Journal of Informatics Engineering (JUTIF) Vol. 6, No. 6, December 2025, Page. 5333-5348
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4465

Labeling Optimization and Hybrid CNN Model in Sentiment Analysis of Movie Reviews with Slang Handling

Alfin Nur Aziz Saputra*[1], Rujianto Eko Saputro[2], Dhanar Intan Surya Saputra[3]

[1,2]Department of Computer Science Universitas Amikom Purwokerto, Indonesia
[3]Department of Informatics Universitas Amikom Purwokerto, Indonesia

Email: [1]lordsaputra@gmail.com

Received : Mar 14, 2025; Revised : Jun 5, 2025; Accepted : Jun 5, 2025; Published : Dec 22, 2025

Abstract

This research focuses on the development of a hybrid Convolutional Neural Network (CNN) model for sentiment analysis of movie comments, specifically designed to overcome the challenges of handling nonstandard language and slang. Slang is often an obstacle in sentiment analysis due to its non-standard nature and is difficult to recognize by traditional algorithms. By utilizing an kamusalay as a data preprocessing step, this research successfully converts slang words into standardized forms, thus improving the quality of data used in modeling. The data was collected through YouTube Data API on the comments of the movie “Pengabdi Setan 2: Communion” and processed using tokenization, stemming, stopwords removal, and TF-IDF feature extraction techniques. The hybrid model combines machine learning algorithms such as Naive Bayes, Logistic Regression, and Random Forest with CNN's ability to extract complex spatial patterns from text data. The evaluation results show that this model is able to achieve up to 95% accuracy, with consistently high precision, recall, and F1-score. This approach not only improves the accuracy of sentiment analysis, but also provides an effective solution for handling non-standard language variations, making it relevant for application in digital opinion analysis on social media.

Keywords : *alay dictionary, hybrid CNN, movie commentary, sentiment analysis, slang, TF-IDF*

This work is an open access article and licensed under a Creative Commons Attribution-Non Commercial 4.0 International License

1. INTRODUCTION

The increase in online discussion activities through social media platforms and film review sites is an increasingly relevant phenomenon in the context of modern communication [1]. Social media has become the main means for individuals to share opinions, experiences, and information, including in the context of films. In this case, platforms such as Instagram, Twitter, and YouTube not only function as communication tools, but also as a space for deeper discussions about films, both in terms of criticism and recommendations. Research shows that the use of social media can increase user engagement and expand the reach of discussions, which in turn can influence viewers' decisions in choosing which films to watch [2]-[3].

One important aspect of online discussions is the ability of social media platforms to facilitate interaction between users. For example, research [4] shows that social media can be used to enhance communication with customers, where information about films and promotions can be conveyed effectively. Further more, *electronic word of mouth*(e-WoM) on social media has been shown to have a significant influence on movie ticket purchasing decisions. Research shows that e-WoM activities, such as*tweet* And *retweet*, can be an important indicator of a film's popularity, which in turn influences the audience's interest in watching the film [2].

In addition, the importance of social media literacy in the context of film discussions cannot be ignored. Users need to have a good understanding of how to use social media effectively to participate in constructive discussions. Research shows that socialization and training in the use of social media

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4465

can improve users' knowledge and skills in interacting on the platform [5]-[6]. Thus, increasing social media literacy can contribute to the quality of discussions that occur on the platform, thereby creating a more positive and informative environment for all users.

In the context of films, film review platforms such as IMDb and *Rotten Tomatoes* also play an important role in enhancing online discussion activities. Reviews and ratings from users can influence audiences' perceptions of a particular film, as well as encourage further discussion on social media. Research shows that users often refer to film reviews before making a decision to watch, thus creating a relationship between review platforms and social media in the context of film discussion [3]. Thus, collaboration between social media platforms and film review sites can create a richer ecosystem for film discussion, where users can share opinions and obtain more diverse information.

Natural Language Processing *(NLP*) plays an important role in sentiment analysis, especially in the context of movie reviews. The application of machine learning techniques in sentiment analysis has been explored extensively, with studies demonstrating the effectiveness of various algorithms in classifying sentiment from movie reviews. For example, study [7] emphasizes the development of a robust sentiment analysis model specifically designed for movie review data, highlighting the importance of text preprocessing and feature identification to achieve high classification accuracy.

The increasing activity of online discussions about films is often colored by the use of non-standard language, including slang, which can complicate the process of automatic sentiment analysis. In this context, the use of informal language and slang terms in film comments on social media is a challenge for sentiment analysis algorithms. Research shows that non-standard language, such as slang, can reduce the accuracy of natural language processing used to analyze sentiment [8]. This is due to variations in the use of non-standardized language, which is often unrecognizable by automated systems.

One relevant example is the study [9] which identified the use of slang in the film "*Rampage*". This study shows that *slang* used in film dialogue reflects the broader culture and social context, and often does not conform to the formal language commonly used in analysis. When film commentary on social media contains slang, this can make it difficult to accurately classify sentiment, as algorithms may not be able to understand the meaning contained in these terms.

Furthermore, research [10] shows that comments generated by users on digital platforms often reflect a diversity of opinions and different language styles, which can include pejorative language and slang. In this context, automated sentiment analysis must be able to handle the heterogeneity of language used by users. Limitations in understanding the context and nuances of slang can result in errors in interpreting sentiment, which in turn can affect the overall analysis results [11].

In addition, the use of non-standard language in film commentary can also create challenges in terms of data collection. Research by Duque et al. shows that commentary taken from digital platforms often contains non-standard language, reflecting the various experiences and perspectives of viewers [12]. This suggests that analysis conducted without considering the social and cultural context behind the use of such language may produce inaccurate conclusions.

