JURNAL INFORMATIKA, Vol. 11 No. 2 October 2025, Page 81-89
ISSN (Print) 2355-6579 | ISSN (Online) 2528-2247
DOI: https://doi.org/10.31294/inf.v12i2.25897

Performance Evaluation of LSTM and GRU Models for Movie Genre Classification Based on Subtitle Dialogs Using Augmented Data and Cross-Validation

Ni Luh Putu Yonita Putri Utami[1*], Desy Purnami Singgih Putri[2], Ni Kadek Dwi Rusjayanthi[3]

[1,2,3] Universitas Udayana
Kampus Bukit Jimbaran Street, Badung 80361, Bali, Indonesia

Correspondence e-mail: yonitaputriutami26@gmail.com

Submission:	Revision:	Acceptance:	Available Online:
21-05-2025	29-06-2025	01-07-2025	02-10-2025

Abstract

This study aims to evaluate and compare the performance of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models in classifying movie genres based on subtitle dialogs. To address data imbalance across genres, data augmentation was applied to create balanced datasets with 500 and 700 samples per genre, in addition to the original dataset. The classification models were built using Word2Vec for word embedding, followed by LSTM and GRU architectures with a single hidden layer and dropout regularization. Model performance was assessed using accuracy and further validated through 5-fold cross-validation. The best test accuracy was achieved with the dataset containing 700 samples per genre, reaching 91% for LSTM and 92% for GRU. Cross-validation showed stable performance with average accuracies of 0.68 for LSTM and 0.67 for GRU. A paired t-test analysis yielded a p-value of 0.341, indicating no statistically significant difference between the two models. These findings suggest that both LSTM and GRU are effective for genre classification based on subtitle dialogs. The use of data augmentation is a key contribution of this study, enabling improved model performance on underrepresented genres. This research supports the development of automated movie recommendation systems that utilize subtitle-based genre prediction.

Keywords: LSTM, GRU, Data Augmentation

1. Introduction

Films are one of the most widely consumed forms of entertainment, and the film industry has experienced rapid growth over the past few decades. Every year, thousands of movies with different genres are produced. Every movie has a genre that serves to distinguish the type of movie. Common movie genres are action, comedy, drama, fantasy, horror, mystery, romance, sci-fi and thriller. Films can have more than one main genre and other secondary genres, which in some parts are a combination of different genres (Rajput & Grover, 2022). Film genres play a very important role in determining audience preferences and offer filmmakers guidance for plot and production (Mangolin et al., 2022). The variety of movie genres available often confuses viewers to distinguish one genre from another. Automatic classification of movie genres is still a challenge in the field of natural language processing and text analysis (Akbar et al., 2025).

People in Indonesia tend to be more interested in movies that come from abroad (Azka et al., 2024). Subtitles can be used to better understand the message contained in the film, in the form of subtitles and dubbing. Subtitles make it easier for the audience to understand the plot by reading the subtitles or listening to the conversations in the movie in the language in which the movie is shown.

The source of information that can be used to classify movie genres is subtitle data. Subtitles contain dialog transcriptions and other important descriptions that can be used for movie genre analysis and classification. Deep learning methods have emerged in recent years as a popular and effective approach for performing natural language processing tasks, including text classification (Alzoubi et al., 2024).

Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are two types of recurrent neural network (RNN) models that are particularly adept at processing sequential data, such as subtitle texts, in addition to various deep learning approaches (Wijaya et al., 2024). LSTM, introduced by (Hochreiter & Schmidhuber, 1997)

was developed to address the vanishing gradient problem in traditional RNNs by enabling the capture of long-term dependencies. GRU, proposed by (Cho et al., 2014), is a simplified variant of LSTM that reduces computational complexity while maintaining competitive performance in sequence modeling tasks. LSTM and GRU have been demonstrated to possess the capacity to identify patterns and long-term dependencies within text, a capability that can facilitate the extraction of features relevant to movie genres. While both models have been shown to be effective in handling long-term dependencies, a direct comparison of their performance in classifying movie genres using subtitle dialogues has yet to be thoroughly explored. This makes it a compelling subject for further research (Akbar et al., 2025).

Previous research has explored subtitle-based genre classification using traditional machine learning algorithms. Nikhil Kumar Rajput and Bhavya Ahuja Grover investigated movie genre recognition using English subtitle data. They used 964 movies belonging to six genres—action, fantasy, horror, romance, sports, and war—for movie genre recognition. The genres were classified using six algorithms, including logistic regression, support vector machine, naïve Bayes classifier, decision tree, neural network, and k-nearest neighbor. The .srt file is tokenized to predict the genre of the subtitle dialogue. The results obtained were an average accuracy ranging from 70% to 80%. The neural network achieved a similar level of accuracy, followed by logistic regression. The K-Nearest Neighbor (KNN) model achieved an accuracy of 77% (Rajput & Grover, 2022).

