Sentiment Analysis on Indonesia-English Code-Mixed Data

1st Hilal Ramadhan Utomo
School of Computing
Telkom University
Bandung, Indonesia
hilalramadhan@student.telkomuniversity.ac.id

2nd Ade Romadhony
School of Computing
Telkom University
Bandung, Indonesia
aderomadhony@telkomuniversity.ac.id

***Abstract*—Social media users nowadays tend to use code-mixed language to express their opinion. The users of social media has exponentially risen in some countries like Indonesia, it has given rise to large volumes of code-mixed data, in which users use more than one language in a single text. Data with code-mixed is often noisy and most importantly, the monolingual model usually does not work well on it. This has been a challenge for Natural Language Processing (NLP) for processing and analyzing the data. In this work, we conduct experiment of sentiment analysis on English-Indonesian code-mixed data. The approach that is by utilizing a multilingual pre-trained model, mBERT. By analyzing the sentiment analysis models' predictions, we may assess how effectively the model can adjust to the implicit noises inherent in code-mixed data. The classification model's performance was tested using batch size and epochs parameters to discover and obtain the highest accuracy. The experimental result shows that the highest accuracy we obtained from the mBERT model that is trained with our dataset obtained was 76%, with 16 batch size and epochs used 7.**

***Index Terms*—Code-mixed, Sentiment analysis, mBERT, Natural Language Processing**

I. Introduction

In Indonesia, there is an actress named Cinta Laura Kiehl or well known as Cinta Laura. She was well known for her English accent when she speaks bahasa Indonesia, also because she has a mixed race Indonesia-Germany makes it difficult for her to speaks Indonesian fluently, so she always speaks mix between Indonesia-English when speaking. While the style of that code-mixing remains iconic in Indonesia even 10 years after the appearance of Cinta Laura, code-mixing phenomenon is common in multilingual countries like Indonesia. The practice of incorporating linguistic constructions from one language, such as words and phrases, into another is known as code-mixing. [1].

Social media such as Facebook and Twitter were famous over the past decade. Social media users have exponentially risen in some countries like Indonesia. This has also led to the big amount of code-mixing usage on each platform. It is a common phenomenon that occurs in multilingual communities, and it can take many different forms. For example, a person might use words or phrases from one language in the midst of speaking or writing in another language, or they might switch back and forth between languages within a single sentence or phrase. Code-mixing can also involve combining elements of different language varieties, such as combining standard and colloquial forms of a language or mixing dialects. Code-mixing is a natural and common part of multilingual communication, and it can serve various purposes, such as emphasizing a point, expressing solidarity with a particular group, or adding emphasis or nuance to a message. This has been a challenge for Natural Language Processing (NLP) for processing and analyzing the data. Code-mixed text is a challenge for the NLP practitioners community [2].

In the field of sentiment analysis and transliteration, code-mixing is now popular. It is challenging to use the right method to extract the correct sentiment from code-mixed data [3]. The social text data that is Hindi-English code-mixed have been studied by Vijay et al. [4]. Code-mixing text inherits the vocabulary and grammar of multiple languages and often forms new structures based on the user preference. This poses a challenge to sentiment analysis, as traditional semantic analysis approaches do not capture the meaning of sentences. The lack of annotated data available for sentiment analysis model, the presence of multiple languages in a single text document, and the need to handle language-specific characteristics, such as idioms and collocations also limits progress in this area. There are several techniques to do sentiment analysis on code-mixed data, such as Word2Vec, FastText, Convolutional Neural Network (CNN), multilingual BERT (mBERT), etc.

In this study, we focus on the code-mixed Indonesian-English text. Code-mixing has been referred to as intrasentential code-mixed [5]. The two or more languages were mixed between sentences to make a whole sentence, and can be merged with affixes to make the sentence more meaningful. There has been research for code-mixed Indonesia-English that was conducted in 2020 using word embedding, and the best result accuracy for the study is 67.27% for Code-Mixed Embedding [6]. Based on that study, Code-Mixed word embeddings also has good potential also for other NLP tasks that require cross-lingual or multilingual word embeddings, in this paper we explore more about code-mixed sentiment analysis using multilingual BERT

(mBERT).

II. Methodology

The purpose of this study is to build a sentiment analysis of Indonesian-English Code-Mixed dataset from social media, specifically Twitter using mBERT. The description of the sentiment analysis system steps can be seen in Figure1.

