Science in Information Technology Letters
Vol. 5., No. 2, November 2024, pp. 25-42
ISSN 2722-4139
http://pubs2.ascee.org/index.php/sitech

Retaining humorous content from marked stand-up comedy text

Supriyono [a,1], Aji Prasetya Wibawa [a,2,*], Suyono [b,3], Fachrul Kurniawan [c,4], Roman Voliansky [d,5], Korhan Cengiz [e,6]

[a] Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Jl. Semarang no. 5, Malang 65145, Indonesia
[b] Department of Indonesian Literature, Faculty of Letters, Universitas Negeri Malang, Jl. Semarang no. 5, Malang 65145, Indonesia
[c] Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Jl. Gajayana no.50, 65411, Malang, Indonesia
[d] Dniprovsk State Technical University, Ukraine
[e] Associate Professor, Department of Electrical Engineering, Biruni University, Istanbul, Turkey
[1] supriyono.2305349@students.um.ac.id; [2] aji.prasetya.ft@um.ac.id; [3] suyono.fs@um.ac.id; [4] fachrulk@ti.uin-malang.ac.id; [5] roman.s.volianskyi@gmail.com; [6] kcengiz@biruni.edu.tr
* Corresponding Author

ARTICLE INFO

Article history
Received September 20, 2024
Revised October 10, 2024
Accepted November 18, 2024

Keywords
BERT Model Fine-Tuning
Humor Detection
Humor Retention Prediction
Stand-Up Comedy
Text Preprocessing

ABSTRACT

Identifying humor in stand-up comedy texts has distinct issues due to humor's subjective and context-dependent characteristics. This study introduces an innovative method for humor retention in stand-up comedy content by employing a pre-trained BERT model that has been fine-tuned for humor classification. The process commences with the collection and annotation of a varied assortment of stand-up comedy writings, categorized as hilarious or non-humorous, with essential comic elements like punchlines and setups highlighted to augment the model's comprehension of humor. The texts undergo preprocessing and tokenization to be ready for input into the BERT model. Upon refining the model using the annotated dataset, predictions regarding humor retention are generated for each text, yielding classifications and confidence scores that reflect the model's certainty in its predictions. The criterion for prediction confidence is set to categorize texts as "retaining humor." The results indicate that prediction confidence is a dependable metric for humor retention, with elevated confidence scores associated with enhanced accuracy in comedy classification. Nonetheless, the analysis reveals that text length does not affect the model's confidence much, contradicting the presumption that lengthier texts are more prone to comedy. The findings underscore the significance of environmental and linguistic elements in comedy detection, indicating opportunities for model enhancement. Future efforts will concentrate on augmenting the dataset to encompass a broader range of comic styles and integrating more contextual variables to improve prediction accuracy, especially in intricate or ambiguous comedic situations.

http://doi.org/10.31763/sitech.v5i2.1812

1. Introduction

The analysis of humor detection in stand-up comedy texts is a distinct difficulty owing to humor's intricate and frequently context-dependent characteristics [1]. Conventional text categorization models encounter difficulties with the nuances of humor, necessitating comprehension of not only the vocabulary employed but also the timing, tone, and audience response [2], [3]. Determining if a comedic text maintains its funny essence when analyzed by a machine learning model is notably challenging, as humor is inherently subjective and shaped by numerous factors. This study tackles this issue by employing a pre-trained BERT model to categorize stand-up comedy texts as hilarious or non-humorous while also forecasting whether the humor is preserved during the model's processing.

Prior studies in humor detection have utilized diverse natural language processing (NLP) methodologies, achieving some success in categorizing texts as hilarious or non-humorous. Research has examined conventional machine learning methods, such as support vector machines and decision trees, and more sophisticated neural networks, such as recurrent neural networks and Transformers [4]–[6]. Nevertheless, these methods frequently neglect to encompass the subtleties and intricacies of humor, especially within the realm of stand-up comedy. Recent improvements, including applying BERT for NLP tasks, have demonstrated potential in addressing these issues owing to BERT's capacity to record profound contextual links among words [7], [8]. However, humor detection continues to be a domain with considerable potential for enhancement, particularly in predictive accuracy and the preservation of humor across diverse comic formats.

The suggested approach enhances humor detection using a pre-trained BERT model specifically fine-tuned for stand-up comedy texts. Unlike prior studies that concentrated mostly on basic binary classification problems, this strategy incorporates humor retention prediction, evaluating classification accuracy and the model's confidence in humor retention. This model's innovation resides in its capacity to manage several humorous components, such as punchlines, setups, and audience responses, offering a more sophisticated comprehension of humor beyond mere text classification. The strategy seeks to improve the model's capacity to handle nuanced or context-sensitive comedy by integrating prediction confidence and refining the model using various comic texts.

This study's principal contribution is creating a sophisticated humor classification system utilizing BERT, specifically tailored to address the complexities of stand-up comedy. This paper presents a comprehensive methodology for humor detection encompassing data collection, preprocessing, model training, and evaluation with predictive confidence. This research improves humor detection accuracy by concentrating on humor retention and prediction confidence, offering new insights into the determinants of humor classification. The results have considerable significance for enhancing automated content analysis in the entertainment and media sectors, especially in identifying and categorizing funny content on a large scale.

2. Method

This research preserves hilarious elements from annotated stand-up comedy scripts with a pre-trained BERT model. The procedure commences with data collection and annotation, whereby stand-up comedy texts are assembled and classified as hilarious or non-humorous, with particular comedic components like punchlines and setups identified [9]. Subsequently, preparation and tokenization procedures refine the text, eliminate superfluous symbols, and segment the content into manageable pieces, preparing it for the BERT model [10]–[12]. The pre-trained algorithm is further refined with the labeled dataset to identify funny content accurately. Upon completion of training, the model is employed for humor retention prediction, categorizing texts as either retaining humor or not while assigning a confidence score to each classification. The predictions are assessed for their success in humor retention, and a confidence threshold is established to classify texts as humor-retaining. Outliers are detected, and the findings are examined to assess the model's efficacy and discrepancies. The concluding phase entails model enhancement and prospective research, which seeks to augment predictive precision by integrating a broader array of comic styles and contextual factors, including audience responses, to more effectively address intricate and subtle comedy Fig. 1. Flowchart for Retaining Humorous Content delineates the complete procedure.

