Sentiment Analysis of Students Reviews on Campus Curriculum for the Workplace Using Bidirectional Encoder Representations from Transformers (BERT) Algorithm

Henderi[*1], Ilamsyah[2]
[1]Master's Program in Department of Informatics Engineering, Raharja University, Indonesia
[2]Department of Computer Systems, Raharja University, Indonesia
E-mail : [*1]**henderi@raharja.info**, [2]ilamsyah@raharja.info

Abstract

Higher education curricula play a crucial role in preparing graduates for the workforce; however, a gap often exists between theoretical instruction and the practical demands of industry. This study evaluates curriculum effectiveness through sentiment analysis of students reviews using the Bidirectional Encoder Representations from Transformers (BERT) model. A total of 250 students reviews were collected via open-ended questionnaires and processed using Python with libraries such as NLTK, Transformers, and Pandas for text cleaning, normalization, and tokenization. Twenty percent of the data were manually labeled, achieving a Cohen's Kappa coefficient of 0.8405, indicating excellent inter-annotator agreement. The IndoBERT model—implemented using the Torch and Scikit-learn libraries—was trained to classify sentiments into positive, negative, and neutral categories. Results show that 82.7% of reviews were positive, praising the curriculum's relevance to job readiness and technical skills; 12.7% were negative, highlighting insufficient practical content and soft skills development; and 4.5% were neutral. This research demonstrates that BERT-based sentiment analysis using Python is effective for curriculum evaluation, providing data-driven insights to help academic institutions enhance curriculum alignment with industry needs.

Keywords *— Sentiment Analysis, Curriculum, BERT, Python*

1. INTRODUCTION

Higher education curricula play a strategic role in preparing graduates to meet the challenges of the workforce. However, many educational institutions struggle to design curricula that are truly aligned with the ever-evolving needs of industry. A key challenge lies in the gap between theoretical knowledge taught in universities and the practical skills required in professional settings. Numerous graduates report that the material they learned is not fully applicable in real-world work environments, often necessitating additional time to adapt or even requiring supplementary training after graduation. Furthermore, rapid technological advancements compel higher education institutions to continuously update their curricula to maintain relevance. Therefore, it is essential for academic institutions to evaluate the effectiveness of their curricula by understanding how students perceive the connection between the education they received and the demands of the workplace.

According to research conducted by Suryono [1], the percentage of job–education mismatch among workers varies across provinces in Java. Yogyakarta exhibits the lowest mismatch rate at 6.2%. In contrast, Central Java and East Java have not shown improved employment opportunities for higher education graduates, consistently experiencing rising levels of job–education mismatch. Data from Statistics Indonesia (Badan Pusat Statistik, BPS) through the National Labor Force Survey (Survei Angkatan Kerja Nasional, Sakernas) indicate that the phenomenon of horizontal mismatch—where workers are employed in fields unrelated to their educational background—is notably significant in Indonesia. For instance, a Sakernas-based study from February 2022 [2] reported that approximately 33.5% of higher education graduates in Indonesia experience horizontal mismatch. This figure underscores a major challenge in aligning graduate competencies with labor market demands.

One effective approach to evaluating curriculum effectiveness is sentiment analysis of students reviews. Such reviews offer direct insights into the relevance of academic content to industry needs and highlight areas requiring improvement. In this study, sentiment analysis is performed using the Bidirectional Encoder Representations from Transformers (BERT) model, an artificial intelligence–based method that captures linguistic context more accurately than traditional natural language processing techniques. Furthermore, this sentiment analysis application is developed using the Python programming language, leveraging several libraries that enable flexible and adaptive management of tasks in response to evolving requirements during development.

Numerous studies have investigated the application of BERT in aspect-based sentiment analysis (ABSA). Hoang et al. [3] and Ansar et al. [4] highlight BERT's effectiveness for this task, with Ansar introducing an improved aspect extraction technique to boost both efficiency and accuracy. Sun et al. [5] further refine BERT's capabilities by incorporating deep contextual features and by constructing auxiliary sentences, respectively. Karimi [6] proposes specialized modules along with a gating mechanism to enhance BERT's performance in both aspect extraction and sentiment classification. Meanwhile, Lakshmidevi [7] integrates BERT for aspect extraction and pairs it with various machine learning classifiers for sentiment analysis, attaining high accuracy on standard benchmark datasets.

Several key issues underpin this research, including the relevance of higher education curricula to the evolving demands of industry, the application of artificial intelligence (AI) technologies in sentiment analysis of Indonesian-language text, and the critical role of students feedback as an indicator of curriculum effectiveness. The mismatch between education and employment also carries significant economic consequences, such as reduced workforce productivity and rising levels of disguised unemployment [8]. Therefore, this study is not only academically relevant but also holds practical implications for policymakers in education and labor sectors.