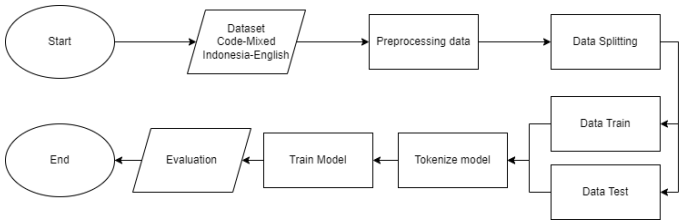

Fig. 1. System Overview

Figure 1 shows our proposed System overview. Starting from the dataset, we obtain 900 Indonesia-English code-mixed Tweets in total from previous research [7], [8]. Before analyzing the sentiment of Indonesian-English code-mixed dataset, we conduct several pre-processing steps before train and evaluate the classification model. We preprocess the available dataset by splitting it into 80% for train data and 20% for test data respectively.

A. Dataset and Annotation

Most research on social media texts to date has focused on English, while the majority of these texts are now in languages other than English. [9]. In this study, we concentrate on the Indonesia, a nation with 718 languages that has over than 200 million people, while mostly Indonesia and English are used for nation-wide communication [10]. Hence, Indonesians are multilingual by necessity and adaptation, and mix languages in social media contexts. We get 900 Tweets in total from the code-mixed corpus for the experiment. 825 Indonesia-English code-mixed Tweets were gathered for the dataset that is utilized in this study were obtained from [7] and 13170 Indonesian mix with Indonesia-English Tweets from [8]. We fetch all the data that is obtained from [7] for 825 Tweets and fetch code-mixed Tweets randomly from [8] for 75 Tweets.

In terms of natural language understanding, sentiment analysis can be considered an important subfield of semantic analysis because its goal is to recognize the topics people are talking about and their feelings about those topics [11]. The gold standard for labeling the constructed Tweets for the annotator can be seen in table I.

Based on the table I, the task of the annotators is to label all the data into positive and negative based on the Tweets constructed. The annotators were asked to label the data based on the number of positive or negative words in a Tweets which is marked with the following words example below.

Indonesia

TABLE I
Annotators Task

Label	Description
Positive	There is an explicit or implicit clue in the Tweets, suggesting the writer is in a positive state or can be based on the text if most of the text showed a non-negative word.
Negative	There is an explicit or implicit clue in the Tweets, suggesting the writer is in a negative state or can be based on the text if most of the text showed a negative word.

- **Positive:** *aktif, donasi, semangat, senang, enak, lucu, lancar, etc.*
- **Negative:** *sesat, jahat, kasar, bau, protes, rawan, aneh, etc.*

English

- **Positive:** active, donate, enthusiastic, happy, delicious, funny, smooth, etc.
- **Negative:** perverted, evil, rude, smelly, protest, vulnerable, weird, etc.

In this research, we use three annotators from college students to annotate our dataset. The annotators consist of two male and one female with the various ages 20-25 years(since most Twitter users in Indonesia came from that age [12]). Noise in code-mixed data extracted from social media platforms including abbreviations, spelling variations for the same word, no standard grammar misspellings, emojis, etc [13]. The noise in text will affect the training and the evaluation metric, we preprocess the dataset using Natural Language ToolKit (NLTK) 'stopwords' module and Regular Expression (RE) module. Stopwords module are basically removing the stopwords that do not have strong meaningful connotations for instance, 'will', 'they', 'and', 'a', etc. RE module functions is to check if a particular string matches a given regular expression and then remove the particular string. The statistics of labeled dataset can be seen on table II.

TABLE II
Dataset Statistics

Sentiment	Total reviews (Tweets)
Positive	463
Negative	437
Total	900

Based on the table II, the data are well distributed.We provide some examples from the dataset that were obtained along with the sentiment of Tweets that can be seen in table III.

