BAREKENG: Journal of Mathematics and Its Applications
June 2024 Volume 18 Issue 2 Page 1081–1094
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss2pp1081-1094

OPTIMIZING LONG TEXT CLASSIFICATION PERFORMANCE THROUGH KEYWORD-BASED SENTENCE SELECTION: A CASE STUDY ON ONLINE NEWS CLASSIFICATION FOR INDONESIAN GDP GROWTH-RATE DETECTION

Dinda Pusparahmi Sholawatunnisa [1], **Lya Hulliyyatus Suadaa** [2*]

[1,2] *Statistical Computing Department, Politeknik Statistika STIS*
Jl. Otto Iskandardinata No. 64C, Jakarta, 13330, Indonesia

*Corresponding author's e-mail: ** **lya@stis.ac.id**

Article History:

Received: 1st January 2024
Revised: 1st February 2024
Accepted: 24th March 2024
Published: 1st June 2024

Keywords:

Extractive Summarization
Keyword-Based Sentence Selection;
Long-text handling.

ABSTRACT

Efficiently managing lengthy textual data, particularly in online news, is crucial for enhancing the performance of long text classification. This study explores innovative approaches to streamline the Gross Domestic Product (GDP) computation process by utilizing modern data analytics, Natural Language Processing (NLP), and online news sources. Leveraging online news data introduces real-time information, which promises to improve the accuracy and timeliness of economic indicators like GDP. However, handling the complexity of extensive textual data poses a challenge, demanding advanced NLP techniques. This research shifts from traditional word-weight-based methods to keyword-based extractive summarization techniques. These tailored approaches ensure that selected sentences align precisely with specific keywords relevant to the research case, such as GDP growth rate detection. The study emphasizes the necessity of adapting summarization methods to capture information in unique research contexts effectively. Classification results show that the implementation of sentence selection significantly improves classification accuracy. Specifically, there was an average accuracy increase of 0.0226 for machine learning and 0.0164 for transfer learning models. Additionally, in terms of computational efficiency, sentence selection also accelerates processing time during hyperparameter tuning and fine-tuning, as observed using the same computational resources.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License.

How to cite this article:

D. P. Sholawatunnisa and L. H. Suadaa., "OPTIMIZING LONG TEXT CLASSIFICATION PERFORMANCE THROUGH KEYWORD-BASED SENTENCES SELECTION: A CASE STUDY ON ONLINE NEWS CLASSIFICATION FOR INDONESIAN GDP GROWTH-RATE DETECTION," *BAREKENG: J. Math. & App.*, vol. 18, iss. 2, pp. 1081-1094, June, 2024.

Copyright © 2024 Author(s)
Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*
Research Article • **Open Access**

1. INTRODUCTION

Gross Domestic Product (GDP) is pivotal for planning policy decisions, supporting research, and guiding economic strategies. However, one of the central issues in GDP computation is the time-consuming data aggregation and verification process. while standardized procedures exist for calculating GDP, such as those outlined for GDP computation using the Standard Classification of Indonesia Business Fields (KBLI), the process still involves multiple data streams. It requires meticulous collection, validation, and data integration from diverse sources. This complexity often leads to delays in producing GDP figures. This delay has broader implications for policymakers, analysts, and stakeholders who rely on timely economic indicators for decision-making, policy formulation, and investment strategies. Moreover, as economies become more interconnected and dynamic, the time lag between data collection and GDP estimation may result in inaccuracies due to rapidly changing economic conditions. Despite the commendable efforts in handling these intricate calculations, the data release timeframe still emphasizes the need for innovative advancements and technological implementations to refine and streamline the existing operational processes.

In light of these challenges, research is needed to explore innovative approaches and strategies that leverage data processing, automation, and technological advancements to streamline the GDP computation process. By harnessing the potential of modern data analytics, Natural Language Processing (NLP), and unconventional data sources such as online news, detecting fluctuations in GDP becomes achievable **[1]**. This advancement promises to introduce an innovative breakthrough in GDP movement detection. Simultaneously, this detection can be utilized as supplementary data to complement the official statistics published by BPS-Statistics Indonesia. Integrating unconventional data sources, particularly online news, into economic analysis introduces a promising innovation in the field. Several studies have demonstrated encouraging results in pattern detection through analyzing online news content. Noteworthy among these is the work by Chowdhury et al. (2014) **[2]**, wherein sentiment analysis of online news is utilized to predict stock price patterns. Similarly, Duarte et al. (2020) **[3]** focus on detecting price declines, aiming to enable anticipation and prevention of financial losses. In a broader context within the financial industry, Kumar and Ravi (2016) **[4]** conducted a comprehensive review of text mining applications. These studies collectively underline the growing significance of leveraging online news data for predictive analytics within the economy.

These online news sources offer a wealth of real-time information that has the potential to enhance the accuracy and timeliness of economic indicators like GDP. Online news, as an information medium, has a distinct advantage due to its unrestricted reach across space and time. This capability allows it to transcend geographical and temporal boundaries, providing access to information anytime and anywhere, as long as an internet connection is available **[5]**. This characteristic underscores the fundamental shift in information dissemination facilitated by the digital age, where traditional limitations related to print media, broadcasting, and physical distribution no longer apply. However, this strategy presents a challenge alongside its promise due to the complexity of handling lengthy textual data. The voluminous and unstructured format of online news content demands sophisticated NLP techniques to extract meaningful insights effectively **[6]**. Overcoming this challenge is pivotal for unleashing the full potential of this innovative approach and ensuring that it seamlessly integrates into the official statistical analysis, as released by authoritative institutions like BPS-Statistics Indonesia.

