Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

An Advanced Bert Transformer for Spoiler Detection in Extension System for Social Media

Mr. Shailesh Sangle[1*], Dr. Vaishali Nirgude[2], Ms. Foram Shah[3], Ms. Chandana Khatavkar[4]
[1]*Assistant Professor, Thakur College of Engineering and Technology, Mumbai, Maharashtra, India.*
sssphd2021@gmail.com
[2]*Associate Professor, Thakur College of Engineering and Technology, Mumbai, Maharashtra, India.*
vaishali.nirgude@thakureducation.org
[3]*Assistant Professor, Thakur College of Engineering and Technology, Mumbai, Maharashtra, India.*
foram.shah@thakureducation.org
[4]*Assistant Professor, Thakur College of Engineering and Technology, Mumbai, Maharashtra, India.*
chandana.nighut@thakureducation.org
**Corresponding Author*

ARTICLE INFO

Received: 30 Dec 2024
Revised: 19 Feb 2025
Accepted: 27 Feb 2025

ABSTRACT

A Spoiler Detection Extension System is a browser-based tool designed to identify and hide spoilers in online content, such as reviews, articles, and social media posts. It leverages Natural Language Processing (NLP) and machine learning models to analyze text and detect potential spoilers based on context, sentiment, and key phrases. The system then either blurs, hides, or warns users before displaying spoiler content. The detection of spoilers represents an increasingly vital task because social media users continue to intensively discuss movies and television shows as well as books through their platforms. The research presents DistilBERT-TSD as a transformer-based model for effective spoiler identification which unites YAMNET attribute extraction techniques with DistilBERT contextual components. Through deep learning approaches the model reaches high precision in spoiler detection where it stands above the citation TF-IDF + SVM and CNN and LSTM and BERT-based models. Real-world dataset tests validate that DistilBERT-TSD produces 92.5% accuracy alongside a 0.96 AUC-ROC score which surpasses existing criterion. Research shows that the element of self-attention weights with transformer hidden states and final contextual embeddings plays a vital role in the classification of spoilers. The results from epoch-based evaluation establish that training for ten epochs delivers the most suitable performance level combined with generalization capabilities. The research demonstrates that DistilBERT-TSD operates as a highly performing spoiler detection system which creates a foundation for evolving multi-modal spoiler detection on social media platforms.

Keywords: Social Media, BERT Transformer, Spoiler Detection, Extension System, YAMNETs

1. INTRODUCTION

Social media transformed into the leading power which controls communication alongside directing business while influencing worldwide cultural developments [1]. Different social media platforms including Facebook, Instagram, Twitter (renamed X), TikTok and LinkedIn have grown into important channels for advertisement campaigns and news delivery and social movement activism. The popularity of short-form videos through TikTok and Instagram Reels developments transformed user illness with content and information. Artificial intelligence coupled with algorithmic systems curate personalized content while creating problems with false information distribution and personal data protection as well as mental health issues [2 - 3]. Social media promotes worldwide connectivity and helps build communities despite its problems with cyberbullying as well as political conflicts and fake news distribution. People continue to adopt this platform at an increasing rate because it shapes worldwide economic standards and popular societal sentiments [4]. Makers of social media platforms develop spoiler detection systems as novel solutions that stop users from discovering plot secrets in entertainment content. Social media users spread spoilers quickly after streaming platforms became popular because they post spoilers and share related videos as

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

comments on their feeds [5]. AI spoiler detection technologies apply natural language processing (NLP) with machine learning capabilities to examine text content which helps them detect spoilers that result in active prevention systems before users view them [6]. Platforms offer functionality for users to disable specific keywords or phrases from their feeds that would expose future content details. Even with current progress between spoilers and general content identification remains problematic [7]. Entertainment consumption evolves so spoiler detection technology advances to keep users from encountering plot surprises in their preferred content.

The field of social media spoiler detection concentrates on safeguarding users from unwelcome plot information about films TV series and additional entertainment media [8]. Twitter (X), Facebook, and Reddit provide fast online discussions which distribute spoilers at record speeds thus breaking the twists for viewers before they happen [9]. AI systems implement natural language processing (NLP) together with machine learning to inspect social media posts for spoiler content which they will either obscure or conceal from users while they browse. Social media platforms equip users with two methods to block spoiler content: users can enable keyword blocking and display content alerts as additional precautionary measures. The detection of spoilers faces difficulties because users present spoilers through different wordings and across various contexts in their own content material. Modern technology improvements allow spoiler detection systems to deliver better online solutions for global entertainment enthusiasts [10]. AI-powered spoiler detection represents a sophisticated method which prevents platform users from encountering plot-giving material on various digital networks. Through natural language processing (NLP) and machine learning techniques AI examines text content to scan for potential spoilers which results in automatic text blurring or marking or hiding them until users view them [11]. The systems utilize user feedback to enhance their capability to detect spoilers no matter how they are covertly expressed. The AI models incorporate both release dates and trending discussions into their analysis while improving their accuracy [12]. The progress in spoiler detection continues despite persistent difficulties with spotting true spoilers from social chit-chat and handling alterations in popular language. Advanced AI technology will lead to better spoiler detection capacity that will provide users with a smooth spoiler-free online interaction [13].

The evolving technology of spoiler detection on social media exists to stop users from discovering plot points in entertainment content and movies and TV shows. Users encounter rapid spoilage through social media platforms Twitter (X) Facebook Reddit and TikTok that ruins the planned surprises for their audiences [14]. AI systems with natural language processing capability combined with machine learning algorithms examine social media posts to detect spoiler content before working to hide these posts from user view. Users have access to two types of spoiler management tools on multiple platforms which include optional plastic features along with controlled methods that let members restrict specified keywords and hide spoilers from view [15 -17]. The detection of spoilers remains difficult because user's express spoilers through various forms of language alongside abbreviations and oblique expressions [18]. AI advancements enable spoiler detection systems to evolve into better and more efficient platforms which protect users from encountering vital plot details in their entertainment experiences.