Nathania Novenrodumetasa et al. conducted a study analyzing movie genres based on subtitle data using Random Forest and Naïve Bayes algorithms. The study aimed to analyze movie genres based on subtitle data by comparing the two algorithms. The Random Forest algorithm was more accurate than the Naïve Bayes algorithm. The Random Forest algorithm had an accuracy of 0.841, while the Naïve Bayes algorithm had an accuracy of 0.682 (Novenrodumetasa et al., 2023). However, these studies did not explore the capabilities of deep learning models such as LSTM and GRU in this context.

To address this gap, the present study conducts a comparative evaluation of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models for movie genre classification based on subtitle dialogs. A notable contribution of this research lies in the use of data augmentation to overcome class imbalance, combined with a 5-fold cross-validation strategy to ensure robust and reliable model evaluation. By integrating deep learning architectures with systematic preprocessing and validation techniques, this study offers valuable insights into the effectiveness of sequential models for text-based genre classification, and supports the advancement of intelligent systems for automated movie recommendation.

2. Research Methods

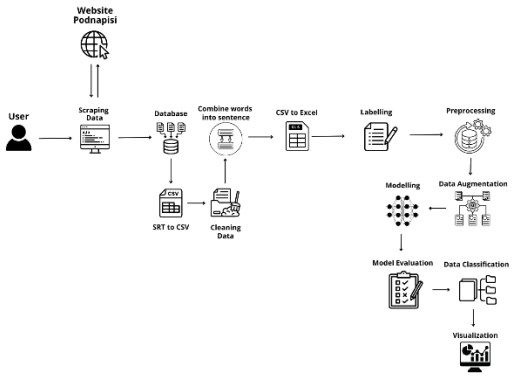


Figure 1. Research methods

Figure 1 shows a systematic research method for evaluating the performance of LSTM and GRU models in classifying movie genres using subtitle dialogue data. The process begins with the collection of subtitle data using web scraping from the Podnapisi.net website, followed by data conversion and cleaning to create a suitable dataset structure. After data cleaning, sentences are merged and exported to Excel format to facilitate manual genre labeling. The labeled data then undergoes preprocessing and augmentation phases to ensure the quality and balance of the distribution between genres. After data preparation is complete, modeling is performed using the LSTM and GRU architectures, followed by an evaluation of the model performance. The final classification results are visualized to support the analysis and conclusions of the study. A detailed explanation of each phase of the method is provided in the following subsections.

2.1 Web Scraping

Web scraping is a technique used to retrieve data or information from a website using a markup language such as HTML or XHTML. Information can be in the form of text, links, video, audio, or documents (Kusumo & Somya, 2022).

2.2 Preprocessing

Preprocessing is the first phase of raw data processing to create clean, consistent data for machine learning processing. Preprocessing techniques include case correction, sanitization, tokenization, and stop word removal. Case correction converts all letters in the text to lowercase. Cleaning involves removing irrelevant

or distracting elements such as punctuation, numbers, special characters, or meaningless words from data (Purnomo & Syafarina, 2024). Tokenization involves breaking text into smaller units called tokens. These are usually words or phrases. Stop word removal removes frequently occurring words that are not essential to the analysis, such as "and," "or," "which," and similar terms(Salam et al., 2023).

2.3 Data Augmentation

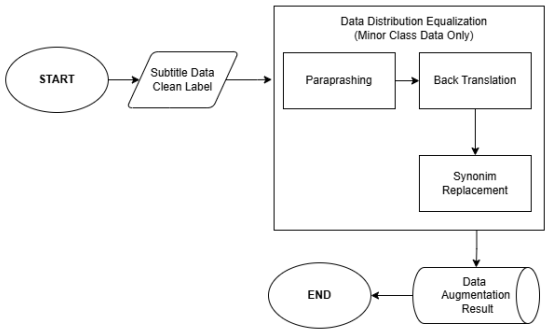


Figure 2. Data Augmentation Flow

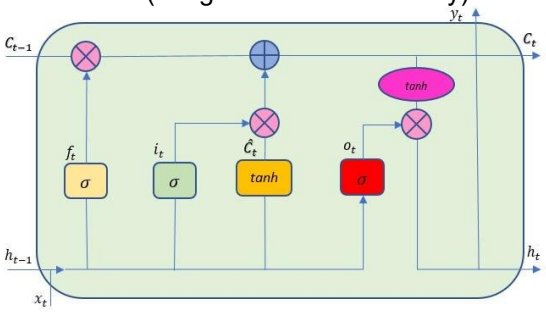
Data augmentation is the process of creating new data based on existing data to enhance the size and variety of a dataset, particularly for training machine learning models(Shorten et al., 2021). The goal is to improve model performance, prevent overfitting, and address class imbalances (Du et al., 2023). As shown in Figure 2, the augmentation techniques used are paraprashing, back translation and synonym replacement. Augmentation techniques include paraphrasing, changing the overall structure of a sentence while maintaining its original meaning (Beddiar et al., 2021). Back translation refers to the translation of text into another language and its subsequent conversion back to the source language (Ibrahim et al., 2024). Label distribution adjustment is a technique for adjusting the data set between classes by adding data samples to the minority class until the target size is reached (Du et al., 2023). Synonym replacement is a technique for replacing specific words with their synonyms using a dictionary or thesaurus (Nifanto & Nurhopipah, 2024).

