Journal of
Information Systems Engineering
and Business Intelligence

Vol.10, No.1, February 2024
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Sentiment Analysis on a Large Indonesian Product Review Dataset

Ade Romadhony [1)*], Said Al Faraby [2)], Rita Rismala [3)], Untari Novia Wisesti [4)], Anditya Arifianto [5)]

[1)2)3)4)5)]*School of Computing, Telkom University, Bandung, Indonesia*

[1)]aderomadhony@telkomuniversity.ac.id, [2)]saidalfaraby@telkomuniversity.ac.id, [3)]ritaris@telkomuniversity.ac.id, [4)]untarinw@telkomuniversity.ac.id, [5)]anditya@telkomuniversity.ac.id

Abstract

Background: The publicly available large dataset plays an important role in the development of the natural language processing/computational linguistic research field. However, up to now, there are only a few large Indonesian language datasets accessible for research purposes, including sentiment analysis datasets, where sentiment analysis is considered the most popular task.
Objective: The objective of this work is to present sentiment analysis on a large Indonesian product review dataset, employing various features and methods. Two tasks have been implemented: classifying reviews into three classes (positive, negative, neutral), and predicting ratings.
Methods: Sentiment analysis was conducted on the FDReview dataset, comprising over 700,000 reviews. The analysis treated sentiment as a classification problem, employing the following methods: Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), LSTM, and BiLSTM.
Result: The experimental results indicate that in the comparison of performance using conventional methods, MNB outperformed SVM in rating prediction, whereas SVM exhibited better performance in the review classification task. Additionally, the results demonstrate that the BiLSTM method outperformed all other methods in both tasks. Furthermore, this study includes experiments conducted on balanced and unbalanced small-sized sample datasets.
Conclusion: Analysis of the experimental results revealed that the deep learning-based method performed better only in the large dataset setting. Results from the small balanced dataset indicate that conventional machine learning methods exhibit competitive performance compared to deep learning approaches.

Keywords: Indonesian review dataset, Large dataset, Rating prediction, Sentiment analysis

Article history: Received 31 July 2023, first decision 22 December 2023, accepted 26 February 2024, available online 28 February 2024

I. INTRODUCTION

Sentiment analysis is an important task in the natural language processing and text mining domain, with numerous real-world applications such as product review analysis and political opinion mining. Studies on sentiment analysis in English have been ongoing since the last decade and remain a primary topic in text classification tasks [1][2], while research on sentiment analysis in the Indonesian language emerged as a trend later. Early studies on English sentiment analysis primarily focused on formal text. However, with increased public access to the Internet and the proliferation of user-generated content, research in this field has shifted from formal to informal text. Approaches to sentiment analysis range from conventional machine learning-based methods to deep learning-based methods [3]-[4]. The availability of large review/rating datasets has spurred research in this field, exemplified by popular datasets such as the IMdB dataset [5], MovieLens dataset [6] and Amazon product review dataset [7][8]. However, most of these datasets are in English, limiting research opportunities in sentiment analysis for languages other than English. One notable exception is the Large-Scale Arabic Review (LABR) dataset [9], which contains review text in a non-English language.

Several sentiment analysis studies have been conducted on Indonesian text, covering various domains such as the restaurant domain [10], [11] , the movie domain [12], [13], application review [14], transportation service review [15], hotel review [16], [17], and beauty product review [18], [19]-[20]. However, most available datasets are small, containing only hundreds to thousands of reviews. This limitation hampers research in sentiment analysis for

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.10.1.167-178

Indonesian, particularly when a large dataset is required. Only one previous study has employed a large Indonesian review dataset, focusing on a recommender system using the collaborative filtering method [20], which also shares the same data source as our study. However, while the previous study utilized a large review dataset to discuss recommender systems, our study focuses on sentiment analysis tasks.

In this study, we present an investigation into the Female Daily review (FDReview) dataset. While there have been prior studies on sentiment analysis of beauty product reviews [18], [19], [20], these studies utilized smaller datasets. Conducting sentiment analysis on a much larger dataset could support a broader understanding of the market compared to analysis on smaller datasets. Market analysis of the beauty industry in Indonesia plays a crucial role since this industry significantly contributes to Indonesia's economic improvement[21].

In this work, following LABR [9], we conducted two experiments on sentiment analysis: sentiment classification and rating prediction. Sentiment classification involves categorizing user reviews into sentiment polarity tasks, wherein this study, we utilize three classes: positive, neutral, and negative. Rating prediction entails identifying user ratings based on their reviews, utilizing a five-star rating scale (1-5) in this study. We employed both conventional machine learning methods and deep learning-based methods to investigate which approach is best suited for large beauty product review case studies. Despite the current trend favoring deep learning-based approaches in sentiment analysis tasks due to their excellent performance [22], several studies have revealed that conventional machine learning outperforms deep learning-based methods. For instance, a study on sentiment analysis of Lithuanian text found that the traditional machine learning method, Multinomial Naïve Bayes (MNB), exhibited better performance than the deep learning-based method (Convolutional Neural Network/CNN) [22]. Similarly, another sentiment analysis study on a book review dataset also revealed that neural network methods outperformed deep learning-based methods [23]

We employed the following conventional machine learning methods: Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM), as well as deep learning-based classification methods: LSTM and BiLSTM. MNB and SVM are among the best conventional machine learning approaches for sentiment analysis [22], [24], [25], while LSTM and BiLSTM are also popularly used as deep learning-based methods due to their performance over textual data [26] [27].

FDReview is a large skincare and makeup product dataset (containing more than 700,000 reviews), and to the best of our knowledge, no other sentiment analysis study on Indonesian text has been conducted using a dataset as large as the one utilized in this study. The dataset comprises more than 700,000 reviews from over 60,000 users. The reviews were primarily gathered from online forum members and are believed to be honest reflections from the users.

