Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.4, August 2025, pp. 1578~1586
e-ISSN: 2548-6861

Few-Shot Learning for Classifying Genuine and Bot Comments on YouTube Using Transformer Models

Nahdah Fikriah Nst [1]*, Defry Hamdhana [2], Mukti Qamal [3]***
* Department of Informatics, Universitas Malikussaleh, Lhokseumawe, Indonesia
* Department of Information Technology, Universitas Malikussaleh, Lhokseumawe, Indonesia
nahdah.210170244@unimal.ac.id [1], defryhamdhana@unimal.ac.id [2], mukti.qamal@unimal.ac.id [3]

Article Info

Article history:

Received 2025-06-30
Revised 2025-07-03
Accepted 2025-07-09

Keyword:

Few-Shot Learning (FSL),
Transformers,
DistilBERT,
Comment Bot,
Text Classification,
YouTube,
Natural Language Processing, Web Applications.

ABSTRACT

This study aims to develop a comment classification system on the YouTube platform to distinguish between real accounts and bot accounts, addressing the challenge of limited labeled data through a few-shot learning approach. The issue of bot accounts masquerading as real users in comment sections is becoming increasingly prevalent and has the potential to spread spam, misinformation, and influence public opinion. In this study, a Transformer-based model, DistilBERT, is used, which is known for its efficiency in understanding natural language context. The model is trained in a few-shot scenario (N5 to N50) using a very limited amount of training data. Testing results show that the model maintains high and stable performance even with minimal data (N5), achieving an F1-score above 0.90. In addition, this system is implemented into a web application using Flask to enable direct and interactive comment detection. The main contribution of this research is the proof that the combination of few-shot learning and the DistilBERT model can provide a practical and efficient solution for classifying YouTube bot account comments even with limited data conditions, as well as providing a replicable approach for similar problems on other digital platforms.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

In today's digital age, social media has become the primary means for people to interact, exchange opinions, and share information openly. One of the most popular and widely used platforms in Indonesia is YouTube, which allows users to upload comments, express opinions, and interact directly with creators and other users. With millions of comments uploaded every day, YouTube's comment section has evolved into a dense and dynamic digital public space [1]

However, the openness of this interactive space also presents significant challenges, particularly regarding the misuse of the comment section by bot accounts. Bots are automated programs that can generate large amounts of content, often inserting spam, misleading links, or manipulative messages that disrupt the user experience and undermine the platform's credibility [2], [3]. Such bots contribute to the spread of misinformation and manipulation of public opinion, which is increasingly difficult to detect as their language patterns increasingly resemble those of human users [4].

More than just a medium of communication, YouTube comments also reflect the sentiments, behavioral tendencies, and opinion dynamics of its users. In this context, these comments can be considered unstructured text that contains important insights for analyzing social dynamics in the digital space [5]. The large and continuously flowing amount of data opens up opportunities for researchers to conduct in-depth analysis using computational linguistics techniques.

To effectively process such data, researchers utilize Machine Learning (ML) approaches, which are a branch of Artificial Intelligence (AI) and function to optimize system performance through the analysis of sample data or historical data [6]. According to [7], ML plays a central role in recognizing patterns from digital text, particularly in Natural Language Processing (NLP) tasks such as text classification, sentiment analysis, and spam filtering. This capability makes ML highly relevant for analyzing informal and varied

language structures, such as those found in YouTube comments.

However, most ML models still heavily rely on large amounts of labeled training data to function effectively. This poses a challenge in real-world applications, where labeled data is often limited, imbalanced, or costly to obtain. On the other hand, Few-Shot Learning (FSL) has shown promising results in low-resource classification tasks, such as rumor detection [8] and COVID-19-related tweet analysis [9].

Previous studies have discussed bot detection on platforms like Twitter and Facebook using traditional learning methods [10], [11]. However, research specifically applying few-shot learning approaches to classify comments on the YouTube platform remains limited. Despite this, YouTube is one of the platforms with a vast and complex comment ecosystem, requiring more adaptive and efficient classification approaches. This gap is the primary focus of this research.

To address this gap, this study explicitly aims to develop a YouTube comment classification system that can distinguish between comments from genuine accounts and bot accounts, using a Transformer-based few-shot learning approach, specifically DistilBERT. The model will be tested in several limited data scenarios (N5 to N50) and will ultimately be implemented into an interactive web application that can be used directly. This research is expected to contribute to the development of more efficient, lightweight, and adaptive NLP solutions for limited data conditions.

II. METHODS

This article discusses the theories underlying research into the classification of genuine and bot comments on youtube. Each section presents important concepts such as comment types, account characteristics, classification, and the role of YouTube. It also reviews methods such as data mining, machine learning, and few-shot learning (FSL), as well as the transformers architecture with key components such as self-attention and multi-head attention. Finally, it describes the use of web scraping, flask, and model evaluation to measure the performance of the classification system.