The paraphrasing was performed using the ramsrigouthamg/t5_paraphraser model based on the T5 architecture. Back-translation used Helsinki-NLP models for English–German–English conversion. For synonym replacement, sentences were tokenized, and replaceable words were identified using the WordNet dictionary. Augmentation was applied only to training data, targeting underrepresented genres to reach 500 and 700 samples per class.

2.4 Modelling

Modeling is the process of mathematically representing patterns or relationships discovered by an algorithm in data. Machine learning models serve as mathematical representations of patterns found in data and are used to make predictions or classify new, previously unknown data. When building a model, aspects such as overfitting, generalization, and data imbalance must be considered to ensure that the resulting model has optimal performance(Sarker, 2021).

2.4.1 LSTM (Long Short-Term Memory)



Source: (Sari et al., 2019)

Figure 3. LSTM Architecture

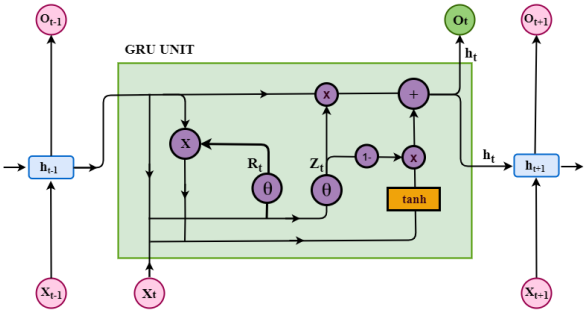
As shown in Figure 3, the LSTM architecture consists of several components. LSTM (Long Short-Term Memory) is an artificial neural network (RNN) architecture specifically designed to solve the problems of missing and erratic gradients that occur in deep learning models. LSTMs are able to store and retain important information over long periods of time through special internal memory mechanisms. This makes them ideal for natural language processing, including text classification.

LSTM is comprised of a structural composition involving cell states and multiple gates. The forget gate determines which information from the preceding cell state should be discarded. The input gate is responsible for the management of new information additions to the cell state. The cell state is then updated by combining the output of the forget gate with the incoming input. As stated by (Sari et al., 2019) the output gate is responsible for determining the information that is transmitted as the hidden state output.

2.4.2 GRU (Gated Recurrent Unit)

As shown in Figure 4, the GRU architecture is composed of several key components. The GRU (Gated Recurrent Unit) is a recurrent neural network (RNN) architecture that has been shown to overcome the vanishing gradient problem and process sequential data efficiently. GRU is a model that is less complex than Long Short-Term

Memory (LSTM), but is still able to capture long-term dependencies in data. This model regulates the dissemination of information through two primary gates: the reset gate and the update gate



Source: (Bibi et al., 2020)

Figure 4. GRU Architecture

The reset gate modulates the extent to which information from the previous hidden state is discarded during the computation of the new hidden state. At the same time, the update gate controls the extent to which the previous hidden state is retained and the degree to which new information is integrated (Bibi et al., 2020).

2.5 Model Evaluation

Model evaluation is used to objectively and more precisely measure and validate the performance of machine learning models. Evaluation is performed using cross-validation to avoid the problem of overfitting. This occurs when the model overfits the training data and is therefore less able to generalize to new, unprecedented data. Cross-validation involves splitting the dataset into multiple folds. The model is then trained and tested alternately on these folds (Wijiyanto et al., 2024).

2.6 Classification

Classification is the process of predicting or grouping data into specific categories or classes based on the characteristics or attributes of the data (Pamungkas et al., 2019).

3. Results and Discussion

3.1 Web Scraping

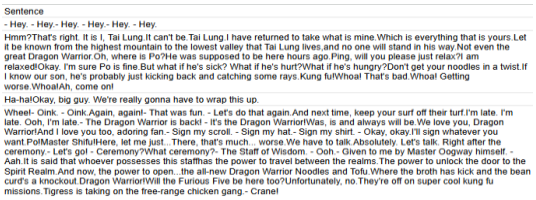
The collected data included 400 subtitle dialog files. The dataset consists of three columns: start, end, and text. Table 1 shows an example of data successfully collected through web scraping on the Podnapisi website. The dataset is then used for the next process.

Table 1. Contoh Dataset

Start	End	Text
0:01:28	0:01:29	- Hey. - Hey.
0:01:32	0:01:34	- Hey. - Hey.
0:01:36	0:01:37	- Hey. - Hey.
0:01:57	0:01:58	Hmm?
0:02:00	0:02:05	That's right. It is I, Tai Lung.