The common approach for handling lengthy text data such as online news is through text summarization, which aims to balance computational efficiency and enhance classification performance. In many cases, to retain information while summarizing, these methods often rely on word importance determined through word weighting within a document. Identifying word importance is crucial as it helps extract key elements, enhancing the quality of summaries. Domeniconi et al. (2015) show that methods like TF-IDF and advanced techniques such as neural network-based embeddings significantly contribute to summary quality. Another breakthrough in sentence selection, demonstrated by Fiok et al. (2020) **[6]** as "Text Guide," employs features and their importance from a pre-trained machine learning classifier to select relevant text fragments.

However, the method based on word weight principles to identify keywords for sentence selection is irrelevant for application in a case study involving specific and significant keywords. Therefore, adjustments and optimizations are needed to ensure that the selected sentences in the summary align with the relevant keywords specific to the research case. To address this requirement of highly specific terms directly related to the unique research context, the sentence selection method adopted a keyword-based extractive summarization approach instead of relying on word-weight-based techniques. A method with a similar principle is SWAP-NET, an extractive summarization method that utilizes keywords to extract essential sentences. This method employs a different approach by utilizing a neural sequence-to-sequence model **[8]**. Another relevant method is keywords-guided abstractive sentence summarization **[9]**. This method improves performance by adapting a multi-task learning framework to extract keywords and generate text summaries simultaneously. This methodology is designed to prioritize important information during the encoding process using keyword references. This study focuses on detecting GDP growth rates in temporal categories such as month-to-month, quarter-to-quarter, year-on-year, and cumulative-to-cumulative. These temporal categories introduce unique characteristics, and therefore, it is essential to tailor the summarization approach to capture and convey information related to the investigated case effectively.

This study introduces innovative approaches for GDP growth-rate detection using modern data analytics, NLP, and online news sources. Machine learning models and pre-trained language transformers are employed, including Random Forest, SVM, Naïve Bayes, IndoBERT, and MultilingualBERT (mBERT). Each model offers unique strengths for analyzing textual data and detecting patterns in GDP movements. For instance, the Random Forest classifier constructs decision trees for accurate classification. At the same time, SVM excels in identifying optimal feature combinations, making it robust for high-dimensional text data. Naïve Bayes leverages probabilistic principles and performs well in text classification tasks despite its simplifying assumptions. IndoBERT and mBERT, designed for NLP tasks, demonstrate superior performance in capturing nuanced language features relevant to GDP growth-rate detection. Leveraging these models aims to enhance the accuracy and timeliness of economic indicators like GDP, benefiting decision-making and policy formulation processes.

This research focuses on a case study of the online news platform Detik.com. Its potential for applicability in other case studies needs a thoughtful adaptation process. Adapting for similar cases involves focusing on the underlying framework of the online news platform. This requires a closer examination of the mechanisms used for online news scraping, which may need adjustments based on the specific functionalities of the target website. Furthermore, the keywords employed in this study case are inherently transferrable and can be readily applied to other online news platforms with minimal to no further adaptation. This implies that researchers can confidently expand their data sources to encompass a broader spectrum of online news platforms, thus enhancing the versatility and generalizability of the research findings.

2. RESEARCH METHODS

2.1 The Scope of The Study

The primary focus of this research is to design a simplified approach for handling lengthy text data, particularly within the context of a specific research problem. In this study, the specific research context is detecting GDP growth rates through the classification of online news. The approach employed to address lengthy online news text is a sentence selection method that utilizes a keyword-based mechanism. These keywords have been identified as related to the categories of GDP growth rates being detected. The objective is to streamline the handling of lengthy text data while effectively capturing and summarizing essential information for analysis within the context of GDP growth rate detection. Simultaneously, the objective involves minimizing computational costs and processing time while enhancing classification performance as evaluated through relevant metrics.

2.2 Timeframe of The Study

The study's timeframe is during the third quarter (Q3) of 2022. The researchers deliberately selected the quarterly data release timeframe to address the concern regarding potential revised news versions instead of opting for shorter periods, such as monthly updates. This strategic choice accommodates news revisions within the dataset, which often occur shortly after initial publications. Analysis of the dataset revealed instances of redundant titles with distinct links, suggesting the presence of multiple news posts, with one potentially being a revision of the other. To ensure the inclusion of revised content, the data cleansing and preprocessing methods focused on eliminating redundancy based on unique news links rather than relying solely on titles. This approach aimed to retain initial publications and subsequent revisions in the dataset, thereby enhancing its comprehensiveness and reliability. Additionally, the choice of the quarterly period considered the substantial volume of raw news data obtained, totaling 13,286 news articles based on the keywords used. A manual validation process was implemented to ensure the validity of this news data, which needed to encompass economic activities within the GDP framework. This process involved subject matter expert reviews conducted through rigorous discussions to assess the validity of each news article. It's worth noting that, when conducting this research, specific procedures for automating this validation process had yet to be developed, given the resource constraints and the need for timely analysis. Strategically, the quarterly reporting period was chosen among various GDP release options, including monthly, semester, and annual releases. Specifically, the third quarter was selected as it represented the most recent data available during the research period. Regarding data adequacy, the evaluation phase yielded highly favorable results, with a robust metric score of 80% based on the classification of 1,586 news articles.