In an effort to improve the accuracy of sentiment analysis, it is important to develop methods that can handle non-standard language and *slang*. Research [8] highlights the importance of developing translation tools. *Slang* which can help in understanding and analyzing comments containing non-standard language. Thus, the development of technology that is able to recognize and interpret slang in the context of films can help in improving the quality of sentiment analysis that is carried out automatically.

On the other hand, it is also important to consider the impact of the use of non-standard language in film discussions. Research [13] shows that user-generated comments can provide valuable insights into the film-watching experience, even though they are often delivered in non-standard language.

Vol. 6, No. 6, December 2025, Page. 5333-5348
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4465

Therefore, while the use of slang can complicate sentiment analysis, it can also provide a richer context for how viewers respond to films.

Sentiment analysis plays a vital role in understanding user perceptions of a film, which is highly relevant for marketing and decision-making purposes in the film industry. With the increasing use of social media and review platforms, sentiment analysis can provide deep insights into how audiences respond to a particular film. Research shows that sentiment analysis can be used to identify users' opinions, emotions, and attitudes towards a film, which in turn can influence marketing strategies and production decisions [14].

In sentiment analysis there is a crucial stage that determines the accuracy of the analysis results, namely the process *Labeling*. *Labeling* usually done after the data collection stage, where data obtained from sources such as Twitter or other platforms are categorized into sentiment classes, such as positive, negative, or neutral. Zaelani et al. explained that in sentiment analysis of *tweet*, the labeling process is carried out to classify *tweet* into three categories of sentiment, which are then followed by stages *text processing* and weighting using TF-IDF [15]. In addition, other studies have shown that in film sentiment analysis, the process *labeling* conducted after data collection through *web scraping*, where the data has been processed through the level *preprocessing* then labeled for further analysis using the Naive Bayes method [16]. Thus, the process *labeling* systematic and structured analysis is essential to improve the reliability and validity of sentiment analysis.

Previous research has shown that sentiment analysis can provide valuable insights into an individual's perception of a service or product. For example, research using the algorithm *Long Short-Term Memory* (LSTM) to evaluate lecturers' teaching in higher education showed 91.08% accuracy in classifying student sentiment [17].

The advancement of artificial intelligence (AI) technology in natural language processing has had a significant impact on sentiment analysis, especially in the context of movies. With the application of deep learning algorithms such as LSTM and CNN, the accuracy in analyzing movie reviews has increased drastically, as shown by studies comparing various classification methods [18]. In addition, models based on *transformer*, such as BERT, have been shown to be effective in capturing emotional nuances in text, improving sentiment analysis capabilities [19].

Conventional methods in sentiment analysis often face limitations when dealing with non-standard text, including the use of slang terms commonly found in movie comments. This non-standard text, which often includes slang terms and informal language variations, can be difficult for traditional algorithms such as *Naive Bayes* and *Support Vector Machine* (SVM) to classify sentiment accurately. Research shows that conventional models *hybrid* CNN shows significant potential in understanding public opinion. Methods such as *Convolutional Neural Networks* (CNN) and *Long Short-Term Memory* (LSTM) can be combined to improve accuracy in sentiment classification. Research by Al Omari et al. showed that a hybrid CNN-LSTM model can effectively classify sentiment by leveraging the strengths of each architecture [20]. In addition, other approaches such as *Naive Bayes* and *Support Vector Machine* (SVM) has also been applied to sentiment analysis on social media platforms such as Twitter, demonstrating the diversity of methods that can be used in this research [14]-[16]. Research using BERT also highlights the importance of understanding context in sentiment analysis, which can be integrated with models *hybrid* for better results [21]. Thus, the combination of these techniques can provide deeper insights into movie sentiment across platforms.

With the background of the problems that have been explained, this study aims to develop a hybrid CNN model that can overcome the challenges in analyzing the sentiment of film comments containing non-standard language, including slang. This study integrates an kamusalay as a preprocessing step that is specifically designed to handle *slang words*, allowing the model to be more accurate in identifying words that have different meanings depending on the social and cultural context, thereby increasing the

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4465

effectiveness of sentiment analysis. After the handling stage *slang words*, the labeling stage is carried out to convert the processed movie review data into a numerical representation using the TF-IDF vectorizer technique, which takes into account the weight of important words in the overall context. To improve the accuracy of sentiment classification, this study implements a hybrid approach using three machine learning algorithms, namely *Naive Bayes*, *Logistic Regression*, And *Random Forest*, combined with CNN architecture. Thus, this hybrid model combines the traditional strengths of machine learning in classifying data with the ability of CNN to handle deeper relationships between features. This model is trained using data that has been divided to learn from various patterns present in movie comments, both simple and more complex, resulting in a more reliable and effective model in analyzing sentiment from movie comments containing informal language and *slang*.

2. RESEARCH METHODS

In Figure 1, this research method illustrates the structured steps in formulating the research problem and explains the proposed method in detail.

