy, M. Tubishat, and O. Alfandi, “A sentiment analysis approach for understanding users’ perception of metaverse marketplace,” *Intell. Syst. Appl.*, vol. 22, p. 200362, Jun. 2024, doi: 10.1016/j.iswa.2024.200362.

[9] A. A. Raza, A. Habib, J. Ashraf, B. Shah, and F. Moreira, “Semantic Orientation of Crosslingual Sentiments: Employment of Lexicon and Dictionaries,” *IEEE Access*, vol. 11, pp. 7617–7629, 2023, doi: 10.1109/ACCESS.2023.3238207.

[10] G. Kontonatsios *et al.*, “FABSA: An aspect-based sentiment analysis dataset of user reviews,” *Neurocomputing*, vol. 562, p. 126867, Dec. 2023, doi: 10.1016/j.neucom.2023.126867.

[11] T. Zhou, Y. Shen, K. Chen, and Q. Cao, “Hierarchical dual graph convolutional network for aspect-based sentiment analysis,” *Knowl.-Based Syst.*, vol. 276, p. 110740, Sep. 2023, doi: 10.1016/j.knosys.2023.110740.

[12] H. Qin, G. Chen, Y. Tian, and Y. Song, “Improving Federated Learning for Aspect-based Sentiment Analysis via Topic Memories,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3942–3954. doi: 10.18653/v1/2021.emnlp-main.321.

[13] R. Dutta, N. Das, M. Majumder, and B. Jana, “Aspect based sentiment analysis using multi-criteria decision-making and deep learning under COVID-19 pandemic in India,” *CAAI Trans. Intell. Technol.*, vol. 8, no. 1, pp. 219–234, Mar. 2023, doi: 10.1049/cit2.12144.

[14] W. Zheng, H. Jin, Y. Zhang, X. Fu, and X. Tao, “Aspect-Level Sentiment Classification Based on Auto-Adaptive Model Transfer,” *IEEE Access*, vol. 11, pp. 34990–34998, 2023, doi: 10.1109/ACCESS.2023.3265473.

[15] W. Ahmad, H. U. Khan, T. Iqbal, and S. Iqbal, “Attention-Based Multi-Channel Gated Recurrent Neural Networks: A Novel Feature-

Centric Approach for Aspect-Based Sentiment Classification," *IEEE Access*, vol. 11, pp. 54408–54427, 2023, doi: 10.1109/ACCESS.2023.3281889.

[16] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.

[17] S. Salmi, R. van der Mei, S. Mérelle, and S. Bhulai, "Topic modeling for conversations for mental health helplines with utterance embedding," *Telemat. Inform. Rep.*, vol. 13, p. 100126, Mar. 2024, doi: 10.1016/j.teler.2024.100126.

[18] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[19] M. N. P. Ma'ady, A. F. A. Rahim, T. S. N. Syahda, A. F. Rizqi, and M. C. A. Ratna, "Malaysia Citizen Sentiment on Government Response Towards Covid-19 Disaster Management: Using LDA-based Topic Visualization on Twitter," *Seventh Inf. Syst. Int. Conf. ISICO 2023*, vol. 234, pp. 561–569, Jan. 2024, doi: 10.1016/j.procs.2024.03.040.

[20] S. E. Uthirapathy and D. Sandanam, "Topic Modelling and Opinion Analysis On Climate Change Twitter Data Using LDA And BERT Model.," *Int. Conf. Mach. Learn. Data Eng.*, vol. 218, pp. 908–917, Jan. 2023, doi: 10.1016/j.procs.2023.01.071.

[21] M. Husna, L. P. Purba, M. E. Rinaldy, and A. R. Lubis, "Predictive Analytics for IMDb Top TV Ratings: A Linear Regression Approach to the Data of Top 250 IMDb TV Shows," vol. 8, no. 1.

[22] A. Meddeb and L. B. Romdhane, "Using Topic Modeling and Word Embedding for Topic Extraction in Twitter," *Knowl.-Based Intell. Inf. Eng. Syst. Proc. 26th Int. Conf. KES2022*, vol. 207, pp. 790–799, Jan. 2022, doi: 10.1016/j.procs.2022.09.134.

[23] R. Rani and D. K. Lobiyal, "Performance evaluation of text-mining models with Hindi stopwords lists," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, Part A, pp. 2771–2786, Jun. 2022, doi: 10.1016/j.jksuci.2020.03.003.

[24] A. Nzeyimana, "Morphological disambiguation from stemming data," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 4649–4660. doi: 10.18653/v1/2020.coling-main.409.

[25] P. Yang, Y. Yao, and H. Zhou, "Leveraging Global and Local Topic Popularities for LDA-Based Document Clustering," *IEEE Access*, vol. 8, pp. 24734–24745, 2020, doi: 10.1109/ACCESS.2020.2969525.

[26] Y. Zhang and L. Zhang, "Movie Recommendation Algorithm Based on Sentiment Analysis and LDA," *8th Int. Conf. Inf. Technol. Quant. Manag. ITQM 2020 2021 Dev. Glob. Digit. Econ. COVID-19*, vol. 199, pp. 871–878, Jan. 2022, doi: 10.1016/j.procs.2022.01.109.

[27] Y. Li, Q. He, and L. Yang, "Part-of-speech based label update network for aspect sentiment triplet extraction," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 1, p. 101908, Jan. 2024, doi: 10.1016/j.jksuci.2023.101908.

[28] G. Z. Nabiilah, S. Y. Prasetyo, Z. N. Izdihar, and A. S. Girsang, "BERT base model for toxic comment analysis on Indonesian social media," *7th Int. Conf. Comput. Sci. Comput. Intell. 2022*, vol. 216, pp. 714–721, Jan. 2023, doi: 10.1016/j.procs.2022.12.188.