2.3 Data Source

The primary data source for this research is the online news portal Detik.com during the third quarter (Q3) of 2022. Detik.com is a reputable Indonesian news website that covers a wide range of topics, including economic news, making it a suitable source for collecting online news articles related to GDP growth rates during this specific time frame. Detik.com encompasses various channels, including DetikFinance, which specializes in economic matters, and region-specific channels like DetikSumut for North Sumatra and DetikJabar for West Java. As reported by SimilarWeb, from July to September 2022, Detik.com attracted a substantial audience of 516.7 million visitors, achieving the top-ranking position within the News and Media Publisher category [10]. In line with these findings, Brajawidagda et al. (2017), in their study utilizing online news from the same source, selected Detik.com as a primary data source. This choice was supported by Detik.com's status as one of Indonesia's two most prominent online news platforms, as indicated by alexa.com. It's worth noting that Indonesia, one of the world's most populous democracies, upholds press freedom through Law 40/1999. As a result, the press maintains a significant degree of impartiality and operates without the constraints of official censorship. This democratic context accentuates the trustworthiness and dependability of Detik.com as a research source [11].

2.4 Data Collection Method

News articles from detik.com for the specified period are collected through web scraping. Custom scripts are developed to navigate the website, extract article URLs, and retrieve each article's full content. The web scraping process is executed using Python 3 programming language, employing the Requests and Selenium libraries. The web scraping process considers the structure and layout of detik.com, ensuring that articles are collected in a structured format. To create a comprehensive dataset, the scraper collects key metadata, such as the publication date, article title, source, and the article's full content. To ensure the comprehensive representation of every industry sector within the dataset, the selected keywords correspond to the distinctive terms associated with 17 distinct industrial sectors defined within the Standard Classification of Indonesian Business Fields (Klasifikasi Baku Lapangan Usaha Indonesia or KBLI) as shown in Table 1. This alignment has enabled the categorization and identification of data for each specific industry sector within the specified timeframe.

2.5 News Data Labelling

The collected online news data undergoes a manual filtering process to separate relevant news data from irrelevant ones regarding GDP and KBLI categories identified during the web scraping stage. Data labeling for GDP growth rate classification is performed by an odd number of annotators, following labeling guidelines prepared by the researchers and validated by a subject matter expert and practitioner in GDP computation from BPS-Statistics Indonesia. The final labels are determined using the majority vote principle. As a prerequisite, news data labeling is conducted by annotators with fundamental knowledge of the Gross Domestic Product. In this research, contributing annotators are Polytechnic of Statistics STIS students who have completed courses in SNA (System of National Accounts) and GDP. For the detection of GDP growth rates in online news classification, there are three labels: quartal-to-quartal, year-on-year, and cumulative-to-cumulative. However, the news narrative does not specify the GDP growth rate information. In that case, annotators mark the news data with the " not specified " label. During the analysis phase, assumptions will be applied based on the recommendation given by the subject matter expert, assuming it to be the annual

GDP growth rate. The reliability of the annotators' labeling results is assessed using Krippendorf's alpha measure.

2.6 Online News Data Preprocessing

The preprocessing procedures include the following steps:

2.6.1 Case Folding & Data Cleaning

In the initial data preprocessing phase, the text undergoes standardization by converting all letters to lowercase. This ensures consistency throughout the text, facilitating easier processing and analysis. Additionally, redundant news articles, punctuation, and excessive spaces are removed to enhance the quality and readability of the data. These steps are crucial for preparing the text for further analysis and modeling in subsequent stages of the study. Eliminating unnecessary elements and focusing on relevant content helps improve the efficiency and accuracy of subsequent natural language processing tasks.

2.6.2 Stemming & Stopwords Removal

In this phase, specifically tailored for machine learning input, affixes are stripped from words to derive their root forms. Frequent words with minimal semantic value, such as conjunctions and prepositions, are eliminated [12]. The PorterStemmer library is utilized to execute this task efficiently. By employing this approach, the text is refined to focus on essential semantic content, enhancing the effectiveness of subsequent machine learning algorithms in processing and analyzing the data.

2.6.3 Tokenization

During this preprocessing stage, the text document is segmented into smaller units known as tokens, each representing a word within the machine-learning algorithm [13]. In the context of pre-trained transformers, these tokens often represent subwords. The NLTK library is employed for tokenization throughout this phase. These preprocessing steps collectively contribute to refining the online news data, ensuring its uniformity and compatibility for subsequent analysis and classification tasks. By standardizing text, eliminating noise, and segmenting it into manageable units, these procedures enhance the accuracy and efficacy of the data within the framework of machine learning algorithms and NLP techniques.

Table 1. Industrial Sectors Defined within the Standard Classification of Indonesian Business Fields

No.	Category	Industrial Sectors
1	A	Agriculture, Forestry, and Fisheries
2	B	Mining and Quarrying
3	C	Manufacturing Industry
4	D	Electricity and Gas Supply
5	E	Water Supply, Waste Management, and Recycling
6	F	Construction
7	G	Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles
8	H	Transportation and Warehousing
9	I	Accommodation and Food and Beverage Service Activities
10	J	Information and Communication
11	K	Financial and Insurance Activities
12	L	Real Estate
13	M, N	Business Services
14	O	Government Administration, Defense, and Mandatory Social Security
15	P	Education Services
16	Q	Health and Social Activities
17	RSTU	Other Services

Source: Quarterly Indonesian GDP 2017-2021, BPS Publication

2.7 Feature Selection

Feature selection determines which data attributes are most relevant for the classification task. The news data is transformed into vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) weighting method, computed with the sci-kit learn (sklearn) library. The Term Frequency (TF) component in sklearn is computed according to Equation (1), which is expressed as follows [14]:

$$TF_{(t,d)} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad (1)$$

Where:
$f_{t',d}$ is the frequency of the term *t* in the document *d*.
$\sum_{t' \in d} f_{t',d}$ represents the total frequency of term *t* across all terms in the document *d*.