To support this research, several prior studies have been conducted in related areas. For instance, implemented a BERT model to analyze sentiment in user reviews of the PeduliLindungi application, demonstrating the model's effectiveness in classifying sentiment in Indonesian text. Additionally, evaluated graduate readiness for workforce demands and emphasized that higher education curricula are a crucial element in producing job-ready,

competitive graduates[9]. Moreover, it found that sentiment analysis of Coursera app reviews on the Google Play Store across three countries revealed an overall positive trend, with the United States showing the highest proportion of positive comments[10]. The study demonstrated that an ensemble method combining Naive Bayes, SVM, and KNN algorithms effectively improved classification accuracy, yielding more comprehensive insights into user perceptions. Another study concluded that a study program's success can be measured by the extent to which its graduates benefit society; however, tracer study results indicated that curriculum alignment with workforce needs remains suboptimal relative to graduate profiles[11]. Another research applied sentiment analysis to understand public perception of the Merdeka Belajar (Freedom to Learn) curriculum implementation[12]. The research found that clustering public opinions on Twitter regarding the Merdeka Belajar–Kampus Merdeka (MBKM) program yielded four distinct groups, with negative sentiment predominating—highlighting specific aspects of the MBKM program requiring improvement to enhance participant satisfaction[13].

Recent studies on detecting fake reviews using BERT architecture have yielded encouraging outcomes. Several investigations have confirmed that BERT-based models are highly effective at accurately spotting deceptive reviews in diverse sectors such as hospitality, dining, and healthcare. Deepa et al. [14] offered an in-depth overview of BERT's structure and its performance in sentiment analysis, emphasizing its strength in capturing contextual nuances. Nevertheless, Shih [15] and Deng et al. [16] pointed out certain limitations of BERT, especially its difficulty with implicitly evaluative language and the need for enhanced model supervision. To address these issues, Verma et al. [17] introduced hybrid approaches integrating BERT—He enhanced accuracy in sentiment analysis of movie reviews, while Verma combined BERT with XGBoost for analyzing reviews generated by ChatGPT.These prior studies provide a strong foundation for exploring the use of the BERT model in analyzing students sentiment regarding the effectiveness of higher education curricula in preparing graduates for the workforce. By leveraging an AI-driven approach and developing a flexible application using the Python programming language along with libraries such as pandas, torch, transformers, scikit-learn, NLTK, pyspellchecker, and TensorFlow, this research aims to deliver more accurate, data-driven recommendations to academic institutions for designing curricula that better align with industry requirements.

2. RESEARCH METHOD

This study employs a sentiment analysis approach to evaluate students perceptions regarding the impact of their university curriculum on their preparedness for the workforce. The research process involves several key stages: collecting students review data, preprocessing textual data, training a Bidirectional Encoder Representations from Transformers (BERT) model, and analyzing the results to classify sentiments into positive, negative, and neutral categories. Furthermore, the sentiment analysis implementation is carried out using the Python programming language to ensure flexibility and efficiency in task management. The steps of the research methodology are outlined as follows:



Figure 1. Research Procedure

The following are the methodological steps employed in this study:

1) Data Collection
The initial stage of sentiment analysis involves gathering textual data from various sources such as social media, user reviews, surveys, or discussion forums. This data consists of raw text containing individuals' opinions or sentiments regarding a specific topic—in this case, students perceptions of their university curriculum's relevance to workplace readiness.

2) Data Preprocessing
Collected data is typically unstructured and requires cleaning before modeling. This step includes removing special characters, eliminating stopwords, applying stemming, performing tokenization, and normalizing the text to enhance its suitability for subsequent computational analysis.

3) Modeling
This phase employs machine learning algorithms or artificial intelligence models—specifically, the Bidirectional Encoder Representations from Transformers (BERT)—to analyze sentiment in the preprocessed text. The model is trained and evaluated on labeled data to classify sentiments into three categories: positive, negative, or neutral.

4) Results Analysis
The final stage involves interpreting and evaluating the model's output to extract meaningful insights. Classified sentiment data is visualized through charts, graphs, or summary reports to support data-driven decision-making, particularly in refining academic curricula based on students feedback.

2.1. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a language representation model introduced by Devlin et al. in 2018 [18]. It is designed to understand the context of words in a sentence bidirectionally—simultaneously considering contextual information from both the left and right sides of a word. This approach differs from earlier models, which typically processed text sequentially in one direction (either left-to-right or right-to-left). BERT is pre-trained using two primary tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP).