The purpose of this research is to create a method for classifying YouTube comments that can distinguish between comments from bot accounts and real accounts. The methodology of this research is few-shot learning with a transformer-based DistilBERT model. Starting with the process of collecting comment data through the API, followed by manual labeling, pre-processing, training the model with limited data, evaluating model performance, and implementing the system in the form of a web application.

The few shot learning method is trained using a model with a relatively small amount of data per class, ranging from 5 (N5) to 50 (N50), in accordance with the approach used in this research. thus resulting in the ability of the model to perform categorization using a small amount of data sources. Based on the tests conducted, it was found that the model maintained stability even when the amount of data was increased (from N15 to N50) to the extent that it was able to achieve good performance even under training conditions with very minimal data (N5). In addition, a web-based application featuring an interactive and responsive interface has integrated the system well.

System Workflow The classification process in the Flask-based system follows a sequential flow:

- The user inputs a YouTube comment via a text box or uploads a .csv file.
- The comment(s) are sent to the Flask back-end, where the text is preprocessed using the same normalization steps used during training (such as lowercasing, punctuation cleaning, and emoji handling).
- The cleaned text is tokenized using DistilBertTokenizerFast and converted into tensors compatible with the model input.
- The DistilBERTForSequenceClassification model generates a prediction, outputting logits that are interpreted as labels: "Genuine Account" or "Bot Account."
- The prediction is returned to the front-end interface and displayed in real time.

This system architecture enables fast and efficient inference even on standard CPUs and can be seamlessly integrated into real-time comment moderation or sentiment analysis workflows.

Through the research flow scheme in the figure above, Kaggel and Google Colab are used in this research as data preprocessing and model implementation on few-shot learning for the categorization of real account comments and YouTube bot accounts using the Trasnformers model.

A. Few-shot learning

Model training in this study was conducted using the Few-Shot Learning (FSL) approach, which is a training method designed to allow models to learn from a very limited amount of data. This paradigm belongs to the realm of meta-learning, where models are trained to be able to recognize patterns from new data despite having only a few examples per class [12]. FSL is especially important in cases such as the classification of bot and genuine account comments on YouTube, which often do not have a large amount of balanced or abundant data.

In this study, the few-shot scenario is divided into several levels of the amount of data per class, namely N5, N10, N15, and N50. This means that the model is only trained with 5, 10, 15, and 50 data samples per label (real or bot) respectively. The goal is to evaluate how well the model can recognize patterns with extreme data limitations. This reflects real-world situations where manual annotation data is often limited and expensive to collect.

Each training scenario was repeated with consistent parameters so that results between scenarios could be fairly compared. The training parameters used were 5 epochs, batch size 8, and the default learning rate of the AdamW optimizer. By keeping the parameters fixed, the evaluation of the model's

performance focuses more on the amount of data used, rather than on changes in the hyperparameters.

Each training scenario was repeated with consistent parameters so that the results between scenarios could be compared fairly. The training parameters used were 5 epochs, batch size 8, and the default learning rate of the AdamW optimizer. By keeping the parameters fixed, the evaluation of model performance focuses more on the amount of data used, rather than on changes in hyperparameters.

This few-shot concept utilizes transfer learning, which is a strategy of using pre-trained models that have previously absorbed knowledge from large datasets, then adapted to specific tasks using a small amount of new data [6]. In this study, the DistilBERT model pre-trained on the Indonesian language corpus was reused for the comment classification task, which had a limited amount of data.

Through this approach, the research attempted to measure the efficiency and adaptability of the model in real data-poor conditions. The results of various few-shot scenarios provide an idea of how far the model can generalize prior knowledge to new data that is only available in small amounts. Evaluations were conducted on each scenario to see performance trends and identify the optimal point between the amount of data and the classification results obtained

B. Data Collection

The dataset used in this study was collected using the YouTube Data API v3. Comments were retrieved from a public YouTube video published by the KompasTV channel titled “Roy Suryo CS Tetap Tenang! Penyelidikan Polisi soal Ijazah Palsu Jokowi Sudah 90 Persen”. Link di bawah: (https://www.youtube.com/watch?v=A0THLldlxgs&t=38s)
A total of 300 comments were successfully extracted using a custom-built Python script and stored in .csv format for further processing.

The labeling process was conducted manually by the researcher, based on reference criteria from related studies [2], [4], [13], [14], [15]. Label assignment relied on analyzing textual patterns such as:

- Repetitive sentence structures
- Presence of spam, promotional links, or irrelevant hashtags
- Unnatural or robotic language
- Use of emojis or personal tone

These criteria were used to distinguish between comments from genuine users and those likely generated by bot accounts. No account-level metadata (e.g., subscriber count, channel age) was used, as such data was not available through the API. All annotations were carried out independently to maintain consistency. Out of the 300 collected comments:

- 162 comments (54%) were labeled as from genuine users
- 138 comments (46%) were labeled as from bot accounts

This balanced distribution provides a solid foundation for few-shot learning experiments. The labeled dataset was then used for training the DistilBERT model under various few-shot scenarios (N5, N10, N15, N50), and for performance evaluation using metrics such as accuracy, precision, recall, and F1-score.