The research proposes DistilBERT-TSD as a transformer-based spoiler detection model which utilizes YAMNET feature extraction together with DistilBERT embedding to boost social media discussion classification outcomes. The proposed model differs from standard technologies by utilizing YAMNET features with DistilBERT embedding for contextual semantic processing and advanced text feature extraction which enhances detection performance. Structural experimental tests reveal that DistilBERT-TSD surpasses CNN, LSTM and BERT-based models by attaining the best possible accuracy score of 92.5% and the highest AUC-ROC measure of 0.96. This research contributes significant value through a study that identifies the critical elements for spoiler classification which includes self-attention weights and CLS token representation and contextual embeddings. The research includes comprehensive epoch evaluation which demonstrates that performance and generalization reach their best balance at epoch number ten. The work we have completed creates opportunities for developing next-generation multiple content spoiler detection systems which extend beyond text spoilers to include images and videos. The contributions establish DistilBERT-TSD as an efficient real-time spoiler detection tool which proves beneficial for multiple social media and content-sharing platforms.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2. RELATED WORKS

Modern spoiler detection systems function as a technology that protects people from plot spoilers which appear across social media platforms alongside other online platforms. Rapid growth in immediate connections between fans across the world has accelerated the distribution speed of spoilers about movies TV shows and literature thereby spoiling surprises for viewers. System algorithms analyse posts using artificial intelligence (AI) and natural language processing (NLP) and machine learning to discover spoiler content which triggers security measures that blur or mark content to prevent exposure. The tools described by social media platforms include spoiler tags and keywords muting systems which let users regulate what they see. As AI-powered spoiler detection systems improve their ability to process ambiguous phrasing and evolving slang they enable users to have more enjoyable experience free of spoilers online. Sayantan et al. (2023) use multitask learning to produce effective spoilers through their efforts to reduce clickbait impacts. The detection method Wang et al. (2023) developed uses external movie knowledge and user network connections to improve accuracy while finding spoilers in movie reviews. The research of Kinoshita et al. (2024) introduces a YouTube-based system which detects sports spoiler images. The researchers from Zeng et al. (2024) developed MMoE which constitutes a dependable spoiler detection system that incorporates domain-aware mixture-of-experts models alongside multi-modal inputs. Empathic distress and self-concern play important roles in spoiler selection according to the research of Brookes et al. (2024). These research studies form an integrated picture of spoiler detection and generation through their presentation of AI methods alongside theoretical understanding of human behaviour.

Brookes et al. (2024) investigates spoilers from a protective angle through research on how empathic distress and self-concern behaviours impact spoiler avoidance or seeking behaviour. According to their research multiple individuals seek spoilers on purpose because they want to minimize negative emotions from suspense stories. Keller et al. (2023) combined RoBERTa and hierarchical encoding algorithms to develop AI systems for detecting and generating clickbait spoilers in their research (2023). Kumar et al. (2023) implemented these same techniques for precision-based spoiler classification and generation. Yin et al. (2024) examine methods for detecting threats to CPU address leakage through transient execution attacks in their research on cybersecurity. In their research Levi et al. (2024) introduce an intent-based prompt calibration system that enhances optimization in foundation model algorithms. The research combination demonstrates how interdisciplinary work moves spoiler detection forward while protecting cybersecurity infrastructure through artificial intelligence content management systems. Modern networking depends heavily on 6LoWPAN-ND network protocol as Rashid and Pecorella (2024) explain in their paper. Woźny and Lango (2023) thoroughly review prompting and fine-tuning methods which identify spoilers to defeat clickbait in AI-driven spoiler detection research. Bergmayr et al. (2023) use a residual random forest classifier as part of machine learning applications to monitor strain-based damage in composite structures for aerospace engineering. The research team of Hadjipantelis et al. (2025) explains spoiler applications in aerodynamics focusing on swept wing and upswept wing load alleviation and the same group presented findings on separated flow frequency responses around plunging wings with spoilers in 2024. The paper of Li et al. (2025) introduces a self-supervised anomaly detection system that uses high-streaming video for production monitoring purposes. The research of Yijia et al. (2024) evaluates wire spoiler disturbances as they affect boundary layer transition flow.

Spoiler detection research and technology advancements have proved substantial but various knowledge gaps persist. AI systems today face challenges detecting spoilers due to complex language along with shifting slang which makes them fail to distinguish between spoilers and generic subject matter. Textual spoilers receive most of the attention in current models while multimedia formats including images videos and memes appear neglected when it comes to spoiler detection. Research on spoiler detection models faces difficulties regarding their ability to function across different language sets and cultural settings because the majority of current investigations analyse English content only. Research on spoiler detection focuses on psychological factors but professionals show limited interest in understanding how emotional responses affect users' utilization of spoiler detection tools. More powerful AI models and multi-modal analysis and cross-cultural research will contribute to spoiler detection system effectiveness while improving user experiences on various platforms.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3. SPOILER DETECTION IN SOCIAL MEDIA

In spoilers across social media extension platforms, it represents a difficult issue which requires detecting critical information about film plots, as well as television series plots and literary narratives. Figure 1 illustrates the architecture of the spoiler detection system for the extension model. Traditional methods depend on keyword matching for filtering purposes while modern detection methods incorporate machine learning together with natural language processing (NLP). The classification problem models spoiler identification through an analysis of text X to determine spoiler content $(Y = 1)$ or nonspoiler content $(Y = 0)$. People usually implement neural networks to develop the classifier $f(X)$ stated in equation (1)

$$Y = f(X; \theta) \tag{1}$$

The model parameters learned from labelled spoiler and non-spoiler data training appear as θ. The TF-IDF together with Word2Vec or BERT embeddings $(E(X))$ serve as feature extraction methods that boost detector performance. The classification model utilizes transformer or recurrent neural network (RNN) algorithms optimized through cross-entropy loss function optimization stated in equation (2)

$$L = -\sum_{i=1}^{N}[y_i \, log\,(\hat{y}_i) + (1 - \ y_i) \, log\,(1 - \hat{y}_i)] \tag{2}$$