3.2 Combine Words into Sentences

In this phase, words in the "Text" column are combined to form sentences. Word combination occurs within 10 seconds of text entry.The purpose of sentence merging is to make the dialogue text more coherent and thus facilitate data analysis. Figure 4 shows an example of the results of sentence merging.


Sentence

- Hey. - Hey.- Hey. - Hey.- Hey. - Hey.

Hmm?That's right. It is I, Tai Lung.It can't be.Tai Lung.I have returned to take what is mine.Which is everything that is yours.Let it be known from the highest mountain to the lowest valley that Tai Lung lives,and no one will stand in his way.Not even the great Dragon Warrior.Oh, where is Po?He was supposed to be here hours ago.Ping, will you please just relax?I am relaxed!Okay. I'm sure Po is fine.But what if he's sick? What if he's hurt?What if he's hungry?Don't get your noodles in a twist.If I know our son, he's probably just kicking back and catching some rays.Kung fu!Whoa! That's bad.Whoa! Getting worse.Whoa!Ah, come on!

Ha-ha!Okay, big guy. We're really gonna have to wrap this up.

Wheel- Oink. - Oink.Again, again!- That was fun. - Let's do that again.And next time, keep your surf off their turf.I'm late. I'm late. Ooh, I'm late.- The Dragon Warrior is back! - It's the Dragon Warrior!Was, is and always will be.We love you, Dragon Warrior!And I love you too, adoring fan - Sign my scroll. - Sign my hat - Sign my shirt. - Okay, okay.I'll sign whatever you want.Po!Master Shifu!Here, let me just...There, that's much... worse.We have to talk.Absolutely. Let's talk. Right after the ceremony.- Let's go! - Ceremony?What ceremony?- The Staff of Wisdom. - Ooh.- Given to me by Master Oogway himself. - Aah.It is said that whoever possesses this staffhas the power to travel between the realms.The power to unlock the door to the Spirit Realm.And now, the power to open...the all-new Dragon Warrior Noodles and Tofu.Where the broth has kick and the bean curd's a knockout.Dragon Warrior!Will the Furious Five be here too?Unfortunately, no.They're off on super cool kung fu missions.Tigress is taking on the free-range chicken gang.- Crane!

Figure 5. Sentence merging results

3.3 Labelling

Labeling is performed for sentences from the sentence merging process. The labeling process is performed manually for each sentence line in the subtitle dialog file. The labeling of the subtitle data is divided into nine categories or classes: action, comedy, drama, fantasy, horror, mystery, romance, science fiction, and thriller. The labeling phase serves to prepare training data with subtitle dialogs that already have genre labels

3.4 Preprocessing

Preprocessing involves several processes: case conversion, sanitization, tokenization, and stop word removal. Case conversion converts all letters to lowercase. Sanitization cleans and tidies the data format. Tokenization breaks words into tokens. Stop word removal removes words that have no meaning in the stop word dictionary used.

3.5 Data Augmentation

The data augmentation techniques used include paraphrasing, back translation, label alignment, and synonym replacement. Table 2 shows the data augmentation results.

Table 2. Example of dataset

Genre	Without Augmentation	500 data Augmentation	700 Data Augmentation
Drama	1276	1276	1276
Thriller	529	529	700
Comedy	262	500	700
Horror	257	500	700
Action	236	500	700
Romance	225	500	700
Mystery	219	500	700
Sci-fi	206	500	700
Fantasy	185	500	700

3.6 Model Structure Planning

The LSTM and GRU models are based on the same parameters. The models have an identical architecture in terms of layer structure and number of units used, which allows for a fair evaluation of model performance. Table 3 shows a comparison of the parameters of the LSTM and GRU models.

Table 3. Comparison of LSTM and GRU Parametrs

Components	LSTM	GRU
Lapisan Embedding	embedding_layer (Word2Vec, trainable=False)	embedding_layer (Word2Vec, trainable=False)
RNN Layer 1	LSTM(128, return_sequences =True)	GRU(128, return_sequences= True)
RNN Layer 2	LSTM(64, return_sequences =False)	GRU(64, return_sequences= False)
Dropout 1	Dropout(0.2)	Dropout(0.2)
Dense Hidden Layer	Dense(64, activation='relu', kernel_regularizer =l2(0.001))	Dense(64, activation='relu', kernel_regularizer=l 2(0.001))
Dropout 2	Dropout(0.2)	Dropout(0.2)
Output Layer	Dense(len(genre_ labels), activation='softma x')	Dense(len(genre_la bels), activation='softmax')
Loss Function	categorical_cross entropy	categorical_crosse ntropy
Optimizer	Adam(learning_ra te=0.001)	Adam(learning_rate =0.001)
Evaluation Metric	accuracy	accuracy

3.7 Modelling

After the LSTM and GRU models are created, they are trained. The goal of training is to enable the model to learn patterns from the data in order to make predictions or make decisions automatically. Model training is performed using three scenarios: In the first scenario, data is used without data augmentation. In the second scenario, the data is augmented to 500 data points, and in the third scenario, to 700 data points. Table 4 shows the training results of the LSTM and GRU models.