TABLE III
DATASET SAMPLE

Code-Mixed Tweet	Translation	Sentiment Label
All the best Team Indonesia, bawa pulang juara!!!	All the best team Indonesia, bring home the champion	Positive
Its kinda weird while you chat with someone your not really past. Terus dia entah pura pura atau emang gatau dah beneran apa engga lupa dan ngapus kontak...Emosi.	It's kinda weird while you chat with someone you're not really past. Then he's either pretending or he doesn't know if it's real or not, he forgets and deletes contacts Emotions.	Negative
Apa yg bisa tidak disukai about this product :))	What is not to be like about this product	Positive

B. Model Architecture

The multi-layer bidirectional transformer encoder used in the model architecture of BERT is based on the initial implementation published in [14]. .Multilingual-BERT (mBERT) is a version of BERT [15], Multilingual BERT or mBERT is an open-source language representation model, mBERT trained on and usable with 104 languages. As a result, mBERT is able to comprehend all 104 languages while also understanding how each one relates to the others. mBERT recognizes when the information in many languages is semantically similar but it is not a translation technology that can comprehend all languages.

mBERT has been chiefly used for code-mixed data, research about the mBERT performance using an English-German pre-trained model was conducted in 2021 using the variety of BERT, mBERT got the scores from English-to-German around 27.8% and from German-to-English around 34.01%, this proven to be adequate to facilitate cross-lingual learning [16].

Figure 2 shows the model of mBERT architecture. mBERT model is trained using a technique called "masked language modeling", which involves predicting the missing tokens in a sentence based on the context provided by the rest of the sentence. 15% of the token places are randomly selected by the training data generator for prediction. [17]. This training process enables mBERT to learn contextual relationships between words and to capture the meaning of a word based on the words that come before and after it.

C. Evaluation

As an evaluation issue, sentiment analysis uses metrics such as confusion matrix, precision, recall, F-score, and accuracy for the evaluation, that are used to assess a variety of machine learning, rule-based, and deep learning methods. Also, average measures such as macro, micro, and weighted F-1 values.

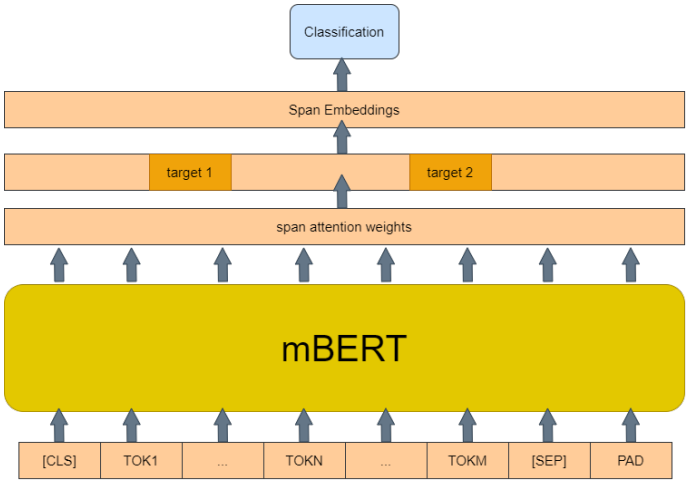

Fig. 2. mBERT architecture

Confusion Matrix is a kind of method used to evaluate the classification model's performance on datasets to get to know which are the true and false of the predicted label [18]. The four measurements, namely true positive (TP), true negative (TN), false positive (FP), and false negative (FN), as visualized in Table V [19].

TABLE IV
CONFUSION MATRIX

		Predicted Label	
		Positive	Negative
Actual Label	Positive	TP	FN
	Negative	FP	TN

In this study, our dataset has been evaluated through the use of Indonesia-English code-mixed sentiment analysis utilizing with Multilingual-BERT (mBERT) text classification experiments. We illustrate four evaluation metrics, including recall, accuracy, precision, and F1-measure as shown in equation 1, 2, 3 and 4 are used to measure the performance of the classification model in this research, with accuracy as the main objective.

In general, the accuracy metric calculates the proportion of accurate forecasts to all occurrences that were considered. [17]. One of the often used measures for evaluating classification issues is accuracy, which is calculated as the total number of correct predictions divided by the total number of predictions made for the dataset, by using the equation (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The percentage of positive patterns that are correctly classified is measured by recall. [17]. When False Negative is more important than False Positive, recall will be a valuable metric, by using the equation (2).

$$Recall(r) = \frac{TP}{TP + TN} \tag{2}$$

F Score, The harmonic mean of the recall(r) and precision(p) values is represented by this metric. [17]. It provides a synthesis of the Precision and Recall measurements. It is at its highest when Precision and Recall are equal, by using the equation (3).

$$F - score = \frac{2 * p * r}{p + r} \tag{3}$$

The positive measure in a positive class that are accurately predicted from the total anticipated patterns are measured by precision. [17]. Precision explains how many of the instances that were accurately predicted ultimately turned out to be positive. Precision is useful in the cases of False Positive is a higher concern than False Negatives by using the equation (4).