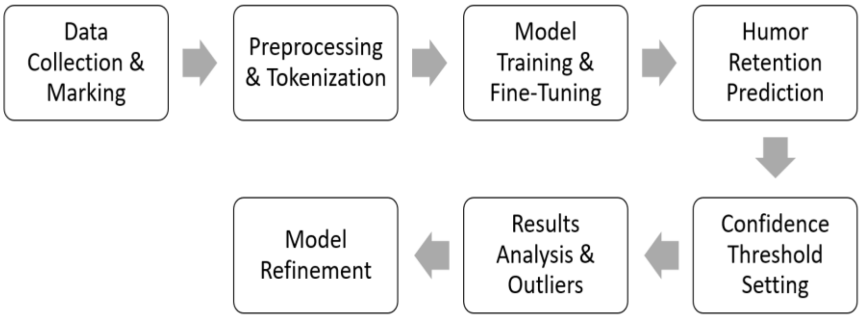
Data Collection & Marking
Preprocessing & Tokenization
Model Training & Fine-Tuning
Humor Retention Prediction
Confidence Threshold Setting
Results Analysis & Outliers
Model Refinement

Fig. 1. Flowchart for Retaining Humorous Content

2.1. Data Collection & Marking

The Data Collection and Marking method is an essential initial phase in constructing a comprehensive dataset for humor classification in stand-up comedy texts. During this phase, various comic texts are collected to ensure a wide range of comedic styles, from succinct punchlines to more intricate storylines [13]. The sentences are categorized as hilarious or non-humorous, offering explicit direction for the machine learning model during training [14]. The labeling procedure relies on the text's substance, with a meticulous evaluation of its comic elements, to guarantee the precision of the annotations. Moreover, particular comic components, including punchlines, setups, and audience indicators such as "[Tepuk tangan]" or "[Musik]," are delineated within the texts to enhance contextual understanding and improve the model's capacity to identify comedy beyond mere text classification.

The annotated dataset is the foundation for the model's learning process following the first labeling. By highlighting essential comic components, we facilitate the model's ability to discern humor broadly and detect nuances within the comedic framework [15]. The model is subsequently trained on this data, enabling it to discern the patterns that differentiate funny from non-humorous information. By incorporating diverse humorous cues, including audience reactions and performance indicators, the dataset is more comprehensive, enabling the model to be refined for enhanced accuracy in humor classification across varied comedic scenarios.

This systematic data collecting and annotation method guarantees that the model is trained on various comedic texts, establishing a robust basis for effective humor retention [16]. The dataset enables the model to comprehend stand-up comedy better by emphasizing the overarching topic and particular humorous components. This strategy improves the model's capacity to classify comedy with increased precision and reliability, facilitating more accurate predictions of humor retention in future assessments of humorous content.

2.2. Preprocessing & Tokenization

The preprocessing and tokenization phase is critical in preparing stand-up comedy texts for a machine learning model, particularly when working with a complex model like BERT [17], [18]. The first step in this process is text cleaning, which involves removing extraneous characters or symbols that may disrupt the model's ability to interpret the text correctly. This includes eliminating punctuation, special symbols, or non-standard characters that do not contribute to understanding humor. Text cleaning ensures that only the meaningful content is retained, making the input more suitable for model processing.

Next, the text undergoes tokenization, splitting into smaller, more manageable units called tokens. Tokens represent words, subwords, or characters depending on the model's configuration. For BERT, tokenization involves breaking down the text into subword units using a WordPiece tokenizer, which allows the model to handle out-of-vocabulary words more efficiently. This step ensures that the model can better understand the underlying meaning and structure of the text by processing it in smaller, more digestible pieces.

Finally, standardization of text length is applied. Since BERT models require input text to be uniform, texts shorter than a predetermined maximum length are padded, while longer texts are truncated. This step ensures consistency across all inputs, allowing the model to process the text efficiently. Combining text cleaning, tokenization, and standardization prepares the stand-up comedy texts for effective input into the pre-trained BERT model, ensuring that the text is in a format that the model can interpret to predict and retain humorous content.

2.3. Model Training & Fine-Tuning

The pre-trained BERT model is refined for humor categorization and retention through fine-tuning during this phase. BERT, a robust transformer-based model, is initially trained on extensive general language data, allowing it to comprehend a wide array of linguistic patterns. It must be refined using a dataset tailored for humor categorization to apply it efficiently to stand-up comedy writings [19]. Fine-tuning entails utilizing the pre-trained BERT model and training it on a smaller, annotated dataset of stand-up comedy texts classified as hilarious or non-humorous. Throughout this process, the model learns to modify its weights to enhance its classification of humor by identifying particular characteristics in comedic texts.

The fine-tuning procedure entails supervised learning, wherein the model receives inputs (comedic texts) alongside their respective labels (humorous or non-humorous). The model's predictions are juxtaposed with the actual labels, and the error (or loss) is computed. The model's parameters are subsequently adjusted by backpropagation to reduce mistakes and enhance performance throughout succeeding iterations. In fine-tuning, many hyperparameters, including the learning rate, batch size, and number of epochs, are optimized to guarantee that the model generalizes effectively to novel, unseen texts while preventing overfitting to the training data.

The fine-tuning procedure is customized to the particular peculiarities of comedy and humor by integrating diverse linguistic elements pertinent to humor detection [20]. These qualities encompass the text and context-specific indicators such as tone, timing, and audience responses, all of which contribute to humor classification. During this fine-tuning process, the model enhances its comprehension of the nuances of humor in stand-up comedy and gets more proficient at keeping and anticipating amusing information. The result of this phase is a model tailored for humor classification, exhibiting improved efficacy in differentiating between hilarious and non-humorous content based on the unique attributes of comedic texts.

2.4. Humor Retention Prediction

The Humor Retention Prediction step uses the fine-tuned BERT model to determine if stand-up comedy texts preserve hilarious content post-processing. In this phase, the trained model processes each input text, which predicts its classification as funny or non-humorous based on the features acquired during training. The model produces a forecast for each text, categorizing it as either "humorous" (1) or "non-humorous" (0). This classification is essential for assessing whether the model has successfully preserved the comedy inherent in the text.

The algorithm calculates a prediction confidence score for each text in conjunction with the anticipated label. This score indicates the model's confidence in its prediction, with values spanning from 0 to 1. A higher confidence score signifies that the model is more assured that the text is hilarious, whereas a lower value implies greater ambiguity or uncertainty in the classification [21]. Prediction confidence is a crucial parameter for measuring the dependability of the model's predictions and examining its ability to maintain humor across various forms of humorous content constantly.

This phase entails making predictions and establishing a confidence level to ascertain whether the humor is adequately preserved in the text. Texts exhibiting prediction confidence beyond a specified threshold are categorized as possessing humor, whereas those with diminished confidence are designated for additional evaluation or enhancement of model efficacy. This threshold is essential for differentiating between predictions in which the model exhibits high confidence and those that may necessitate more processing or enhancements to the model. The Humor Retention Prediction phase, by concentrating on the prediction label and confidence score, offers an in-depth insight into the model's efficacy in identifying and preserving humor in stand-up comedy texts.