The Inverse Document Frequency (IDF) formula in sklearn incorporates a constant value of 1 in the numerator and denominator to prevent division by zero, this can be represented as shown in Equation (2):

$$IDF_{(t)} = log\frac{1+n}{1+df(t)} + 1 \quad (2)$$

Where:
n is the number of training documents used.
$df(t)$ is the number of training documents containing the term *t*.

This TF-IDF vectorization process is fundamental for transforming news data into a numerical format suitable for machine learning algorithms. It assigns weights to terms based on frequency and importance within the dataset, facilitating accurate classification of GDP growth rate information.

2.8 Development of Classification Models for GDP Growth-Rate Detection

This section introduces the machine learning classification models and pre-trained language model transformers for detecting GDP movements and GDP growth rates using online news. Two transformer-based language models, IndoBERT and MultilingualBERT, designed for natural language processing in the Indonesian context, are employed. Below, a concise explanation of the models used is provided:

2.8.1 Random Forest

The Random Forest classifier is an ensemble learning algorithm designed for classification tasks [15]. It operates by constructing a collection of individual decision trees, each independently categorizing class labels based on its analysis. These individual trees are combined to formulate predictions in the Random Forest ensemble. One notable advantage of Random Forest lies in its utilization of a substantial number of decision trees [16]. This multitude of trees contributes to an enhancement in the accuracy of classification results. It is particularly beneficial when dealing with complex datasets. Moreover, Random Forest demonstrates proficiency in handling datasets with many input variables. It can effectively manage scenarios where datasets suffer from class imbalance issues, thus maintaining the balance between errors [17].

2.8.2 Support Vector Machine

SVM is grounded in the fundamental principle of identifying a hyperplane that effectively separates distinct classes within a dataset. It offers the advantage of determining the optimal feature combination, making it highly robust for high-dimensional data [16]. Often sparse and high-dimensional, text data aligns perfectly with SVM's strengths. This is due to the tendency of certain features to be irrelevant yet potentially correlated. In text data, these features are typically organized into categories that can be linearly divided [18]. SVM is the best classifier for text-based tasks, particularly well-suited for handling large datasets with numerous features [19].

2.8.3 Naïve Bayes Classifier

The Naïve Bayes classifier relies on probabilistic principles and makes strong independent assumptions within the feature space. It leverages Bayesian probability rules to learn parameters from training data [20]. The core concept of Naïve Bayes is its assumption that word occurrences in a document are independent, which contrasts with human language, where word relationships exist. Despite this simplifying assumption, Naïve Bayes performs commendably in text classification tasks [21]. In this study, Multinomial Naïve Bayes models are employed.

2.8.4 IndoBERT

IndoBERT is a pre-trained Transformer-based model derived from BERT and was trained on Indonesian vocabulary using a masked language model with the hugging face framework (Wilie et al., 2020). The training process involved a comprehensive corpus of 220 million words from diverse sources. Wikipedia Indonesia contributed 74 million words. News articles from Kompas and Tempo, provided by Tala et al. (2003), added another linguistic dimension [23]. Liputan6 contributed 55 million words. Indonesian Web Corpus, assembled by Medved & Suchomel (2017), enriched the dataset with 90 million words [24]. To evaluate its performance, IndoBERT was compared to conventional machine learning models, serving as a baseline model, as well as Multilingual-BERT and MalayBERT. In sentiment analysis, IndoBERT outperformed all models with an impressive F1-score of 84.13% [25].

2.8.5 MultilingualBERT (mBERT)

mBERT is a pre-trained BERT model that has been trained in a diverse set of languages, including 102 languages (uncased) and 104 languages (cased), using Wikipedia data [26], [27]. This includes coverage of Bahasa Indonesia, making mBERT a candidate for natural language processing (NLP) tasks in Indonesian. However, it's worth noting that despite its wide linguistic coverage, mBERT may not excel in learning all 104 languages to the same degree. High-resource languages like English tend to perform worse when compared to Monolingual-BERT models of the same capacity [28]. On the other hand, training various Monolingual BERT models for low-resource languages using the same data size can result in lower-quality representations for these languages. Interestingly, this performance difference between Monolingual-BERT and mBERT is not solely attributed to BERT's hyperparameters or model sharing across languages [29]. This study compares Multilingual-BERT (mBERT) with Monolingual-BERT, specifically IndoBERT, for Bahasa Indonesia. This comparison sheds light on how these models perform when dealing with Indonesian text data, considering the nuances of language representation in multilingual and monolingual contexts.

2.9 Mechanism of Sentence Selection

The mechanism of sentence selection involves a specific feature, which the researcher defines as a predefined list comprising keywords sought within each sentence of the text. Each predefined keyword signifies the period of GDP growth rate, which is used as the classification label, specifically quarterly, annually, or cumulative. If none of these predefined keywords appear in any sentence of the text, the algorithm extracts the initial three sentences. The implementation of sentence selection takes place in the early stages, before preprocessing, and employs the NLTK library for sentence tokenization. The selection of the initial three sentences for extraction is determined based on the statistical summary results of the sentence selection process. The information indicates that, on average, the number of sentences containing predefined keywords is around three to four. Extracting the first three sentences of a document for text summarization is commonly known as a "3-lead summary" [30]. This approach of extracting initial sentences aligns with the concept of naive truncation in text summarization, widely applied across various frameworks to address the processing of lengthy text surpassing the 512-token limit.

