Sentiment Analysis of Code-mixed Social Media Data on Philippine UAQTE using Fine-tuned mBERT Model

Lany L. Maceda[1], Arlene A. Satuito[2], Mideth B. Abisado[3]
Computer Science and Information Technology Department, Bicol University, Legazpi City, Philippines[1, 2]
College of Computing and Information Technologies, National University, Manila, Philippines[3]

***Abstract*—The Universal Access to Quality Tertiary Education (UAQTE) marks a significant policy change in the Philippines. While the program's objective is to offer free higher education and tertiary education subsidies to eligible Filipino students, its viability and effectiveness have been subject to scrutiny and continuous evaluation. This study explores the sentiments of Filipinos towards UAQTE. Leveraging a fine-tuned multilingual Bidirectional Encoder Representations from Transformers (mBERT) model, we conducted sentiment analysis on code-mixed data. With minimal preprocessing, our model achieved an accuracy of 80.21% and an F1 score of 81.14%, surpassing previous related studies and confirming its effectiveness in handling code-mixed data. The results reveal that the majority of social media users view UAQTE positively or beneficially. However, negative sentiments highlight concerns related to subsidy delays, alleged fund misuse, and application challenges. Additionally, neutral sentiments center around subsidy-related announcements. These findings provide valuable insights for its key stakeholders involved in the implementation, enhancement, and evaluation of UAQTE.**

Keywords*—*Sentiment analysis; UAQTE; code-mixing; policy-making; multilingual BERT

I. Introduction

Countries worldwide increasingly recognize the importance of higher education for economic competitiveness [1]. This has led to a greater focus on understanding the costs and benefits of education and research, and the need for improved productivity in higher education [2]. Among the 17 Sustainable Development Goals (SDGs) set by the United Nations, the fourth goal specifically focuses on Quality Education. This goal is dedicated to achieving inclusive and equitable access to high-quality education and fostering lifelong learning opportunities for all individuals by the year 2030 [3].

One of the most significant turning points in Philippine education is the proactive efforts of the government to broaden the reach and involvement in higher education by approving the Republic Act 10931, otherwise known as the "Universal Access to Quality Tertiary Education (UAQTE) Act", enacted into law on August 2017. The UAQTE initiative represents a significant legislative effort to ensure widespread access to high-quality tertiary education by providing free tuition and covering other school fees in State Universities and Colleges (SUCs), Local Universities and Colleges (LUCs), and State-Run Technical Vocational Institutions. Specifically, this law established four programs namely Free Higher Education (FHE), Tertiary Education Subsidy (TES), Tulong Dunong Program (TDP), Free Technical and Vocational Training, and Student Loan Program. Its implementation is primarily led by the Unified Student Financial Assistance System for Tertiary Education (UniFAST), an attached agency of the Commission on Higher Education (CHEd) [4].

Although its objectives were clearly defined, its feasibility has been questioned. Concerns have been raised about the design of the law, its implementing rules and regulations, and the adequacy of resources to sustain its programs. Early assessment of the program [5], explained that the absence of clear and prompt guidelines had caused difficulties in service delivery and utilization, particularly in the processing of billing requirements, resulting in delays in reimbursement. Thus, continuous evaluation and improvement are necessary to ensure the program's effectiveness in achieving its goals.

With the changing landscape of information sharing, social media platforms have emerged as vital channels for public discussions and viewpoints on a wide range of social and political matters [6]. As of this writing, the Philippines has a notable digital presence, with 83% of its population being internet users. Filipinos spend approximately four to five hours per day on social media, which is twice the global average of two to three hours [7]. This significant difference highlights the pervasive role of social media in the lives of Filipinos. Consequently, leveraging this wealth of social media data through Natural Language Processing (NLP) techniques, such as sentiment analysis, holds immense potential for enhancing our understanding of public perception and sentiment towards the UAQTE.

Many government-initiated practices tend to adopt a top-down approach to knowledge sharing [8]. This approach often treats local communities as passive recipients rather than active collaborators. This observation highlights the significance of incorporating the sentiments of local communities, in our case, the stakeholders of UAQTE, in the analysis process.