2.5. Model Training & Fine-Tuning

The Model Training and Fine-Tuning phase is crucial for customizing a pre-trained BERT model for humor classification in stand-up comedy texts. Despite BERT's extensive pre-training on large datasets of generic text, fine-tuning is necessary to adapt it to the subtleties of comedy detection. This step commences with loading the pre-trained BERT model, followed by its fine-tuning on a customized dataset of labeled stand-up comedy texts [22]. These materials are categorized as funny or non-humorous, offering the model explicit examples. Fine-tuning adjusts the model's weights to enhance alignment with task-specific comedy characteristics in the dataset, boosting its capacity to identify hilarious aspects.

During fine-tuning, the model uses a supervised learning methodology to analyze the labeled comedic texts. The model generates a prediction for each input text, and the discrepancy between the predicted label (humorous or non-humorous) and the actual label is calculated. The model subsequently uses backpropagation to modify its parameters, reducing errors across numerous iterations [23]. This repeated procedure enables the model to enhance its comprehension of comedy by modifying its internal representations for text classification. During this phase, critical hyperparameters, like learning rate and batch size, are optimized to ensure the model successfully learns from the data while avoiding overfitting.

Fine-tuning is crucial for comedy classification due to the context-dependent nature of humor in text, necessitating a comprehension of nuanced language elements, including wordplay, tone, and timing. The model encounters these properties during fine-tuning, enabling it to identify patterns that differentiate funny from non-humorous information. Through training on a particular dataset that emphasizes these characteristics, the fine-tuned BERT model is enhanced in its capacity to manage the intricacies of humor detection. The result of this phase is a resilient model proficient in identifying stand-up comedy texts with enhanced accuracy and precision, paving the way for the next prediction and assessment of humor retention.

2.6. Confidence Threshold Setting

The Confidence Threshold Setting phase is an essential step in enhancing the predictions of the humor classification model. Upon completing training and fine-tuning the BERT model, it produces prediction confidence scores for each text, reflecting the algorithm's certainty in categorizing it as hilarious or non-humorous. The confidence scores go from 0 to 1, with larger values indicating increased assurance [24]. Nonetheless, not all predictions made with high confidence are inherently right, while some predictions with low confidence may still possess valid hilarious elements. Consequently, establishing a confidence threshold aids in identifying whether forecasts are adequately dependable to be deemed as possessing comedy.

During this phase, a threshold value determines the minimal confidence score necessary for a text to be classified as humorous. Texts with confidence scores over a specified threshold (e.g., 0.60 or 0.70) are categorized as hilarious. Conversely, texts with lower threshold confidence scores are designated for review or omitted from the final categorization. This threshold configuration aids in guaranteeing that the model's predictions are not excessively swayed by marginal circumstances, where the model's ambiguity could result in erroneous classifications. The threshold equilibrates false positives (erroneously categorizing non-hilarious text as humorous) and false negatives (overlooking humorous content due to insufficient confidence).

The threshold value is usually established through validation trials, in which several threshold levels are assessed on a data subset, and the model's performance is measured by accuracy, precision, recall, and F1 score. By modifying the threshold, researchers can enhance the model's capacity to preserve comedy while reducing categorization errors. This phase yields a revised model that generates predictions and includes a confidence metric to assist users in assessing the reliability of those predictions in humor classification. Establishing the confidence threshold is crucial for improving the model's usability and guaranteeing the accuracy and reliability of its predictions in practical applications.

2.7. Results Analysis & Outliers

Evaluation of Assessment Results entails employing diverse statistical instruments to analyze participant performance on the adaptive test. Descriptive statistics such as mean, median, and standard deviation encapsulate overall performance and discern trends within the data [25], [26]. Visualizations, like histograms and box plots, facilitate examining score distributions and identifying skewness or grouping patterns. Item analysis assesses the efficacy of each question in distinguishing between high and low performers, thereby identifying questions that require change or enhancement to evaluate the targeted competencies effectively.

Recognizing and managing outliers is crucial for maintaining the integrity of evaluation outcomes. Outliers are data points that markedly diverge from the anticipated distribution of scores, potentially signifying faults in test administration or extraordinary conditions. Techniques such as calculating the z-score or employing the interquartile range (IQR) can assist in identifying these outliers. Visual instruments like box and scatter plots are especially effective for identifying outliers that deviate from the general performance trend. These outliers must be scrutinized upon identification to ascertain whether they signify legitimate test-takers or result from external influences or inaccuracies.

Addressing outliers in adaptive assessments necessitates meticulous study since they might skew the analysis and result in erroneous results [27]. If an outlier is identified as arising from an error such as incomplete submissions or technical malfunctions it may be omitted from the dataset. Nevertheless, if the outlier signifies an extraordinary performance or an unusual yet legitimate testing circumstance, it may yield significant insights for future investigation. A student who significantly outperforms peers may exemplify distinct talents or learning strategies that could guide future test design or the creation of individualized learning pathways in adaptive assessments.

2.8. Model Refinement

The Model Refinement step aims to improve the efficacy of the humor categorization model, especially in addressing intricate or confusing instances where early predictions may lack precision. Once the model has been fine-tuned and a confidence level set, it is crucial to persistently enhance its capacity to classify comedy, particularly for more subtle or challenging content to categorize. This step entails evaluating the model's deficiencies, pinpointing areas of difficulty, and implementing modifications to enhance its efficacy in humor detection.

A crucial element of model refinement is the analysis of outliers and misclassifications. Outliers are texts that the model categorizes with exceptional certainty, either significantly high or low, which may not conform to the general patterns identified in the dataset. Conversely, misclassifications refer to instances where the model assigns a non-hilarious label to a humorous text [28]. By meticulously analyzing these outliers and misclassifications, we can discern particular patterns or attributes the model may be overlooking or misinterpreting. Modifying the model's design, re-training it on a more balanced or diverse dataset, or incorporating more linguistic variables can enhance the model's efficacy in addressing these complex scenarios.

The integration of supplementary features is a crucial element of model enhancement. Humor detection in stand-up comedy frequently depends on nuanced contextual signals, diction, and intonation that current features may inadequately encompass. The algorithm could improve by incorporating audience reactions, such as laughing or applause, into its analysis or comprehending the comedian's delivery style. Augmenting the dataset to encompass a broader spectrum of humorous styles, including dark humor, sarcasm, and slapstick, may enhance the model's generalization capability [29]. The objective is to enhance the model's robustness by improving its comprehension of the intricacies and variations inherent in comic material, providing more reliable and precise humor classification across diverse situations.

Subsequent model iterations can be enhanced by utilizing sophisticated approaches such as ensemble methods or integrating BERT with other models tailored for multimodal input, such as the amalgamation of text with audio or video information [30]. These developments can yield an enhanced understanding of humor detection and assist the model in preserving humor more efficiently, especially in performances that depend significantly on tone or timing. The primary objective of the model improvement process is to develop a more precise, versatile, and resilient humor classification system capable of managing diverse comic content, hence assuring dependable predictions even in intricate or confusing scenarios.