In MLM, a portion of the input tokens is randomly masked, and the model is tasked with predicting the original masked words based on their surrounding context. In NSP, the model learns to determine whether two given sentences appear consecutively in the original text (i.e., whether they form a coherent pair). The combination of these two pre-training objectives enables BERT to capture deep, contextualized representations of language [19],

significantly enhancing its performance across a wide range of natural language processing (NLP) tasks—including text classification, sentiment analysis, and question answering. In the context of data annotation for training or evaluating models like BERT, it is essential to assess inter-annotator consistency to ensure data quality and reliability. A widely used metric for this purpose is Cohen's Kappa coefficient. Cohen's Kappa measures the level of agreement between two annotators while accounting for the possibility of agreement occurring by chance. The coefficient ranges from –1 to 1. 1 indicates perfect agreement, 0 implies agreement equivalent to random chance.

Negative values suggest less agreement than expected by chance. Common interpretative guidelines for Cohen's Kappa are as follows: < 0: Poor agreement, 0.01–0.20: Slight agreement, 0.21–0.40: Fair agreement, 0.41–0.60: Moderate agreement, 0.61–0.80: Substantial agreement, 0.81–1.00: Almost perfect agreement. Applying Cohen's Kappa in data annotation evaluation helps ensure that the labeled dataset used to fine-tune models like BERT is both high-quality and consistent, thereby improving model performance in downstream NLP tasks. Cohen's Kappa is calculated using the following formula:

$$k = \frac{Po - Pe}{1 - Pe}$$

Where *Po* (Observed Agreement) is the proportion of actual agreement between two annotators. It is calculated by dividing the number of observations where both annotators assigned the same label by the total number of observations. This is determined using the formula:

$$Po = \frac{Sum\ of\ the\ main\ diagonal}{Total\ Observations}$$

Pe (Expected Agreement) is the probability that the two annotators would assign the same label by chance alone. It is computed by summing, across all categories, the product of the marginal row proportion and the marginal column proportion for each category. The formula is:

$$Pe = \Sigma \left(\frac{Total\ Row}{Total\ Observations}\ x\ \frac{Total\ Column}{Total\ Observatins} \right)$$

Once the values of *Po* (Observed Agreement) and *Pe* (Expected Agreement) are obtained, the Cohen's Kappa coefficient (κ) can be calculated. The resulting value is interpreted according to the following table:

Table 1. Interpretation of Cohen's Kappa

Nilai κ	Tingkat Kesepakatan
< 0.00	No agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

The primary strength of BERT lies in its ability to deeply understand word context through its bidirectional approach. Unlike earlier unidirectional models that process text sequentially in only one direction, BERT simultaneously considers both left and right contextual information around each word[15]. This enables BERT to achieve superior performance across a wide range of natural language processing (NLP) tasks. Furthermore, BERT's flexible architecture allows for straightforward adaptation to diverse NLP applications, such as sentiment analysis, text classification, and question answering through fine-tuning with relatively small task-specific datasets.

3. RESEARCH RESULTS AND DISCUSSION

The data used in this study consist of textual reviews from students regarding the relevance of their academic curriculum to the workplace. The data were collected through open-ended questionnaires distributed to students from various study programs via platforms such as Google Forms. The questions covered their perceptions of the curriculum, the skills they have acquired, and how well their education aligns with current job market demands. A minimum of 150 reviews was targeted to ensure adequate representativeness. The inclusion criteria specified that responses must come from active students, ensuring that the feedback reflects the most up-to-date implementation of the curriculum.

Table 2. Questionnaire Results Dataset

created_date	text
01/01/2025	The curriculum really helped me prepare for work in the tech industry.
02/01/2025	It's just theory, Tho the practical aspects are really lacking.
03/01/2025	It's not bad, but it could be more relevant to industry needs.
04/01/2025	The course material is okay, but it's not up-to-date with new technologies.
05/01/2025	The good curriculum has given me more confidence in my work.
……..	…….
25/04/2025	I use many of the skills I learned from college in my work..

The collected text data is then processed to ensure its quality before being analyzed using the BERT model. The pre-processing steps include:

3.1. Text Cleaning

This stage aims to remove special characters (e.g., excessive punctuation, emoticons), URLs, and irrelevant words such as stopwords. The cleaning process is implemented in Python using libraries including Natural Language Toolkit (NLTK) and Transformers. Stopwords—common words like *"dan"* (and), *"atau"* (or), and *"di"* (in)—typically carry little semantic value in text analysis and are therefore filtered out. To perform this step, the NLTK stopwords corpus is first downloaded via pip and the nltk.download('stopwords') command. The Indonesian stopwords list is then loaded from nltk.corpus.stopwords. Customization is also supported, allowing researchers to add or remove specific terms based on the context of the study. Each word in the input text is compared against this stopwords list, and only non-stopwords are retained for further processing. Upon successful execution of the text cleaning pipeline, the output will resemble the following format:

Table 3. Text Cleaning Process Results

created_date	text	email	text_cleaned
2025-01-01	Good curriculum, Tho lacking practical experience	student1@gmail.com	Good curriculum, less practice
2025-01-02	Just theory, no practice https://example.com	student2@yahoo.com	Just theory, no practice
…….	……	……..	……..