2.10 Model Evaluation: Assessing Performance

The GDP growth rate detection models are evaluated on the testing data through a 5-fold cross-validation process. K-fold cross-validation is a well-established method for model evaluation, involving the division of data into training and test sets, which are then distributed into a specified number of folds [31]. Several key metrics are employed to assess the model's performance for text classification, including accuracy, precision, recall, and F1 score. These metrics are calculated based on the information provided by the confusion matrix, as illustrated in Table 2, a vital tool in evaluating the model's effectiveness in distinguishing between different GDP movement categories.

Table 2. Confusion Matrix in The Context of Binary Classification for A GDP Movement Detection Case Study.

		Predicted	
True	Label	Down	Up
	Down	TN (true negatives)	FP (false positives)
	Up	FN (false negatives)	TP (true positives)

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1 - score = \frac{2 \times (recall \times precision)}{(recall + precision)} \quad (6)$$

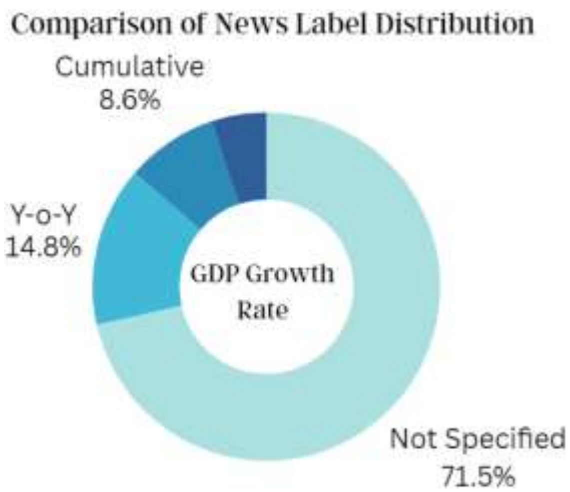
Key metrics such as accuracy, precision, recall, and F1 score are utilized to evaluate model performance, derived from the confusion matrix illustrated in **Table 2**. Accuracy provides an overall measure of the model's correctness, while precision quantifies its ability to identify positive cases correctly. Recall focuses on the model's capacity to detect all actual positives, and the F1 score balances precision and recall, which is particularly useful for imbalanced datasets. These metrics collectively offer insights into the model's effectiveness in distinguishing between different GDP movement categories, ensuring robust evaluation of its performance.

3. RESULTS AND DISCUSSION

3.1 Building the Online News Dataset

The data collection process, conducted through web scraping from the Detik.com news portal, yielded 13,286 records. After filtering out irrelevant entries, the dataset was refined to 1,586 news articles for further analysis. The selection of relevant articles was guided by the consideration that the percentage of relevant news articles within each industry sector was notably low, falling below 25%. This prompted the utilization of a sampling method during the news selection process. The chosen sampling technique was circular systematic random sampling, aiming to obtain a hundred relevant news articles for each industry sector. Despite the method's application, some industry sectors fell short of the relevant news article target following the selection process. This occurred primarily in less popular industry sectors, including categories B, E, L, MN, and O.

The analyzed news data pertains to GDP growth rate, classified such as "not specified," "year-on-year," "cumulative-to-cumulative," and "quarter-to-quarter." The labeling process by annotators resulted in an imbalanced data distribution. However, this issue of imbalanced data was not addressed in the study to ensure that the representation of economic news aligns with real-world proportions. The reliability of the data labeling was assessed using Krippendorf alpha, yielding a value of 0.490. This low alpha value indicates a lower level of reliability in labeling the GDP growth rate. Further analysis revealed that some news articles contained multiple types of growth periods or insufficient time-related information in their narratives, contributing to the low labeling reliability. This issue in labeling is also reflected in the unequal distribution of the various labels, as illustrated in **Figure 1**.



082199074334**Figure 1.** **Comparison of News Label Distribution.**

3.2 GDP Growth Rate Detection through Online News Classification without Sentence Selection

Hyperparameter tuning was conducted on all machine learning models to enhance model performance. The classification evaluation results for GDP growth rates using machine learning are presented in **Table 3**. Among the three compared models, the SVM model demonstrates the best performance with an accuracy of 0.7811. While the achieved accuracy is commendable, there is still potential for further improvement. One contributing factor to this accuracy level is the prevalence of news articles that do not specify the period for GDP growth rates, resulting in data imbalance. Following discussions with subject matter experts, it is assumed that news articles with unspecified growth rate periods refer to annual GDP growth rates.

Following this, classification using transfer learning with pre-trained transformer-based language models was conducted to determine if there was an improvement in classification performance. The results of GDP growth rate classification evaluation using pre-trained Transformers models are presented in Table 4. The pre-trained language model Transformers that performed the best was IndoBERT-base, with an accuracy of 0.7899. Based on recall, F1 score, and accuracy, the IndoBERT-base outperforms the best baseline model, SVM. However, the precision of the IndoBERT-base is still lower than that of SVM. Table 6 also reveals that IndoBERT-large delivered poor GDP growth rate classification results. This indicates that IndoBERT-large might not be suitable for handling imbalanced data classification. Nevertheless, further analysis with model parameter adjustments and more diverse data is needed to validate these findings comprehensively.