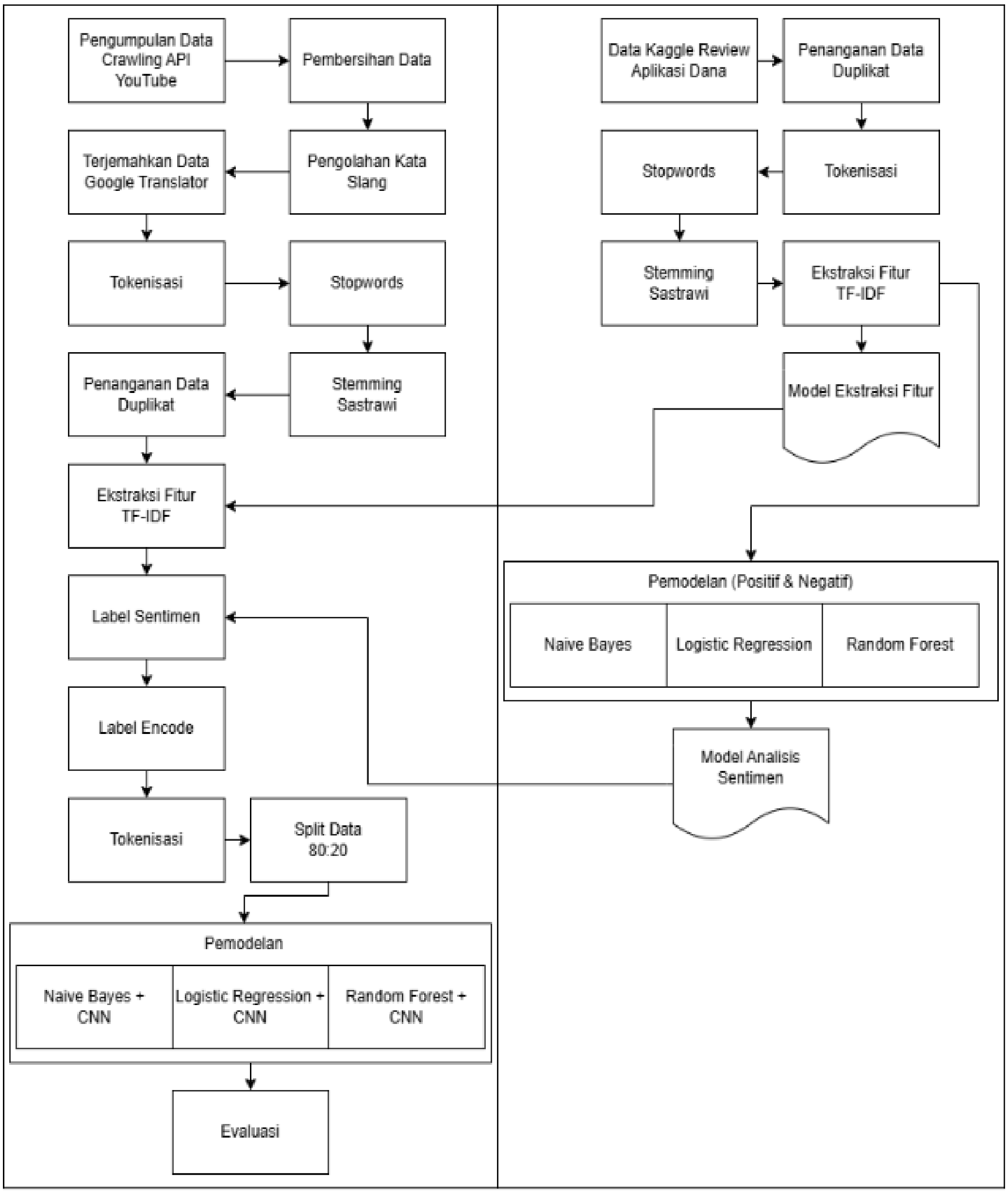

Figure 1. Research Method

Journal of Informatics Engineering (JUTIF) Vol. 6, No. 6, December 2025, Page. 5333-5348
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4465

2.1. Data Collection

Film review data collection methods *Devil's Servant 2: Communion* through the YouTube API crawling technique is done by utilizing the YouTube Data API service. The process begins by activating the YouTube API through *Google Cloud Platform* (GCP) and download the API credentials in JSON format for use in scripts *crawling*.

2.2. Data Cleansing

The text data cleaning process is carried out to ensure that the text in the dataset is cleaner, more consistent, and ready to be used in further analysis or processing. The first step is to convert all text to lowercase (*lowercase*) so that the difference between upper and lower case does not affect the analysis. Next, punctuation such as periods, commas, or exclamation marks are removed using regular expressions to simplify the text. Numbers are also removed because they are often irrelevant in the context of text analysis. Additionally, characters that repeat more than twice are converted to a single repetition to avoid redundant text. Emojis that have no textual meaning are also removed using an emoji library. Finally, empty lines that may have formed after the cleaning process, such as from text consisting entirely of deleted characters or emojis, are removed to ensure that the dataset only contains meaningful text.

2.3. Word Processing Slang

Methods used to process words *slang* in the text aims to replace non-standard words with more formal standard words. This process begins by creating a dictionary containing pairs of slang words and their standard words [22]. The data for these word pairs is taken from the reference file kamusalay.csv and processed into a dictionary, so that each word *slang* can be matched with its standard equivalent. Next, the text in the dataset is split into tokens (words) using space separators to facilitate the word replacement process. After the text is split, each word is checked for its existence in the dictionary *slang*. If the word is found, it is replaced with the standard equivalent from the dictionary; if not found, the word is kept as is. After all words have been checked and replaced, these tokens are recombined into a complete sentence.

2.4. Translate Data (*Google Translator*)

This data translation method aims to convert text into Indonesian to make it easier to understand. The process begins by automatically detecting the original language of the text, then translating it into Indonesian using a Google Translate-based translation tool. Each text is translated one by one, and if an error occurs in the process, the original text is still used but converted to lowercase. This method is applied to all rows of text in the dataset, resulting in a new column containing the translated text. In this way, the data becomes more relevant and suitable for analysis needs in Indonesian.

2.5. Tokenization

In this method, each sentence is broken down into a list of words to make data processing easier. This process is carried out on a column of text that has been translated into Indonesian, so that each line of text turns into a collection of words that are easier to analyze. The results of this process are stored in a new form, where long text is replaced with a list of words. In this way, the data becomes more organized and ready for the next processing step.