Our contribution in this work are as follows: 1) Providing analysis on a large Indonesian product review dataset (consisting of hundreds of thousands of reviews) 2) Offering a baseline system for sentiment analysis and rating prediction tasks by presenting experiments and analyses of those tasks on a large Indonesian product review dataset, specifically in the beauty product domain. Additionally, we conducted experiments on smaller-sized datasets to explore the gaps in sentiment analysis performance across different dataset sizes. According to prior studies investigating the impact of dataset sizes on supervised sentiment analysis tasks, experiments on larger datasets yield higher performance [22-23]

II. Related Work

There are few Indonesian review datasets that have been used in sentiment analysis tasks. An early study on sentiment analysis of Indonesian review text was conducted by [12] , using translated movie reviews from the Internet Movie Database (www.imdb.com). The experimental results on translated reviews showed slightly lower accuracy compared to the performance on the source text in English.

Other research that utilized genuine corpus in the Indonesian language includes sentiment analysis on mobile banking reviews [14], aspect-based sentiment analysis on restaurant review [10], and sentiment analysis on movie review [13]. The mobile banking reviews used by [14] were presented in informal Indonesian language, consisting of 50 positive and 50 negative reviews. The movie review dataset mentioned in [13] consists of 783 positive reviews and 418 negative reviews, while the restaurant reviews explored in [10] consist of 992 sentences for training and 383 sentences for evaluation. When checking the Nusa Catalogue using the keyword "sentiment analysis," we obtain 13 listed datasets, with 9 of them containing datasets in the Indonesian language [28]. The only dataset listed on the Nusa Catalogue that contains text in both Indonesian and English is NusaX [29]. Moreover, based on the domain, most publicly available datasets listed on the Nusa Catalogue are in the hotel review domain [30], automotive domain [31], restaurant domain [29], application domain, and general domain including text from social media [32], [33]. Given all the sentiment analysis datasets from user-generated content on the Nusa Catalogue, the Indonesian dataset with the highest number of texts is SmSA [34], which contains 12,760 sentences.

A study on rating prediction of review text from the Female Daily review website has been conducted by [20]. However, it only involved a small part of the whole dataset, consisting of 688 reviews. All the review datasets in the Indonesian language that were studied in a sentiment analysis task can be considered as small-sized datasets, especially when compared to the available English review datasets. In this paper, we attempt to address the lack of study in sentiment analysis using a large Indonesian review dataset.

As document classification tasks, sentiment classification and rating prediction have been widely addressed as supervised classification problems. Early studies on both tasks involved conventional machine learning methods such as Naive Bayes, MNB, Bernoulli Naive Bayes, Logistic Regression, and SVM [35], while recent studies have shown that deep learning-based methods are more popular.

A study by [4] provided a survey of deep learning-based methods applied in sentiment analysis and summarized that many deep learning techniques have achieved state-of-the-art results on various sentiment analysis tasks. They classified the sentiment analysis task into several categories, including document-level sentiment analysis, which we explore in this study. Previous studies on document-level sentiment analysis have used various classification methods and word representations. The classification methods include CNN, LSTM, and GRU-based sequence encoders, while word representations generally fall into two types: Bag of Words (BoW) and word embedding.

III. METHODS

In this work, before conducting experiments on sentiment analysis, we performed exploratory data analysis on the dataset. Based on our analysis, we conducted preprocessing to prepare the input for sentiment classification and rating prediction. After preprocessing, the dataset was split into training and testing sets. We defined several settings for both sentiment analysis and rating prediction tasks, based on the features and classification methods. Lastly, after completing the classification, we performed evaluation and result analysis. Fig. 1 shows the proposed system process flow.

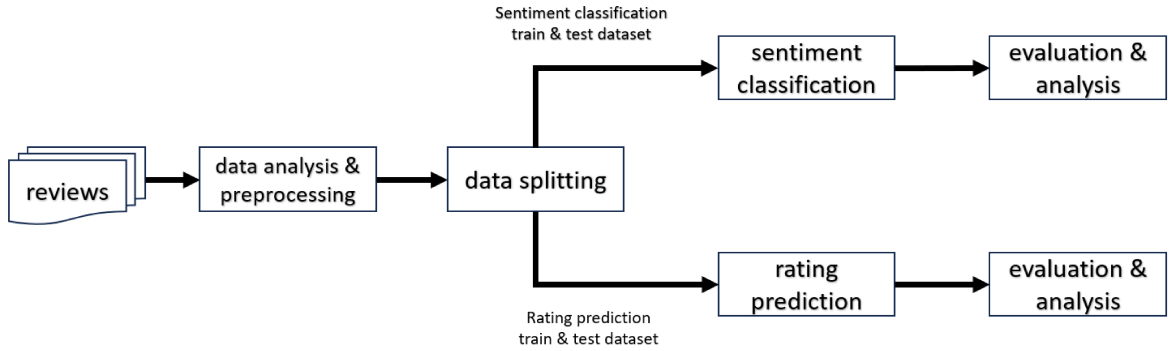

Fig. 1 System overview

A. Data Analysis & Preprocessing

FDReview (https://reviews.femaledaily.com) is a large skincare and makeup product dataset, and to the best of our knowledge, no other sentiment analysis study has been conducted using this dataset. The dataset consists of more than 700,000 reviews from over 60,000 users. The reviews were mostly submitted by members of the Female Daily community and are believed to be honest user reviews. The following are the dataset attributes of FDReview: user personal information, review timestamp, product category, product ID and name, review text, overall rating, packaging rating, price rating, and repurchase intention. The main features that we explore in this study are the review text and overall rating. Originally, the overall rating was presented as a floating-point number (0.0-5.0); then, we rounded the overall rating to an integer number. Furthermore, we mapped the overall rating to three sentiment classes that will be used in sentiment classification tasks. The sentiment classes are positive, neutral, and negative, where the positive class represents reviews with a rating of 4-5, the neutral class represents reviews with a rating of 3, and the negative class represents reviews with a rating of 1-2.

To analyze dataset characteristics, we adopted several procedures from text classification works: class distribution analysis, class statistical analysis, and analysis of the most frequent words/n-grams in each class. The presence of an imbalanced dataset is not uncommon in a review dataset, which usually contains more positive sentiments than neutral or negative sentiments [36]. Information on the dataset length of each class and the most frequent words in each class could support the selection of the most suitable features for text classification [37].