The findings of this study aim to contribute to a deeper understanding of public sentiments surrounding UAQTE and can serve as inputs for decision-makers at CHEd, UniFAST, educational institutions, and other stakeholders involved in implementing and evaluating the UAQTE program. To the best

Philippine Commission on Higher Education (CHEd) Leading the Advancement of Knowledge in Agriculture and Science (LAKAS) Project No. 2021-007, eParticipation 2.1: Harnessing Natural Language Processing (NLP) for Community Participation.

of our knowledge, this study is the first to evaluate UAQTE sentiments using sentiment analysis on code-mixed social media data. Through the application of the state-of-the-art multilingual Bidirectional Encoder Representations from Transformers (mBERT) [9], we also aim to test its efficacy in identifying prevailing neutral, positive, and negative sentiments expressed towards UAQTE.

In this introduction, we have provided an overview of the research focus and rationale for conducting sentiment analysis on social media data related to the UAQTE. Section II discusses related works on education and utilizing code-mixed data for sentiment analysis. Section III covers the methodology, including data collection, preprocessing, and fine-tuning of mBERT. In Section IV, we present the results on the effectiveness of mBERT in code-mixed sentiment analysis and discuss public sentiment distribution towards UAQTE. Section V concludes with key insights and implications, while Section VI outlines the limitations and future research directions to enhance sentiment analysis in code-mixed data.

II. Related Works

Sentiment Analysis involves analyzing a sequence of words to uncover the underlying emotional tone and gain insights into the attitudes, opinions, and emotions conveyed in online mentions [10]. This type of analysis is used in various fields, including business and marketing, politics, health, and social policy, to allow policymakers to formulate informed adjustments to new user-centric rules and regulations [11]. In this section, we will discuss the application of sentiment analysis in the field of education and its findings when this technique is applied to code-mixed data settings.

A. Sentiment Analysis on Education

In the context of education, [12] utilized a range of machine learning techniques, including Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), Random Forest (RF), and Multilayer Perception classifier. By exploring and comparing different SA models, the study successfully determined the effective approaches for analyzing student's classroom feedback that would enhance the quality of teaching in higher education institutions. Moreover, they highlighted the potential of social media platforms like Twitter and Facebook as valuable sources for gathering information and extracting opinions pertaining to students' learning experiences. The study of [13] aimed to understand the drivers of success for higher education institutions (HEIs) in the online realm. They conducted text mining and sentiment analysis on online reviews from various business schools. The findings revealed that HEIs can enhance their online attractiveness by offering financial support for students' cost of living, providing courses in English, and cultivating an international environment. These factors were identified as influential in shaping the perceptions and preferences of international students seeking to study abroad. Similarly, [14] observed through an analysis of social media posts that a predominant expression of negativity exists regarding the K to 12 Program in the Philippines, indicating a pressing need for its implementers to enhance its implementation measures.

B. Sentiment Analysis on Code-Mixed Data

Code-mixing is a linguistic phenomenon observed among multilingual individuals who prefer using their native language over English to express information [15]. Code-mixed text, whether spoken or written, is prevalent in multilingual societies including the Philippines [16]. Social media platforms like Facebook, Twitter, and online forums are common sources of code-mixed content. As stated in [17], sentiment analysis of monolingual text has been extensively studied, but code-mixing introduces additional complexity in analyzing the text's sentiment due to its non-standard writing style.

As presented in the review of [18], the sentiment analysis of code-mixed and switched English with Indian languages achieved a range of 0.39 to 0.77 F1 Score wherein SVM, NB, and RF were the most used Machine Learning classifiers.

In the local setting, [19] utilized SentiWordNet 3.0 and FilCon, English and Filipino lexicons respectively to generate initial sentiment classifications. NB and SVM hybrid models were trained using these sentiment labels. To handle TagLish comments, a Code-Switching Point Detection Module was employed to separate English and non-English words, which were then processed using the appropriate lexicons. The lexicon module achieved an overall accuracy of 55%, while the Naive Bayes before Support Vector Machines hybrid model achieved accuracies of 55% overall, 58% for English, 46% for Filipino, and 57% for a mix of Tagalog and English or TagLish. The Support Vector Machines before Naive Bayes hybrid model achieved accuracies of 61% overall, 70% for English, 54% for Filipino, and 54% for TagLish.