2.6. Stopwords

Deletion process *stop words* is an important step in text data processing to remove common words that do not provide much information value, such as "which," "in," or "and." In this method, each word

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4465

in the tokenized text is compared to a list of *stop words* in Indonesian. Words included in this list are removed from the text to leave more relevant words. This process helps clarify the main content of the text and makes further analysis easier. After *stop words* removed, the text column contains a cleaner list of words and focuses on the information that is important for further processing.

2.7. Stemming

Process *stemming* is a step to simplify words into their basic form. In this code, a tool is used to change words such as "walking" to "road" or "promising" to "promise." The goal is so that words that have similar meanings can be recognized as the same form. This process helps organize text data into a more organized and understandable form, making it easier to analyze in the next step. In this way, various variations of words that may appear in the data can be grouped more efficiently.

2.8. Duplicate Data Handling

Handling duplicate data aims to avoid unnecessary repetition of information in the dataset. In this case, if there are rows that have the same text, only the first row will be retained, while the rest will be deleted. This is important because duplicate data can cause inaccurate analysis results. By removing duplicates, we ensure that each piece of information in the dataset is unique, thus facilitating the analysis process and providing more precise results.

2.9. Feature Extraction

The feature extraction process is carried out using secondary data in the form of Dana application user reviews obtained from the Kaggle platform. This data was chosen because it reflects public opinion widely and has high relevance for sentiment analysis or understanding user experience. The first step in this process is to handle duplicate data, where identical or similar entries are removed to maintain data quality and avoid bias due to repetition. Next, the cleaned data is processed through tokenization, which is breaking down text into individual word units to facilitate further analysis. The next stage is the removal of stopwords, which are words that appear frequently but do not have significant information value, such as "adalah", "saya", or "dengan". After that, stemming is carried out using the Sastrawi library to change words to their basic form so as to reduce the variation of words that have the same meaning.

The final stage is feature extraction using the method *Term Frequency-Inverse Document Frequency* (TF-IDF). This method is used to measure the level of importance of each word in a document by calculating the frequency of occurrence of the word (*Term Frequency*) and compare it with the appearance of the word in the entire document (*Inverse Document Frequency*). By using secondary data from Kaggle, which has been curated and has a wide coverage, the TF-IDF process becomes more efficient and effective. The end result is a numerical representation of the text that reflects the relevance of each word in the context of the document, making it ready to be applied in machine learning models or primary data analysis.

2.10. Sentiment Labels

In the labeling stage, new review data that has gone through preprocessing is applied to *TF-IDF vectorizer* the same to ensure consistency of feature representation. Machine learning models, such as *Naive Bayes*, *Logistic Regression*, And *Random Forest*, trained using data that has been represented in the form of TF-IDF vectors. The results of this transformation are then fed into a model that has been trained to predict sentiment labels. Each model learns patterns in the data to distinguish between positive and negative sentiment.

2.11. *Label Encode*

Label encoding is a technique used to convert category labels into numbers so that they can be processed by machine learning models. In the code above, the labels in the review data are converted into numeric values using *Label Encoder*. This process converts categories such as positive and negative into numbers, for example negative becomes 0 and positive becomes 1.

2.12. Tokenization

Tokenization is the process of converting text into a form that can be understood by a computer. In this case, tokenization breaks sentences into smaller words or chunks and numbers each word based on the order in which it appears. That way, each word that appears in the text will be represented by a certain number. After that, the processed text is adjusted in length, so that each text has the same length to facilitate further data processing.

2.13. Split Data

After the data is processed with tokenization, the next step is to divide the data into two parts: training data (*training*) and test data (*testing*). This division is important to train the model on part of the data (80%) and test its performance on data that was not used during training (20%). In this way, the model can be evaluated to avoid *overfitting* and ensure that it can generalize well to new data.

2.14. Modeling

At the modeling stage, this study implements an approach *hybrid* by using three machine learning algorithms, namely *Naive Bayes*, *Logistic Regression*, And *Random Forest*, with architecture *Convolutional Neural Network* (CNN). Hybridization is done to utilize the advantages of each machine learning algorithm in sentiment classification and CNN's ability to extract spatial features from text data that has been converted into numeric representations. In this process, the algorithm *Naive Bayes* [23], *Logistic Regression*, And *Random Forest* used as an initial classification layer, while CNN acts as a boosting layer that captures more complex patterns from the data. The model *hybrid* it is trained using the shared data.

2.15. Evaluation

In the model evaluation stage, several metrics are used to assess the model's performance in predicting text sentiment. Among them are using *classification report* which provides detailed information about the accuracy, precision (*precision*), *recall*, and F1 score for each category (positive and negative) [24]. These metrics help understand the extent to which the model is able to correctly recognize positive and negative sentiments.

3. RESULTS AND DISCUSSION

3.1. Data Collection

The data used in this study only includes the named comment texts *textDisplay*, as seen in the table. These comment texts are obtained from user reviews retrieved through the YouTube API, where the comments include various expressions and opinions from viewers towards the film "Pengabdi Setan 2: Communion". Each extracted comment contains review text that may include slang words, informal expressions, and emojis, which are then processed to ensure their cleanliness and consistency. The processed comment texts will be used as the main data in sentiment analysis. This data, which was initially obtained in JSON format, is then stored in a structured format such as CSV to facilitate further analysis, as shown in Table 1.