Since user review datasets are mostly written in informal language, we performed several preprocessing steps to transform the text into a more standardized form. We followed a similar approach to a prior study that performed word normalization, converting words written in informal language style to a formal one [38]. The normalization process was performed based on several words listed in a self-defined dictionary. We defined a particular dictionary to perform word normalization since each dataset has different characteristics, including the words contained in the review text. We also performed stopword removal based on a self-defined stopword list to remove several words that, based on analysis, do not correlate with the class characteristics, following a study on sentiment analysis [39]

After word normalization, the next step is feature extraction to prepare the input for classification with conventional machine learning methods. Features such as bag-of-words (BoW)/count unigram, BoW/count bigram, BoW/count trigram, tf-idf unigram, tf-idf bigram, and tf-idf trigram are among the popular features used in sentiment analysis tasks [40], [41]

B. Data Splitting

We conducted experiments using the dataset on two tasks: sentiment classification and rating prediction. The goal of sentiment classification is to predict the review polarity, with the positive class denoted by ratings 4 or 5, the neutral class denoted by rating 3, and the negative class denoted by ratings 1 or 2. The goal of rating prediction is to predict the rating class (1-5).

For both sentiment classification and rating prediction tasks, we split the dataset into training and test sets using an 80:20 ratio with stratified shuffling split method. An 80:20 ratio for splitting the dataset into train and test datasets is a common practice in data splitting; for example, this ratio was explored in a document-level text classification study[42].

Furthermore, besides conducting the experiment on the original large dataset, we also generated two types of small datasets containing 15,000 reviews each to investigate the impact of classification performance based on different dataset sizes and different class distributions (balanced vs unbalanced). Several studies have investigated the impact of dataset size and class distribution on classification performance [43], [44]. We generated two types of small datasets: unbalanced, representing the same rating distribution as the large dataset, and balanced, which has 300 reviews for each rating. We selected the first 300 reviews of each rating in each product category. The small dataset contains reviews from the top 10 categories (ten categories with the highest review numbers): *Lipstick, Facial Wash, Toner, Wash-Off, Serum & Essence, Face, Sun Protection, BB & CC Cream, Lipbalm & Treatments, and Scrub & Exfoliator*.

C. Classification: Sentiment Classification and Rating Prediction

In terms of classification methods, we followed prior works on text classification tasks that involved conventional machine learning methods and deep learning-based methods [22], [23], [45] to investigate which approach gives better performance in the case study. We chose the MNB and SVM methods since both are widely used in sentiment analysis problems and have shown good performance. On the other hand, we chose LSTM and BiLSTM (RNN-based) methods since these approaches are designed to view text as a sequence of words [24], which is suitable for the dataset characteristics as we are working with text data, and both methods usually exhibit good performance in sentiment analysis problems.

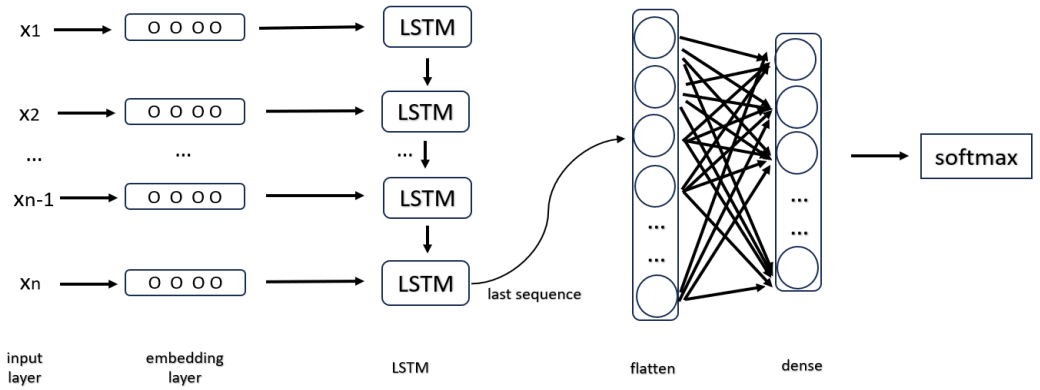

Fig. 2 Illustration of LSTM architecture

In this work, we performed sentiment classification and rating prediction with LSTM and BiLSTM using the following settings: an embedding layer with 64 dimensions, an LSTM layer with 64 units, a dense layer with 64 outputs and ReLU activation, and finally a softmax layer. The only additional layer for BiLSTM is the bidirectional layer. Fig. 2 shows an illustration of LSTM architecture and Fig. 3 shows an illustration of BiLSTM architecture.

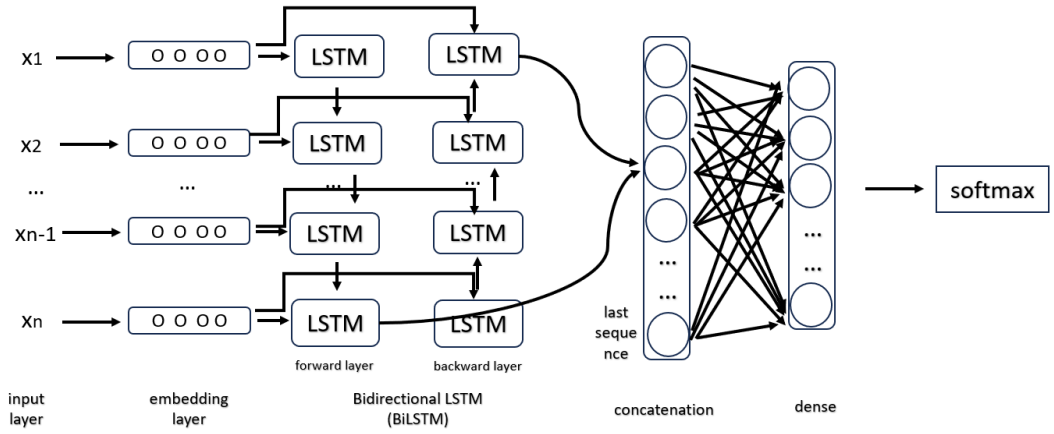

Fig. 3 Illustration of BiLSTM architecture

D. Evaluation and Analysis

To evaluate model performance on sentiment classification and rating prediction tasks, we adopt the following widely used evaluation metrics for multi-class text classification: accuracy, precision, recall, and F-1 score [46], [47]. Accuracy is calculated based on (1), where it represents the ratio of number of items that were classified correctly and numbers of all items.