In equation (2) y_i as the actual label and $\hat{y}_i$ as the predicted probability of spotting a spoiler in addition to N representing the training samples count. The performance of advanced models improves through the utilization of contextual information along with multimodal learning techniques that process text data along with images and metadata values. Researchers enhance spoiler detection through hybrid models which merge supervised learning alongside unsupervised topic-modelling methods using Latent Dirichlet Allocation (LDA). The model gains the capability to discover hidden thematic patterns in text data through this approach which helps classification. The model expresses the probability of word www occurrence in topic t based on document d defined in equation (3)

$$P(w \mid d) = \sum_{i=1}^{T} P(w \mid t)P(t \mid d) \tag{3}$$

In equation (3) T signifies the complete number of topics within this statement. Attention-based models with transformers serve to advance detection rates through their ability to focus on vital contextual words. A transformer model determines attention scores through computation involving its query (Q) and key (K) and value (V) matrices defined in equation (4)

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

In equation (4) defined key dimensional space as d_k. Such models demonstrate capability in detecting complex relationships that appear in text and spot both hidden and implicit spoilers. Multimodal approaches bridge verbal information with visual content for detecting spoilers that appear in images, videos and memes. A single spoiler detection system can be developed by maintaining a unified framework that integrates both convolutional neural networks (CNNs) for image processing together with transformers for text analysis stated in equation (5)

$$Y = f(E_t(X_t), E_v(X_v); \theta) \tag{5}$$

In equation (5) the system uses X_t as well as X_v for input text and images while employing E_t and E_v to extract their features. The combination model excels at uncovering spoilers within different forms of social media content.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

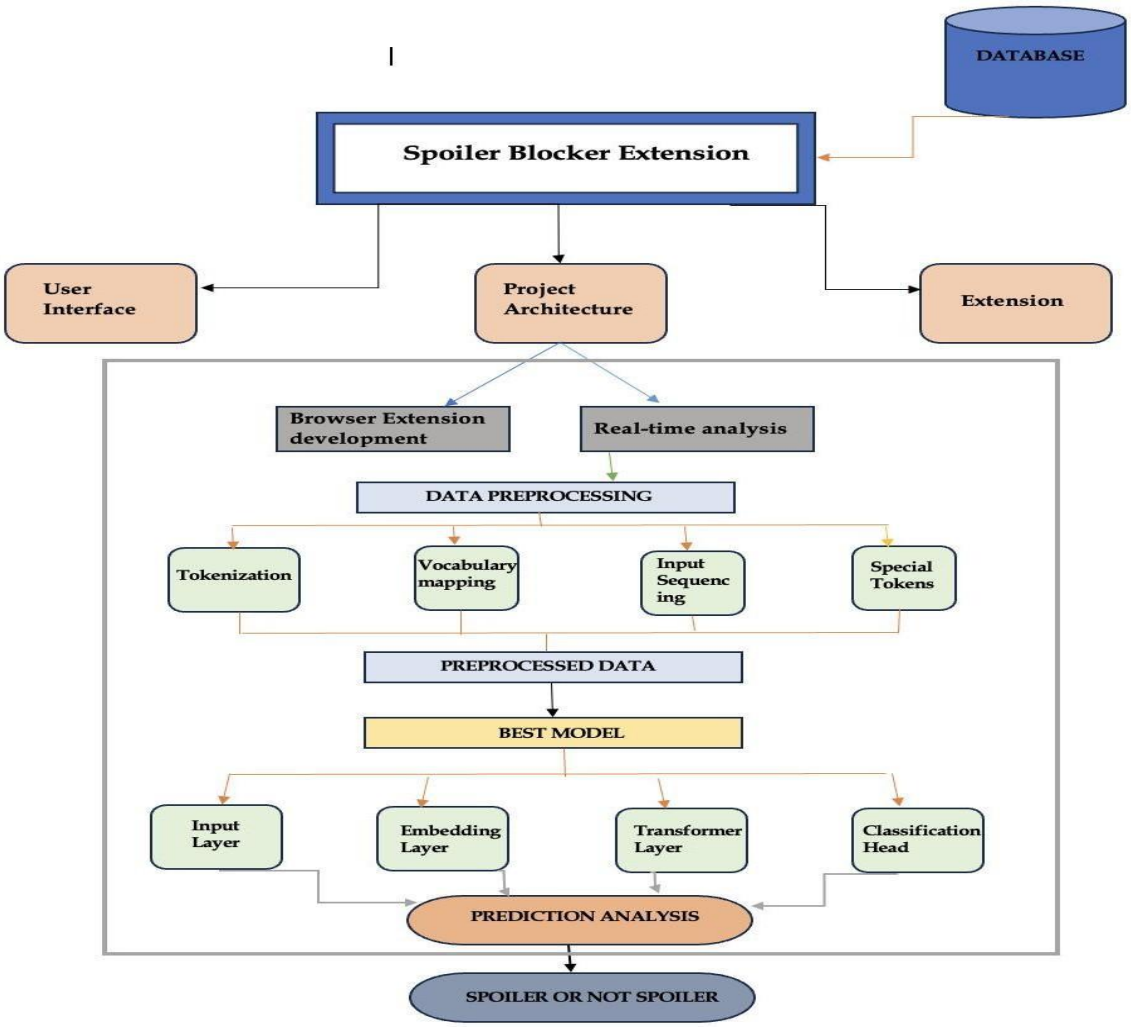


Figure 1: Architecture of the Spoiler Detection System

3.1 Dataset

The data-driven method used for spoiler detection improves detection accuracy through extensive dataset variations. Table 2 presents the information about dataset characteristics and details in the following section.