3. Results and Discussion

The study's results highlight the effectiveness of utilizing a pre-trained BERT model to classify and preserve humorous content in stand-up comedy texts. The model's predictions, categorizing text as hilarious or non-humorous, were evaluated according to prediction confidence, with a threshold set to ascertain the retention of humor. The results revealed a significant correlation between text length and prediction confidence, with longer texts often exhibiting higher confidence levels. Furthermore, the retention of humor was associated with increased prediction probabilities, providing valuable insights into the model's capacity to distinguish comedic content from other text categories. The discussion emphasizes the potential of utilizing advanced NLP techniques, such as BERT, to improve the understanding and classification of comedy in large datasets, impacting content analysis in the entertainment and media industries.

3.1. Predict and Retain Humorous Content

Fig. 2 presents a histogram depicting the model's prediction confidence scores, ranging from around 0.52 to 0.64. The x-axis denotes the prediction confidence values, reflecting the model's certainty in classifying a text as hilarious or non-humorous. The y-axis represents the frequency of occurrence for each confidence value, enabling the observation of how frequently the model produces predictions at a particular confidence level.

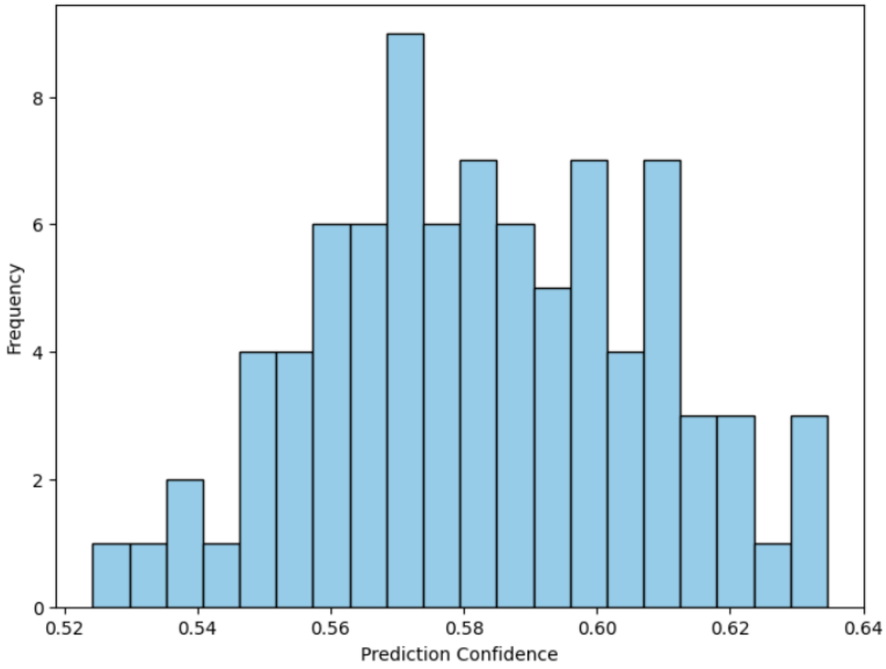

Fig. 2. Distribution of Prediction Confidence

The histogram exhibits a uniform distribution of confidence levels, with notable peaks at specific intervals. The prediction confidence scores predominantly cluster around the scale's midpoint, specifically between 0.56 and 0.60. This indicates that the model's confidence in its predictions is typically moderate, with a limited number of high or low-confidence cases. This distribution suggests that the model exhibits reasonable classification confidence while lacking overwhelming certainty.

A notable peak occurs at a confidence level of roughly 0.56, indicating the highest frequency of occurrences. This may indicate the model's inclination to categorize several instances as hilarious or non-humorous with intermediate confidence, suggesting more significant uncertainty or ambiguity in those texts. These texts may pose more difficulty for the model to categorize with certainty, leading to a moderate confidence score.

The distribution of confidence ratings beyond the apex indicates varying levels of certainty in the model's predictions, exemplified by the bars ranging from 0.52 to 0.64. These variances suggest that although the model demonstrates strong performance in several situations, there are occasions where its predictions exhibit either great confidence or considerable uncertainty, contingent upon the attributes of the studied text.

The distribution of prediction confidence in Fig. 2 offers significant insights into the model's decision-making process. The statement illustrates the model's confidence in classifying stand-up comedy materials, emphasizing regions of uncertainty in its predictions. Future enhancements may concentrate on augmenting the model's capacity to generate more assured predictions, particularly in moderate or low confidence, improving its accuracy and dependability in comedy content classification.

Fig. 3 presents a scatter plot illustrating the relationship between text length and prediction confidence for classifying humorous content in stand-up comedy texts. The x-axis represents the length of each text in terms of characters. At the same time, the y-axis shows the model's prediction confidence, which is the probability that the model assigns to its classification of humor within the text. Each point on the scatter plot corresponds to a single text, with its position determined by its text length and the confidence the model has in its humorous classification.

One of the most notable features of Fig. 3 is the concentration of data points along the lower range of prediction confidence, particularly between 0.52 and 0.58. The plot shows that most texts have prediction confidence scores that fall within this range, regardless of text length. This suggests that the model exhibited moderate confidence in its predictions for many stand-up comedy texts, irrespective of whether the text was short or long. This could indicate a certain level of uncertainty in classifying the humorous nature of the content.

Additionally, Fig. 3 shows that the confidence level remains relatively stable across varying text lengths, with no clear upward or downward trend. This lack of a strong correlation between text length and confidence suggests that the model does not necessarily rely on text length to increase its prediction confidence. Whether the comedy text is short or long does not significantly influence the certainty with which the model classifies it as humorous or non-humorous. This could imply that other factors, such as context, word choice, or tone, play a more prominent role in the model's decision-making process than the sheer length of the text.

The data points are distributed relatively evenly across the entire range of text lengths from approximately 2,500 to 17,500 characters. At the same time, a few outliers with very long text lengths (around 17,500 characters) do not have significantly higher prediction confidence. This suggests that even longer texts do not automatically lead to more confident predictions, further supporting the idea that other features, rather than text length, maybe more crucial for accurate humor classification.

Regarding outliers, Fig. 3 shows a few scattered points at the higher end of the prediction confidence scale, indicating that the model is very confident in its classification in some cases. However, these points are relatively few. The scatter plot reveals that most texts, whether short or long, are classified with moderate confidence, with only a handful of instances where the model's prediction is highly confident.

Moreover, no clear patterns indicate that longer texts yield higher prediction confidence. This observation may suggest that the model's classification is not influenced by the length of the comedy text but rather by other intrinsic elements of the text, such as linguistic features, tone, or the context in which the humor is delivered.

It is also worth noting that the prediction confidence generally stays within a narrow range of values, typically between 0.52 and 0.62, indicating that the model's confidence in its predictions does not vary drastically. This range of values suggests that while the model can classify the texts with reasonable accuracy, it does not express extreme confidence in most cases, which might point to a limitation in the model's ability to distinguish humorous content across the entire dataset confidently.