3.2. Normalization

After obtaining cleaned text, the next step is normalization, which involves converting all characters to lowercase, expanding common Indonesian abbreviations (e.g., *"tdk"* → *"tidak"*), and correcting simple spelling errors. A predefined dictionary is used to handle frequent informal abbreviations, as shown in the following Python code snippet:

```
abbreviation_dict = {
   "nope": "no",
   "pass": "how",
   "coz": "because",
   "so": "very",
   "dont": "don’t",
   “nah”:”no”,
   “No way”:”not”,
   “Tho”:”but”,
   “yet”:”no”
}
```

Each word in the cleaned text is checked against this dictionary, and if a match is found, it is replaced with its standardized form. For automatic spelling correction of typographical errors (e.g., *"teory"* → *"teori"*), the pyspellchecker library is employed. This library uses statistical language models to suggest the most probable correct spelling based on word frequency and context. If the normalization process is successfully executed, it produces output such as the following:

Table 4. Results of the Normalization Process

created_date	text	email	text_cleaned	text_normalized
2025-01-01	Good curriculum, tho lacking practical experience	student1@gmail.com	Good curriculum less practice	Good curriculum less practice
2025-01-02	Just theory, yet practice https://example.com	student2@yahoo.com	Just theory yet practice	Just theory no practice

3.3. Tokenization

Tokenization involves splitting the normalized text into linguistic units (tokens)—such as words, subwords, or special symbols—that can be effectively processed by the BERT model. Unlike basic word-based tokenization (e.g., using NLTK), this research employs the IndoBERT tokenizer from the IndoBenchmark library, which is specifically fine-tuned for the

Indonesian language and utilizes WordPiece subword tokenization. This approach is essential for handling out-of-vocabulary words and morphologically rich or informal Indonesian expressions commonly found in alumni reviews. The tokenizer processes the text_normalized column generated during the normalization stage. It converts each input sentence into a sequence of token IDs, along with special tokens such as [CLS] (classification token) and [SEP] (separator token), and applies padding or truncation to ensure uniform input lengths required by the BERT architecture. If the tokenization process is successfully executed, it produces output similar to the following:

Table 5. Results of the Tokenization Process

created_date	text_normalized	email	tokens
2025-01-01	Good curriculum, tho lacking practical experience	student1@gmail.com	[‘Good’, ‘curriculum’, ‘less’, ‘practice’]
2025-01-02	Just theory, yet practice https://example.com	student2@yahoo.com	[‘Just’, ‘theory’, ‘no’, ‘practice’]

3.4. Manual Labeling (Annotation)

A small subset of the dataset—approximately 20% (capped at 50 samples to ensure manageability)—was manually labeled by two independent annotators to serve as the initial training data. Each review was assigned one of three sentiment labels: positive, negative, or neutral, based on the expressed opinion regarding curriculum relevance to the workplace. To ensure annotation reliability, inter-annotator agreement was measured using Cohen’s Kappa coefficient, with a target value close to 0.8, indicating substantial to almost perfect agreement. This step is critical for producing high-quality labeled data that accurately reflects alumni sentiment. The labeling process was conducted in two stages: Data preparation and manual annotation and Computation of Cohen’s Kappa. The following Python script snippet was used to select the sample for labeling:

```
sample_size = min(50, int(0.2 * len(df)))
df_sample = df.sample(n=sample_size, random_state=42)
print(f" Number of reviews selected for labeling: {sample_size}")
```

This generated a CSV file named dataset_for_labeling.csv, containing the normalized reviews to be annotated. Each annotator independently filled in their sentiment labels in separate columns (e.g., label_annotator1, label_annotator2).

Table 6. Example of Manual Label Filling

text_normalized	label_anotator1	label_anotator2
Good curriculum, tho lacking practical experience	netral	negative
Just theory, yet practice https://example.com	negative	negative
……	……	……

Next, the level of consistency between the two annotators who labeled the text dataset was measured using Cohen’s Kappa. The coefficient is calculated using the following formula:

$$k = \frac{Po - Pe}{1 - Pe}$$

The calculation begins by constructing a contingency table (also known as a confusion matrix) based on the labels provided by label_annotator1 and label_annotator2. Since the sentiment categories are positive, negative, and neutral, the resulting table is a 3×3 matrix.