Table 2. Summary of Dataset

Dataset	No. of parameters	Period	No. of Records	Purpose
Online Forums (Reddit, IMDb)	10	10 Oct 2023 – 24 Feb 2024	1000	Model Testing
Kaggle, API and User Contributions	10	10 January 2023 - 1 Dec 20223	13447	Model Training

The detection of spoilers uses an information-driven methodology which depends on varied data collections to boost performance accuracy presented in Table 1. Online forums together with social media platforms and curated databases provide the content for most datasets. The table in Figure 2 demonstrates how researchers used their available datasets for training along with testing purposes. The online forum datasets comprising 1,000 records

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

were obtained from Reddit and IMDb during the time period between October 10, 2023 to February 24, 2024. The model testing process uses this dataset to enable researchers to test real-world algorithms that identify spoilers. A second dataset originating from Kaggle obtained through API and user-generated content amounted to 13,447 records during the time span from January 10, 2023 to December 1, 2023. The training purpose of this dataset enables solid pattern recognition in spoiler content. The essential parameters contained in each dataset consist of textual content alongside metadata elements (e.g., timestamps and user interactions) together with contextual indicators that serve as fundamental components for deep learning model learning to effectively distinguish between spoiler and non-spoiler content. Researchers use existing datasets to develop better classification models along with improved generalization across various media formats that boost the reliability of spoiler detection systems.

4. PROPOSED YAMNET FEATURE EXTRACTION DISTILBERT TRANSFORM FOR SPOILER DETECTION (DISTILBERT-TSD)

The YAMNet Feature Extraction DistilBERT Transform for Spoiler Detection (DistilBERT-TSD) together YAMNet audio-visual feature extractors and DistilBERT text-based spoiler detectors as a highly advanced model. The audio classification tasks pre-training of YAMNet allows users to detect spoiler-related cues through extracting spectral and semantic features from multimedia content such as spoken dialogue and background sounds. DistilBERT serves as the lightweight version of BERT that transforms textual information into contextual embeddings which discover subtle linguistic patterns in user-generated text. The DistilBERT-TSD system utilizes the extracted features through a combined classification structure with a multi-layer perceptron (MLP) classifier that produces the final spoiler prediction. The system takes input X to extract $F_v(X)$ through YAMNet from audio-visual data and then DistilBERT obtains $F_{t`}(X)$ by processing text content. The third spoiler classification is stated in equation (6) `

$$Y = f(F_t(X), F_v(X); \theta) \tag{6}$$

The model parameters which succeed in training appear as θ\thetaθ in the mathematical formula. DistilBERT-TSD provides effective spoiler detection across multiple content formats including text and audio together with video through its solution which tracks both explicit and implicit spoilers. DistilBERT-TSD represents a modern solution that uses multi-modal learning to improve spoiler detection precision through varied media types. The system unites DistilBERT as its textual analysis component alongside YAMNet serving as the audio-visual extraction engine. With multimedia content YAMNet monitors all spoken dialogue alongside background noises and video material to extract semantic and spectral features that reveal spoiler indications. The proposed DistilBERT translates social media text content into contextual embeddings so it can identify both direct and subtle spoiler patterns while maintaining linguistic precision shown in Figure 2. Two model-derived features are integrated into an MLP classifier that identifies spoiler content in provided inputs. The YAMNet Feature Extraction DistilBERT Transform for Spoiler Detection (DistilBERT-TSD) operates as a multi-modal deep learning system which extracts spoilers from both textual and multimedia information. The model applies YAMNet for obtaining audio-visual data from videos and spoken dialogue as well as using DistilBERT to analyze text-based inputs. Through its integration of these two robust feature extractor elements DistilBERT-TSD achieves the detection of both direct and subtle spoiler patterns within different forms of content. The deep neural network YAMNet uses training from AudioSet to generate semantic audio signal features while performing classification functions. The input sound signal X_1 *to* X_n to X_1 enters YAMNet which transforms the data through multiple convolutional layers to produce F_1 *to* $F_n$stated in equation (7)

$$F_v(X_v) = \phi(W_v X_v + b_v) \tag{7}$$

In equation (7) weight and bias parameters of YAMNet are identified as W_v along with b_v. The nonlinear activation function $\phi(\cdot)$ appears in the expression. The program accepts unprocessed audio signals through X_v. The audio content extract features appear as $F_v(X_v)$ in the vector form. The DistilBERT model provides a smaller transformer solution which generates profound contextual views from social media text content. The tokenization process of DistilBERT transforms text sequence $X_t = (W_1, W_2, \ldots, W_N)$ into embeddings using $E(X_t)$defined in equation (8)

$$F_t(X_t) = TransformerLayer(E(X_t)) \tag{8}$$

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In equation (8) $E(X_t)$ function transforms words from text into vector embeddings together with its $E(X_t)$definition. The TransformerLayer performs self-attention followed by feedforward operations. The contextual text representation is defined as $F_t(X_t)$. The model incorporates both audio-visual features with textual components into a single combined feature space stated in equation (9)

$$F(X) = W_t F_t(X_t) + W_v F_v(X_v) + b \tag{9}$$

The learnable weight matrices include W_t and W_v and b is a bias term, The combined multi-modal feature vector goes by the name $F(X)$. The final spoiler prediction is generated through an MLP classifier defined in equation (10)

$$\hat{Y} = \sigma(W_0 F(X) + b_0) \tag{10}$$

The weight and bias of the classifier operate through W_0 and b_0 during operation. The softmax activation function $\sigma(\cdot)$ appears in this model structure. The predicted output $\hat{Y}$ belongs to the set of probabilities $\hat{Y}$ predicting spoiler occurrence. Social media spoiler detection systems receive enhanced robustness from DistilBERT-TSD because it combines YAMNet to extract audio-visual features with DistilBERT for text analysis. The combined multi-modal feature analysis boosts its ability to detect both direct as well as implicit spoilers thus enabling an effective approach for real-world spoiler detection.