TABLE I
DATA SET

No	Publishe dat	Authordis playname	Textdisplay	Like Count
1.	2025-05-10T07:29:01Z	@Anindita Uswatun	Main bisa dari HP atau PC di **PLUTO88**	0
2.	2025-05-10T07:28:01Z	@PoerWanto-r8n	Penguasa yg punya hukum 😂😂😂😂😂 Menguak kebenaran itu adalah perjuangan tapi ingatlah siapa yg kita lawan monopoli baru di mainkan	0
3.	2025-05-10T07:21:36Z	@WahyuningsihElif	**PLUTO88** situs slot paling rame dan terpercaya	0
4.	2025-05-10T07:12:47Z	@tokzio7717	Agak lucu, katanya sdh 90% namun ijazah baru akan ditunjukkan.. lalu yg 90% pengujian trhdp objek apa?????	0
5.	2025-05-08T13:19:13Z	@radityaarimbawa2618	Logika aja, kalaupun itu palsu gak mungkin juga ada yang berani bilang itu palsu karna ini menyangkut nama baik indonesia 😊	63
6.	2025-05-10T07:02:58Z	@makjangdj4634	Tinggal nanya UGM, kl sampai sudah dijawab asli masih bilang palsu berarti semua ijazah yang dikeluarkan UGM palsu menurut Roy Suryo. UGM juga bisa menuntut ini, termasuk alumni lainnya	0
7.	2025-05-10T06:55:28Z	@muyi8057	Selama ini banyak polisi sujud kepada Mulyono. bisa jadi uji forensik polri tidak netral	0
8.	2025-05-10T06:52:17Z	@nursholeh758	Haha.. polisi lagi yg cek.. coba dari FBI Susah nyogok nya klo pake dolar 😂😂😂	0
9.	2025-05-10T06:49:54Z	@AmeliaSyafitri-n6n	Slot dengan RTP tinggi cuma di **PLUTO88**	0
10.	2025-05-08T13:06:16Z	@mudjiati1458	Setelah ini Roy Suryo juga harus menunjukkan keaslian ijasahnya.kami bangsa Indonesia juga pingin tahu	17

C. Data Labeling

TABLE II
LABELING DATA

No	Publishedat	Authordisplayname	Textdisplay	Like Count	Label
1.	2025-05-10T07:29:01Z	@AnindaUswatun	Main bisa dari HP atau PC di **PLUTO88**	0	0
2.	2025-05-10T07:28:01Z	@PoerWanto-r8n	Penguasa yg punya hukum 😂😂😂😂😂 Menguak kebenaran itu adalah perjuangan tapi ingatlah siapa yg kita lawan monopoli baru di mainkan	0	1
3.	2025-05-10T07:21:36Z	@WahyuningsihElif	**PLUTO88** situs slot paling rame dan terpercaya	0	0
4.	2025-05-10T07:12:47Z	@tokzio7717	Agak lucu, katanya sdh 90% namun ijazah baru akan ditunjukkan.. lalu yg 90% pengujian trhdp objek apa?????	0	0
5.	2025-05-08T13:19:13Z	@radityaarimbawa2618	Logika aja, kalaupun itu palsu gak mungkin juga ada yang berani bilang itu palsu karna ini menyangkut nama baik indonesia 😅	63	1
6.	2025-05-10T07:02:58Z	@makjangdj4634	Tinggal nanya UGM, kl sampai sudah dijawab asli masih bilang palsu berarti semua ijazah yang dikeluarkan UGM palsu menurut Roy Suryo. UGM juga bisa menuntut ini, termasuk alumni lainnya	0	1
7.	2025-05-10T06:55:28Z	@muyi8057	Selama ini banyak polisi sujud kepada Mulyono. bisa jadi uji forensik polri tidak netral	0	1
8.	2025-05-10T06:52:17Z	@nursholeh758	Haha.. polisi lagi yg cek.. coba dari FBI Susah nyogok nya klo pake dolar 😂😂😂	0	0
9.	2025-05-10T06:49:54Z	@AmeliaSyafitri-n6n	Slot dengan RTP tinggi cuma di **PLUTO88**	0	0
10.	2025-05-08T13:06:16Z	@mudjiati1458	Setelah ini Roy Suryo juga harus menunjukkan keaslian ijasahnya.kami bangsa Indonesia juga pingin tahu	17	1

D. Preprocesing Data

Before YouTube comments are used in model training, a pre-processing step is performed to clean the data. Raw comments usually contain nonstandard words, slang, random symbols, and repeated letters such as “baguuuusss”. If left unchecked, this can confuse the model in understanding the meaning of the text. Therefore, normalization is performed to simplify word variations to standard forms such as “bangettt” to “really”.