In [20], it was determined that existing bilingual embedding techniques were not suitable for processing code-mixed text. They emphasized the necessity of developing multilingual word embeddings specifically designed for code-mixed text processing. In [21], a study was conducted on code-mixed Persian-English data to perform sentiment analysis. They argued that the uniqueness of analyzing code-mixed data lies in the fact that the presence of English words within the Persian text can significantly impact the emotional expressions conveyed. This poses a challenge for Persian-only sentiment analysis models, as they may struggle to accurately interpret and generate correct outputs due to the mixed language context. They utilized Yandex and dictionary-based translation techniques to translate the code-mixed words present in the text. Additionally, pre-trained BERT embeddings, specifically mBERT was employed to represent the data which achieved an accuracy of 66.17% and an F1 score of 63.66%, surpassing the performance of the baseline models namely Naive Bayes and Random Forest methods. The effectiveness of mBERT in this setting was also assessed in a comparative study conducted by [22] on Hindi-English code-mixed data. The study revealed that mBERT outperformed ALBERT, vanilla BERT, and RoBERTa models, except for HingBERT-based models, which were specifically trained on Hindi-English code-mixed data.

III. Methodology

In this section, we present the methodology employed to collect, preprocess, and annotate the data, perform the fine-tuning of the mBERT model, and evaluate the sentiment analysis results.

A. Data Collection

To capture information preceding the official implementation of UAQTE, data collection was conducted from April 1, 2017, to April 10, 2023. The process involved employing scraper libraries, enabling the systematic and organized retrieval of relevant data.

1) Facebook: We collected a total of 10,443 textual data from the posts that were set as public and official regional groups of CHEd. These groups, which include Ilocos Region (Region 1), MIMAROPA (Region 4-B), Bicol Region (Region 5), Western Visayas (Region 6), Central Visayas (Region 7), Eastern Visayas (Region 8), Zamboanga Peninsula (Region 9), Northern Mindanao (Region 10), Davao Region (Region 11), CARAGA (Region 13), National Capital Region (NCR), Cordillera Administrative Region (CAR), were specifically established to facilitate discussions, address concerns, and provide a platform for beneficiaries and stakeholders of the UAQTE. These regional groups served as valuable sources of opinions and insights related to UAQTE.

2) Twitter: To collect a relevant dataset, we utilized specific keywords and hashtags such as “ched unifast”, “tdp grantee”, “tes grantee”, “#PINASkolar”, and “#TertiaryEducationSubsidy”. We were able to collect 1,112 raw tweets.

3) YouTube: In addition to collecting data from social media platforms like Facebook and Twitter, data was also collected from YouTube comments due to its significant presence as a popular video-sharing platform. To gather relevant content, YouTube videos were manually searched using specific keywords such as “CHED UniFAST”, “UniFAST TES”, “Libreng Edukasyon UniFAST”, and “RA 10931 Free Tuition”. The selection process involved identifying videos with high view counts, ensuring that the chosen content resonated with a wide audience and potentially represented popular sentiments and discussions surrounding the topic of interest. From the 49 videos, we were able to collect 1,777 raw YouTube comments.

B. Data Preprocessing

A total of 13,332 data points were collected for the purpose of data preprocessing. Since posts, tweets, or comments originate from various users and reflect individual opinions, URLs were the key identifier used to remove duplicates. Furthermore, posts that were unrelated to the implementation of UAQTE such as presidential election campaign-related content, underwent manual inspections and were excluded from the dataset. It is important to note that for YouTube data, the collection date was recorded and used to determine the timeframe of the content, whether it was posted months, days, or years ago. As mentioned in [23], BERT has shown optimal performance with little to no preprocessing. Therefore, a few preprocessing techniques were applied, including spelling corrections (e.g., “dhil” to “dahil” (English: because), “cguro” to “siguro” (English: maybe), normalization of special characters, removal of white spaces, punctuations and symbols, and the conversion of emojis to their textual representation.

C. Data Annotation

From the preprocessed data, 4,650 (50%) data points were manually labeled and subjected to validation by experts. This annotated data serves as the benchmark and reference for sentiment classification. Following the approach adopted from [24], the data were classified into three categories: neutral, negative, and positive encoded as numbers 0, 1, and 2 respectively.