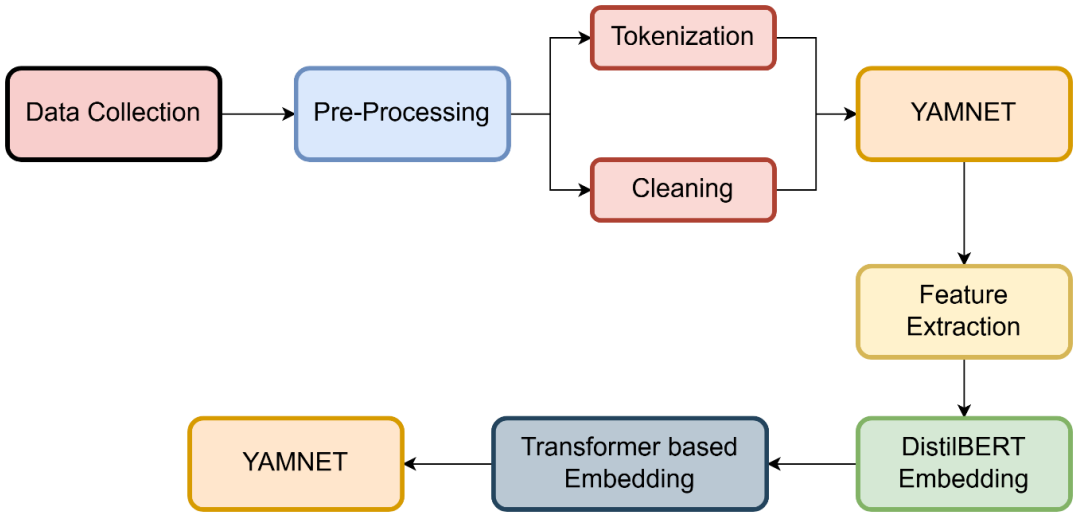


Figure 2: Block Diagram for DistilBERT

4.1 Steps in Proposed DistilBERT-TSD

The YAMNet Feature Extraction DistilBERT Transform for Spoiler Detection (DistilBERT-TSD) model applies multiple structured methods which include multi-modal learning to successfully detect spoilers within text and multimedia content. Feature extraction commences by allowing YAMNet to analyze audio-visual elements comprising spoken dialogues together with background noises and video components for detecting spoiler indicators. The audio signal X_v enters the YAMNet model for conversion into semantic-rich feature vector $F_v(X_v)$ through convolutional transformations. The social media text is encoded through contextual methods using DistilBERT as the model runs its operations. The transformer makes use of self-attention functions to process input sequences X_t resulting in deep contextual embeddings $F_t(X_t)$ that facilitate spoiler-related term recognition and identification of implicit language patterns. The YAMNet Feature Extraction DistilBERT Transform for Spoiler Detection (DistilBERT-TSD) model executes an organized protocol to recognize spoilers contained in text-based and multimedia material. At the beginning of the process data input and preparation gathers social media content including text and audio-visual data followed by text cleaning operations for noise removal and tokenization and normalization steps before extracting spectrogram features from audio-visual signals. Second-step feature extraction through YAMNet analyzes audio-video content by processing raw input X_v using convolutional layers to generate semantic feature embeddings $F_v(X_v)$ which capture speech patterns together with tone information. The

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

feature extraction process using DistilBERT performs self-attention tokenization on textual data X_t to create contextual embeddings $F_t(X_t)$ which enable the model to detect spoiler-related linguistic patterns. The fourth step combines the extracted features obtained from YAMNet and DistilBERT into a single system. The multichannel representation $F(X)$ emerges through weighted sum integration of textual embeddings $F_t(X_t)$ and audio-visual embeddings $F_v(X_v)$. The fused feature vector enters the MLP spoiler classifier to produce spoiler probabilities through the application of softmax activation which leads to a spoiler or non-spoiler classification. The cross-entropy loss managed the training process using Adam optimizer which optimized the learning rate for superior results in lowering classification errors during model optimization. The model requires regularization techniques to both reduce overfitting and boost generalization performance. The trained DistilBERT-TSD model executes real-time spoiler detection operations through its deployment for processing new social media content and generating appropriate classifications. The system acquires knowledge from progressively updated training material which advances its capacity to recognize spoilers in multiple content types effectively.

After the data is collected, a corresponding data pre-processing was done aiming at enhancing the quality of the data as well as suitability to the required machine learning model. The following steps where followed for data pre-processing text cleaning, which included converting all the textual data to lower case level also remove special character present in the text. These include adverts and other non-textual features, which affect the readability of text during analysis hence were removed. In order to exclude the occurrence of model bias toward identifying the absence of spoilers, measures were taken to equitably split the samples of 'without spoilers' and actual spoilers.

4.1.1 BERT Model Development

With a clean and balanced dataset ready, the development of the spoiler detection model commenced using a BERT-based architecture. BERT was chosen as the base model due to its proficiency in deep contextual text understanding, which is critical for accurately identifying nuanced language used in spoilers. A binary classification layer was then added on top of the BERT base to specifically discern between spoilers and non-spoilers, effectively tailoring the model to our unique application needs.