Table 4. Training results of the LSTM and GRU models

Model	Scenario	Accuracy	Loss
LSTM	I	30.95%	1.6356
	II	94.18%	0.1862
	III	97.05%	0.1139
GRU	I	50.17%	0.9856
	II	96.18%	0.1110
	III	98.37%	0.0611

The performance of the LSTM and GRU models improves with data augmentation. In the first scenario without augmentation, both LSTM and GRU showed low accuracy of 30.95% and 50.17%, respectively. However, when 500 and 700 data points were added to the augmentation data in the second and third scenarios, respectively, the accuracy increased significantly. The LSTM model achieved the highest accuracy of 97.05% with a loss of 0.1139, while GRU achieved an accuracy of 98.37% with the lowest loss of 0.0611. This demonstrates that adding training data can improve the model's pattern recognition, and GRU tends to perform better than LSTM in all three training scenarios.

Table 5. Test results of the LSTM and GRU models

Model	Skenario	Testing Accuracy
LSTM	I	17%
	II	89%
	III	91%
GRU	I	21%
	II	86%
	III	92%

Table 5 shows the test accuracy of the LSTM and GRU models. The performance of both models improved with increasing training data through augmentation. In the first scenario (without augmentation), the LSTM model achieved only 17% accuracy, while GRU performed slightly better with 21% accuracy. After augmenting data by 500 (Scenario II) and 700 data sets (Scenario III), the model accuracy improved significantly. LSTM achieved 89% accuracy in Scenario II and 91% in Scenario III. At the same time, GRU performed better with 86% accuracy in Scenario II and 92% in Scenario III. These results support the findings from the training phase that data augmentation plays an important role in improving model generalization and that GRU consistently performs slightly better than LSTM in classifying genres based on subtitle data.

These results outperform several previous studies that applied traditional machine learning approaches to the same task. For example, (Rajput & Grover, 2022) reported genre classification accuracies ranging from 70% to 80% using logistic regression, support vector machines, and neural networks on subtitle data from 964 movies spanning six genres. Their best-performing model, K-Nearest Neighbors (KNN), achieved an accuracy of 77%. Similarly, (Novenrodumetasa et al., 2023) demonstrated that the Random Forest algorithm outperformed Naïve Bayes, achieving 84.1% and 68.2% accuracy, respectively.

In comparison, the deep learning models utilized in this study, GRU demonstrate superior classification performance, achieving up to 92% accuracy when supported by data augmentation strategies such as paraphrasing, back-translation, and synonym replacement. These findings

underscore the advantages of sequence-based architectures in capturing contextual and temporal patterns in subtitle dialogs, and emphasize the importance of addressing class imbalance through augmentation of underrepresented genres.

3.8 Model Evaluation

Model evaluation is performed using cross-validation to obtain more accurate and objective performance estimates, especially with limited data sets. A further model evaluation using cross-validation serves to generally measure the consistency of model performance. Cross-validation is performed using k-fold cross-validation (k=5) on 700 augmented data sets.

Table 6. Results of cross-validation

Number of folds	LSTM Accuracy	GRU Accuracy	Average Difference
1	0.69	0.64	0.05
2	0.69	0.67	0.02
3	0.71	0.72	-0.01
4	0.66	0.67	-0.01
5	0.66	0.65	0.01

Table 6 shows the test accuracy comparison result from the five-fold cross-validation between the LSTM and GRU models. It can be seen that the performance difference between the two models is relatively small. In the first case, LSTM outperforms GRU with a difference of 0.05. In the second case, the difference is even smaller, at 0.02. Interestingly, in the third and fourth cases, GRU even slightly outperforms LSTM with a negative difference of -0.01, while in the fifth case, the difference is also small at 0.01. The overall average difference between the two models is 0.012, which shows that the performance of LSTM is generally slightly higher than that of GRU, but the difference is very small and within a range that can be considered practically insignificant.

The paired t-test is performed to test whether the performance difference between LSTM and GRU models is statistically significant. The reason for using the paired t-test is that each cross-validation uses the same training and validation data for both models. Cross-validation is performed with random mixing but is controlled by random_state = 42, so that the data distribution remains consistent and the model evaluation results can be fairly compared. The paired t-test is used to compare two different conditions (in this case, LSTM vs. GRU) on the same sample. The goal is to reduce the inter-sample variation that can confound the comparison results

Table 7 shows the result of the paired t-test calculation with a p-value of 0.34185. This p-value is well above the general significance threshold of 0.05 and indicates that there is no statistically significant difference between the performance of the LSTM and GRU models in classifying movie genres based on subtitle dialogue. The t-statistic of 1.078 and the standard error of 0.011 also support this conclusion. The average difference of 0.012 and the standard deviation of the difference of 0.0249 indicate that the performance difference is very small and fluctuating. Thus, both LSTM and GRU exhibit statistically equivalent performance in the context of the dataset and experimental scenarios used.