Table 3. The Classification Performance of the Machine Learning Model for GDP Growth Rate

Model	Label	Precision	Recall	F1-Score	Accuracy
SVM	Not Specified	0.8182	0.9777	0.8907	0.7811
	Cumulative	0.5000	0.0719	0.1183	
	Q-to-Q	0.6600	0.2733	0.3585	
	Y-o-Y	0.4911	0.3076	0.3752	
	macro avg	0.6173	0.4076	0.4357	
Random Forest	Not Specified	0.7899	0.9896	0.8783	0.7610
	Cumulative	0.3667	0.0530	0.0901	
	Q-to-Q	0.4667	0.0983	0.1612	
	Y-o-Y	0.4442	0.1727	0.2438	
	macro avg	0.5169	0.3284	0.3434	
Multinomial Naïve Bayes	Not Specified	0.7959	0.9687	0.8737	0.7597
	Cumulative	0.2400	0.0515	0.0821	
	Q-to-Q	0.4600	0.1217	0.1774	
	Y-o-Y	0.4680	0.2633	0.3325	
	macro avg	0.4910	0.3513	0.3665	

Table 4. The Classification Performance of The Pre-Trained Language Model Transformers

Model	Label	Precision	Recall	F1-Score	Accuracy
IndoBERT-base	Not Specified	0.8896	0.9242	0.9055	0.7899
	Cumulative	0.3809	0.4055	0.3832	
	Q-to-Q	0.6155	0.6783	0.6133	
	Y-o-Y	0.4972	0.3264	0.3759	
	macro avg	0.5958	0.5836	0.5694	
IndoBERT-large	Not Specified	0.7895	0.9793	0.8716	0.7597
	Cumulative	0.1000	0.0167	0.0286	
	Q-to-Q	0.2000	0.1700	0.1833	
	Y-o-Y	0.1910	0.2182	0.2036	
	macro avg	0.3201	0.3460	0.3218	
Multilingual BERT-uncased	Not Specified	0.8768	0.9395	0.9069	0.7887
	Cumulative	0.4212	0.3131	0.3487	
	Q-to-Q	0.5667	0.3483	0.4214	
	Y-o-Y	0.4702	0.4067	0.4333	
	macro avg	0.5837	0.5019	0.5276	

3.3. Application of Sentence Selection in GDP Growth Rate Detection

Following the performance observed in the first experiment for GDP growth rate classification, both using machine learning and transfer learning from pre-trained language models transformers, which did not yield significant metric improvements, the researchers identified the potential for further method development to optimize performance. Thus, the study continued with a second experiment involving modifications to the input text data through sentence selection mechanisms. Data modification for news articles involved selecting only the sentences that indicated GDP growth rate periods as input for the classification model. This approach was chosen due to the sparse distribution of GDP growth rate period contexts within news text data, which is often extensive, posing challenges for the model to effectively learn crucial features relevant to the classification context. By utilizing sentences as the unit of data, the context of news information was preserved. This facilitated effective learning by the pre-trained language model Transformers, which rely on an attention mechanism.

In the second experiment, the classification models remained the same as in the first experiment, with IndoBERT-base chosen as the state-of-the-art (SOTA) method for Indonesian text classification, supported by the first experiment's results that indicated IndoBERT-Base as the best-performing model for the research data. The baseline model used was the SVM classification model, which was identified as the best machine learning model in the first experiment. To obtain important text features, sentence selection leveraged a search method for identified keywords. Out of a total of 1,586 news articles, 159 articles did not contain any of the identified keywords. This occurred due to several possibilities, including the GDP growth rate mentioned in the most recent economic news (temporal information at the time of release), the GDP growth rate information being implicitly stated, or the news data being classified as "unspecified" labels. On average, four sentences were selected for each article, while the average length of an article was 25 sentences. The minimum number of sentences extracted through sentence selection was one sentence. If none were found based on the identified keywords, the article would have the first three sentences as input. There was no maximum limit for the number of sentences selected. However, the maximum number of sentences selected through sentence selection was 31 sentences.

The results of the sentence selection application will be used as input for the classification model and tested for performance in the second experiment. Experiment (a) will compare GDP growth rate classification performance with sentence selection application and unmodified full-text news data on the baseline model. The performance of the classification results using the SVM machine learning model with input data modified by sentence selection (Experiment I.a) will be compared with the performance of the same model using input data of unmodified full-text news (Experiment II.a), as indicated in Table 5. In terms of computational efficiency, observations comparing the processing time of SVM (Experiment a) and IndoBERT (Experiment b) were conducted on the Google Colaboratory platform with specifications including an Intel Xeon CPU @ 2.30 GHz, 13 GB RAM, Tesla K80 accelerator, and 12 GB GDDR5 VRAM.

Table 5. Comparison of Classification Performance After Sentence Selection with The Baseline Model.

Experiment	Label	Precision	Recall	F1 Score	Support	Accuracy	Hyperparameter tuning (s)
I.a (SVM, with Sentences Selection)	Not Specified	0.8410	0.9774	0.9040	117	0.8038	1172.6
	Cumulative	0.6595	0.2002	0.2828	12		
	Q-to-Q	0.7400	0.3950	0.4979	9		
	Y-o-Y	0.5238	0.3649	0.4251	22		
	macro avg	0.6911	0.4844	0.5275	159		
	weighted avg	0.7835	0.8038	0.7706	159		
II.a (SVM, without Sentences Selection)	Not Specified	0.8182	0.9777	0.8907	117	0.7811	4900.8
	Cumulative	0.5000	0.0719	0.1183	12		
	Q-to-Q	0.6600	0.2733	0.3585	9		
	Y-o-Y	0.4911	0.3076	0.3752	22		
	macro avg	0.6173	0.4076	0.4357	159		
	weighted avg	0.7457	0.7811	0.7352	159		
Improvement In Classification performance & Processing Time	Not Specified	0.0228	-0.0002	0.0133		0.0226	-3728.2
	Cumulative	0.1595	0.1283	0.1644			
	Q-to-Q	0.0800	0.1217	0.1395			
	Y-o-Y	0.0328	0.0573	0.0499			
	macro avg	0.0738	0.0767	0.0918			
	weighted avg	0.0378	0.0226	0.0354			