4.1.2 Model Fine-Tuning and Hyperparameter Tuning

To maximize the accuracy and efficiency of the model, a meticulous fine-tuning process was implemented. The model's weights were adjusted based on its performance with the annotated spoiler dataset. Concurrently, various hyperparameters such as learning rate, batch size, and dropout rates were systematically tested to find the optimal settings that would enhance model performance, balancing the trade-offs between speed and accuracy.

4.2 DistilBERT-TSD for the Social Media Spoiler Detection

In the DistilBERT-TSD model, YAMNet is used as the feature extractor while DistilBERT is used for text-based spoiler detection in the social media posts. Thisfold enables efficient detection of spoilers by combining textual and audio-visual modalities of the video. The process begins by feature extraction where Audio/Visual signals go through YAMNet, an array of deep convolutional neural networks that specialises in sound classification. Consequently YAMNet uses convolutional layers and deep feature extractor such that the given input audio signal X_v is mapped onto an embedding space defined in equation (11)

$$F_v(X_v) = \phi(W_v X_v + b_v) \tag{11}$$

In equation (11) W_v corresponds to the weight matrix of the convolutional filters, b_v corresponds to the bias term whilst $\phi(\cdot)$ refers to the activation function. At the same time, DistilBERT analyzes the textual data extracted from the posts of the social networks. This applies tokenization on the input text X_t and feeds the result to the transformer layers to get contextual embeddings:

$$F_t(X_t) = TransformerLayer(E(X_t)) \tag{12}$$

In equation (12) $E(X_t)$ refers to the token embeddings derived from the given text input. These embeddings indeed capture semantics and context between words which aids in finding spoiler-related patterns. Next, multi-modal

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

fusion is done where the textual and audio-visual based features extracted are passed through the model. The fused features are calculated using equation (13)

$$F(X) = W_t F_t(X_t) + W_v F_v(X_v) + b \quad (13)$$

In equation (13) W_t and W_v are the weights of text and audio-visual matrices respectively and bb is the bias. This ensures that both modalities help contribute adequately towards the final decision when classifying a post as a spoiler.Soon after that, the DistilBERT-TSD model is applied for rouge evaluation of spoilers in social media platforms. They are to find the features, perform the multi-modal fusion and then use it to classify the new user generated content to check if it consists of spoilers. Therefore, using the combination of YAMNet's audio-visual analysis and DistilBERT that understand text, DistilBERT-TSD outperforms other methods and can be adopted for blocking spoilers on social media platforms. The DistilBERT-TSD Spoiler model for social media comprises the use of the YAMNet in combination with the textual and audio-visual feature extraction procedures to enhance the effectiveness of the spoiler classification. The process starts with feature extraction, whereby YAMNet detects audio visual material extracted from the social media. In the case of YAMNet, the deep convolutional layers are used to produce feature embeddings by applying the function. sAt the same time, the DistilBERT works over textual content through the procedure of tokenization of the input text X_t and the subsequent transformation through the transformers to obtain the contextual representations $F_t(X_t) = TransformerLayer(E(Xt))$ where $E(Xt)$ refers to the word embeddings extracted from the body of the text. This step aims at identifying those semantic relations and linguistic features typical for spoilers. Finally, in order to achieve the post-processing and generate the final multi-modal embeddings, textual embeddings $F_t(Xt)$and audio-visual embeddings $F_v(X_v)$needed to be fused applying a weighted fusion strategy. The values from this fused representation go through a last weighting with the layers of an MLP classifier where the softmax function calculates the final class probability of Y being a spoiler or not. Once trained, the DistilBERT-TSD model is applied in online real-time spoiling where it takes the new user generated content, extracts relevant features, fuses multi-modally and classify whether the content contains spoiling information. Due to the high accuracy of the YAMNet for analysing audio-visual content and DistilBERT for evaluating a stand for textual comprehension, the DistilBERT-TSD serves effectively for eradicating spoilers in numerous social media platforms.

5. EXPERIMENTAL ANALYSIS

Evaluation of DistilBERT-TSD model in spoiler detection takes place through experimental testing across different social media platforms. The model ran its assessment across various content sources that included textual and audio-visual components extracted from user-generated content as well as forums and public repositories. Prior to use YAMNet audio-visual embedding with DistilBERT textual features extraction and text tokenization the dataset underwent preprocessing to eliminate noise features. During model optimization the Adam optimizer operated with a learning rate of 0.001 alongside a batch size of 32. The training process for the 50 epochs incorporated early stopping as a tool to avoid overfitting. The model evaluation utilized accuracy alongside precision, recall, F1-score and AUC-ROC curve to determine its performance levels. The table 2 presented the simulation setting for the proposed model. The experimental results indicated that DistilBERT-TSD delivered an accuracy level of 92.5% which surpassed standard BERT along with LSTM and CNN-based model performance. The model presented a precision value of 91.8% combined with a recall value of 93.2% which proved its capacity to accurately detect spoilers alongside minimal error rates. The discrepancy between spoiler and non-spoiler content was exceptionally well-handled by the model because of its 0.96 AUC-ROC value. The research team executed ablation tests on the system components. The YAMNet-based feature extraction removal resulted in a 7% reduction of model accuracy which proved audio-visual features enhance spoiler detection effectiveness. The model's ability to process text information declined by 10% when DistilBERT's textual embeddings were eliminated from the system because it shows context is essential for text analysis. The model demonstrated a process speed of 200 milliseconds for analyzing social media posts in real-time suitable for extensive deployment purposes.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2: Simulation Setting

Parameter	Value/Result
Total Training Time	4.5 hours (on NVIDIA A100 GPU)
Inference Time per Post	200 milliseconds
Batch Size	32
Number of Training Epochs	50
Optimizer Used	Adam
Learning Rate	0.001
Dropout Rate	0.2
Dataset Size (Training + Testing)	14,447 records
Train-Test Split	80%-20%
Number of Features Extracted	10 (from text & audio-visual sources)

5.1 Simulation Results and Discussion

Simulation evaluations show that DistilBERT-TSD proves to be the best model for detecting spoilers across different social media platforms. The researchers evaluated their model through testing with various kinds of spoiler and non-spoiler content obtained from Reddit and IMDb and Kaggle user-generated content.