$$\text{Accuracy} = \frac{TP+TN}{\text{TP+TN+FP+FN}} \quad (1)$$

While precision denotes the number of correctly predicted positive class items compared to the predicted positive class items, and recall denotes the number of correctly predicted positive class items compared to all positive class items. The formulas for these metrics are shown in (2) and (3).

$$\text{Precision} = \frac{TP}{\text{TP+FP}} \quad (2)$$

$$\text{Recall} = \frac{TP}{\text{TP+FN}} \quad (3)$$

TP (True Positive) represents the number of correctly predicted positive class items, FP (False Positive) represents the number of items predicted as the positive class while the actual class is non-positive, and FN (False Negative) represents the number of items predicted as the non-positive class while the actual class is positive. On the other hand, the F-1 score is the harmonic mean of precision and recall (4).

$$\text{F1-score} = \frac{2*precision*recall}{\text{precision+recall}} \quad (4)$$

IV. RESULT

In this section, we present experimental results on exploratory data analysis, preprocessing, and two tasks of sentiment analysis: sentiment classification and rating prediction. We conducted experiments using the Python language and several libraries that support the functionalities we needed for each experimental process. For example, we used:1) The NLTK library to obtain the most frequent unigrams/bigrams/trigrams; 2) The scikit-learn library for feature extraction (obtaining unigrams/bigrams/trigrams) and conventional machine learning methods for classification; 3) The TensorFlow library to perform classification using deep learning-based methods.

A. Exploratory Data Analysis and Preprocessing

Table I displays the basic statistics of the FDReview dataset, and Fig. 1 illustrates the rating distribution. It is evident that the rating distribution is imbalanced, with positive reviews (rating 4 and 5) dominating over neutral and negative

reviews. Despite the original dataset containing an imbalanced number of class items, in this work, we generated a smaller balanced dataset to explore the differences in characteristics between imbalanced and balanced datasets.

We investigated the reviews submitted from September 2013 to May 2019, which consisted of 729,061 reviews from 64,317 users who shared their reviews on a total of 27,744 products. Although most of the review text was written in informal Indonesian with a few words in English, there were several review texts that were fully written in English. We identified the review texts that were mostly (>50% of words) in Indonesian and kept the text to be processed further. We employed a Python library, langid, to perform text language identification [48]. Several review texts might contain English words; for example, the following review text: "*ih pokoknya ini* ***must have masker*** *bangetttt buat aku yang kulitnya* ***acne prone*** *n berminyak ini*" ("Anyway, this is really a must have mask for me who have acne prone and oily skin"), however it is still identified as an Indonesian text.

TABLE 1
FDREVIEW BASIC STATISTIC

Statistics	Number
Number of reviews	702,634
Number of reviewers	63,440
Average reviews/user	11.07
Number of products	27,744
Average positive review text length	365.97
Average neutral review text length	341.55
Average negative review text length	336.65

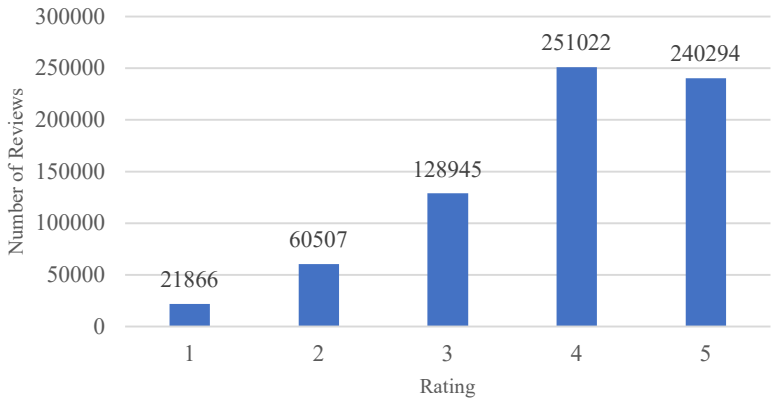

Fig. 4 Review counts per rating

Since this study focuses specifically on the relationship between review text and rating, we only consider review text and review rating as observation objects, despite the dataset containing additional information such as packaging rating, price rating, and repurchase intention. The original ratings from users were presented as continuous numbers; therefore, we converted the continuous rating to a discrete form within the 1-5 range. After removing review texts written mostly in English, deleting duplicate reviews, and eliminating incomplete reviews (e.g., those lacking product ID information or having a rating of 0), the total number of reviews is 702,634. Table 1 presents several important facts about the dataset, while Figure 1 illustrates the distribution of reviews for each rating. Based on the numbers in Figure 1, we can conclude that the dataset is unbalanced, with a significantly higher number of positive reviews than negative reviews.

Before performing the sentiment classification task, we conducted several preprocessing steps on the review text: a) casefolding; b) punctuation removal; c) word normalization; d) stopword removal. We defined several words in the stopword list based on the observation that these words carry no polarity information. The defined stopwords include: *ini* (this), *itu* (that), *sih, nya, dan* (and), *yang/yg* (which), *di* (in), *dengan* (with), *ke* (to). Additionally, we created a list for word normalization, as shown in Table 2. The objective of word normalization is to ensure that words with the same meaning are represented by a single string. For example, users often write the word *aku*/I with the following informal words: *aq*, *gw*, then we replace the word *aq* and *gw* with *aku* (I). Therefore, all words mean I will be represented only by single word, *aku*. We chose the word normalization list based on occurrence frequency and also the polarity information. For example, we include the word that shows negation such as *ga/gk/gak* (not) in the list, since it has important effect on the review sentiment.