Table 7. Contingency Table

Label Anotator	Positif	Negatif	Netral	Total Baris
Positive	22	0	0	22
Negative	0	18	1	19
Neutral	0	4	5	9
Total column	22	22	6	50

Next, determine the proportion of actual agreement (Po) by dividing the number of main diagonals (agreements) by the total observations.

$$Po = \frac{Number\ of\ main\ diagonals}{Total\ Observations}$$

From the contingency table:
Number of main diagonals = 22 + 18 + 5 = 45
Total observations = 50
Then :

$$Po = \frac{45}{50} = 0.9$$

The random agreement proportion (Pe) is calculated by adding the products of the marginal proportions of rows and columns for each category with the equation:

$$Pe = \Sigma \left(\frac{Total\ Rows}{Total\ Observations}\ x\ \frac{Total\ column}{Total\ Observations} \right)$$

Calculation for Each Category:
Positive
Row Proportion $= \frac{22}{50}$
Column Proportion $= \frac{22}{50}$
Hasil perkalian $= \frac{22}{50}\ x\ \frac{22}{50} = 0{,}2116$

Using the same method as above, the proportion value for the negative category was obtained = 0.1672 and the proportion for the positive category = 0.0216. So the Pe value was obtained.

Pe = 0,2116 + 0,1672 + 0,0216 = 0,4004
Cohen's Kappa coefficient is calculated based on the Po and Pe values.

$$k = \frac{0{,}9 - 0{,}4004}{1 - 0{,}4004} = 0{,}833$$

Based on Cohen's Kappa interpretation scale:

- $0.8 \leq \kappa \leq 1.0$: Very good agreement.
- $0.6 \leq \kappa < 0.8$: Fair agreement.
- $\kappa < 0.6$: Low agreement..

The calculation results show that:

$\kappa \approx 0.8405$

This indicates that the inter-annotator agreement is very good.

To facilitate the calculation of Cohen's Kappa coefficient, the pandas and sklearn.metrics libraries in Python can be used. The process begins by loading the manually labeled dataset from a CSV file using pandas. The script then extracts the labels provided by the two annotators into two separate lists: labels1 for Annotator 1 and labels2 for Annotator 2. These lists serve as input to the cohen_kappa_score function from sklearn.metrics, which computes the Kappa score—accounting for agreement occurring by chance. The resulting score is printed with three decimal places to ensure precision. Following the Kappa calculation, the script evaluates the level of annotation consistency based on the obtained value.

```
import pandas as pd
from sklearn.metrics import cohen_kappa_score
df_labeled = pd.read_csv('dataset_labeled.csv')
print("Data after labeled:")
print(df_labeled[['text_normalized','label_anotator1', 'label_anotator2']].head())
labels1 = df_labeled['label_anotator1'].tolist()
labels2 = df_labeled['label_anotator2'].tolist()
kappa = cohen_kappa_score(labels1, labels2)
print(f"\nCohen's Kappa Score: {kappa:.3f}")
……..
```

Once high inter-annotator consistency (e.g., $\kappa = 0.8405$) is confirmed, the next phase proceeds to BERT-based modeling. The modeling stage uses the 50 manually labeled reviews as the initial training set to fine-tune the IndoBERT model—a BERT variant pre-trained specifically for the Indonesian language. After fine-tuning, the model is applied to classify sentiment across the remaining unlabeled reviews in the dataset. Before training, it is essential to ensure that the following Python libraries are installed, transformers, torch, pandas, scikit-learn. Model training benefits significantly from GPU acceleration (e.g., using Google Colab's free GPU runtime) to reduce computation time. While CPU execution is possible, it will be considerably slower. Upon successful training, the fine-tuned model is saved as a sentiment model directory containing the following files: config.json – model architecture configuration, model.safetensors (or pytorch_model.bin) – trained model weights, tokenizer_config.json and vocab.txt – tokenizer-related files. These artifacts enable future inference, deployment, or further evaluation of alumni sentiment at scale, providing data-driven insights for curriculum improvement.

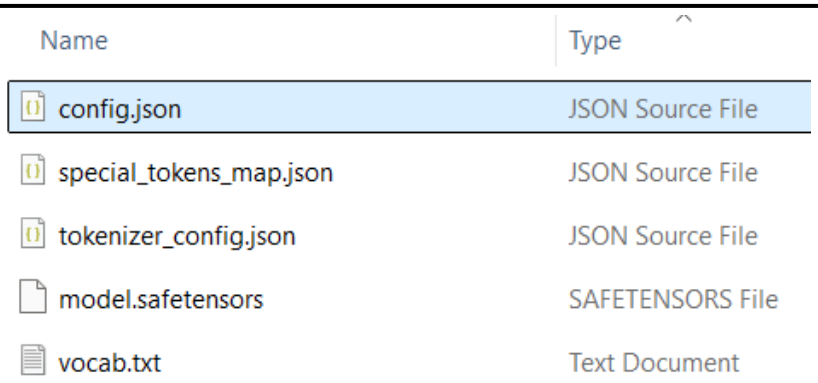


Figure 2. Files Generated by the Modeling Process

The JSON file stores the model's architectural parameters—such as the number of hidden layers (num_hidden_layers), embedding size (hidden_size), number of classification labels (num_labels), and other configuration details—enabling the model to be reloaded later with identical settings without needing to recall specific training configurations. The .safetensors file contains all trained model parameters, including embedding weights, transformer layer weights, and the final classification layer. The .safetensors format is specifically designed to be secure—it prevents the execution of malicious code that could be embedded in traditional binary weight files (e.g., .bin). Additionally, it offers faster loading times and smaller file sizes compared to older formats.