After normalization, the text is broken into small pieces or tokens (tokenization). This process makes modeling easier as each piece can be analyzed individually. In this study, emojis were retained because they are considered an important part of user expressions that can reinforce the context of emotions, such as 😂 for laughter or 😢 for sadness [16].

The next step is to remove stopwords, which are common words such as “which”, ‘in’, or “and” that do not provide important meaning. After that, stemming is performed to convert words to their basic form, so that the model can recognize words such as “menonton”, “ditonton”, and “penonton” as one basic meaning: “tonton”.

After all stages are completed, the processed text will be cleaner and ready to be analyzed. For example, a comment such as “Wahh keren banget videonya 😂🔥” will become “keren banget video 😂🔥”, with emojis retained to keep the nuances of expression.

Overall, this process not only cleans the data, but also preserves the emotional elements that are important in digital communication. This is believed to improve the model's understanding of social context and sentiment in comments [17].

E. Model Training

The model used in this research is DistilBERTForSequenceClassification, a lightweight variant of BERT developed by Hugging Face. DistilBERT has a more efficient architecture while still maintaining competitive performance in understanding natural language context. The selection of DistilBERT itself is based on the consideration of computational efficiency and better training speed compared to standard BERT.

The training process begins with the tokenization stage, which converts the pre-processed text into tokens that match the input format of the DistilBERT model using DistilBertTokenizerFast. Next, the data is divided into two parts with a ratio of 80:20, where 80% is used as training data and the remaining 20% as testing data.

The model was trained using the Trainer API of the Hugging Face library which makes it easy to manage the training process, including automatically calculating the evaluation metrics. The optimizer used was AdamW, which is a version of the Adam optimizer customized for model-transformer training.

Training was conducted in various few-shot learning scenarios with varying amounts of data per class, namely N5, N10, N15, and N50. In this way, the performance of the model is tested based on how well it can learn from very limited data. An N-shot, two-way classification configuration was used to train the model, where N is the number of training instances in each class. Four scenarios (N5, N10, N15, and N50) and two classes (genuine and bot accounts) were used to simulate limited-data conditions. The training parameters used are as follows:

- Epoch: 5
- Batch size: 8
- Learning rate: default from AdamW

The purpose of this experiment is to evaluate the performance of the model in a few-shot learning scenario, which is training the model with a minimal amount of data to test its efficiency in learning and recognizing patterns from YouTube comments.

F. Feature Extraction

The feature extraction process is carried out by utilizing DistilBERT's built-in tokenizer, DistilBertTokenizerFast, which is tasked with converting comment text into numerical tokens so that it can be understood by the model. This tokenizer not only breaks the sentence into small parts (tokens), but also adds important elements such as [CLS], [SEP], and attention masks that help the model understand the overall sentence structure.

This process is like "translating" human language into machine language. An example sentence like "This comment seems to be from a fake account" will be converted into a series of numbers that represent the words. These numbers are then learned by the model to detect patterns and meanings behind the comments.

DistilBERT itself is a transformer-based classification model that is designed to be more concise. It can still capture the meaning and context of sentences quite well, but it is much lighter and faster-characteristics that are very supportive of web-based systems that require high performance.

DistilBERT is trained using the knowledge distillation method, where large models like BERT act as teachers that transfer knowledge to smaller versions. The structure of DistilBERT is made simpler with fewer layers, but still effective in understanding sentence context [18].

G. Transformers Model Training

Model training in this study uses a Transformers-based approach with the help of the Trainer API from the Hugging Face library. This API simplifies the training process, from data partitioning to the evaluation of metrics such as accuracy, precision, and F1-score. The model is optimized using AdamW, an optimizer specifically designed for transformer models to be more stable and efficient in weight adjustment.

Three important mechanisms in the encoder are self-attention, bidirectional attention, and multi-head attention. Self-attention allows each token to consider all other tokens in a sentence [19]. Bidirectional attention enables the model to view context from two directions, left and right [20]. Multi-head attention processes multiple patterns in parallel and combines them for richer results [21].

The combination of these three mechanisms enables the model to understand text deeply. This is crucial in distinguishing between genuine comments and bots, which often have subtle differences in structure and language style. With full-context-based training, the model becomes more accurate in classifying comments appropriately.

H. Model Evaluasi

The evaluation is based on the commonly used classification metrics of accuracy, precision, recall, and f1-score. The main focus of the evaluation was on two scenarios: N5 and N15. The evaluation data was obtained from testing the model on test data that was not included in the training data.