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

III. Result and Discussion

A. Testing result

We use a pre-trained model that is available in Huggingface called "bert-base-multilingual-cased" for training the Indonesia-English Code-mixed dataset that we obtained. The pre-training process substantially adheres to the existing literature on language model pre-training [15].

The testing result of this model is separated into two indexes used, which are weighted according to the label positive and negative. The result can be seen in Figure 3, it shows the obtained results between actual label and the prediction. The blue color used in the bar-plot indicates the actual values, while the orange value indicates the right predicted values by the program.

Based on Figure 3 we could see that the program could predict the positive value for 43 values over 56 of the actual values and the program could predict the negative value for 25 values over 34 of the actual values.

B. Evaluation

In order to discover and obtain the accuracy of the classification, evaluation of the model is needed after the predicted values of each index are obtained, the confusion matrix is used to evaluate the testing result; it distributes the actual predicted values included.

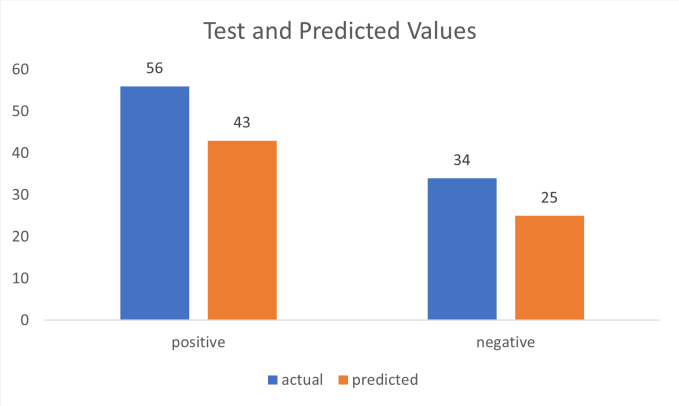

Fig. 3. Comparison between actual and predicted values

TABLE V
Confusion Matrix Data Tested on mBERT Model

		Predicted Label	
		Positive	Negative
Actual Label	Positive	25	9
	Negative	13	43

Table V shows the confusion matrix that is obtained from the 90 data that is tested on our mBERT model. The confusion matrix shows that the model predict 25 positive Tweets and 43 negative Tweets data correct. Also, predict 9 positive Tweets and 13 negative Tweets data wrong from 90 test data that is collected.

TABLE VI
Classification Report

	Precision	**Recall**	**F1-Score**	**Support**
Negative	0.66	0.74	0.69	34
Positive	0.83	0.77	0.80	56
Accuracy			0.76	90
Macro avg	0.74	0.75	0.75	90
Weighted avg	0.76	0.76	0.76	90

We do some experiments and try different settings on the model such as epoch and batch size, and come out with the best optimal record with 16 batch size and 7.0 epoch number for our dataset. The precision that we obtain from each label is 66% for negative labels and 83% for positive labels as shown on table VI.

Based on Figure 4, we conduct some experiment to the model to get most optimum parameter. We also use several batch

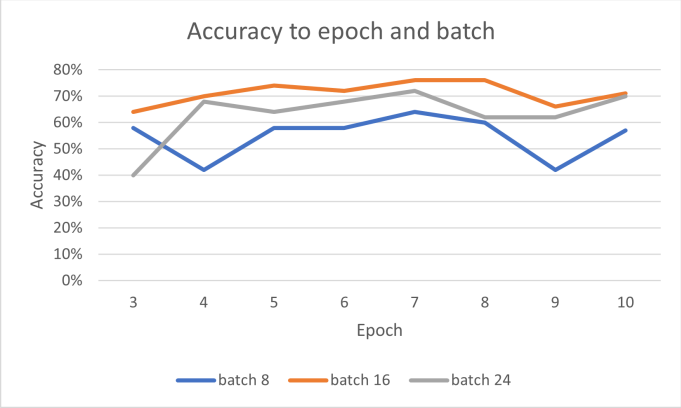

Fig. 4. Comparison between accuracy to epoch and batch size

and epoch settings to see how the model with our dataset react. The result of overall model accuracy that we get after conducting several parameter based on figure 4 is 76% with 90 test data, with the most optimum batch and epoch is 16 and 7 respectively. The model is stable and had the best accuracy is on batch 16, while batch 8 and 24 were unstable when we increase the epoch and it did not reach the optimum accuracy of the model. By looking at figure 5, the training loss is for model with 16 batch and 7 epoch by the optimum parameter setting. Training loss is decreasing by the time epoch value increase. This means that at that point, that the training model is not overfit or underfit the dataset.