The .txt file (often named vocab.txt) typically includes the tokenizer's vocabulary, a list of all subword units or tokens recognized by the IndoBERT tokenizer. In some cases, additional .txt files may also contain training logs or user notes, but the core vocabulary file is essential for consistent tokenization during inference. After fine-tuning, the trained IndoBERT model was used to classify the remaining 200 unlabeled reviews. The predictions were saved in a new file named dataset_predicted.csv. The following Python script performs this inference step:

```
df_unlabeled = pd.read_csv('dataset_tokenized.csv')
texts_unlabeled = df_unlabeled['text_normalized'].tolist()
encodings = tokenizer(texts_unlabeled, padding=True, truncation=True, max_length=128, return_tensors='pt')
with torch.no_grad():
outputs = model(**encodings)
predictions = outputs.logits.argmax(-1).tolist()
predicted_labels = label_encoder.inverse_transform(predictions)
df_unlabeled['predicted_label'] = predicted_labels
df_unlabeled.to_csv('dataset_predicted.csv', index=False)
```

Below is a snippet of the prediction results based on the normalized and tokenized dataset:

Table 8. Prediction Results with BERT

No	text_cleaned	text_normalized	tokens	predicted_label
1	The curriculum helps work in the technology industry	The curriculum helps work in the technology industry	['curriculum', 'helps', 'work', 'industry', 'technology']	positive
2	Only theory, too little practice	Only theory too little practice	['theory', 'only', 'too', 'little', 'practice']	negative
3	Quite relevant to industry needs	Quite relevant to industry needs	['quite', 'relevant', 'industry', 'needs']	positive
4	Course materials are good and updated with technology	Course materials are good and updated with technology	['materials', 'good', 'updated', 'technology']	positive
5	I believe the curriculum works well	I believe the curriculum works well	['I', 'believe', 'curriculum', 'good']	positive
6	Not practical, makes it hard to adapt to work	Not practical makes it hard to adapt to work	['not', 'practical', 'hard', 'adapt', 'work']	negative
7	The curriculum is just standard	The curriculum is just standard	['curriculum', 'just', 'standard']	negative
8	Lots of useful knowledge from campus, thank you lecturers	Lots of useful knowledge from campus, thank you lecturers	['lots', 'useful', 'knowledge', 'campus', 'thank', 'lecturers']	positive
9	Why study outdated theory, the work world needs skills	Why study outdated theory the work world needs skills	['why', 'study', 'outdated', 'theory', 'work', 'needs', 'skills']	negative
10	Quite helpful for internships, influential	Quite helpful for internships influential	['quite', 'helpful', 'internship', 'influential']	positive
11	Top curriculum but confused when entering work	Top curriculum but confused when entering work	['curriculum', 'top', 'directly', 'work', 'confused']	negative
12	Practice and theory are too heavy, make me dizzy	Practice and theory are too heavy make me dizzy	['practice', 'theory', 'heavy', 'make', 'dizzy']	negative
13	The materials are relevant, lecturers understand the work world	The materials are relevant lecturers understand the work world	['materials', 'relevant', 'lecturers', 'understand', 'work', 'world']	positive

The following graph shows the distribution of sentiment in student reviews of the curriculum. The graph shows the distribution of sentiment in three categories: positive, negative, and neutral.