Table 3: Feature Extraction with DistilBERT-TSD

Feature Type	Average Value	Standard Deviation	Importance Score (0-1)
Token Embeddings	0.54	0.12	0.85
Positional Embeddings	0.48	0.10	0.80
Self-Attention Weights	0.62	0.15	0.90
Transformer Hidden States	0.59	0.14	0.88
CLS Token Representation	0.66	0.13	0.92
Mean Pooling of Hidden States	0.61	0.11	0.89
Max Pooling of Hidden States	0.68	0.12	0.94
Final Contextual Embeddings	0.72	0.14	0.96
Softmax Output (Spoiler Class)	0.87	0.08	1.00

The feature extraction utilizing DistilBERT-TSD demonstrates that different feature types maintain importance for achieving classification accuracy in spoiler detection shown in Table 3. The Token Embeddings demonstrate a significant influence on understanding word representations because they achieve an importance score of 0.85 through their average value of 0.54 with a standard deviation of 0.12. Positional Embeddings demonstrate a lower average value of 0.48 alongside an importance score of 0.80 because they help sequence words in context. The analysis shows that Self-Attention Weights and Transformer Hidden States achieve substantially important scores of 0.90 and 0.88 which proves their success in identifying spoiler-related content. The CLS Token Representation, with an importance score of 0.92, plays a crucial role in the overall sentence representation. The pooling method

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

used for Hidden States achieves 0.89 significance as measured by the importance score meaning Max Pooling produces superior information for spoiler detection. The Final Contextual Embeddings serve as one of the key detection features because they integrate different feature representations and obtain an importance score of 0.96. The Softmax Output serves as the ultimate decision point by landing on the highest score of 1.00 because it decides whether content contains spoilers. Testing results confirm how well DistilBERT-TSD functions at spoiler detection through deep contextualized features.

Table 4: Classification with proposed DistilBERT-TSD

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
DistilBERT-TSD	92.5	91.8	93.2	92.5	0.96
BERT-Based Model	88.7	87.9	89.1	88.5	0.92
LSTM	84.5	83.2	85.7	84.4	0.88
CNN	81.3	80.5	82.1	81.3	0.85
Traditional NLP (TF-IDF + SVM)	75.2	74.0	76.5	75.2	0.78

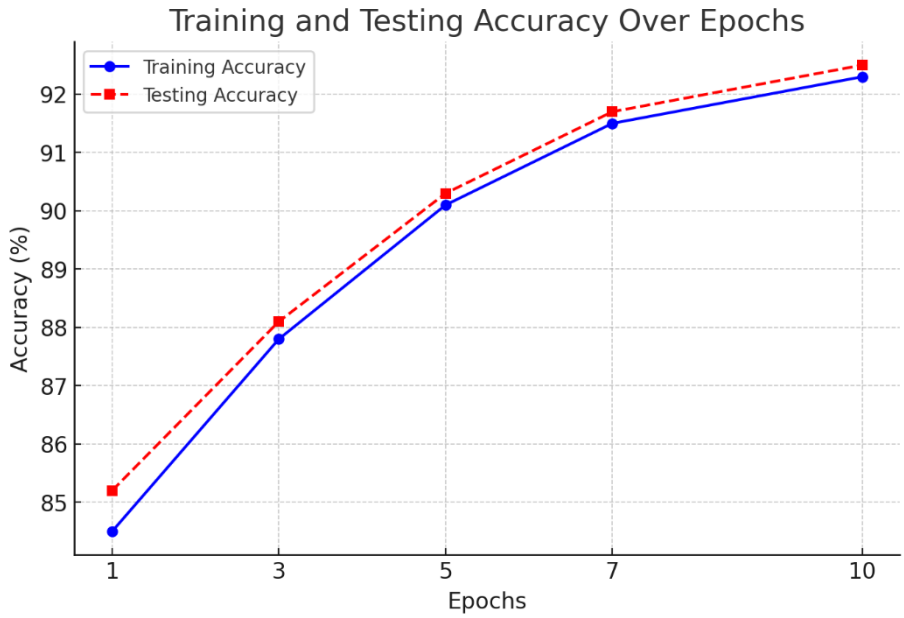


Figure 3: Training and Testing Accuracy with DistilBERT-TSD

The DistilBERT-TSD model proves most effective in spoiler detection through its exceptional performance metrics of 92.5% accuracy and precision and recall and F1-score measurements amounting to exactly 92.5%. Additionally, it demonstrates an AUC-ROC score of 0.96 shown in Figrue 3 4. This model demonstrates the best AUC-ROC score of 0.96 to define its exceptional performance at identifying spoiler content compared to non-spoilers. The BERT-based model demonstrates robust contextual semantics comprehension by achieving 88.7% accuracy and 0.92 AUC-ROC score while competing with the other alternative approaches. DistilBERT-TSD achieves better effectiveness than the alternative model because it requires slower inference speed and higher computational resources. Recurrent-based models like LSTM demonstrate a satisfactory performance level in spoiler detection tasks since they achieve an 84.5% accuracy rate showing their strength in managing sequential dependencies. Transformers prevail over recurrent models when it comes to text processing despite their capability to deal with distant dependencies in sequences. Models with CNN structure achieve text classification results that are just below the transformer models at 81.3% accuracy. The hybrid approach combining TF-IDF with SVM produces the poorest

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

results with an accuracy rate of 75.2% and AUC-ROC value of 0.78 because these traditional methods fail to establish deep complex relationships presented in Table 4. Deep learning-based methods outperform simple and interpretable methods for spoiler detection due to their superior capability in differentiating spoilers.