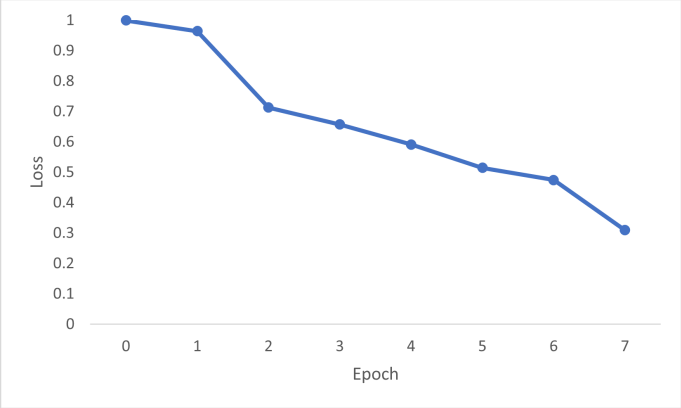

Fig. 5. Train Loss with 16 batch and 7 epoch

Based on our analysis, the program cannot determine the tweet that has an implicit meaning on the tweet. We provide some example that predicted wrong by the program.

- *saya lebih parah nder setangan tangan sampai ke punggung tangan ngelopek kering begitu rajinin pakai vaseline yang petroleum or whatever jelly nder it works well on me so far*

Translation: I have more severe hand attacks to the back of my hands peeling dry so diligently use vaseline which is petroleum or whatever jelly sender it works well on me so far

Actual sentiment: Positive

Predicted sentiment: Negative

- *sa aku sudah tau kamu kan jutek sama orang baru you know me so well berarti lah yah ya hahaha jutek sama jaga jarak memang beda tipis tebal*

Translation: sa I know that you don't care about new people you know me so well that means yeah hahaha being tough and keeping your distance is a little bit different

Actual sentiment: Negative

Predicted sentiment: Positive

- *keterpurukan pengetahuan politik ini yang kita alami disumbang juga oleh masyarakat yang sudah well educated tentang politik tidak menjalankan fungsi sosialnya sebagai bagian dari yang mencerdaskan bangsa*

Translation: The downturn in political knowledge that we are experiencing is also contributed by people who are already well educated about politics not carrying out their social functions as part of educating the nation.

Actual sentiment: Positive

Predicted sentiment: Negative

- *hari ini izin tidak masuk kantor karena sakit tapi karena memang lagi hectic di kantor jadinya tetap saja kerja terus saya pikir lebih enak kerja remote gini sih sepertinya saya lelah sama kantornya bukan kerjaannya tapi lelah sama cara kerjanya kantor juga deng resign or not to resign*

Translation: Today permission is absent from the office due to illness but because it's really hectic at the office so it keeps working I think it's better to work remotely like this I think I'm tired of the office not the job but tired of how the office works too with resigning or not to resign

Actual sentiment: Negative

Predicted sentiment: Positive

Based on the example above, one of the reason why the proposed model give wrong prediction is because there is some implicit meaning in the tweet, so that the token that is gathered by the program to predict the result is what was not expected. We analyze that several words are occurred more frequently than other words in positive Tweets, the word *sudah* (done) can be found in 89 positive Tweets and 27 negative Tweets, or *pikir* (think) can be found in 18 positive Tweets and 5 negative Tweets. Similar condition found in negative Tweets. However, not all Tweets that contains many positive Tweets has positive sentiment. Based on the previous false prediction, the first tweet mislabeled is since the first three words in the sentence are found in negative labeled Tweets. The word *parah* (critical)

can be found in three positive Tweets and 10 negative Tweets and the word *ngelopek* (peel off) found in 0 positive Tweet and five negative Tweets. This also occured in the second example, the word *sudah* (done) can be found mostly on positive Tweets and the word *tau* (know) found in 33 positive Tweets and 7 negative Tweets.

The possible explanation why the model mislabels the Tweets is because the number of positive or negative words in a Tweets might affect the prediction. As seen in table VI and table V, the total that the program predicted wrong is 22 data over 90 test data which means the mBERT model is able to obtain high overall accuracy on the positive and negative labels with 76% with the precision that we obtain from each label is 66% for negative labels and 83% for positive labels. This shows that the program has good training data so it could predict the data besides the Tweets that have an implicit meaning.