Based on the evaluation metrics detailed in Table 6, the implementation of sentence selection on input news data has led to improvements across almost all evaluation metrics. Only one evaluation metric shows a slight decrease, with a relatively small value compared to the increases obtained in other evaluation metrics, which is -0.0002. On average, there is an increase of 0.0226 points in accuracy after applying sentence selection to the SVM model input. The accuracy of the SVM model with input data processed through sentence selection reaches 0.8038. In addition to the measurable increase in classification performance indicated by evaluation metrics, the advantage of applying sentence selection to the classification input is also demonstrated by faster data processing times. In machine learning-based classification, the data processing phase that takes a significant amount of time is hyperparameter tuning. The application of sentence selection successfully reduced the time required for hyperparameter tuning from 4900.8 seconds to 1172.6 seconds on the same device. Experiment (b) will compare GDP growth rate classification performance with sentence selection application and unmodified full-text news data in the transfer learning method from pre-trained language model Transformers. The performance of the classification results using transfer learning from the IndoBERT-base model with input data modified by sentence selection (Experiment I.b) will be compared with the performance of the same model using input data of unmodified full-text news (Experiment II.b), as shown in **Table 6**.

Table 6. Comparison of Pretrained Tranformers' Classification Performance with the Sentence Selection.

Experiment	Label	Precision	Recall	F1Score	Support	Accuracy	Hyperparameter tuning (s)
I.b (IndoBERT-base, with Sentences Selection)	Not Specified	0.8938	0.9472	0.9187	117	0.8063	783.0
	Cumulative	0.4165	0.4318	0.4054	12		
	Q-to-Q	0.6443	0.5867	0.5937	9		
	Y-o-Y	0.5392	0.3464	0.4126	22		
	macro avg	0.6234	0.5780	0.5826	159		
	weighted avg	0.8010	0.8063	0.7951	159		
II.b (IndoBERT-base, without Sentences Selection)	Not Specified	0.8896	0.9242	0.9055	117	0.7899	1411.4
	Cumulative	0.3809	0.4055	0.3832	12		
	Q-to-Q	0.6155	0.6783	0.6133	9		
	Y-o-Y	0.4972	0.3264	0.3759	22		
	macro avg	0.5958	0.5836	0.5694	159		
	weighted avg	0.7877	0.7899	0.7804	159		
Improvement In Classification performance & Processing Time	Not Specified	0.0042	0.0230	0.0132		0.0164	-628.4
	Cumulative	0.0356	0.0263	0.0222			
	Q-to-Q	0.0288	-0.0916	-0.0196			
	Y-o-Y	0.0420	0.0200	0.0367			
	macro avg	0.0276	-0.0056	0.0132			
	weighted avg	0.0133	0.0164	0.0147			

According to the evaluation metrics detailed in **Table 6**, implementing sentence selection for the news data input led to improvements in almost all evaluation metrics. However, a few evaluation metrics showed a decline, specifically the recall and f1-score metrics for the "Q-to-Q" label, indicating decreases of 0.0916 and 0.0196, respectively. An average increase in accuracy of 0.0164 points was observed after applying sentence selection to the IndoBERT-base model's input. The accuracy of the IndoBERT-base model with input data resulting from sentence selection reached 0.8063. The magnitude of the accuracy improvement in the IndoBERT-base model was not as significant as the improvement seen in the SVM model. Nevertheless, based on the accuracy values obtained after the application of sentence selection, the classification of GDP growth rates using the transfer learning method from the pre-trained language model Transformers proved to perform better than the baseline model using machine learning methods.

In addition to the measurable performance improvement, as indicated by evaluation metrics, the advantage of applying sentence selection to the classification input is also demonstrated by faster data processing times. In classification using the transfer learning method, the data processing stage that consumes a significant amount of time is the fine-tuning phase. Implementing sentence selection successfully reduced the time required for the fine-tuning phase from the previous 1411.4 seconds to 783 seconds when performed on the same hardware.

4. CONCLUSIONS

Several conclusions can be drawn regarding the research objectives based on the research findings presented. The study successfully constructed an online news dataset to classify GDP growth rates. This dataset exhibited an imbalance in labels. However, the research should have addressed this issue to ensure that the economic news aligns with real-world representation. Sentence selection was implemented to classify GDP growth rates to extract essential features from news articles, improving performance in long-text classification tasks with specific contextual requirements. This approach ensures that not all information from news articles is necessary for classification. The results of the experiments demonstrated a notable enhancement in performance and a reduction in processing time for both machine learning and transfer learning. These conclusions highlight the effectiveness of sentence selection in improving the classification of GDP growth rates from online news articles, leading to better accuracy and faster processing times for both machine learning and transfer learning models

REFERENCES

[1] F. Khairani, A. Kurnia, M. N. Aidi, and S. Pramana, "Predictions of Indonesia Economic Phenomena Based on Online News Using Random Forest," *SinkrOn*, vol. 7, no. 2, pp. 532–540, Apr. 2022, doi: 10.33395/sinkron.v7i2.11401.

[2] S. G. Chowdhury, S. Routh, and S. Chakrabarti, "News Analytics and Sentiment Analysis to Predict Stock Price Trends." [Online]. Available: www.ijcsit.com

[3] J. J. Duarte, S. Montenegro González, and J. C. Cruz, "Predicting Stock Price Falls Using News Data: Evidence from the Brazilian Market," *Comput Econ*, vol. 57, no. 1, pp. 311–340, Jan. 2021, doi: 10.1007/s10614-020-10060-y.