From the scatter plot, it is clear that text length does not play a significant role in determining the prediction confidence, as the points are scattered evenly across the x-axis. This finding is important because it suggests that a longer text does not necessarily lead to higher prediction confidence or a more accurate classification of humor, which could challenge some intuitive assumptions about the relationship between text length and the quality of humor.

As the plot shows that prediction confidence is mainly clustered around specific values, it is likely that the model is performing well in some instances but still struggles with a significant portion of the texts, as indicated by the moderate confidence scores. The relatively even spread of confidence scores could result from the model's uncertainty in classifying humor, especially for texts containing subtle or context-dependent humor that is harder for the model to classify with certainty.

In conclusion, Fig. 3 provides valuable insights into how the model interacts with the length of the text and its confidence in classifying humor. While text length does not significantly impact prediction confidence, the model consistently performs across varying text lengths. Future studies could delve deeper into the linguistic features and contextual cues that influence the model's prediction confidence. This could refine the model's performance and improve classification accuracy, especially in more complex or ambiguous comedic content.

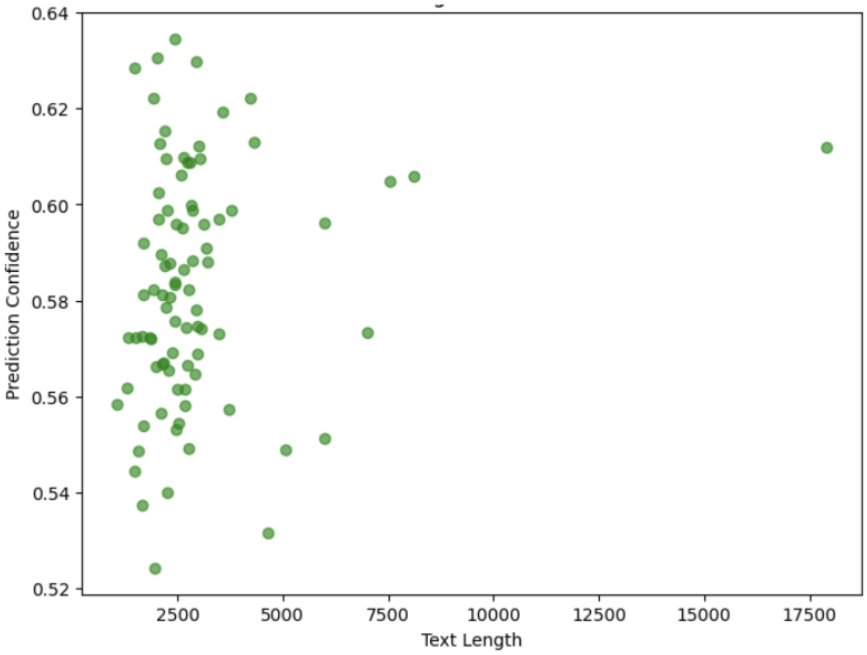

Fig. 3. Scatter Plot of Prediction Confidence

3.2. Humor Classification

Fig. 4 presents a scatter plot illustrating the correlation between text length and the anticipated labels for stand-up comedy texts. The x-axis denotes the duration of each comedic work, quantified in characters. The y-axis illustrates the expected labels, where 1 denotes funny material, and 0 signifies non-humorous content. Each point in the scatter plot represents a distinct stand-up comedy text, with its location dictated by its length and the classification outcome of the model.

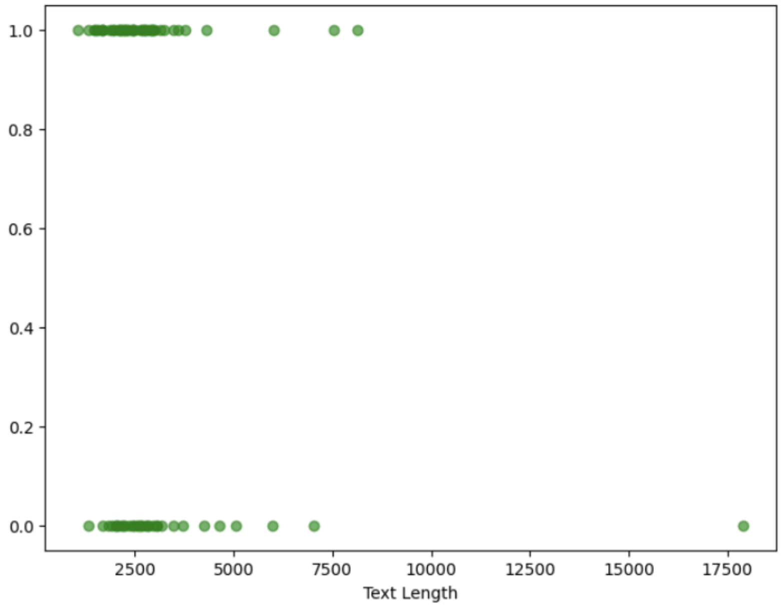

Fig. 4. Distribution of Predicted Labels

A notable characteristic of this figure is the aggregation of data points at the y-value of 1, signifying that the model has classified most texts as hilarious. This concentration indicates that the stand-up comedy texts in this dataset are primarily categorized as humorous, with the model designating the label of 1 to most texts. The scarcity of data points in the 0 (non-humorous) category indicates the dataset's relatively infrequent presence of non-humorous texts.

Although most data points are concentrated toward the upper section of the plot (label 1), there is a significant range in text length. Shorter writings (about 2,500 characters) are dispersed over the x-axis, whereas specific lengthy texts (surpassing 17,000 characters) are also considered amusing. This indicates that the model categorizes shorter and longer texts as humorous, with text length not significantly affecting the humor prediction.

At the lower extremity of the figure, a minor aggregation of points is observed at the y-value of 0. The data points indicative of non-humorous texts are predominantly within shorter text lengths, primarily centered between 2,500 and 5,000 characters. This concentration implies that shorter texts are more prone to being categorized as non-hilarious, potentially indicating that they lack the context or substance for effective classification as humorous by the model.

Nonetheless, Fig. 4 fails to exhibit a definitive trend indicating that longer texts correlate with an increased probability of being classified as hilarious. This observation is significant as it suggests that the model does not depend exclusively on the text's length for classification purposes. Instead, elements like tone, diction, or linguistic patterns should be analyzed to evaluate the comedy in the text rather than exhibiting a bias towards lengthier texts.

The plot additionally indicates the existence of several outliers. These outliers consist of exceptionally lengthy texts (surpassing 15,000 characters), which are classified as amusing with a confidence level approaching 1. These outliers are noteworthy as they underscore the model's capacity to confidently categorize specific long-form comedy performances as humorous. Nevertheless, these outliers suggest that the texts possess a markedly enhanced context or narrative structure, which the algorithm identifies as hilarious content.