TABLE 2
WORD NORMALIZATION LIST

Source word	Normalized word
aq/gw	*aku* / I
bgs	*bagus* / good
bgt	*banget* / very
blm	*belum* / not yet
bs	*bisa* / can
ga/gk/gak	*tidak* / not
gt	*gitu* / no meaning expression
hrs	*harus* / must
Jd	*jadi* / so
jlk	*jelek* / bad
karna/krn	*karena* / because
kl	*kalau* / if
lg	*lagi* / again
mhl	*mahal* / expensive
mmg	*memang* / really
pake/pk	*pakai* / use
plg	*paling* / the most
tp	*tapi* / but
yg	*yang* / which

We also conducted an analysis on the most frequent words that appeared in each class. Table 3 presents a comparison of the 20 most frequent words/unigrams in positive, neutral, and negative reviews, while Tables 4 and 5 show the comparison of bigrams and trigrams. Based on the comparison of the most frequent unigrams, bigrams, and trigrams in positive, neutral, and negative reviews, we found that differences between sentiment classes are more pronounced in bigrams and trigrams. It is difficult to distinguish between two different sentiment classes based solely on the top frequent unigrams.

TABLE 3
TOP 20 UNIGRAMS ON EACH SENTIMENT CLASS

Class	Top 20 unigram
positive	*aku*/I, *banget*/ very, *tidak*/not, *pakai*/use, *tapi*/but, *jadi*/become, *karena*/because, *juga*/also, *bikin*/make, *buat*/for, *sama*/same, *kulit*/skin, *suka*/like, *ada*/exist, *muka*/face, *lagi*/again, *gitu*/that's it, *kalau*/if, *bisa*/can, *lebih*/more
neutral	*aku*/I, *tidak*/not, *tapi*/but, *pakai*/use, *jadi*/become, *banget*/very, *karena*/because, *aja*/just, *juga*/also, *bikin*/make, *ada*/exist, *gitu*/that's, *buat*/to, *untuk*/for, *suka*/like, *sama*/same, *kulit*/skin, *kalau*/if, *beli*/buy, *lebih*/more
negative	*aku*/I, *tidak*/not, *pakai*/use, *banget*/very, *tapi*/but, *jadi*/become, *karena*/because, *cocok*/suitable, *sama*/same, *beli*/buy, *bikin*/make, *muka*/face, *ada*/exist, *produk*/product, *jerawat*/acne, *lagi*/again, *aja*/only, *kulit*/skin, *gitu*/that's, *juga*/also

TABLE 4
TOP 20 BIGRAMS ON EACH SENTIMENT CLASS

Class	Top 20 bigram
positive	*aku pakai*/ I use, *tidak bikin*/will not affect, *suka banget*/really like, *aku suka*/I like, *banget sama*/really with, *bagus banget*/really good, *enak banget*/really good, *make up*/make up, *tapi tidak*/but not, *banget buat*/really for
	tidak ada/not exist, *tidak terlalu*/not really, *tahan lama*/durable, *kulit aku*/my skin, *bener bener*/really, *sama sekali*/at all, *bikin kulit*/affect skin, *aku tidak*/I am not, *cocok banget*/really fit, *pertama kali*/first time
neutral	*aku pakai*/I use, *aku tidak*/I am not, *tidak terlalu*/not really, *tidak bikin*/do not affect, *tidak ada*/not exist, *aku suka*/I like, *beli karena*/buy because, *tapi tidak*/but not, *biasa aja*/so so, *tidak cocok*/not fit, *tapi aku*/but I
	kurang suka/like less, *tahan lama*/durable, *kulit aku*/my skin, *suka sama*/like with, *aku beli*/I buy, *make up*/make up, *muka aku*/my face, *menurut aku*/I think, *bikin kulit*/affect skin
negative	*tidak cocok*/not fit, *aku pakai*/I use, *aku tidak*/I am not, *beli karena*/buy because, *sama sekali*/at all, *tidak ada*/not exist, *muka aku*/my face, *cocok sama*/fit with, *kulit aku*/my skin, *aku beli*/I buy, *malah bikin*/even make
	pertama kali/first time, *aku coba*/I try, *bikin muka*/affect face, *tapi tidak*/but not, *big no*/big no, *muncul jerawat*/acne appears, *pakai produk*/use product, *sama produk*/with product, *muka jadi*/face becomes

We employed unigram/bigram/trigram features for classification with conventional machine learning methods. We utilized the Python scikit-learn library to extract bag-of-words (BoW)/count and TF-IDF unigram/bigram/trigram features. Meanwhile, for classification with deep learning-based methods, we used the Tokenizer from the TensorFlow library.

TABLE 5
TOP 20 TRIGRAMS ON EACH SENTIMENT CLASS

Class	Top 20 trigram
positive	*suka banget sama*/really like with, *aku suka banget*/I really like, *tidak bikin kering*/not make skin dry, *banget sama produk*/really with product, *cocok banget buat*/really fit for, *tidak bikin kulit*/not make skin, *tidak bikin bibir*/not make lip
	bikin bibir kering/make lip dry, *tidak bikin muka*/not make face, *tapi tidak bikin*/but not make, *wanginya enak banget*/smells really nice, *aku pakai buat*/I use for, *pertama kali pakai*/first time using, *cinta banget sama*/really love with, *biasanya aku pakai*/usually I use, *bagus banget buat*/really good for, *sama sekali tidak*/absolutely not, *tidak bikin breakout*/not make breakout, *wanginya juga enak*/also smells nice, *paling aku suka*/I like most
neutral	*aku kurang suka*/I like less, *tidak ada efek*/no effect, *suka banget sama*/really like with, *tidak tahan lama*/not long lasting, *kurang suka sama*/do not really like, *aku suka banget*/I really like, *aku tidak suka*/I do not like,
	bikin bibir kering/make lip dry, *aku beli karena*/I buy because, *aku lebih suka*/I prefer, *so so aja*/just so so, *tapi aku tidak*/but I am not, *tidak bikin kering*/not make dry, *aku tidak cocok*/I do not fit, *tidak bikin breakout*/not make breakout, *pertama kali pakai*/first time use, *biasa aja tidak*/so so not, *aku pakai buat*/I use for, *tidak ada perubahan*/no changes, *tapi lama kelamaan*/but over time
negative	*aku tidak cocok*/I do not fit, *tidak cocok sama*/not fit with, *sama sekali tidak*/at all not, *tidak cocok aku*/me not fit, *tidak cocok pakai*/not fit use, *tidak ada efek*/no effect, *pertama kali pakai*/first time use,
	aku tidak suka/I do not like, *repurchase big no*/repurchase big no, *cocok sama produk*/fit with product, *ternyata tidak cocok*/turns out not fit, *muka aku jadi*/my face becomes, *bikin muka aku*/make my face, *aku beli karena*/I buy because, *suka banget sama*/really like with, *tidak cocok banget*/really not fit, *tidak akan repurchase*/will not repurchase, *cocok sama sekali*/fit at all, *tidak ada perubahan*/no changes, *aku pakai lagi*/I use again