Positive sentiment encompasses expressions of satisfaction, happiness, admiration, interest, and gratitude. Neutral sentiment pertains to statements that do not exhibit any discernible sentiment, conveying information or facts. Negative sentiment includes expressions of dissatisfaction, anger, disappointment, sarcasm, mockery, and frustration regarding the UAQTE implementation. Table I shows the sample of annotated and validated data points.

TABLE I. SAMPLES OF ANNOTATED DATA

Content	Label
CHED Chairman Popoy De Vera discusses the guidelines on the new degree programs approved for limited face-to-face classes.	0
Pero ba’t wala pong budget ang mga 2021-2022 applicants new grantees? Saan po napunta ang budget? (English: But why do the 2021-2022 applicants who are new grantees have no budget? Where did the budget go?)	1
Thank you so much CHED- UNIFAST Tertiary Education Subsidy.. And to ACLC College of Bukidnon ..	2

D. Fine-Tuning of Multilingual BERT

The BERT model’s architecture is built upon the Transformer framework [25]. When provided with a sequence of up to 512 tokens as input, BERT generates a representation of the entire sequence, which can comprise one or two segments. The first token of the sequence, denoted as [CLS], holds a special classification embedding, and another special token, [SEP], is used to separate segments. To represent the entire sequence for text classification tasks, BERT leverages the final hidden state of the first token [CLS]. A classifier is added on top of BERT to predict the probability or sentiment of the label. During the training process, BERT is fine-tuned based on the task-specific training dataset. Fig. 1 illustrates the BERT fine-tuning model architecture for sentiment classification, adapted from [26], although in our case, sentiments were classified into three sentiment classes, namely neutral, positive, and negative.

While BERT has been pre-trained on English corpus, a multilingual version of BERT (mBERT) [9] has been trained jointly on Wikipedia on 104 languages including English, Tagalog, Waray, and Cebuano, hence, considering the code-mixed nature of the dataset, we downloaded the BERT multilingual base model (cased) from Hugging Face via ktrain [27], a lightweight wrapper for the deep learning library TensorFlow Keras on Google Colaboratory that provides

NVIDIA Tesla T4 12 GB of RAM, and Intel® Xeon® Processor.

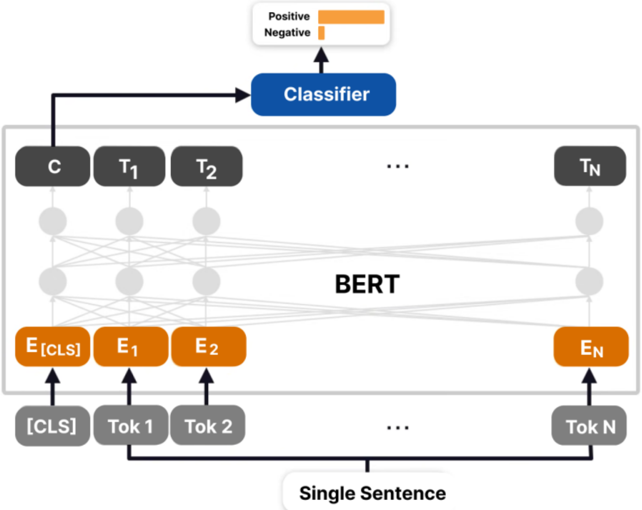

Fig. 1. Fine-tuning BERT for sentiment classification [26].

Among the recommended values of learning rates when fine-tuning on specific tasks were 2e-5 [28], 5e-5, and 3e-5 using batch sizes 16 and 32 at 4 epochs, [29]. Further, a maximum length of 160 was set considering the 95th percentile value obtained from the descriptive statistics of the word lengths, to ensure a manageable memory size and sequence length. To address the class imbalance, class weights were applied to the three classes.

E. Model Evaluation

To evaluate the performance of the sentiment analysis models, two key metrics were utilized: F1 score and accuracy. Subsequently, the best-generated model will be utilized to predict the sentiment classes of the remaining data.

IV. Results and Discussion

The fine-tuned mBERT model exhibited varying levels of performance across different combinations of batch size and learning rate. Table II presents the accuracy and F1 scores achieved by each model configuration.