N5 evaluation results:

TABLE III
EVALUATION RESULT N5

	Precision	**Recall**	**F1-score**
Kelas 0 (Bot)	0.89	0.80	0.84
Kelas 1 (Asli)	0.89	0.94	0.92
Accuracy			0.89

Manual calculation based on confusion matrix N5:

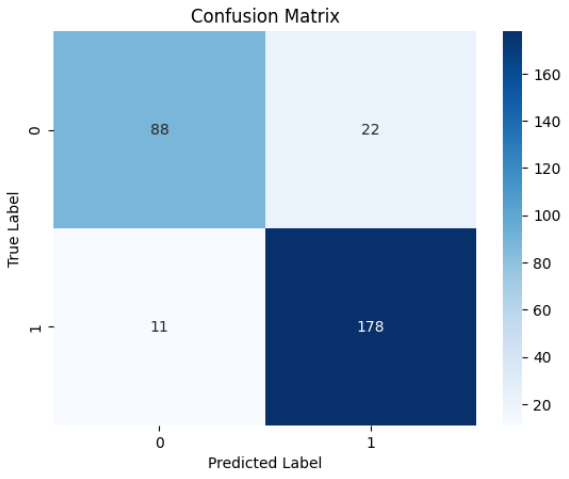

Figure 1 Confusion Matrix N5

True Positive (TP): 178 ,True Negative (TN): 88,
False Positive (FP): 22, False Negative (FN): 11.

$$\text{Accuracy} = \frac{TP + TN}{Total} = \frac{178 + 78}{299} = 0{,}89$$

$$\textit{Precision} = \frac{TP}{TP + FP} = \frac{178}{178 + 22} = 0{,}89$$

$$\textit{Recall} = \frac{TP}{TP + FN} = \frac{178}{178 + 11} = 0{,}92$$

$$\textit{F1 Score} = 2 \times \frac{(Precision \times Recall)}{Precision + Recall}$$

$$= 2 \times \frac{(Precision \times Recall)}{Precision + Recall} = 0{,}92$$

N15 evaluation results:

TABLE IV
EVALUATION RESULT N15

	Precision	Recall	F1-score
Kelas 0 (Bot)	0.86	0.83	0.84
Kelas 1 (Asli)	0.90	0.92	0.91
Accuracy			0.89

Manual calculation based on confusion matrix N15:

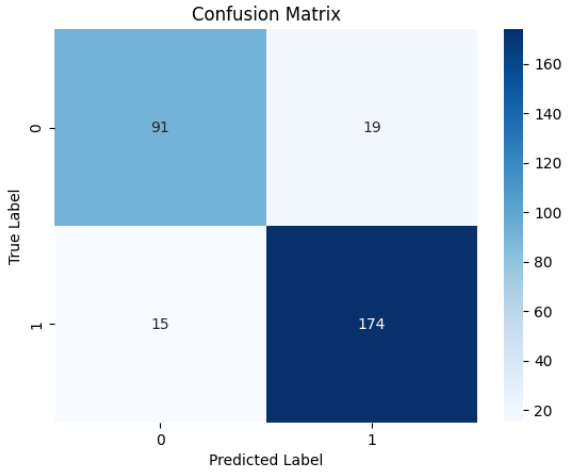

Figure 2 Confulation Matrix N15

True Positive (TP): 174 ,True Negative (TN): 91,
False Positive (FP): 19, False Negative (FN): 15.

$$\text{Accuracy} = \frac{TP + TN}{Total} = \frac{174 + 91}{299} = 0{,}89$$

$$\textit{Precision} = \frac{TP}{TP + FP} = \frac{174}{174 + 19} = 0{,}90$$

$$\textit{Recall} = \frac{TP}{TP + FN} = \frac{174}{174 + 15} = 0{,}92$$

$$\textit{F1 Score} = 2 \times \frac{(Precision \times Recall)}{Precision + Recall}$$

$$= 2 \times \frac{(0.90 \times 0{,}92)}{0{,}90 + 0{,}92} = 0{,}91$$

Since each scenario was executed only once, statistical measures such as standard deviation or confidence intervals could not be calculated. This is acknowledged as a limitation, and future studies are encouraged to include repeated runs to capture performance

I. Visualisasi Hasil Evaluasi

Because of the small dataset, each N-shot experiment was conducted just once, and results are presented using a single train-test split. For a more thorough performance review, future research might involve averaging outcomes across several episodes.

The evaluation results for various scenarios (N5-N50) are displayed in a bar chart. Each chart shows the scores for accuracy, precision, recall, and F1-score.
In general, the graphs show that:

- The highest F1-score is obtained in scenarios N5 and N15
- The model remains consistent and stable up to N50
- Accuracy tends to be above 0.87 in all scenarios

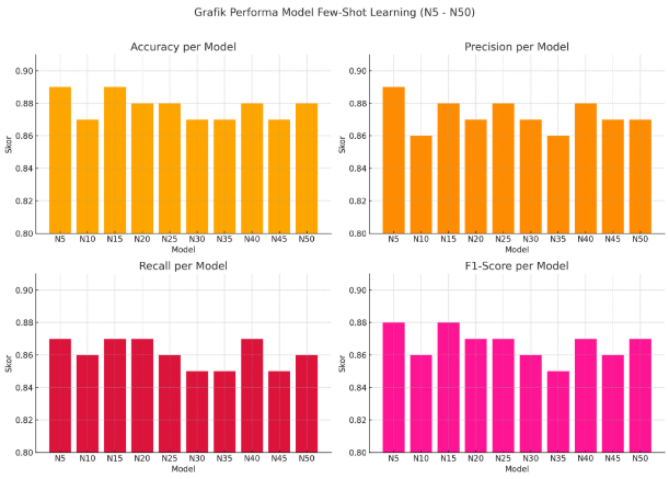

Figure 3 Shows a graph of the model's performance on each data size.