B. Sentiment Classification and Rating Prediction

Based on the experiments conducted for sentiment analysis and rating prediction tasks, we evaluated the model performance using the following metrics: accuracy, F-1 score, precision, and recall. Table 6 displays the overall sentiment classification accuracy on three types of datasets using the conventional machine learning method. The results indicate that the SVM method with TF-IDF features achieved higher accuracy compared to the Multinomial Naive Bayes (MNB) method with count features. Additionally, the results demonstrate that the inclusion of bigram and trigram features did not significantly affect the overall accuracy.

TABLE 6
SENTIMENT CLASSIFICATION AND RATING PREDICTION ACCURACY

	Features	Large		Small-Balanced		Small-Unbalanced	
		SVM-Tf Idf	MNB-Count	SVM-Tf Idf	MNB-Count	SVM-Tf Idf	MNB-Count
Sentiment Classification	unigram	0.73	0.71	0.66	0.62	0.73	0.69
	unigram+bigram	0.73	0.71	0.66	0.62	0.73	0.69
	unigram+bigram+trigram	0.73	0.71	0.67	0.62	0.73	0.69
Rating Prediction	unigram	0.48	0.49	0.46	0.47	0.48	0.48
	unigram+bigram	0.48	0.49	0.44	0.47	0.49	0.49
	unigram+bigram+trigram	0.48	0.49	0.45	0.45	0.48	0.48

Table 6 also displays the accuracy of the rating prediction task. As expected, the results show lower accuracy compared to the sentiment classification task, since the rating prediction task is more challenging due to having more classes. Additionally, the results indicate that the MNB classification method performs slightly better, with the unigram feature yielding the best accuracy. Despite the initial observation suggesting that bigram and trigram features may provide more valuable information for distinguishing between positive and negative classes, the results from both the sentiment classification and rating prediction tasks show no significant improvement when using bigram or trigram features.

TABLE 7
SENTIMENT CLASSIFICATION F1 SCORES: CONVENTIONAL METHODS

Features	SVM -Tf Idf			MNB - Count		
	positive class	neutral class	negative class	positive class	neutral class	negative class
unigram	0.84	0.01	0.37	0.84	0.4	0.5
unigram+bigram	0.84	0.01	0.36	0.84	0.4	0.5
unigram+bigram+trigram	0.84	0.01	0.37	0.84	0.4	0.5

We also evaluated the F1 score for the positive, neutral, and negative classes. Table 7 presents the F1 scores for these three classes using conventional classification methods. It is worth noting that the F1 scores for the neutral and negative classes are lower compared to those for the positive classes. The lower F1 scores for the neutral and negative

classes can be attributed to the lower number of neutral and negative reviews, making them more challenging to classify accurately.

TABLE 8
SENTIMENT CLASSIFICATION & RATING PREDICTION PERFORMANCE COMPARISON: CONVENTIONAL VS DEEP LEARNING BASED METHODS

		Small-balanced dataset				Small-unbalanced dataset				Large dataset			
		Prec.	Recall	F1-score	Acc.	Prec.	Recall	F1-score	Acc.	Prec.	Recall	F1-score	Acc.
Sentiment Classification	Best Conventional Method (SVM)	**0.64**	**0.67**	0.63	**0.67**	**0.68**	**0.73**	0.67	**0.73**	0.7	0.73	0.63	0.73
	LSTM	0.44	0.4	0.23	0.4	0.49	0.7	0.58	0.7	**0.77**	**0.79**	0.77	**0.79**
	BiLSTM	0.63	0.64	**0.64**	0.64	**0.68**	0.72	**0.69**	0.72	**0.77**	**0.79**	**0.78**	**0.79**
Rating Prediction	Best Conventional Method (MNB)	**0.46**	**0.47**	**0.46**	**0.47**	0.46	**0.49**	0.46	**0.49**	0.5	0.49	0.5	0.49
	LSTM	0.23	0.2	0.07	0.2	0.39	0.36	0.2	0.36	0.54	0.55	0.54	0.55
	BiLSTM	**0.46**	0.46	**0.46**	0.46	**0.48**	**0.49**	**0.48**	**0.49**	**0.56**	**0.57**	**0.56**	**0.57**

Tables 8 present the performance comparison of conventional learning methods and deep learning-based methods on sentiment classification and rating prediction tasks. We can conclude that the BiLSTM method exhibits the best performance, and that the deep learning-based methods (both LSTM and BiLSTM) outperform the conventional methods only when used with large datasets. Interestingly, the performance of LSTM on small balanced and unbalanced datasets is even lower than that of the conventional methods, while the performance of BiLSTM is comparable to it

V. DISCUSSION

Based on the experimental results presented in the previous section, we can confirm that the deep learning-based approach does not always outperform conventional machine learning approaches. In the small dataset setting, both balanced and imbalanced datasets, conventional machine learning approaches show comparable performance with the deep learning-based approach. We argue that the performance of the deep learning-based approach is highly correlated with dataset size, consistent with the widely known finding that deep learning-based approaches can yield excellent results under conditions of a large training dataset.

The model performance on the sentiment analysis task exhibits better performance than that of the rating prediction task model. This difference in performance is attributed to the disparity in the number of classes. A model with more classes possesses more complex characteristics compared to a model with fewer classes.

Based on a prior study on rating prediction that also utilized datasets from the same website, the experimental results showed an average accuracy of all classes (rating 1-5) to be 38.62% [20]. This study employed the Naïve Bayes algorithm as a classifier and count-based features. Although the study was conducted on a much smaller dataset size (688 reviews) compared to the dataset size used in our study, the performance on rating prediction was lower than the rating prediction accuracy achieved in this study (average accuracy of 49% on a large dataset, using the same classifier, Naïve Bayes). Two other studies discussed the sentiment aspect from datasets from the same source as our study, but the sentiment analysis studied was based on several aspects (product aspects such as price, scent) [18-19]. Therefore, we cannot fairly compare the system performance with our system performance.