Table 7. Results of the T-statistic calculation

P-Value	0,341850962
Average difference	0,012
Standard deviation difference	0,0248997992
Standard error	0,01113552873
T-Statistics	1,077631812

Cross-validation using 5-fold was performed for more rigorous validation, the performance of the GRU model decreased. A key limitation of this study is the significant difference between the testing accuracy achieved through the train-test split (92%) and the average accuracy obtained via 5-fold cross-validation (67%). This discrepancy may be explained by several factors. Firstly, the training and testing splits probably resulted in a testing set with a distribution closely aligned to the training data. This enabled the model to recognise similar patterns and thus attain artificially high performance. Secondly, inconsistent application of data augmentation across genres may have affected the model's generalisability, particularly for genres such as drama and thriller, which received no augmentation and consequently lacked contextual variation in their input samples. Thirdly, the high performance observed in the train-test evaluation may indicate overfitting, whereby the model becomes overly tailored to the training data and fails to generalise effectively to new instances. In contrast, cross-validation provides a more stringent and representative evaluation framework by partitioning the data into five distinct folds, offering a more robust estimation of the model's generalisation capability. Taken together, these findings suggest the need for further refinements, specifically in addressing class imbalance through more equitable augmentation, and in enhancing the representational diversity of underrepresented genres.

3.9 Classification

Genre classification with new data was performed using the GRU model trained in the third scenario with an 80:20 test split. Cross-validation results showed that the performance of LSTM and GRU was equivalent. Therefore, the

GRU model trained in the third scenario was selected for classification.

	Judul	Tahun	Genre Prediksi GRU
0	14 Days Girlfriend intlo	2025.0	[action, comedy, drama, fantasy, horror, myste...
1	1888.0	2023.0	[action, comedy, drama, fantasy, horror, myste...
2	A Brothers Bond	2024.0	[action, comedy, drama, fantasy, horror, myste...
3	A Different Man	2024.0	[action, comedy, drama, fantasy, horror, myste...
4	A Family Affair	2024.0	[comedy, drama, fantasy, horror, mystery, roma...
...	...	...	...
302	monster summer	2024.0	[comedy, drama, fantasy, horror, mystery, roma...
303	my old ass	2024.0	[action, comedy, drama, fantasy, horror, roman...
304	paris christmas waltz	2023.0	[action, comedy, drama, fantasy, horror, myste...
305	remnant	2024.0	[action, comedy, drama, fantasy, horror, roman...
306	saturday night	2024.0	[comedy, drama, fantasy, horror, mystery, roma...

307 rows × 3 columns

Figure 6. Classification results

Figure 6 shows the result of movie genre classification using the previously trained GRU model. The classification results consist of three columns: "Title," "Year," and "Genre." Each row of data represents a movie with its release year and a list of genres predicted by analyzing the movie's subtitles. The genres are displayed in a list format consisting of the categories Action, Comedy, Drama, Fantasy, Horror, Mystery, Romance, Science Fiction, and Thriller. The table contains 307 rows, meaning that 307 movies were analyzed using the GRU model. After genre classification, the percentage of genres in each movie is determined.

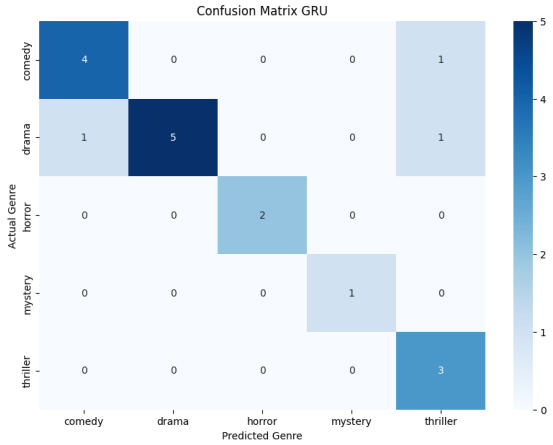


Figure 7. Confusion Matrix of Classification Results

Figure 7 presents the confusion matrix for the GRU model on the first experimental classification. As shown in the figure, the drama genre achieved the highest prediction accuracy, with 5 out of 7 instances correctly classified. This was followed by comedy, with 4 correct predictions, and thriller, with 3 correctly classified samples. The horror and mystery genres obtained 2 and 1 correct predictions, respectively. Several misclassifications occurred for example, one drama instance was misclassified as comedy, and one comedy instance was misclassified as thriller suggesting contextual overlaps and ambiguity between genres with similar dialog patterns. These errors indicate that certain genres may share linguistic features that challenge the model's ability to draw clear distinctions, especially in the absence of sufficient data diversity.