III. RESULT AND DISCUSSION

A web application built with Flask is then used to implement the trained model. Through the application of the trained model, the app performs classification after the user manually enters comments or imports them from a .csv file. The main features of the app include comment input, automatic classification of comments as genuine accounts or bots, as well as informative display of results. Comments can

be uploaded by the user manually or through a .csv file. Clear labels are used to display classification results, while graphs or summary tables are used to present performance metrics including accuracy, precision, and f1 score.

Through several main pages that support the overall functionality of the system, the user interface of the web application is kept simple and easy to use.

1. Homepage View

The Home page, which is the first page, provides a brief overview of the application and its features.
description: if you click try now it will go to the comment detection section

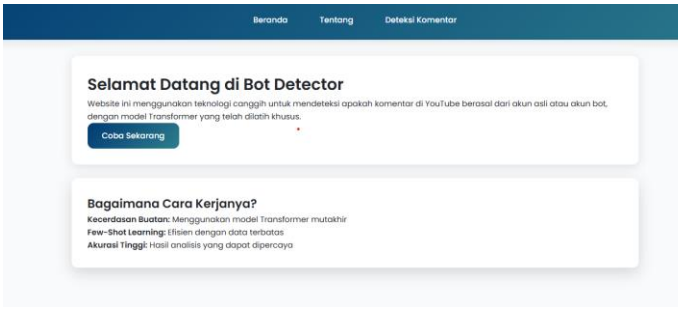

Figure 4 Homepage View

2. About view

Web the About page provides an overview explaining the background, objectives, and what methods are used in the system development process.

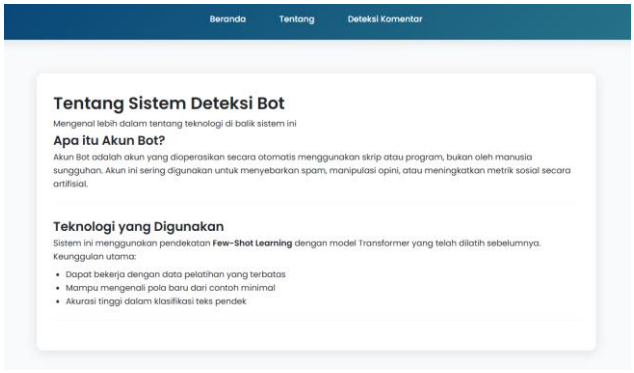

Figure 5 About View

3. Detection View

The Comment Detection Web page, where users can upload comments and start the classification process, is the main feature. After that, the list of comments that have been categorized based on their classification labels will display on two different pages, namely the results derived from the original comment classification and the bot comment classification results, so that each can present a list of comments according to their classification labels.
description:

1. After entering the next youtube video title
2. Entering the new comment, click analyze now and the results will come out in the comment analysis results web view section.

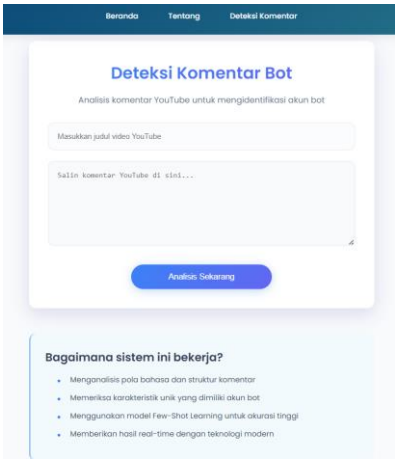

Figure 6 Detection View

4. Original Account Analysis Result

This page displays the Comment Analysis Results that have been processed by the transformers model in order to get the results of comments that have been entered into "bot comment detection" and analyze the results are "Original".

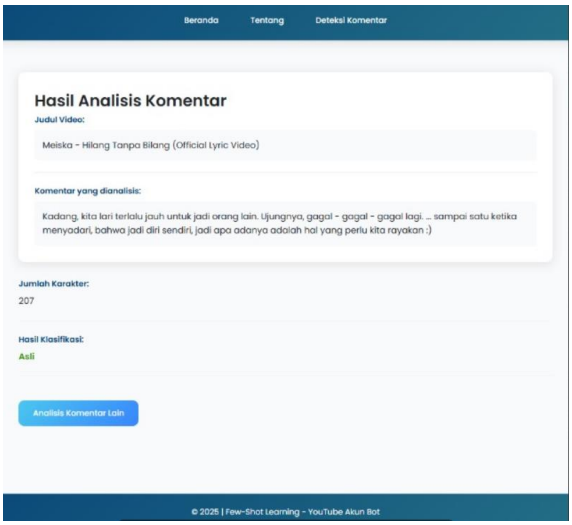

Figure 7 Original Account Analysis Result

5. Bot Account Analysis Result

This page displays the Comment Analysis Results that have been processed by the transformers model in order to get the results of comments that have been entered into "bot comment detection" and analyze the results are "bots".