Journal of Informatics Engineering (JUTIF) Vol. 6, No. 6, December 2025, Page. 5333-5348
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4465

Table 1. List of Comments on the Film Satan's Servant 2

	textDisplay
0	The cleric keeps dying, he's lost to...
1	Crazy, this is really cool, I've done it before...
2	<a href="https://www.youtube.com/watch ...
3	TACOOT
4	Like that
...	...
3370	This is going to be really cool. All the cast...
3371	I think the answer to the puzzle in S1 is...
3372	It's finished, because the father is finished...
3373	🔥🔥🔥🔥🔥
3374	Cant waittt

3.2. Data Cleansing

After cleaning the comment text using the clean_texts function, the comment text data contained in the column *textDisplay* has been changed to be cleaner and more consistent. The cleaning steps taken include converting text to lowercase, removing punctuation, removing numbers, and removing characters that appear more than twice and emojis. Furthermore, filtering is carried out to remove lines containing links (URLs). The results in Table 2 are comment texts that are more focused on user opinions and expressions without interference from external links.

Table 2. Data Cleaning

	textDisplay
0	Ustadz always dies, he will lose to...
1	crazy this is really cool, I've done it before...
2	takod
3	like that
4	joko anwar is a guarantee ...
...	...
3277	the real master piece
3278	It will be really cool for sure, all the casts are in...
3279	It seems the answer to the riddle in s is not yet ...
3280	It's finished because the father is finished...
3282	cant wait

3.3. Word Processing *Slang*

After normalization is carried out on the column *textDisplay* use a dictionary *slang*, comment text containing the words *slang* has been changed into a more standard word form. This normalization process use s*kamuay.csv* which maps every word *slang* to the corresponding standard words. The results of this normalization in Table 3, produce more understandable and more consistent text, which is ready to be used in further analysis such as sentiment analysis or language modeling.

3.4. Translate Data (*Google Translator*)

The results of applying the translation process to the column *textDisplay* shows that comments that may have previously been mixed in language or contained foreign terms have now been translated into Indonesian. This makes the text more relevant and understandable in the context of the target

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4465

language. With this translation, the data becomes more uniform and suitable for analysis purposes, such as sentiment analysis, opinion clustering, or linguistic studies. This process also helps reveal users' emotions and views in the local context, providing deeper insights into their responses. The results of the transformation are shown in Table 4.

Table 3. Slang Word Processing

	textDisplay
0	Ustadz keeps dying, oh my gosh, he's lost...
1	crazy this is really cool, I've done it before...
2	scared
3	scared
4	joko anwar is a guarantee ...
...	...
3277	the real master piece
3278	It will be really cool for sure, all the casts are in...
3279	looks like the answer to the riddle in s...
3280	It's finished because the father finished...
3282	cant wait

Table 4. Translate Data (Google Translator)

	textDisplay
0	Ustadz keeps dying, oh my gosh, he's lost...
1	crazy this is really cool, I've done it before...
2	scared
3	scared
4	joko anwar is a guarantee ...
...	...
3277	real work
3278	It will be really cool for sure, all the casts are in...
3279	looks like the answer to the riddle in s...
3280	It's finished because the father finished...
3282	can't wait

3.5. Tokenization

The tokenization process applied to the review data successfully broke each comment into a separate list of words, so that each word could be analyzed individually as shown in Table 5.

3.6. *Stopwords*

Deletion process *stopwords* from the tokenized data produces comments that are more focused on the main meaningful words. By eliminating common words such as "and," "which," or "di," text data analysis becomes more efficient because only relevant words are retained. Table 6 shows the data from the processing results *stopwords*.

3.7. *Stemming*

The results of the stemming process using the Sastrawi library produce a transformation of each word in the review text into its basic form. As shown in Table 7, after *stemming*, the words in the comment text are changed to base form, for example in the first line, the original text is changed to base.

Journal of Informatics Engineering (JUTIF) Vol. 6, No. 6, December 2025, Page. 5333-5348
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4465

This process aims to reduce the variation of words with the same meaning, such as the use of suffixed verbs.

Table 5. Tokenization

	textDisplay
0	[sir, cleric, he, died, always, oh my, ...
1	[crazy, this, cool, really, before, already, ...
2	[scared]
3	[scared]
4	[joko, anwar, is, a, guarantee, ...
...	...
3277	[work, real]
3278	[will, be, really, definitely, cool, this, all, ...
3279	[it seems, answer, above, riddle, puzzle, in, s, ...
3280	[it's finished, it's finished, sir, ...
3282	[no, wait, wait]

Tabel 6. Stopwords

	textDisplay
0	[ustadz, nya, mati, mulu, gosh, lost, ...
1	[crazy, cool, really, watch]
2	[scared]
3	[scared]
4	[joko, anwar, guarantee, quality, high, ...
...	...
3277	[work, real]
3278	[cool, really, cast, movie, prequel, ...
3279	[seems, riddle, riddle, s, completely, ...
3280	[done, finished, the, devil's, agreement, ...
3282	[patiently, waiting]

Table 7. Mood

	textDisplay
0	[ustadz, nya, mati, mulu, gosh, lost, ...
1	[crazy, cool, really, watch]
2	[scared]
3	[scared]
4	[joko, anwar, guarantee, quality, high, ...
...	...
3277	[work, real]
3278	[cool, really, cast, movie, prequel, ...
3279	[like, riddle, riddle, s, full, answer, s, ...
3280	[done, finished, promised, devil, ...
3282	[be patient, wait]

After that, the list of resulting words *stemming* converted back into complete sentences to facilitate further interpretation and analysis, as presented in Table 8.

Journal of Informatics Engineering (JUTIF) Vol. 6, No. 6, December 2025, Page. 5333-5348
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4465

Table 8. Sentence Conversion

	textDisplay
0	The ustadz keeps dying, oh my God, he's defeated by the devil
1	crazy cool to watch
2	scared
3	scared
4	joko anwar guarantees high quality support ...
...	...
3277	real work
3278	the cast of the prequel film is really cool...
3279	like a puzzle s full answer s brfeling ...
3280	finished, the devil's promise was fulfilled so that...
3282	wait patiently

3.8. Duplicate Data Handling

The process of handling duplicate data is an important step in ensuring the quality of the dataset for further analysis. From the preprocessing results, it was found that there were 83 rows of duplicate data in the initial dataset. These duplicate data were identified based on the column *review*, which contains the review text after going through various preprocessing stages, including stemming and sentence conversion. For example, in Table 8, it can be seen that text such as *"scared"* appears more than once in the third and fourth rows. After the deduplication process is performed, only one entry from each text is retained, while the rest are deleted, as shown in Table 9. In this table, the third row containing*"scared"*are retained, while duplicate entries are removed.