[4] B. S. Kumar and V. Ravi, "A survey of the applications of text mining in financial domain," *Knowl Based Syst*, vol. 114, pp. 128–147, Dec. 2016, doi: 10.1016/j.knosys.2016.10.003.

[5] C. Retno, Y. 1∗, and H. Setiawan, "EDUKATIF: JURNAL ILMU PENDIDIKAN Analisis Framing dan Diksi Berita pada Media Online Detik Travel dan CNN Indonesia Sebagai Bahan Ajar Teks Berita," *Jurnal Ilmu Pendidikan*, vol. 4, pp. 803–814, 2022, doi: 10.31004/edukatif.v4i1.1859.

[6] K. Fiok *et al.*, "Text Guide: Improving the Quality of Long Text Classification by a Text Selection Method Based on Feature Importance," *IEEE Access*, vol. 9, pp. 105439–105450, 2021, doi: 10.1109/ACCESS.2021.3099758.

[7] G. Domeniconi, G. Moro, R. Pasolini, and C. Sartori, "A study on term weighting for text categorization: A novel supervised variant of tf.idf," in *DATA 2015 - 4th International Conference on Data Management Technologies and Applications, Proceedings*, SciTePress, 2015, pp. 26–37. doi: 10.5220/0005511900260037.

[8] A. Jadhav and V. Rajan, "Extractive Summarization with SWAP-NET: Sentences and Words from Alternating Pointer Networks."

[9] H. Li, J. Zhu, J. Zhang, C. Zong, and X. He, "Keywords-Guided Abstractive Sentence Summarization." [Online]. Available: www.aaai.org

[10] SimilarWeb, "Detik.com Website Traffic Rank Juli-September 2022," https://www.similarweb.com/website/detik.com/#overview.

[11] U. Brajawidagda, C. G. Reddick, and A. T. Chatfield, "Urban resilience in extreme events: Analyzing online news and Twitter use during the 2016 Jakarta terror attack," *Information Polity*, vol. 22, no. 2–3, pp. 159–177, 2017, doi: 10.3233/IP-170410.

[12] S. Vijayarani and M. P. Research Scholar, "Preprocessing Techniques for Text Mining-An Overview."

[13] V. S and J. R, "Text Mining: open Source Tokenization Tools – An Analysis," *Advanced Computational Intelligence: An International Journal (ACII)*, vol. 3, no. 1, pp. 37–47, Jan. 2016, doi: 10.5121/acii.2016.3104.

[14] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization."

[15] D. Liparas, Y. Hacohen-Kerner, A. Moumtzidou, S. Vrochidis, and I. Kompatsiaris, "LNCS 8849 - News Articles Classification Using Random Forests and Weighted Multimodal Features," 2014. [Online]. Available: http://www.bbc.com/news/business-25445906

[16] Universitas Diponegoro. Faculty of Science and Mathematics. Department of Informatics, Institute of Electrical and Electronics Engineers. Indonesia Section, Universitas Diponegoro. IEEE Student Branch, and Institute of Electrical and Electronics Engineers, *ICICoS 2019 : the 3rd International Conference on Informatics and Computational Sciences : proceedings : October 29th -30th, 2019, Semarang, Central Java, Indonesia*.

[17] M. Panda and M. R. Patra, "EVALUATING MACHINE LEARNING ALGORITHMS FOR DETECTING NETWORK INTRUSIONS," 2009.

[18] A. E. Brouwer and W. H. Haemers, "Spectra of graphs-Monograph," 2011.

[19] C. C. Aggarwal and C. X. Zhai, "A survey of text classification algorithms," in *Mining Text Data*, vol. 9781461432234, Springer US, 2012, pp. 163–222. doi: 10.1007/978-1-4614-3223-4_6.

[20] T. B. Shahi and A. K. Pant, "Nepali news classification using Naïve Bayes, Support Vector Machines and Neural Networks," in *Proceedings - 2018 International Conference on Communication, Information and Computing Technology, ICCICT 2018*, Institute of Electrical and Electronics Engineers Inc., Mar. 2018, pp. 1–5. doi: 10.1109/ICCICT.2018.8325883.

[21] U. Suleymanov, S. Rustamov, M. Zulfugarov, O. Orujov, N. Musayev, and A. Alizade, "Empirical Study of Online News Classification Using Machine Learning Approaches," in *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*, IEEE, Oct. 2018, pp. 1–6. doi: 10.1109/ICAICT.2018.8747012.

[22] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.05387

[23] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia."

[24] Medved M and Suchomel V, "Indonesian web corpus (idWac)," *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (FAL), Faculty of Mathematics and Physics, Charles University*, 2017.

[25] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," Online. [Online]. Available: https://huggingface.co/

[26] S. Wu and M. Dredze, "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT." [Online]. Available: https://github.com/

[27] Y. Wang, W. Che, J. Guo, Y. Liu, and T. Liu, “Cross-Lingual BERT Transformation for Zero-Shot Dependency Parsing,” Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.06775

[28] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” [Online]. Available: https://github.com/tensorflow/tensor2tensor

[29] S. Wu and M. Dredze, “Are All Languages Created Equal in Multilingual BERT?,” May 2020, [Online]. Available: http://arxiv.org/abs/2005.09093

[30] R. Nallapati, F. Zhai, and B. Zhou, “SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents,” Nov. 2016, [Online]. Available: http://arxiv.org/abs/1611.04230

[31] P. Refaeilzadeh, L. Tang, and H. Liu, “C Cross-Validation.”