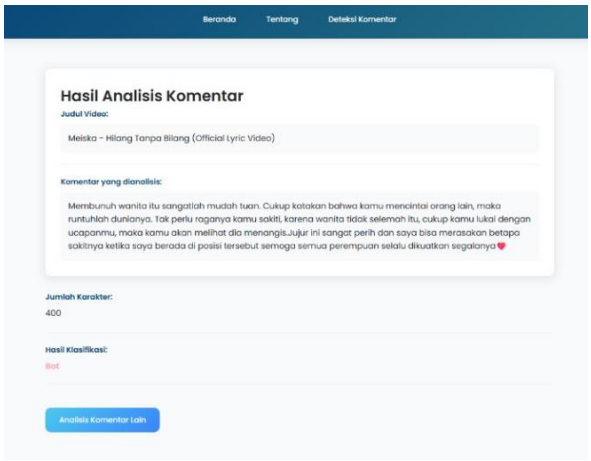

Figure 8 Bot Account Analysis Result

6. Discussion

This section discusses the two main problem formulations addressed in this research. It first focuses on the classification of YouTube account comments using the few-shot learning technique. Five to fifty samples per class (N5 to N50) was the relatively tiny quantity of data used in this study to train the model. Despite these drawbacks, the DistilBERT model was able to efficiently classify comments by learning patterns from the textual input in conjunction with methodical pre-processing processes. With continuously excellent accuracy and F1-scores, the experimental results validated the viability of using few-shot learning for short-text categorization in a low-resource environment.

Second, the transformer-based approach proved to have powerful performance capabilities. Especially in the N5 and N15 situations, DistilBERT shown a high degree of reliability in differentiating between comments from real people and those from bot accounts. This suggests that text classification problems requiring complicated, noisy, informal input with few labels are a good fit for the model architecture and the selected training approach.

But it's crucial to remember that the manual categorization process might include subjective biases, particularly when dealing with remarks that are unclear or deceptive. For example, human-written communications that were repetitive or snarky might have been mistakenly identified as bot-generated. These labeling discrepancies may have an impact on the model's learning and generalization to new data. Incorporating active learning techniques or inter-annotator agreement could enhance label reliability in future research.

Finally, a Flask-based web application successfully included the trained model. This illustrates the study's usefulness by enabling users to categorize YouTube comments in real time. As a result, the study provides a functional prototype that can support social media analysis and content moderation in addition to an academic viewpoint on effective few-shot learning for NLP.

Here I don't compare few-shot learning with other methods but I use few-shot learning because with a little data to train the DistilBERT model can understand it.

IV. CONCLUSION

This study demonstrates that the *Few-shot learning* approach based on Transformer models, specifically DistilBERT, can be effectively utilized to classify comments from genuine and bot accounts on YouTube, even with a very limited amount of labeled training data. This finding highlights the potential of Natural Language Processing (NLP) techniques to remain accurate and efficient in low-resource scenarios, which has long been a challenge in AI-based language understanding systems.The main conclusions of this study are as follows:

1. DistilBERT, applied under *Few-shot* scenarios (N5 to N50), showed optimal performance particularly at N5 and N15, achieving F1-scores above 0.90. This indicates the model's strong capability to learn and generalize from a minimal number of training examples.
2. The implementation process involved comprehensive stages, including text preprocessing, model training, performance evaluation, and deployment into an interactive web application using Flask.
3. The model effectively distinguished between genuine and bot-generated comments, making it a promising tool for supporting comment moderation tasks on digital platforms.
4. The developed web application successfully integrated the classification model into a user-friendly system, enabling real-time and interactive bot detection, and demonstrating its potential for real-world deployment.
5. From a scientific perspective, this research contributes to the advancement of adaptive and lightweight NLP methods, proving that pre-trained models like DistilBERT can be fine-tuned using *Few-shot learning* for text classification tasks in data-scarce environments. This approach offers a practical and scalable solution for future applications in AI-based misinformation detection, spam filtering, and bot identification across various digital platforms.

REFERENCES

[1] R. Qonita, and Laily Rosidah, and Fahmi, "Pengaruh Youtube Terhadap Kemampuan Interaksi Sosial Anak Usia 5-6 Tahun," *Indones. J. Early Child. J. Dunia Anak Usia Dini*, vol. 5, no. 1, pp. 197–206, 2023, doi: 10.35473/ijec.v5i1.2054.

[2] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, 2016, doi: 10.1145/2818717.