Table 9. Data After Duplicate Data Handling

	textDisplay
0	The ustadz keeps dying, oh my God, he's defeated by the devil
1	crazy cool to watch
2	scared
4	joko anwar guarantees high quality support ...
5	willing to let the character die like that, dear...
...	...
3276	wow, you're so patient, bomb
3277	real work
3278	the cast of the prequel film is really cool...
3279	like a puzzle s full answer s brfeling ...
3280	finished, the devil's promise was fulfilled so that...

3.9. Feature Extraction

In the feature extraction stage, text data is converted into numeric representation using the TF-IDF method. This process aims to calculate the weight of each word in the document based on its frequency of occurrence and reduce the influence of common words that are less meaningful. The cleaned dataset is first processed by deleting empty data (NaN) in the column *content*. Next, text features are extracted using *Tf df Vectorizer* with a maximum parameter of 1000 features, which aims to reduce the complexity of data dimensions. The result is a TF-IDF matrix containing the weight values of each word in the document in the form of an array, which is then used as an independent feature (*X*), while

Journal of Informatics Engineering (JUTIF) Vol. 6, No. 6, December 2025, Page. 5333-5348
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4465

the target variable (*and*) are pre-defined sentiment labels. This process ensures that the text data is ready to be used in the modeling stage.

3.10. Sentiment Labels

The results of the evaluation of the performance of the classification models show significant differences between the algorithms *Naive Bayes*, *Logistic Regression*, And *Random Forest* based on the metrics used, namely accuracy, precision, *recall*, And *F1-score*. *Naive Bayes*, although fast and simple, has an accuracy of 79% as shown in the table, with a precision of 78%, *recall* 79%, and *F1-score* 76%. This algorithm shows limitations in handling data because it tends to assume independence between features, which may not be fully applicable in the dataset used. *Logistic Regression* showed improved performance with 82% accuracy, 82% precision, 82% recall, and 81% F1-score, reflecting a model that is more capable of modeling linear relationships between features and targets. However, the best performance was obtained from *Random Forest* with the highest accuracy, namely 93%, which is supported by precision, *recall*, And *F1-score* which all reached 93%. *Random Forest* excels because of its ability to handle complex features and reduce *overfitting* by combining the results of several decision trees. Based on these results, *Random Forest* can be considered as the most reliable algorithm for classification on this dataset, as shown in the performance evaluation table 10.

Table 10. Sentiment Label Model Evaluation Results

Matrix	Naive Bayes	Logistic Regression	Random Forest
Accuracy	79	82	93
Precision	78	82	93
Recall	79	82	93
F-1 Score	76	81	93

After the model is trained, the next step is to use it to predict sentiment labels on primary data. The prediction results of the three algorithms used are: *Naive Bayes*, *Logistic Regression*, And *Random Forest*, indicating a consistent sentiment distribution. Based on Table 11, the distribution of the number of labels for each algorithm produces the same number of reviews for each sentiment label. A total of 1961 reviews are labeled POSITIVE, while 1068 reviews are labeled NEGATIVE. This consistency indicates that the three algorithms have good ability in detecting sentiment patterns uniformly in the dataset.

Table 11. Distribution of Number of Labels

Sentiment Labels	**Naive Bayes**	**Logistic Regression**	**Random Forest**
Positive	1961	1961	1961
Negative	1068	1068	1068

3.11. Label Encode

The process of encoding sentiment labels using *Label Encoder* successfully converted sentiment data into a numeric format that is easier to use in machine learning models. As seen in Table 12, the sentiment label NEGATIVE has been encoded as 0, while POSITIVE has been encoded as 1. After the encoding process, the data distribution remains consistent with the initial data, namely there are 1961 data with label 0 (NEGATIVE) and 1068 data with label 1 (POSITIVE). This result shows that even though the data has gone through a transformation process, the proportion of sentiment data distribution has not changed, which ensures that data integrity is maintained.

Table 12. Label Encode

Sentiment Labels	Encoded Label	Amount
Negative	0	1961
Positive	1	1068

3.12. Tokenization

The results of text tokenization that have been carried out on three datasets on each algorithm involve several stages that produce a numeric representation of the text. The process begins with the creation of an object *Tokenizer*, which is used to convert each word in the review column of each dataset into a numeric index. These indexes reflect the order in which the word appears in the entire text in the dataset.

3.13. Split Data

The results of data division using the method *train_test_split* shows that the dataset was successfully divided into two main parts, namely training data and testing data, with a proportion of 80% for training and 20% for testing. Based on Table 13, the distribution of the number of data divisions of 2423 data was used to train the model, while the remaining 606 data were used for testing. Data division was carried out randomly using the value *random_state* which remains (42), which ensures that the division results can be replicated in subsequent experiments.