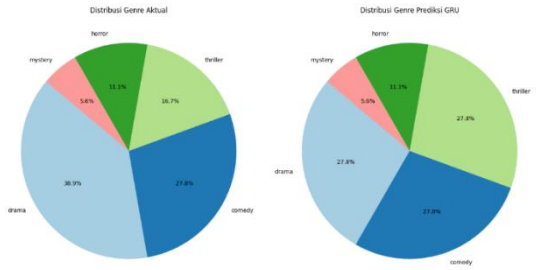


Figure 8. Comparison Of The Actual Genre Distribution With The Predicted Genre Results

Figure 8 compares the actual genre distribution with the genres predicted by the GRU model. The actual genre distribution shows that drama dominates with a share of 38.9%, followed by comedy (27.8%) and thriller (16.7%), while horror and mystery have smaller shares than other genres. The genre distribution predicted by the GRU model appears more balanced, with drama, comedy, and thriller each accounting for 27.8%. This shows that the GRU model tends to provide a more even distribution of predictions between genres, which could indicate that the model has difficulty detecting more dominant genres.

3.10 Comparison Table of Top 3 Actual Genres and Predicted Genres

To further evaluate the performance of the GRU model in genre classification, a comparison was made between the top three predicted genres and the actual genres of two selected films: All Through the Hall and F Marry Kill. This comparison helps determine how well the model can capture the dominant thematic elements of a film based solely on its subtitle dialogue. The table below presents the results of this comparison.

Table 8. Top 3 Actual Genres and Predicted Genres

Film Title	Actual Genres	Top 3 Predicted Genres By GRU
All Through the Hall	Horror, Thriller, Mystery	1. Drama (36.7%) 2. Horror (30.0%) 3. Thriller (23.3%)
F Marry Kill	Comedy, Drama, Romance	1. Comedy (30.3%) 2. Drama (21.2%) 3. Romance(12.1%)

Table 8 comparison of top 3 actual and predicted genre of All Through the Hall and F Marry Kill Movie. For All Through the Hall, the GRU model predicted drama, horror, and thriller

as the top three genres, with drama taking the largest proportion. The actual genres of the film are horror, thriller, and mystery. This shows that the model was able to correctly identify the suspenseful and intense elements of the film through the horror and thriller labels. The presence of drama in the prediction, although not in the official genre list, may indicate the model's sensitivity to emotional or narrative depth reflected in the dialogue.

In the case of F Marry Kill, the model accurately predicted comedy as the dominant genre, which aligns with the actual genre classification. Drama also appeared in the prediction with a significant proportion, likely due to the emotional or interpersonal elements within the film. The third rank is shared by thriller, romance, and horror, each with equal weight (12.1%). This suggests the GRU model is capable of detecting a variety of genre signals from the dialogue, even when they are subtle or overlapping.

Overall, the comparison demonstrates that the GRU model can effectively capture key genre characteristics based on subtitle text, and in many cases, its top predictions align well with the actual genre classifications.

4. Conclusion

This study evaluated the performance of LSTM and GRU models in classifying movie genres using subtitle dialogs. Both models achieved comparable results, with GRU slightly outperforming LSTM in terms of accuracy, although the difference was not statistically significant. A key contribution of this research is the application of data augmentation technique there are paraphrasing, back-translation, and synonym replacement to address class imbalance, which improved the robustness of the models when combined with 5-fold cross-validation.

These findings confirm that LSTM and GRU are suitable for subtitle-based genre classification tasks. Future work may explore multi-label classification, optimization of augmentation strategies, and the use of more advanced models such as transformers to further enhance performance and applicability in real-world systems.

Reference

Akbar, J., Fahmi, H., & Murniati, W. (2025). Multi Label Klasifikasi Genre Film Berdasarkan Sinopsis Menggunakan Metode Long Short-Term Memory (LSTM). *Jurnal Manajemen Informatika & Sistem Informasi (MISI)*, *8*(1). https://doi.org/10.36595/misi.v5i2

Alzoubi, Y. I., Topcu, A. E., Elbasi, E., Buyukyilmaz, M., & Cibikdiken, A. O. (2024). Anticipate Movie Theme from Subtitle: A Deep Learning Approach. *2024 47th International Conference on Telecommunications and Signal Processing, TSP 2024*, 205–210. https://doi.org/10.1109/TSP63128.2024.10605925

Azka, F., Hilyah, N., Hufad, A., Aziz, F., & Kunci, K. (2024). Konten Publikasi Film: Impresi Remaja terhadap Film Indonesia. *Jurnal Gunahumas*, *7*(1), 1–16. https://doi.org/10.17509/ghm.v7i1

Beddiar, D. R., Jahan, M. S., & Oussalah, M. (2021). *Data Expansion using Back Translation and Paraphrasing for Hate Speech Detection*. http://arxiv.org/abs/2106.04681