In contrast to the experiment conducted by [19], which removed stopwords, our approach involves minimal preprocessing of texts and includes stopwords. Additionally, we take into account the textual representation of emojis, setting our study apart from the research conducted by [22]. By considering emojis, we aim to capture a more nuanced understanding of sentiments expressed in social media data, enhancing the contextual analysis capabilities of mBERT. Model 1, with a batch size of 16 and a learning rate of 2e-5, emerged as the best-performing model with 80.21% accuracy and an F1 Score of 81.14%. surpasses the findings of related studies discussed in Section 2 in terms of their F1 Score and Accuracy. Hence, this model was used to classify the remaining data points.

Findings revealed that 2,100 (22.58%) were neutral, 1,647 (17.70%) were negative, and 5,553 (59.70%) were positive, indicating that majority of the social media users view UAQTE as positive or beneficial in the country.

TABLE II. Performance of the Fine-Tuned mBERT Model

Model	Batch Size	Learning Rate	Accuracy	F1 Score
1	16	2e-5	**80.21%**	**81.14%**
2		3e-5	79.42%	80.03%
3		5e-5	79.49%	80.20%
4	32	2e-5	79.78%	80.78%
5		3e-5	76.91%	78.21%
6		5e-5	79.49%	80.65%

A user from Twitter expressed positive feedback towards UAQTE, “Thank you God nakasali ako sa ched-unifast. Malaking tulong na po ito” (English: “Thank you God I was able to be a member of ched-unifast this is helpful”). Neutral contents were composed of announcements and updates regarding subsidies. A Facebook post expressing frustration with CHED reads, “nakakaloka yung CHED beh. patapos na ako, wala pa din yung sa UniFAST hahahaha” (English: “It’s crazy, CHED! I’m almost done with my studies, but I still haven't received anything from UniFAST, hahahaha.”). This post conveys a sense of disappointment and humor, suggesting that the individual has been anticipating support or benefits from UniFAST but has yet to receive them.

As shown in Fig. 2, the sentiments towards the implementation of the UAQTE have shown interesting dynamics over the years.

From 2017 to 2018, the sentiments were characterized by a relatively low number of neutral sentiments, while negative sentiments were more prevalent which can be attributed to the initial challenges in implementing the program as discussed by [5].

In 2019, the distribution of sentiments experienced a significant shift. The number of neutral sentiments increased notably, possibly due to improvements or adjustments made to address previous concerns. Negative sentiments decreased, while positive sentiments remained relatively low, indicating ongoing debates or skepticism surrounding the program.

In year 2020 witnessed a relatively balanced distribution of sentiments across the three categories. This suggests a period of relative stability or reduced controversy, where public opinion was less polarized. However, in 2021, sentiments experienced a sharp rise across all categories. Negative sentiments and positive sentiments increased substantially, while neutral sentiments also reached a significant level. This is likely due to the release of subsidies to beneficiaries, with some expressing gratitude for receiving them while others express disappointment for not receiving them on time. This could be also attributed to the pandemic's impact on the education sector. Inquiries and concerns regarding the bank application of the grantees were also found. The trend continued in 2022 with sentiments reaching even higher levels across all categories. This suggests ongoing discussions, policy developments, or increased public attention toward the program, resulting in a more polarized sentiment landscape. Finally, in 2023, sentiments experienced a notable decline across all categories but experienced a substantial rise in negative sentiments which was likely due to reports of alleged misuse of funds intended for the implementation of the program [30].

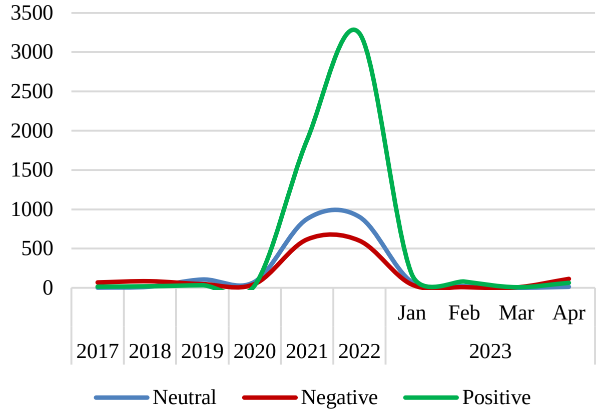

Fig. 2. Trend of social media users' sentiments towards the implementation of UAQTE from 2017 to 2023.

These fluctuations in sentiments over the years highlight the dynamic nature of public opinion and the influence of various factors, such as policy changes, media attention, and public discourse, on the sentiments towards the implementation of the UAQTE.