The limitation of this study lies in the absence of experiments with large pretrained language models, such as BERT [25], IndoBERT [26] [27], mBERT [25], and XLMR [28]. Additionally, we have not explored the entirety of the dataset, which includes code-mixed text (Indonesian-English) and text in languages other than Indonesian (in cases where a review text is fully written in English). We suggest that future studies exploring the FDReview large dataset should investigate classification methods based on large pretrained language models.

VI. CONCLUSIONS

In this study, we presented an analysis of the largest Indonesian review dataset to date and conducted experiments on sentiment classification and rating prediction tasks. We analyzed the dataset characteristics based on sentiment and rating categories, reflected by several keywords found in each category/class. Additionally, we performed experiments on sentiment analysis tasks using two types of machine learning approaches: conventional methods (SVM and MNB) and deep learning-based methods (LSTM and BiLSTM). The experimental results for both tasks indicated that the

deep learning-based method, BiLSTM, outperformed the other methods, with deep learning-based methods showing better performance only in a large dataset setting. Overall, the results suggest that there is still room for improvement in classification performance. We plan to further study the dataset on several tasks, such as aspect-based opinion mining, while also exploring more features and methods, especially other advanced deep learning-based methods such as those involving large pretrained language models (BERT, mBERT, IndoBERT, XLMR, etc).

Author Contributions: *Ade Romadhony*: Conceptualization, Methodology, Writing-Original Draft, Writing-Review & Editing, Supervision. *Said Al Faraby*: Software, Investigation, Data Curation. *Rita Rismala*: Investigation, Data Curation. *Untari Novia Wisesti*: Literature Review. *Anditya Arifianto*: Literature Review, Data Analysis.

All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by Telkom University.

Acknowledgments: Thank you very much to Pak Arman and Mas Bern from Female daily Network who provided access and assisted us regarding the dataset used in this study.

Conflicts of Interest: No other potential conflict of interest relevant to this article was reported

Data Availability: The small size datasets can be accessed at the following link https://drive.google.com/drive/folders/1QpwRiNdTGPJj9ipoZnPE6N7GLPY2iIV1?usp=sharing. For anyone who needs to use the large dataset for research purposes, please contact us.

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ORCID:
Ade Romadhony: https://orcid.org/0000-0002-2930-1689
Said Al Faraby: https://orcid.org/0000-0002-1731-583X
Rita Rismala: https://orcid.org/0000-0002-4674-2400
Untari Novia Wisesti: https://orcid.org/0000-0001-5803-9643
Anditya Arifianto: https://orcid.org/0000-0003-1601-0185

REFERENCES

[1] B. Pang, L. Lee, and others, "Opinion mining and sentiment analysis," *Foundations and Trends® in information retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
[2] B. Liu, *Sentiment analysis and opinion mining*. Springer Nature, 2022.
[3] Z. Madhoushi, A. R. Hamdan, and S. Zainudin, "Sentiment analysis techniques in recent works," in *2015 science and information conference (SAI)*, 2015, pp. 288–291.
[4] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 8, no. 4, p. e1253, 2018.
[5] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.
[6] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *Acm transactions on interactive intelligent systems (tiis)*, vol. 5, no. 4, pp. 1–19, 2015.
[7] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *Proceedings of the 7th ACM conference on Recommender systems*, 2013, pp. 165–172.
[8] V. Nurcahyawati and Z. Mustaffa, "Vader Lexicon and Support Vector Machine Algorithm to Detect Customer Sentiment Orientation.," *Journal of Information Systems Engineering & Business Intelligence*, vol. 9, no. 1, 2023.
[9] M. Aly and A. Atiya, "Labr: A large scale arabic book reviews dataset," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2013, pp. 494–498.
[10] D. Ekawati and M. L. Khodra, "Aspect-based sentiment analysis for Indonesian restaurant reviews," in *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*, 2017, pp. 1–6.
[11] R. A. Laksono, K. R. Sungkono, R. Sarno, and C. S. Wahyuni, "Sentiment analysis of restaurant customer reviews on tripadvisor using na\"\ive bayes," in *2019 12th international conference on information & communication technology and system (ICTS)*, 2019, pp. 49–54.