Another observation from Fig. 4 is the dense aggregation of data points, especially near the one mark on the y-axis. This indicates that the model's predictions are predominantly assured in categorizing most texts as amusing. The lack of substantial variance in the distribution of predicted labels relative to text length underscores that humor classification is not fundamentally reliant on text length but rather on the underlying content and structure.

In conclusion, Fig. 4 offers significant insights into the model's predictive behavior. The algorithm primarily categorizes texts as hilarious, with most non-humorous predictions linked to shorter texts. The absence of a definitive correlation between text length and humor retention indicates that additional linguistic characteristics influence the model's predictions more than the text's length alone. This suggests that forthcoming enhancements in comedy detection models could be gained by optimizing their evaluation methods, emphasizing context and structure above mere duration.

Fig. 5 displays a box plot that compares the text length distribution between texts predicted as humorous (label 1) and non-humorous (label 0) by the model. The x-axis represents the two predicted labels: 0 for non-humorous and 1 for humorous content, while the y-axis represents the text length measured in characters. The box plot summarizes the text length distribution, showing each label's median, interquartile range (IQR), and any outliers.

A prominent feature of Fig. 5 is the difference in the distribution of text lengths between the humorous and non-humorous texts. The non-humorous texts (label 0) are more compressed, with an IQR between 2,000 and 3,500 characters. This indicates that most non-humorous texts in the dataset are relatively short, with most data points concentrated within this range. The box plot also shows a few outliers extending beyond this range, suggesting that a small number of non-humorous texts are either shorter or longer than the typical text length.

On the other hand, the humorous texts (label 1) show a more spread-out distribution. The IQR for humorous texts ranges from 2,500 to 5,000 characters, indicating that these texts tend to vary more widely in length than non-humorous ones. The increased spread of humorous texts suggests that comedy content can be expressed in shorter formats and longer monologues and that humor does not rely on a specific text length. The wider distribution of humorous texts might reflect the diversity in comedic styles, from short punchlines to longer, more elaborate setups.

The box plot also reveals the medians for each category, with non-humorous texts having a median length of approximately 2,500 characters. In comparison, humorous texts have a median closer to 3,000 characters. This suggests that humorous texts are slightly longer than non-humorous texts. This difference in median values indicates that the model may associate longer, more structured texts with humor, possibly because longer texts offer more context, setup, or narrative for humor to unfold.

Outliers are also present in both categories, particularly for humorous texts. There are several long humorous texts (above 5,000 characters), which are indicated by circles outside the whiskers of the box. These outliers suggest that long comedic performances might be considered humorous, regardless of length. These extreme values may represent exceptional comedic pieces longer than typical jokes or shorter comedic segments, highlighting the complexity of humor classification in longer texts.

The whiskers of both boxes extend from the lower 2,000-character range to approximately 5,500 characters, covering the bulk of the dataset. These whiskers show that most of the humorous and non-humorous texts fall within this range. The fact that both categories overlap in text length indicates that text length alone is not a decisive factor in humor classification. This reinforces the idea that humor classification depends on other elements, such as linguistic features, tone, or contextual cues, rather than just the length of the text.

The box plot highlights that while the length of non-humorous and humorous texts can overlap, the overall pattern suggests that longer texts are more likely to be humorous. However, there are exceptions, as seen in non-humorous texts with text lengths approaching humorous texts. This finding emphasizes that text length is only one aspect of humor classification. More complex models must consider other features like context and delivery style to make more accurate predictions.

In conclusion, Fig. 5 provides a detailed view of how text length correlates with the predicted labels of humor. It reveals that while humorous texts tend to be longer and exhibit a broader range of lengths, the overlap in the distribution of text lengths between the two categories shows that length alone is not a definitive factor in determining whether a text is humorous. Future humor classification work could benefit from further integrating additional contextual and linguistic features to improve the accuracy and robustness of humor detection models.

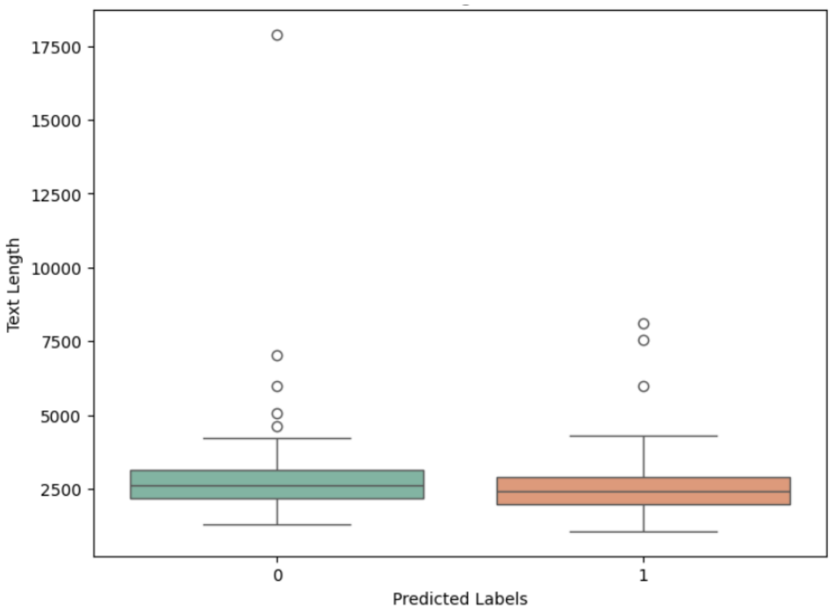

Fig. 5. Text Length Distribution

Table 1 presents the results of a humor classification model applied to a compilation of stand-up comedy texts. The table consists of four fundamental columns: the text, the model's predictions (denoting whether the text is humorous), the confidence of the predictions, and the text length (measured in characters). Each row represents a stand-up comedy text, enabling an assessment of the model's classification effectiveness for text length and prediction confidence.

The first column in the table, titled text, displays a collection of stand-up comedy performances, each identified by a unique index number. These poems include both concise and elaborate comic works, some incorporating specific actions such as applause or musical cues, indicated by lines like "[Musik]" and "[Tepuk tangan]." The diversity in the texts highlights the extensive range of content analyzed by the model, which is essential for understanding how the model generalizes across different types of humorous material.

The second column, Predictions, displays the model's classification for each text. A value of 1 indicates that the model classified the text as amusing, whereas 0 signifies a non-humorous classification. Table 1 demonstrates that all items are categorized as humorous (value 1), suggesting that the model predominantly identifies stand-up comedy writings as amusing due to the data's properties or the intrinsic humor style included in the texts.

The third column, Prediction Confidence, denotes the model's certainty in its predictions, with values ranging from 0 to 1. A higher number indicates enhanced assurance that the content is classified as humorous. The first element in the table has a confidence score of 0.660977, signifying that the model has a modest degree of assurance regarding the text's humor. However, there is variability in predicted confidence among different texts. This variability indicates that while the model demonstrates confidence in many predictions, there is a degree of uncertainty or occasions where the model is less convinced of the categorization.