V. Conclusion

UAQTE has been a notable policy shift in the Philippines. While the program aims to provide free higher education and tertiary education subsidies to eligible Filipino students, its feasibility was questioned and assessed.

This study analyzed the sentiments expressed on social media platforms such as Facebook, Twitter, and YouTube. where code-mixing is prevalent, regarding the implementation of UAQTE policy in the Philippines between April 1, 2017, and April 10, 2023. We used a fine-tuned mBERT model to perform sentiment analysis on the collected data. With minimal preprocessing, mBERT achieved an accuracy and F1 score of 80.21% and 81.14%, respectively. These results outperformed the existing studies, as outlined in Section 2, demonstrating the effectiveness of mBERT in handling code-mixed data for sentiment analysis and revealing a dominant positive sentiment perceived by social media users regarding the implementation of UAQTE.

The prevailing positive sentiments expressed gratitude towards the UAQTE, which suggests that UAQTE has been beneficial for the intended stakeholders. However, the analysis also identified the presence of negative sentiments related to the late distribution of subsidies, alleged misuse of funds, and difficulties in the application of bank accounts such as erroneous online application portal of the beneficiaries. The neutral sentiments mostly involved announcements, news, and updates regarding the release of the subsidies. Addressing these concerns can help enhance the program's effectiveness and ensure that it continues to meet its objectives.

VI. Limitations and Future Work

While this study contributes significant insights, it is essential to acknowledge its limitations. The analysis is limited to social media data and may not fully represent the entire population's sentiments. Moreover, analyzing the sentiment of specific demographic groups, such as students or parents, may provide additional insights that could inform more targeted policies and initiatives. Additionally, exploring the use of cross-lingual language models (XLMs) like XLM-RoBERTa and hybrid BERT models for sentiment analysis may improve the accuracy and reliability of the results with code-mixed data.

Acknowledgment

The researchers would like to thank the Philippine Commission on Higher Education (CHEd) Leading the Advancement of Knowledge in Agriculture and Science (LAKAS) Project No. 2021-007, eParticipation 2.1: Harnessing Natural Language Processing (NLP) for Community Participation, for providing the necessary funds to make this research possible. The researchers are truly grateful for their contribution to the research.

References

[1] M. Krstić, J. A. Filipe, and J. Chavaglia, "Higher education as a determinant of the competitiveness and sustainable development of an economy," Sustainability (Switzerland), vol. 12, no. 16, Aug. 2020, doi: 10.3390/su12166607.

[2] H. Coates, "Productivity in higher education," 2017. [Online]. Available: www.apo-tokyo.org.

[3] Pooja and R. Bhalla, "A review paper on the role of sentiment analysis in quality education," SN Computer Science, vol. 3, no. 6. Springer, Nov. 01, 2022. doi: 10.1007/s42979-022-01366-9.

[4] Official Gazette, "Republic Act No. 10931," Philippines: government of the Philippines, 2017. http://www.officialgazette.gov.ph/2017/08/03/republic-actno-10931/ (accessed Jun. 09, 2023).

[5] M. Kristina et al., "Process evaluation of the universal access to quality tertiary education act (RA 10931): status and prospects for improved implementation." [Online]. Available: https://www.pids.gov.ph.

[6] L. L. Maceda, J. L. Llovido, and T. D. Palaoag, "Corpus analysis of earthquake related tweets through topic modelling," Int J Mach Learn Comput, vol. 7, no. 6, pp. 194–197, Dec. 2017, doi: 10.18178/ijmlc.2017.7.6.645.

[7] J. R. Clapano, "DICT: 83% of Pinoys are internet users, but…," Philstar.com, Jun. 03, 2023. [Online]. Available: https://www.philstar.com/headlines/2023/06/04/2271289/dict-83-pinoys-are-internet-users-but.

[8] J. L. Llovido and T. D. Palaoag, "E-LAHOK: An e-participatory platform for disaster risk reduction and management," in IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, May 2020. doi: 10.1088/1757-899X/803/1/012049.

[9] 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.

[10] K. Bannister, "Sentiment analysis: how does it work? why should we use it?," Brandwatch, Feb. 26, 2018. https://www.brandwatch.com/blog/understanding-sentiment-analysis/.