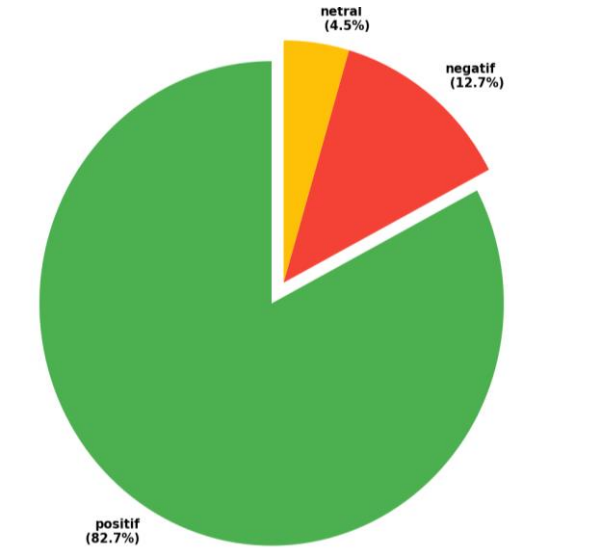


Figure 4. Distribution of Student Review Sentiments

With 82.7% positive reviews — the largest segment occupying the majority of the pie chart — it is evident that the overwhelming majority of reviews are favorable. The second-largest segment, at 12.7%, represents a smaller but notable portion of critical feedback. The smallest segment, only 4.5%, reflects a minimal number of neutral or indifferent reviews. Thus, the majority of students provide positive feedback regarding the curriculum, indicating that it is generally perceived as useful, relevant, and supportive of workforce readiness. Positive reviews often include statements such as "The curriculum helped me succeed in my job" or "Campus materials are awesome." These sentiments suggest that the curriculum demonstrates strength in key areas such as relevance of content, teaching quality, or career preparedness.

Approximately 12.7% of reviews express negative sentiment, signaling dissatisfaction with certain aspects of the curriculum. Negative feedback frequently highlights shortcomings such as *"Too much theory, not enough practice"* or *"Lack of industry-based projects."* This indicates that there are specific areas of the curriculum requiring improvement to better meet student expectations. Although this percentage is relatively small compared to positive sentiment, 12.7% remains statistically significant. The university should evaluate commonly criticized aspects such as insufficient hands on training or lack of soft skills development to enhance curriculum quality and responsiveness to real world demands. Only 4.5% of reviews are neutral, suggesting that a small minority of students hold neither strongly positive nor negative opinions about the curriculum. Neutral reviews tend to be balanced, for example *"It's decent, but could be better"* or *"Materials are okay, but lack adaptation to current technology."* This indicates that while the curriculum is not viewed as poor, it also does not fully meet expectations. These neutral sentiments represent an opportunity for targeted improvement. Students with neutral views may perceive the curriculum as "sufficient," but they sense room for enhancement — such as incorporating more collaborative projects, modern technologies, or industry-aligned case studies.

4. CONCLUSION

This study successfully analyzed student sentiment toward the campus curriculum using the Bidirectional Encoder Representations from Transformers (BERT) model, with a focus on the curriculum's relevance in preparing graduates for the workforce. Based on the analysis of 150 student reviews, the sentiment distribution revealed that 82.7% of reviews were positive, 12.7% were negative, and 4.5% were neutral. The majority of positive reviews (approximately 124 out of 150) indicated that students perceive the curriculum as relevant, beneficial, and supportive of career readiness, particularly in terms of technical skills and lecturer instruction quality. However, the 12.7% negative reviews (about 19 reviews) highlighted key shortcomings, such as insufficient practical content, lack of industry-based projects, and inadequate soft skills training, which pose challenges in meeting student expectations. Meanwhile, the 4.5% neutral reviews (around 7 reviews) suggested that a small group of students view the curriculum as adequate but improvable, citing opportunities such as integration of emerging technologies or more collaborative learning experiences.

Manual labeling of an initial set of 50 reviews yielded a Cohen's Kappa coefficient of 0.8405, indicating almost perfect inter-annotator agreement and confirming the high reliability of the training data used to fine-tune the IndoBERT model. The trained model accurately classified sentiment across the remaining 100 unlabeled reviews (note: total dataset = 150; 50 labeled + 100 unlabeled), producing a sentiment distribution consistent with student perceptions. The implementation using Python and its ecosystem of libraries—including transformers, torch, pandas, scikit-learn, and NLTK—ensured flexibility, efficiency, and reproducibility throughout the pipeline. Overall, this research demonstrates that AI-driven sentiment analysis, particularly using BERT-based models, is an effective and scalable approach for evaluating higher education curricula. It provides data-driven insights that can inform evidence-based academic policy improvements and enhance alignment between educational outcomes and industry demands.

5. SUGGESTED

Based on the findings of this study, universities should consider enhancing the practical dimension of their curricula by incorporating more hands-on training, industry-based projects, and structured internship opportunities to directly address the concerns raised in negative reviews regarding the lack of real-world application. Additionally, soft skills—such as communication, teamwork, critical thinking, and adaptability to emerging technologies—should be systematically integrated into course design, as these competencies are increasingly essential in today's dynamic workplace. While the overwhelming positivity in student sentiment (82.7%) reflects general satisfaction, it may also indicate potential bias in the dataset; therefore, future research should strive for more balanced data collection or employ techniques like oversampling for underrepresented sentiment categories to improve model accuracy and fairness. To ensure the reliability and generalizability of the sentiment analysis model, cross-validation or testing on external datasets from different institutions or platforms is recommended. Furthermore, if review data includes temporal information such as submission dates, longitudinal sentiment analysis could be conducted to monitor shifts in student perceptions over time, enabling institutions to proactively refine their curricula in response to evolving educational and industry needs.