IV. Conclusion

Code-mixed data is essential to multilingual populations communication and culture. Based on evaluation result, conclusions can be made that using mBERT model can be able to predict and classify Indonesia-English code-mixed Tweets by using two labels of sentiment which is positive and negative with the optimum accuracy of 76% with the obtained dataset by using 16 batch size and 7 epochs. For future research, it is highly suggested to provide a good quality dataset and more train data to improve the classification model performance and make a higher accuracy of prediction. The trends of biases in code-mixed data can also be observed through explicable techniques. We believe that code-mixed NLP practitioners, developers, and academics will find great value in our work and explore potential avenues. The incorporation of explicable approaches forges a fresh route for upcoming study and development.

References

[1] S. Poplack and J. Walker, "Pieter muysken, bilingual speech: a typology of code-mixing. cambridge: Cambridge university press, 2000. pp. xvi+306." *Journal of Linguistics*, vol. 39, pp. 678 – 683, 11 2003.

[2] K. Chakma and A. Das, "Cmir: A corpus for evaluation of code mixed information retrieval of hindi-english tweets," *Computacion y Sistemas*, vol. 20, pp. 425–434, 09 2016.

[3] S. Shekhar, D. Sharma, D. Agarwal, and Y. Pathak, "Artificial immune systems-based classification model for code-mixed social media data," *IRBM*, vol. 43, no. 2, pp. 120–129, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1959031820301214

[4] J. Raj, "Machine learning based resourceful clustering with load optimization for wireless sensor networks," *Journal of Ubiquitous Computing and Communication Technologies*, vol. 2, pp. 29–38, 03 2020.

[5] C. Hoffmann, *Introduction to Bilingualism*, ser. Longman Linguistics Library. Taylor & Francis, 2014. [Online]. Available: https://books.google.co.id/books?id=NFugBAAAQBAJ

[6] A. N. Rizal and S. Stymne, "Evaluating word embeddings for Indonesian–English code-mixed text based on synthetic data," in *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*. Marseille, France: European Language Resources Association, May 2020, pp. 26–35. [Online]. Available: https://aclanthology.org/2020.calcs-1.4

[7] A. M. Barik, R. Mahendra, and M. Adriani, "Normalization of Indonesian-English code-mixed Twitter data," in *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 417–424. [Online]. Available: https://aclanthology.org/D19-5554

[8] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in Indonesian Twitter," in *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 46–57. [Online]. Available: https://aclanthology.org/W19-3506

[9] S. Schroeder, "Half of messages on Twitter aren't in English [STATS]," https://mashable.com/archive/half-messages-twitter-english/, 2008, [Online].

[10] G. Alrajafi, "The use of english in indonesia: Status and influence," *SIGEH ELT : Journal of Literature and Linguistics*, vol. 1, pp. 1–10, 03 2021.

[11] J. Zhao, K. Liu, and L. Xu, *Sentiment analysis: mining opinions, sentiments, and emotions*. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ..., 2016.

[12] APJII, "Infografis penetrasi dan pengguna internet indonesia survey 2017. pustaka pelajar asosiasi penyelenggara jasa internet indonesia," Jakarta, 2017.

[13] A. Priyanshu, A. Vardhan, S. Sivakumar, S. Vijay, and N. Chhabra, ""something something hota hai!" an explainable approach towards sentiment analysis on Indian code-mixed data," in *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*. Online: Association for Computational Linguistics, Nov. 2021, pp. 437–444. [Online]. Available: https://aclanthology.org/2021.wnut-1.48

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[16] H. Xu, B. V. Durme, and K. W. Murray, "Bert, mbert, or bibert? A study on contextualized embeddings for neural machine translation," *CoRR*, vol. abs/2109.04588, 2021. [Online]. Available: https://arxiv.org/abs/2109.04588

[17] M. Hossin and S. M.N, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, pp. 01–11, 03 2015.

[18] D. Bužić and J. Dobša, "Lyrics classification using naive bayes," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 1011–1015.

[19] s. M. Piryonesi and T. El-Diraby, "Data analytics in asset management: Cost-effective prediction of the pavement condition," *Journal of Infrastructure Systems*, vol. 26, 01 2020.