[11] G. Manias, A. Mavrogiorgou, A. Kiourtis, C. Symvoulidis, and D. Kyriazis, "Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for

classifying twitter data," Neural Comput Appl, 2023, doi: 10.1007/s00521-023-08629-3.

[12] I. Ali Kandhro, M. Ameen Chhajro, K. Kumar, H. N. Lashari, and U. Khan, "Student feedback sentiment analysis model using various machine learning schemes a review," Indian J Sci Technol, vol. 14, no. 12, pp. 1–9, Apr. 2019, doi: 10.17485/ijst/2019/v12i14/143243.

[13] C. L. Santos, P. Rita, and J. Guerreiro, "Improving international attractiveness of higher education institutions based on text mining and sentiment analysis," International Journal of Educational Management, vol. 32, no. 3, pp. 431–447, 2018, doi: 10.1108/IJEM-01-2017-0027.

[14] F. S. Relucio and T. D. Palaoag, "Sentiment analysis on educational posts from social media,". 2018. doi: 10.1145/3183586.3183604.

[15] N. H. Mahadzir, M. F. Omar, M. N. M. Nawi, A. Salameh, and K. C. Hussin, "Sentiment analysis of code-mixed text: a review," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 3, pp. 2469–2478, Apr. 2021, doi: 10.17762/turcomat.v12i3.1239.

[16] V. Srivastava and M. Singh, "IIT Gandhinagar at SEMEVAL-2020 task 9: code-mixed sentiment classification using candidate sentence generation and selection," arXiv (Cornell University), Jun. 2020, doi: 10.48550/arxiv.2006.14465.

[17] M. Herrera, "TweetTaglish: a dataset for investigating Tagalog-English code-switching," ACL Anthology, Jun. 01, 2022. https://aclanthology.org/2022.lrec-1.225.

[18] G. I. Ahmad, J. Singla, A. Ali, A. A. Reshi, and A. A. Salameh, "Machine learning techniques for sentiment analysis of code-mixed and switched indian social media text corpus - a comprehensive review," International Journal of Advanced Computer Science and Applications, vol. 13, no. 2, Jan. 2022, doi: 10.14569/ijacsa.2022.0130254.

[19] V. Curada, K. C. Javier, G. L. Madamba, R. C. A. Montenegro, and C. Ponay, "Lexicon-based sentiment analysis of professor evaluation students' comments with code switching using Naive Bayes algorithm and Support Vector Machines".

[20] A. Pratapa, M. Choudhury, and S. Sitaram, "Word embeddings for code-mixed language processing." [Online]. Available: https://github.com/lmthang/bivec.

[21] N. Sabri, A. Edalat, and B. Bahrak, "Sentiment analysis of Persian-English code-mixed texts," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2102.12700.

[22] A. Patil, V. Patwardhan, A. Phaltankar, G. Takawane, and R. Joshi, "Comparative study of pre-trained BERT models for code-mixed hindi-english data,". 2023. doi: 10.1109/i2ct57861.2023.10126273.

[23] E. Alzahrani and L. Jololian, "How different text-preprocessing techniques using the BERT model affect the gender profiling of authors," Academy and Industry Research Collaboration Center (AIRCC), Sep. 2021, pp. 01–08. doi: 10.5121/csit.2021.111501.

[24] S. M. Mohammad, "A practical guide to sentiment annotation: challenges and solutions.".

[25] A. Vaswani et al., "Attention is all you need," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.03762.

[26] TensorFlow, "Transfer learning and transformer models (ML tech talks)," YouTube. Jul. 22, 2021. [Online]. Available: https://www.youtube.com/watch?v=LE3NfEULV6k.

[27] A. S. Maiya, "Ktrain: a low-code library for augmented machine learning," 2022. [Online]. Available: https://github.com/Tony607/Chinese_sentiment_analysis

[28] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?," May 2019, [Online]. Available: http://arxiv.org/abs/1905.05583.

[29] I. K. L. Turc, M.-W. Chang, and K. Toutanova, "Well-read students learn better: on the importance of pre-training compact models.".

[30] G. Ombay, "CHED says no misuse of P10-billion fund," GMA News Online, Mar. 22, 2023. [Online]. Available: https://www.gmanetwork.com/news/topstories/nation/864734/ched-says-no-misuse-of-p10-billion-fund/story.