Bibi, I., Akhunzada, A., Malik, J., Iqbal, J., Mussaddiq, A., & Kim, S. (2020). A Dynamic DL-Driven Architecture to Combat Sophisticated Android Malware. *IEEE Access*, *8*, 129600–129612. https://doi.org/10.1109/ACCESS.2020.3009819

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Association for Computational Linguistics*, 1724–1734. https://doi.org/https://doi.org/10.3115/v1/D14-1179

Du, Y., Lavarec, E., & Lalouqette, C. (2023). Text Data Augmentation to Manage Imbalanced Classification: Apply to BERT-based Large Multiclass Classification for Product Sheets. *International Journal of Computational Linguistics (IJCL)*, *14*, 2023–2024. https://www.cscjournals.org/journals/IJCL/description.php

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/https://doi.org/10.1162/neco.1997.9.8.1735

Ibrahim, M. A., Faisal, Sulistiya, Z. D., & Winarto, T. S. Y. (2024). Prompt-Based Data Augmentation with Large Language Models for Indonesian Gender-Based Hate Speech Detection. *Journal of Computer Science*, *20*(8), 819–826. https://doi.org/10.3844/jcssp.2024.819.826

Kusumo, S., & Somya, R. (2022). Penerapan Web Scraping Deskripsi Produk Menggunakan Selenium Python Dan Framework Laravel. *Jurnal Teknik Informatika Dan Sistem Informasi*, *9*(4). https://doi.org/http://dx.doi.org/10.35957/jatisi.v9i4.2727

Mangolin, R. B., Pereira, R. M., Britto, A. S., Silla, C. N., Feltrim, V. D., Bertolini, D., & Costa, Y. M. G. (2022). A multimodal approach for multi-label movie genre classification. *Multimedia Tools and Applications*, *81*(14), 19071–19096. https://doi.org/10.1007/s11042-020-10086-2

Nifanto, S., & Nurhopipah, A. (2024). Balancing Dataset Untuk Klasifikasi Komentar Program Kampus Merdeka Menggunakan Synonym Replacement. *Jurnal Ilmu Komputer*, *17*, 55–64. https://doi.org/http://dx.doi.org/10.24843/JIK.2024.v17.i01.p02

Novenrodumetasa, N., Suarjaya, I. M. A. D., & Raharja, I. M. S. (2023). Analisis Genre Film Berdasarkan Data Subtitle. *JITTER : Jurnal Ilmiah Teknologi Dan Komputer*, *4*(2), 1912. https://doi.org/10.24843/JTRTI.2023.v04.i02.p23

Pamungkas, F. S., Prasetya, B. D., & Kharisudin, I. (2019). Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers Menggunakan Python. *PRISMA, Prosiding Seminar Nasional Matematika*, *3*, 689–694. https://journal.unnes.ac.id/sju/index.php/prisma/

Purnomo, I. I., & Syafarina, G. A. (2024). Analisis Prediktif Dan Preprocessing Untuk Kualitas Buah Apel Pendekatan Machine Learning. *Technologia : Jurnal Ilmiah*, *15*(4), 681. https://doi.org/10.31602/tji.v15i4.15945

Rajput, N. K., & Grover, B. A. (2022). A multi-label movie genre classification scheme based on the movie's subtitles. *Multimedia Tools and Applications*, *81*(22), 32469–32490. https://doi.org/10.1007/s11042-022-12961-6

Salam, R. R., Jamil, M. F., Ibrahim, Y., Rahmaddeni, R., Soni, S., & Herianto, H. (2023). Analisis Sentimen Terhadap Bantuan Langsung Tunai (BLT) Bahan Bakar Minyak (BBM) Menggunakan Support Vector Machine: Sentiment Analysis of Cash Direct Assistance Distribution for Fuel Oil Using Support Vector Machine. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(1), 27-35. https://doi.org/10.57152/malcom.v3i1.590

Sari, W., Rini, D., & Malik, R. (2019). Text Classification Using Long Short-Term Memory. *International Conference on Electrical Engineering and Computer Science (ICECOS)*, 150–155. https://doi.org/https://doi.org/10.1109/ICECOS47637.2019.8984558

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. In *SN Computer Science* (Vol. 2, Issue 3). Springer. https://doi.org/10.1007/s42979-021-00592-x

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text Data Augmentation for Deep Learning. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00492-0

Wijaya, N. N., Setiadi, D. R. I. M., & Muslikh, A. R. (2024). Music-Genre Classification using Bidirectional Long Short-Term Memory and Mel-Frequency Cepstral Coefficients. *Journal of Computing Theories and Applications*, *1*(3), 243–256. https://doi.org/10.62411/jcta.9655

Wijiyanto, W., Pradana, A. I., Sopingi, S., & Atina, V. (2024). Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa. *Jurnal Algoritma*, *21*(1). https://doi.org/10.33364/algoritma/v.21-1.1618