Table 13. Distribution of the Number of Data Divisions

Data Types	Amount	Proportion
Training	2423	80%
Testing	606	20%

3.14. Modeling

This modeling uses an approach *Convolutional Neural Network* (CNN) to classify text data. The process begins by defining the size of the *vocab*(vocabulary) that will be used in the model, which is calculated based on the number of words in the previously processed text data. Then, the CNN model is built using several layers, starting with the layer *embedding* to convert text into a vector representation. Next, a layer is used *Conv 1 D* which serves to extract important features from a sequence of words by using certain filters and kernel sizes. Pooling is done with layers *Max Pooling 1 D* to reduce the data dimension and reduce the model complexity. The final result is processed with layers *Flatten* which converts data into a format that can be accepted by the layer *Dense*, which finally produces a binary classification output, with the activation function *sigmoid*. To train the model, the technique is use d*early stopping*, which allows stopping training if the model does not show improvement on validation data after a few epochs, to prevent overfitting. The model was trained for 10 epochs with a batch size of 64. This approach aims to produce an accurate model in classifying text sentiment based on the available data.

3.15. Evaluation

The results of the model evaluation show very good performance in all three model combinations, namely *Naive Bayes* + CNN (NB+CNN), *Logistic Regression*+ CNN (LR+CNN), and *Random Forest*+ CNN (RF+CNN). All these models have very similar evaluation results with accuracy, precision, *recall*, And *F1-Score*which reaches 94-95%. Based on table 14 Hybrid CNN Model Evaluation Results, the accuracy for the NB+CNN model is 94%, LR+CNN reaches 95%, and RF+CNN obtains an accuracy of 94%. Overall, the LR+CNN model shows a slight advantage with a value of 95% on all evaluation metrics, although the difference is very small when compared to NB+CNN and RF+CNN which each obtained a score of 94%. This shows that the three model combinations have very good performance in binary classification, with almost equal ability to recognize and predict positive and negative classes accurately and balanced. As a note, the use of CNN as a convolutional layer in these three models has a positive impact on model performance, which can identify important features of text data more effectively.

Table 14. Hybrid CNN Model Evaluation Results

Matrix	NB+CNN	LR+CNN	RF+CNN
Accuracy	94	95	94
Precision	94	95	94
Recall	94	95	94
F-1 Score	94	95	94

4. DISCUSSION

The results of this study indicate that the application of the hybrid CNN model in sentiment analysis on film comments provides significant performance with an accuracy level of 94-95%. This model is able to handle challenges that arise due to the use of non-standard language and slang, thanks to the integration of kamusalay in the data preprocessing process. The accuracy in classifying sentiment, both positive and negative, shows the potential of the hybrid algorithm in dealing with the diversity of language styles used by social media users.

These results are in line with previous studies that have shown the effectiveness of hybrid models in analyzing movie review sentiment [25]. However, the approach proposed in this study successfully improves accuracy by adapting more complex data preprocessing strategies, including tokenization, stemming, and stopword removal. This combination ensures that the data used is cleaner and more relevant for analysis.

In comparison, Naive Bayes and Logistic Regression algorithms show good performance, but are not as competent as Random Forest in capturing complex data patterns. In addition, CNN models are proven to be effective in extracting spatial features, which is a major advantage of the hybrid approach**.**

5. CONCLUSION

This research has succeeded in proving that the model *hybrid* CNN with algorithm *Naive Bayes*, *Logistic Regression* and Random Forest, which is integrated with slang word handling through alay dictionary can effectively overcome challenges in film comment sentiment analysis, especially in managing non-standard language and slang. The evaluation results show that this approach produces accuracy, precision, *recall*, And *F1-score*consistently high, making it a promising choice for sentiment analysis in the digital age. These findings not only contribute to the development of sentiment analysis technology, but also offer valuable insights for the film industry in understanding audience perception.

BIBLIOGRAPHY

[1] Y. K. Dwivedi *et al.*, "Setting the future of digital and social media marketing research: Perspectives and research propositions," *Int. J. Inf. Manage.*, vol. 59, no. May 2020, p. 102168, 2021, doi: 10.1016/j.ijinfomgt.2020.102168.
[2] E. P. Fahlevy and M. A. Pribadi, "Pengaruh Electronic Word of Mouth (E-WoM) terhadap

Vol. 6, No. 6, December 2025, Page. 5333-5348
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4465

Keputusan Pembelian Tiket Film Oppenheimer," *Prologia*, vol. 8, no. 1, pp. 143–151, 2024, doi: 10.24912/pr.v8i1.27568.

[3] M. Aldan Nur Zen and A. S. Sitanggang, "Analisis Dampak Sosial Media Dalam Pengembangan Sistem Informasi," *Cerdika J. Ilm. Indones.*, vol. 3, no. 7, pp. 671–682, 2023, doi: 10.59141/cerdika.v3i7.647.

[4] S. O. Pertiaz and S. P. Sunaryo, "Penggunaan Media Sosial oleh PT KAI dalam Meningkatkan Komunikasi dengan Pelanggan," *J. Penelit. Inov.*, vol. 3, no. 2, pp. 385–394, 2023, doi: 10.54082/jupin.169.

[5] S. I. Liestyasari, O. H. Nurcahyono, D. Astutik, and N. Nurhadi, "Literasi Penggunaan Media Sosial Sehat Bagi Forum Anak Surakarta," *Dedik. Community Serv. Reports*, vol. 2, no. 2, pp. 58–65, 2020, doi: 10.20961/dedikasi.v2i2.37834.

[6] L. Rahayuwati, R. H. Permana, and A. Labertha, "Pemutaran Video, Diskusi dan Penggunaan Standing Banner pada Siswa untuk Mencegah Kecanduan Media Sosial," *Media Karya Kesehat.*, vol. 2, no. 1, pp. 85–95, 2019, doi: 10.24198/mkk.v2i1.21667.

[7] A. Gupta, S. Pandey, M. P. Behera, S. Darshana, and A. Dash, "Sentiment Analysis Of Movie Review Using Machine Learning," *Educ. Adm. Theory Pract.*, vol. 30, no. 5, pp. 7344–7354, 2024, doi: 10.53555/kuey.v30i5.4159.