[3] N. Pasieka, M. Kulynych, S. Chupakhina, Y. Romanyshyn, and M. Pasieka, "Harmful effects of fake social media accounts and learning platforms," *CEUR Workshop Proc.*, vol. 2923, pp. 258–271, 2021.

[4] L. H. X. Ng and K. M. Carley, "What is a Social Media Bot? A Global Comparison of Bot and Human Characteristics," pp. 1–18, 2025, doi: 10.1038/s41598-025-96372-1.

[5] D. Hamdana and A. Husna, "Obtaining Elderly Patients' Lifestyle Information from Unstructured Text Sources," *Proc. Malikussaleh Int. Conf. Multidiscip. Stud.*, vol. 3, no. 3, p. 00022, 2023, doi: 10.29103/micoms.v3i.181.

[6] and U. A. K. Amani Aljehani , Syed Hamid Hasan, "Advancing Text Classification : A Systematic Review of Few- Shot Learning Approaches," *Int. J. Comput. Digit. Syst.*, pp. 1–14, 2022.

[7] M. Qamal, D. Hamdhana, and M. Martin, "Sistem Pakar Untuk Mendiagnosa Penyakit Angina Pektoris (Angin Duduk) Dengan Metode Forward Chaining Berbasis Web," *TECHSI - J. Tek. Inform.*, vol. 12, no. 1, p. 86, 2020, doi: 10.29103/techsi.v12i1.2150.

[8] H. yang Lu, C. Fan, X. Song, and W. Fang, "A novel few-shot learning based multimodality fusion model for COVID-9 rumor detection from online social media," *PeerJ Comput. Sci.*, vol. 7, no. 2011, pp. 1–24, 2021, doi: 10.7717/peerj-cs.688.

[9] B. Lwowski and P. Najafirad, "COVID-19 Surveillance through Twitter using Self-Supervised and Few Shot Learning," 2020, doi: 10.18653/v1/2020.nlpcovid19-2.9.

[10] F. Rashif, G. Ihza Perwira Nirvana, M. Alif Noor, and N. Aini Rakhmawati, "Implementasi LDA untuk Pengelompokan Topik Cuitan Akun Bot Twitter bertagar #Covid-19," *Cogito Smart J.* , vol. 7, no. 1, pp. 1–12, 2021.

[11] U. Tunc, E. Atalar, M. S. Gargi, and Z. E. A. And, "Classification of Fake , Bot , and Real Accounts on Instagram Using Machine Learning Makine Öğrenmesi ile Instagram ' da Sahte , Bot ve Gerçek Hesapların Sınıflandırılması," *Politek. Derg.*, vol. 0900, no. 2, pp. 0–13, 2022, doi: 10.2339/politeknik.1136226.

[12] J. S. Rohit Kundu, "Everything you need to know about Few-Shot Learning," digitalocean. Accessed: Nov. 10, 2024. [Online]. Available: https://www.digitalocean.com/community/tutorials/few-shot-learning#how-does-few-shot-learning-work

[13] C. Schneebeli, "Coding Emotion in Computer-Mediated Communication : The Example of YouTube Comments," *Rech. anglaises Nord.*, vol. 51, no. 1, pp. 45–56, 2018, doi: 10.3406/ranam.2018.1563.

[14] S. Yang, S. Park, Y. Jang, and M. Lee, "YTCommentQA: Video Question Answerability in Instructional Videos," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 17, pp. 19359–19367, 2024, doi: 10.1609/aaai.v38i17.29906.

[15] D. O'Callaghan, M. Harrigan, J. Carthy, and P. Cunningham, "Network analysis of recurring YouTube spam campaigns," *ICWSM 2012 - Proc. 6th Int. AAAI Conf. Weblogs Soc. Media*, pp. 531–534, 2012, doi: 10.1609/icwsm.v6i1.14288.

[16] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "TWEETEVAL: Unified benchmark and comparative evaluation for tweet classification," *Find. Assoc. Comput. Linguist. Find. ACL EMNLP 2020*, pp. 1644–1650, 2020, doi: 10.18653/v1/2020.findings-emnlp.148.

[17] M. Gaber, M. Ahmed, and H. Janicke, "Zero Day Ransomware Detection with Pulse: Function Classification with Transformer Models and Assembly Language," *Comput. Secur.*, vol. 148, no. August 2024, p. 104167, 2024, doi: 10.1016/j.cose.2024.104167.

[18] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," pp. 2–6, 2019, [Online]. Available: http://arxiv.org/abs/1910.01108

[19] G. Brauwers and F. Frasincar, "A General Survey on Attention Mechanisms in Deep Learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3279–3298, 2023, doi: 10.1109/TKDE.2021.3126456.

[20] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, no. 1, pp. 111–132, 2022, doi: 10.1016/j.aiopen.2022.10.001.

[21] A. de Santana Correia and E. L. Colombini, "Attention, please! A survey of neural attention models in deep learning," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 6037–6124, 2022, doi: 10.1007/s10462-022-10148-x.