6. REFERENCES

[1] Panji Suryono dan Agus Joko Pitoyo. 2013. Kesesuaian Tingkat Pendidikan Dan Jenis Pekerjaan Pekerja Di Pulau Jawa: Analisis Data Sakernas Tahun 2010. Jurnal Bumi Indonesia

[2] Agil Priyovi Yonanda, Hardius Usman. 2023. Determinan Status Horizontal Mismatch pada Pekerja Lulusan Pendidikan Tinggi di Indonesia. Jurnal Ketenagakerjaan 18(2):142-157

[3] Hoang M, Bihorac OA, Rouces J (2019) Aspect-based sentiment analysis using bert. In Proceedings of the 22nd nordic conference on computational linguistics (pp. 187–196)

[4] Ansar W, Goswami S, Chakrabarti A, Chakraborty B (2021) An efficient methodology for aspect-based sentiment analysis using BERT through refined aspect extraction. J Intell Fuzzy Syst 40(5):9627–9644

[5] Sun Z, Yu H, Song X, Liu R, Yang Y, Zhou D (2020) MobileBERT: A Compact Task-Agnostic BERT for Resource-Limited Devices (arXiv:2004.02984). arXiv. http://arxiv.org/abs/2004.02984

[6] Karimi A, Rossi L, Prati A (2020) Improving bert performance for aspect-based sentiment analysis. arXiv preprint arXiv:2010.11731

[7] Lakshmidevi, Sangram Keshari Swain, M. Vamsikrishna. 2023. A Hybrid Enhancing Aspect-Based Sentiment Analysis with BERT for Aspect Extraction and Diverse ML Classifiers. Conference: 2023 International Conference on Network, Multimedia and Information Technology (NMITCON)

[8] Rahmawati, Siska (2023) Implementasi Algoritma Bert Untuk Analisis Sentimen Ulasan Pengguna Aplikasi Pedulilindungi. Skripsi Thesis, Universitas Teknologi Digital Indonesia.

[9] Muhammad Hanri, Nia Kurnia Sholihah. 2024. Potret Ketidaksesuaian Pendidikan dan Pekerjaan di Indonesia. LPEM FEB UI. Volume 5, Nomor 10, Oktober 2024

[10] Kansha Dianita Pramesti, Nur Indah Meisya, Rizki Amrillah. 2023. Relevansi Lulusan Perguruan Tinggi dengan Dunia Kerja. jurnal pendidikan islam dan sosial agama. Vol. 03 No. 04(Juli2024)

[11] Januar Putra Hidayat, Ida Nurhaida. 2025. Analisis Sentimen pada Ulasan LMS Pembelajaran Menggunakan Metode Natural Language Processing. Vol 10, No 1 (2025)

[12] Fita Fathurokmah, Masran Muin, dan Iin Ulfatul Hasanah. 2023. Persepsi Mahasiswa Program Studi Komunikasi dan Penyiaran Islam dengan Stakeholder terhadap Kesesuaian Kurikulum dalam Dunia Kerja. Dakwah: Jurnal Kajian Dakwah dan Kemasyarakatan, 27 (1), 2023, 88-98

[13] Muhammad Adam Pryono, Satrio Hadi Wijoyo, Fitra Abdurrachman Bachtiar. Analisis Sentimen Terhadap Program Merdeka Belajar Kampus Merdeka Pada Sosial Media Twitter Menggunakan K-Means Clustering, Support Vector Machine (SVM) dan Syntethic Minority Oversampling Technique (SMOTE). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. Vol 8 No 9 (2024): September 2024

[14] Deepa MD (2021) Bidirectional encoder representations from Transformers (BERT) Language model for sentiment analysis task. Turkish J Comput Math Educ (TURCOMAT) 12(7):1708–1721

[15] Shih CF, Tseng YH, Yang CW, Chen PE, Chou HY, Tan LH, Hsieh SK (2021), Oktober Apa yang membingungkan BERT? Evaluasi Linguistik atas Analisis Sentimen terhadap Opini Pelanggan Telekomunikasi. Dalam *Prosiding Konferensi ke-33 tentang Linguistik Komputasional dan Pemrosesan Ucapan (ROCLING 2021)* (hlm. 271–279)

[16] Deng L, Yin T, Li Z, Ge Q (2023) Sentiment analysis of comment data based on BERT-ETextCNN-ELSTM. Electronics 12(13):2910

[17] Verma N, Elbayad M (2024) Merging text transformer models from different initializations. arXiv preprint arXiv:2403.00986

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

[19] Muammar Khadapi, Victor Maruli Pakpahan. 2024. Analisis Sentimen Berbasis Jaringan LSTM dan BERT terhadap Diskusi Twitter tentang Pemilu 2024. Volume 6 Nomor 2 November 2024