[8] K. C. Eleison, S. U. I. Hutahaean, S. C. Tampubolon, T. M. Panggabean, and I. Fitriyaningsih, "An empirical evaluation of phrase-based statistical machine translation for Indonesia slang-word translator," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 25, no. 3, pp. 1803–1813, 2022, doi: 10.11591/ijeecs.v25.i3.pp1803-1813.

[9] U. Tursini and P. I. Fauziah, "Slang Words on Rampage Movie," *Rhizome J. Kaji. Ilmu Ilmu Hum.*, vol. 2, no. 2, pp. 1–8, 2022, doi: 10.56393/rhizome.v2i1.479.

[10] C. Suryanovika and L. N. Affini, "Pejorative Words Indicating Indonesian Hate Speech," *Kata*, vol. 25, no. 1, pp. 53–64, 2023, doi: 10.9744/kata.25.1.53-64.

[11] H. Gong, A. Valido, K. M. Ingram, G. Fanti, S. Bhat, and D. L. Espelage, "Abusive Language Detection in Heterogeneous Contexts: Dataset Collection and the Role of Supervised Attention," *35th AAAI Conf. Artif. Intell. AAAI 2021*, vol. 17A, pp. 14804–14812, 2021, doi: 10.1609/aaai.v35i17.17738.

[12] E. Duque, G. L. de Aguileta, J. M. C. Barbany, and M. J. Foraster, "No Time to Lose Sex-appeal. Love and Attractiveness in the last James Bond Movie 2021," *Masculinities Soc. Chang.*, vol. 11, no. 2, pp. 102–128, 2022, doi: 10.17583/mcs.9777.

[13] F. Toye, K. Seers, and K. Barker, "'It's like she's talking about me' — Exploring the value and potential impact of a YouTube film presenting a qualitative evidence synthesis about chronic pain: An analysis of online comments," *Can. J. Pain*, vol. 4, no. 3, pp. 61–70, 2020, doi: 10.1080/24740527.2020.1785853.

[14] M. Khairudin, A. Sukendar, and A. Somantri, "Analisis Sentimen Film Di Twitter Menggunakan Metode Support Vector Machine," *J. Sains dan Sist. Teknol. Inf.*, vol. 5, no. 1, pp. 97–102, 2023, doi: 10.59811/sandi.v5i1.47.

[15] R. S. A. Al-Zaelani, Y. R. Ramadhan, and M. A. Komara, "Analisis Sentimen Review Produk Motor Honda Pcx Dan Yamaha N-Max Pada Twitter Menggunakan Algoritma Naïve Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 3, pp. 1812–1816, 2023, doi: 10.36040/jati.v7i3.7008.

[16] A. C. Kamilla, N. Priyani, R. Priskila, and V. H. Pranatawijaya, "Analisis Sentimen Film Agak Laen Dengan Kecerdasan Buatan: Text Mining Metode Naïve Bayes Classifier," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 3, pp. 2923–2928, 2024, doi: 10.36040/jati.v8i3.9587.

[17] W. M. Baihaqi and A. Munandar, "Sentiment Analysis of Student Comment on the College Performance Evaluation Questionnaire Using Naïve Bayes and IndoBERT," *JUITA J. Inform.*, vol. 11, no. 2, p. 213, 2023, doi: 10.30595/juita.v11i2.17336.

[18] M. M. Aziz, M. D. Purbalaksono, and A. Adiwijaya, "Method comparison of Naïve Bayes, Logistic Regression, and SVM for Analyzing Movie Reviews," *Build. Informatics, Technol. Sci.*, vol. 4, no. 4, pp. 1714–1720, 2023, doi: 10.47065/bits.v4i4.2644.

[19] S. V. Sureshbhai and T. Nakrani, "A Literature Review : Enhancing Sentiment Analysis of Deep Learning Techniques Using Generative AI Model," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 10, no. 3, pp. 530–540, 2024, doi: 10.32628/cseit24103204.

[20] N. Hicham, S. Karim, and N. Habbat, "Enhancing Arabic Sentiment Analysis in E-Commerce Reviews on Social Media Through a Stacked Ensemble Deep Learning Approach," *Math. Model. Eng. Probl.*, vol. 10, no. 3, pp. 790–798, 2023, doi: 10.18280/mmep.100308.
[21] D. F. Sjoraida, B. W. K. Guna, and D. Yudhakusuma, "Analisis Sentimen Film Dirty Vote Menggunakan BERT (Bidirectional Encoder Representations from Transformers)," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 8, no. 2, pp. 393–404, 2024, doi: 10.35870/jtik.v8i2.1580.
[22] N. Fadhillah and A. Nurweni, "Analyzing the Word Formation of Slang Vocabulary Employed by Fathia Izzati on YouTube," *Abjadia Int. J. Educ.*, vol. 8, no. 2, pp. 215–223, 2023, doi: 10.18860/abj.v8i2.23480.
[23] H. Junianto, P. Arsi, B. A. Kusuma, and D. I. S. Saputra, "Evaluasi Aplikasi Raileo Melalui Analisis Sentimen Ulasan Playstore Dengan Metode Naive Bayes," *SINTECH (Science Inf. Technol. J.*, vol. 7, no. 1, pp. 27–40, 2024, doi: 10.31598/sintechjournal.v7i1.1505.
[24] D. Hardiansyah, R. Z. Abdul, and M. Said, "The Classification Method is Used for Sentiment Analysis in My Telkomsel," vol. 8, no. 2, 2024.
[25] G. Dubey *et al.*, "A Hybrid Convolutional Network and Long Short-Term Memory (HBCNLS) model for Sentiment Analysis on Movie Reviews," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 4, pp. 341–348, 2023, doi: 10.17762/ijritcc.v11i4.6458.