[12] R. Manurung and others, "Machine learning-based sentiment analysis of automatic indonesian translations of english movie reviews," in *Proceedings of the International Conference on Advanced Computational Intelligence and Its Applications (ICACIA)*, 2008, pp. 1–6.
[13] Y. Nurdiansyah, S. Bukhori, and R. Hidayat, "Sentiment analysis system for movie review in Bahasa Indonesia using naive bayes classifier method," in *Journal of Physics: Conference Series*, 2018, p. 12011.
[14] M. A. Fauzi, T. Afirianto, and others, "Improving sentiment analysis of short informal Indonesian product reviews using synonym based feature expansion," *Telkomnika (telecommunication computing electronics and control)*, vol. 16, no. 3, pp. 1345–1350, 2018.
[15] A. R. Prananda and I. Thalib, "Sentiment analysis for customer review: Case study of GO-JEK expansion," *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, no. 1, p. 1, 2020.
[16] T. Sutabri, A. Suryatno, D. Setiadi, and E. S. Negara, "Improving na\"\ive bayes in sentiment analysis for hotel industry in Indonesia," in *2018 Third International Conference on Informatics and Computing (ICIC)*, 2018, pp. 1–6.
[17] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, "Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews," *Procedia Comput Sci*, vol. 179, pp. 728–735, 2021.
[18] C. C. P. Hapsari, W. Astuti, and M. D. Purbolaksono, "Naive Bayes Classifier and Word2Vec for Sentiment Analysis on Bahasa Indonesia Cosmetic Product Reviews," in *2021 International Conference on Data Science and Its Applications (ICoDSA)*, 2021, pp. 22–27.
[19] M. R. Danendra and Y. Sibaroni, "Sentiment Analysis on Beauty Product Reviews using LSTM Method," in *2021 9th International Conference on Information and Communication Technology (ICoICT)*, 2021, pp. 616–620.
[20] D. C. Oktaviana, B. Harjito, and S. W. Sihwi, "Rate Prediction of Cosmetic Product Based on Test Review from Website Female Daily Using Naive Bayes Classifier," *ITSMART: Jurnal Teknologi dan Informasi*, vol. 8, no. 1, pp. 19–25.
[21] T. Hasrudin and U. Sagena, "The Role of Indonesian Government Policy in Shaping the Competitive Landscape of the Southeast Asian Beauty Industry," *Research Horizon*, vol. 3, no. 4, pp. 433–444, 2023.
[22] J. Kapočiūtė-Dzikienė, R. Damaševičius, and M. Woźniak, "Sentiment analysis of lithuanian texts using traditional and deep learning approaches," *Computers*, vol. 8, no. 1, p. 4, 2019.
[23] P. Poomka, N. Kerdprasop, and K. Kerdprasop, "Machine learning versus deep learning performances on the sentiment analysis of product reviews," *Int J Mach Learn Comput*, vol. 11, no. 2, pp. 103–109, 2021.
[24] M. M. Danyal, S. S. Khan, M. Khan, M. B. Ghaffar, B. Khan, and M. Arshad, "Sentiment Analysis Based on Performance of Linear Support Vector Machine and Multinomial Na\"\ive Bayes Using Movie Reviews with Baseline Techniques.," *Journal on Big Data*, vol. 5, 2023.
[25] P. Anastasiou, K. Tzafilkou, D. Karapiperis, and C. Tjortjis, "YouTube Sentiment Analysis on Healthcare Product Campaigns: Combining Lexicons and Machine Learning Models," in *2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 2023, pp. 1–8. doi: 10.1109/IISA59645.2023.10345900.
[26] A. Varshney, Y. Kapoor, A. Thukral, R. Sharma, and P. Bedi, "Performing Sentiment Analysis on Twitter Data Using Deep Learning Models: A Comparative Study," in *Advances in Data and Information Sciences: Proceedings of ICDIS 2021*, Springer, 2022, pp. 371–381.
[27] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *Ieee Access*, vol. 7, pp. 51522–51532, 2019.
[28] S. Cahyawijaya *et al.*, "NusaCrowd: Open Source Initiative for Indonesian NLP Resources," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 13745–13818.
[29] G. Winata *et al.*, "NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 815–834.
[30] A. N. Azhar, M. L. Khodra, and A. P. Sutiono, "Multi-label aspect categorization with convolutional neural networks and extreme gradient boosting," in *2019 International Conference on Electrical Engineering and Informatics (ICEEI)*, 2019, pp. 35–40.
[31] A. Ilmania, S. Cahyawijaya, A. Purwarianti, and others, "Aspect detection and sentiment classification using deep neural network for Indonesian aspect-based sentiment analysis," in *2018 International Conference on Asian Language Processing (IALP)*, 2018, pp. 62–67.
[32] C. Tho, Y. Heryadi, L. Lukas, and A. Wibowo, "Code-mixed sentiment analysis of Indonesian language and Javanese language using Lexicon based approach," in *Journal of Physics: Conference Series*, 2021, p. 12084.
[33] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," *arXiv preprint arXiv:2011.00677*, 2020.
[34] A. Purwarianti and I. A. P. A. Crisdayanti, "Improving bi-lstm performance for indonesian sentiment analysis using paragraph vector," in *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 2019, pp. 1–5.
[35] B. Samal, A. K. Behera, and M. Panda, "Performance analysis of supervised machine learning techniques for sentiment analysis," in *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)*, 2017, pp. 128–133.
[36] Z. Xiao, L. Wang, and J. Y. Du, "Improving the performance of sentiment classification on imbalanced datasets with transfer learning," *IEEE Access*, vol. 7, pp. 28281–28290, 2019.
[37] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian, and A. Alothaim, "Exploratory data analysis and classification of a new Arabic online extremism dataset," *IEEE Access*, vol. 9, pp. 161613–161626, 2021.
[38] A. R. Naradhipa and A. Purwarianti, "Sentiment classification for Indonesian message in social media," in *2012 International Conference on Cloud Computing and Social Networking (ICCCSN)*, 2012, pp. 1–5.
[39] M. Al-Ayyoub, S. B. Essa, and I. Alsmadi, "Lexicon-based sentiment analysis of Arabic tweets," *International Journal of Social Network Mining*, vol. 2, no. 2, pp. 101–114, 2015.
[40] T. Widiyaningtyas, I. A. E. Zaeni, and R. Al Farisi, "Sentiment Analysis Of Hotel Review Using N-Gram And Naive Bayes Methods," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, 2019, pp. 1–5.

[41] T. Hasan and A. Matin, "Extract Sentiment from Customer Reviews: A Better Approach of TF-IDF and BOW-Based Text Classification Using N-Gram Technique," in *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, 2021, pp. 231–244.
[42] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood, and M. T. Sadiq, "Document-level text classification using single-layer multisize filters convolutional neural network," *IEEE Access*, vol. 8, pp. 42689–42707, 2020.
[43] A. Althnian *et al.*, "Impact of dataset size on classification performance: an empirical evaluation in the medical domain," *Applied Sciences*, vol. 11, no. 2, p. 796, 2021.
[44] C. Padurariu and M. E. Breaban, "Dealing with data imbalance in text classification," *Procedia Comput Sci*, vol. 159, pp. 736–745, 2019.
[45] C. N. Kamath, S. S. Bukhari, and A. Dengel, "Comparative study between traditional machine learning and deep learning approaches for text classification," in *Proceedings of the ACM Symposium on Document Engineering 2018*, 2018, pp. 1–11.
[46] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.
[47] D. Nguyen, "Comparing automatic and human evaluation of local explanations for text classification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1069–1078.
[48] M. Lui and T. Baldwin, "langid. py: An off-the-shelf language identification tool," in *Proceedings of the ACL 2012 system demonstrations*, 2012, pp. 25–30.