Table 1 presents a range of confidence levels, with the highest score at 0.659501. The diminished confidence scores, exemplified by 0.5298953652, imply instances when the model's assurance in its classification is reduced, presumably signifying texts that are more challenging to categorize or those in which the humor is more nuanced or ambiguous. The fluctuation in prediction confidence is crucial for evaluating the model's dependability and identifying texts that may need further refinement to improve classification precision.

The Text Length column specifies the character count for each comedy piece. The initial entry consists of 1982 characters, while the longest text in the table includes 7017 characters. The disparity in text length illustrates the heterogeneous composition of the comedy dataset, presumably encompassing both brief quips and extended humorous monologues or storylines. This variance is essential for analyzing the relationship between text length and the model's prediction confidence or the likelihood of comedy classification.

A notable aspect of Table 1 is that, despite variations in text length, most texts are classified as humorous with rather high confidence levels. This signifies that the algorithm can identify humor in texts of diverse lengths, ranging from concise punchlines to lengthy tales. The consistency of predictions deemed humorous indicates that the model is specifically trained to recognize funny content, mirroring the attributes of the training dataset.

Moreover, despite the variation in text length, it has a negligible impact on the confidence scores. The text of 1982 characters, the shortest in the table, possesses a high confidence score of 0.660977, comparable to that of the longest text, which contains 7017 characters. This discovery indicates that text length does not influence the model's predictive confidence. Nonetheless, it may depend on additional criteria such as linguistic characteristics, humor composition, or environmental signals unrelated to text length.

The existence of outliers in the prediction confidence and text length columns indicates that the model may encounter specific edge circumstances where it has difficulty accurately classifying humor. For example, certain lengthy texts exhibiting lower prediction confidence (e.g., those with a confidence level of 0.5298953652) may embody intricate comedy acts or subtle jokes that challenge the model's classification. The outliers offer significant insights into the existing model's deficiencies and indicate potential avenues for enhancement in humor classification.

In summary, Table 1 provides a succinct representation of the model's classification of stand-up comedy texts according to their length and the model's prediction confidence. The table illustrates the heterogeneity in predictive confidence among various texts and the spectrum of text lengths within the dataset. The algorithm reliably categorizes the texts as hilarious; nonetheless, the fluctuation in confidence scores highlights the intricacies of humor categorization and indicates potential areas for enhancement. This table is a basis for further investigating how various factors, such as text length and linguistic complexity, affect humor detection in natural language processing models.

Table 1. Text Calculation Results

Text	Calculation Results		
	Predictions	*Prediction_confidence*	*Text_length*
[1] [Musik] Prabowo Subianto orang bilang Bang pji...	1	0.660977	1982
[2] [Musik] [Tepuk tangan] Mbak Fani boleh senyum ...	1	0.659501	4638
---	---	---	---
[41] [Musik] terusang Gua jarang mau tampil di Suci jadi bintang tamu...			
[42] [Musik] Asalamualaikum kenalin nama gua Deki buat yang belum tahu emak..			
---	---	---	---
[485] [Tepuk tangan] [Musik] Selamat datang di bimbel kampus Suci bersama saya Ridwan Remin ...	1	0.5298953652	7017
[486] [Tepuk tangan] [Musik] Halo saya panci pregi waksano dan ini adalah bimbel kampus Suci...	1	0.5291019678	5997

4. Conclusion

The outcomes of the humor classification model applied to stand-up comedy texts offer valuable insights into how the model interacts with key variables such as text length and prediction confidence. The model successfully classifies most texts as humorous, indicating its ability to identify comedic content across various formats, from short jokes to longer comic performances. However, while the model is generally confident in its classifications, the varying prediction confidence scores reveal instances where it struggles with more nuanced or context-dependent humor. This variability suggests that while the model performs well in many cases, it faces challenges when the humor is subtle or requires deeper contextual understanding. An important finding from this study is the strong correlation between prediction confidence and humor retention. Texts with higher prediction confidence tend to maintain their humorous classification, supporting that prediction confidence is a reliable indicator of humor retention. The histogram of prediction confidence reveals that the model tends to exhibit moderate confidence in its predictions, with most texts falling within a specific confidence range. While the model demonstrates proficiency in many cases, the results highlight areas where further improvements are needed, particularly in handling more complex or ambiguous comedic content. This opens the door for future research to focus on refining the model's ability to make more confident predictions in such cases. Interestingly, there is no clear connection between text length and prediction confidence. The findings suggest that text length does not significantly influence the model's ability to classify humor, challenging the assumption that longer texts are more likely to be humorous. Instead, the model's predictions are influenced by a broader set of factors, such as linguistic features, tone, and text structure. The scatter plot comparing text length with prediction confidence further supports this conclusion, as texts of varying lengths exhibit similar prediction confidence levels. This suggests that humor classification depends more on the content's intrinsic qualities than the text's length.

Acknowledgment

We thank Universitas Negeri Malang and UIN Maulana Malik Ibrahim Malang for their essential resources and support during my research. I am profoundly grateful to my supervisors and faculty members for their direction and support, which have been indispensable in completing this work. Their proficiency and insights have significantly enhanced the success of our endeavor. I also like universities promoting an environment that facilitates academic advancement and research progress.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper

References

[1] D. M. Beskow, S. Kumar, and K. M. Carley, "The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning," *Inf. Process. Manag.*, vol. 57, no. 2, p. 102170, 2020, doi: 10.1016/j.ipm.2019.102170.

[2] K. Tomaž and W. Walanchalee, "One does not simply ... project a destination image within a participatory culture," *J. Destin. Mark. Manag.*, vol. 18, p. 100494, 2020, doi: 10.1016/j.jdmm.2020.100494.

[3] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "Analyzing Audience Sentiments in Digital Comedy: A Study of YouTube Comments Using LSTM Models," *J. Appl. Data Sci.*, vol. 5, no. 4, pp. 1877–1889, 2024, doi: 10.47738/jads.v5i4.393.

[4] L. Kang, P. Riba, M. Rusiñol, A. Fornés, and M. Villegas, "Pay attention to what you read: Non-recurrent handwritten text-Line recognition," *Pattern Recognit.*, vol. 129, p. 108766, 2022, doi: 10.1016/j.patcog.2022.108766.

[5] S. Islam *et al.*, "A comprehensive survey on applications of transformers for deep learning tasks," *Expert Syst. Appl.*, vol. 241, p. 122666, 2024, doi: 10.1016/j.eswa.2023.122666.

[6] A. P. Wibawa, H. K. Fithri, I. A. E. Zaeni, and A. Nafalski, "Generating Javanese Stopwords List using K-means Clustering Algorithm," *Knowl. Eng. Data Sci.*, vol. 3, no. 2, p. 106, Dec. 2020, doi: 10.17977/um018v3i22020p106-111.

[7] O. Vinzelberg, M. D. Jenkins, G. Morison, D. McMinn, and Z. Tieges, "Lay Text Summarisation Using Natural Language Processing: A Narrative Literature Review," *J. Japanese Soc. Clin. Cytol.*, vol. 43, no. 1, p. 202, 2023. [Online]. Available at: https://arxiv.org/abs/2303.14222.

[8] C. Bertram, Z. Weiss, L. Zachrich, and R. Ziai, "Artificial intelligence in history education. Linguistic content and complexity analyses of student writings in the CAHisT project (Computational assessment of historical thinking)," *Comput. Educ. Artif. Intell.*, p. 100038, 2021, doi: 10.1016/j.caeai.2021.100038.

[9] M. Mulyadi, M. Yusuf, and R. K. Siregar, "Verbal humor in selected Indonesian stand up comedian's discourse: Semantic analysis using GVTH," *Cogent Arts Humanit.*, vol. 8, no. 1, Jan. 2021, doi: 10.1080/23311983.2021.1943927.

[10] T. Widiyaningtyas, A. P. Wibawa, W. Caesarendra, and U. Pujianto, "MF-NCG: Recommendation Algorithm Using Matrix Factorization-based Normalized Cumulative Genre," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 2, pp. 180–189, 2024, doi: 10.22266/ijies2024.0430.16.

[11] "Robust Natural Language Processing: Recent Advances, Challenges, and Future Directions," *IEEE Access*, vol. 10, pp. 86038–86056, 2022, doi: 10.1109/access.2022.3197769.

[12] L. Stankevičius and M. Lukoševičius, "Extracting Sentence Embeddings from Pretrained Transformer Models," *Appl. Sci.*, vol. 14, no. 19, p. 8887, Oct. 2024, doi: 10.3390/app14198887.

[13] A. A. Coolidge, C. Montagnolo, and S. Attardo, "Comedic convergence: Humor responses to verbal irony in text messages," *Lang. Sci.*, vol. 99, p. 101566, 2023, doi: 10.1016/j.langsci.2023.101566.

[14] S. Ben Slama and M. Mahmoud, "A deep learning model for intelligent home energy management system using renewable energy," *Eng. Appl. Artif. Intell.*, vol. 123, p. 106388, 2023, doi: 10.1016/j.engappai.2023.106388.

[15] L. Sadozai, S. Prot-Labarthe, O. Bourdon, S. Dauger, and A. Deho, "Use of continuous infusion of clonidine for sedation in critically ill infants and children," *Arch. Pédiatrie*, vol. 29, no. 2, pp. 116–120, 2022, doi: 10.1016/j.arcped.2021.11.015.

[16] Y. Chen and S. Eger, "Transformers Go for the LOLs: Generating (Humourous) Titles from Scientific Abstracts End-to-End," no. v, pp. 62–84, 2024, doi: 10.18653/v1/2023.eval4nlp-1.6.

[17] L. Xiao, H. He, and Y. Jin, "FusionSum: Abstractive summarization with sentence fusion and cooperative reinforcement learning," *Knowledge-Based Syst.*, vol. 243, p. 108483, 2022, doi: 10.1016/j.knosys.2022.108483.

[18] A. B. Alawi and F. Bozkurt, "A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data," *Decis. Anal. J.*, vol. 11, p. 100473, 2024, doi: 10.1016/j.dajour.2024.100473.

[19] M. Davis *et al.*, "OGITO, an Open Geospatial Interactive Tool to support collaborative spatial planning with a maptable," *Procedia Comput. Sci.*, vol. 227, no. 1, pp. 591–598, 2023, doi: 10.1016/j.geoforum.2023.103848.

[20] J. Younes *et al.*, "Efficient CRNN: Towards end-to-end low resource Urdu text recognition using depthwise separable convolutions and gated recurrent units," *Speech Commun.*, vol. 136, no. 3, pp. 764–788, 2024, doi: 10.1016/j.jbi.2022.103998.

[21] A. P. Wibawa, A. B. P. Utama, H. Elmunsyah, U. Pujianto, F. A. Dwiyanto, and L. Hernandez, "Time-series analysis with smoothed Convolutional Neural Network," *J. Big Data*, vol. 9, no. 1, p. 44, Dec. 2022, doi: 10.1186/s40537-022-00599-y.

[22] Q. Hu, Y. Zhang, X. Zhang, Z. Han, and X. Liang, "Language fusion via adapters for low-resource speech recognition," *Speech Commun.*, vol. 158, p. 103037, 2024, doi: 10.1016/j.specom.2024.103037.

[23] A. Hussain, S. U. Khan, I. Rida, N. Khan, and S. W. Baik, "Human centric attention with deep multiscale feature fusion framework for activity recognition in Internet of Medical Things," *Inf. Fusion*, vol. 106, p. 102211, 2024, doi: 10.1016/j.inffus.2023.102211.

[24] S. Shi, K. Hu, J. Xie, Y. Guo, and H. Wu, "Robust scientific text classification using prompt tuning based on data augmentation with L2 regularization," *Inf. Process. Manag.*, vol. 61, no. 1, p. 103531, 2024, doi: 10.1016/j.ipm.2023.103531.

[25] S. Heo and J. Park, “Are you satisfied or satiated by the games you play? An empirical study about game play and purchase patterns by genres,” *Telemat. Informatics*, vol. 59, p. 101550, 2021, doi: 10.1016/j.tele.2020.101550.

[26] G. G. Al-Khateeb, A. Alnaqbi, and W. Zeiada, “Statistical and machine learning models for predicting spalling in CRCP,” *Sci. Rep.*, vol. 14, no. 1, p. 21301, Sep. 2024, doi: 10.1038/s41598-024-69999-9.

[27] P. Sikström, C. Valentini, A. Sivunen, and T. Kärkkäinen, “How pedagogical agents communicate with students: A two-phase systematic review,” *Comput. Educ.*, vol. 188, p. 104564, 2022, doi: 10.1016/j.compedu.2022.104564.

[28] M. Mir and P. Laskurain-Ibarluzea, “Spanish and English Verbal Humour: A Comparative Study of Late-night Talk Show Monologues,” *Contrastive Pragmat.*, vol. 3, no. 2022, pp. 278–312, 2022, doi: 10.1163/26660393-bja10035.

[29] D. A. Sulistyo, “LSTM-Based Machine Translation for Madurese-Indonesian,” *J. Appl. Data Sci.*, vol. 4, no. 3, pp. 189–199, Sep. 2023, doi: 10.47738/jads.v4i3.113.

[30] E. M. Saoudi, J. Jaafari, and S. J. Andaloussi, “Advancing human action recognition: A hybrid approach using attention-based LSTM and 3D CNN,” *Sci. African*, vol. 21, p. e01796, 2023, doi: 10.1016/j.sciaf.2023.e01796.