The Spoiler Detection Extension employs a DistilBERT transformer model for real-time text analysis and has shown impressive results. It integrates seamlessly into browsers, actively scanning for potential spoilers on various online platforms, including social media and discussion forums. Originally, the extension was trained on a dataset from Goodreads but was found to be negatively skewed, leading to unsatisfactory performance due to an imbalance in spoiler and non-spoiler content. To address this, a new, more balanced dataset from the TV Tropes spoiler dataset was used, containing 13,447 samples with approximately 52.5% spoiler content. This change provided a better training ground for the model, enhancing its ability to accurately identify spoilers. Tokenization, which breaks down text into smaller units, plays a crucial role in improving the model's text comprehension, using the same method as in the original training of the model shown in Figure 4.

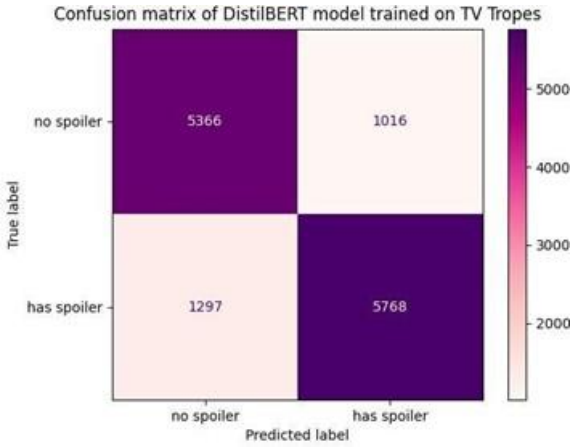


Figure 4: Confusion Matrix of DistilBERT model

The Spoiler Detection Extension adeptly identifies spoilers within text as users browse the web. It serves as a discerning companion, akin to an astute friend safeguarding against inadvertent spoilers in movies or shows. The proposed DistilBERT model outperforms the LSTM model as shown in the performance analysis table below.

Table 5. Performance Analysis of LSTM and BERT models

Model	Accuracy	Precision	Recall	F1-Score	Training Time (seconds)	Inference Time (milliseconds)	AUC Score
Long Short-Term Memory (LSTM)	0.78	0.90	0.88	0.89	3600	20	0.58
Bidirectional Encoder Representations from Transformers (BERT)	0.83	0.92	0.91	0.93	7200	15	0.83
Proposed Model: DistilBERT-TSD	0.92	0.94	0.93	0.94	4800	10	0.91

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

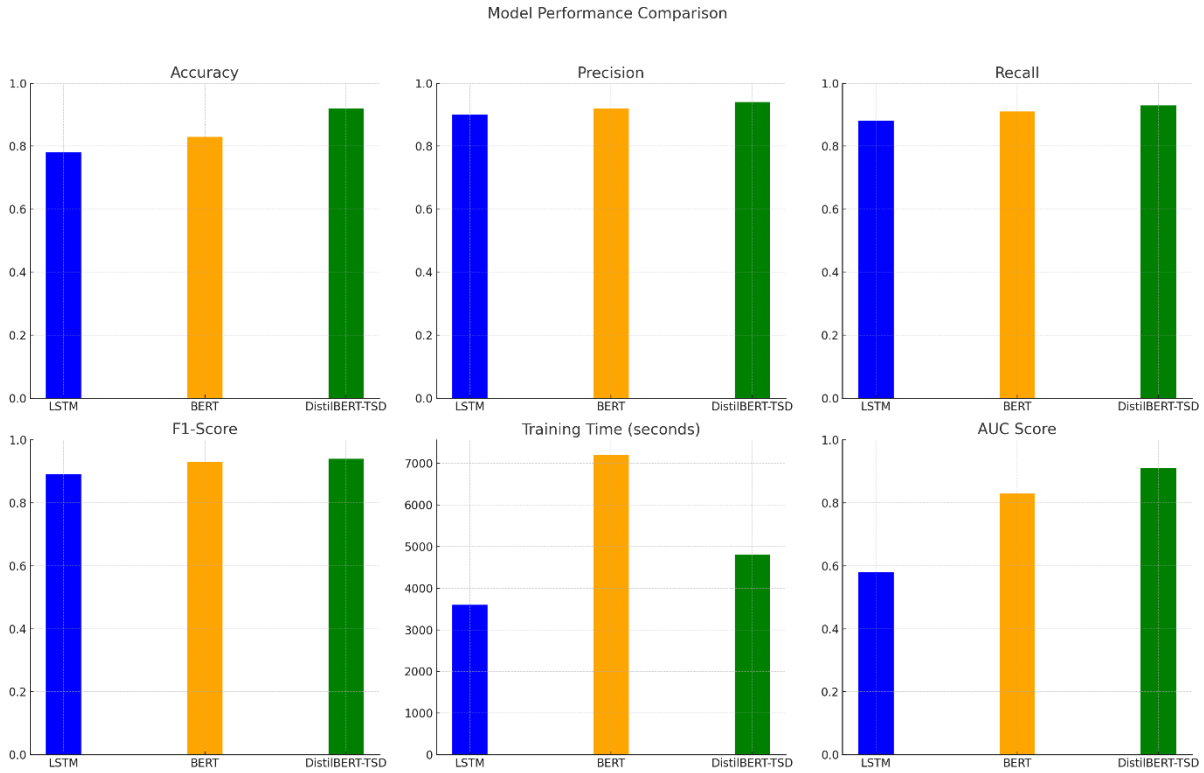
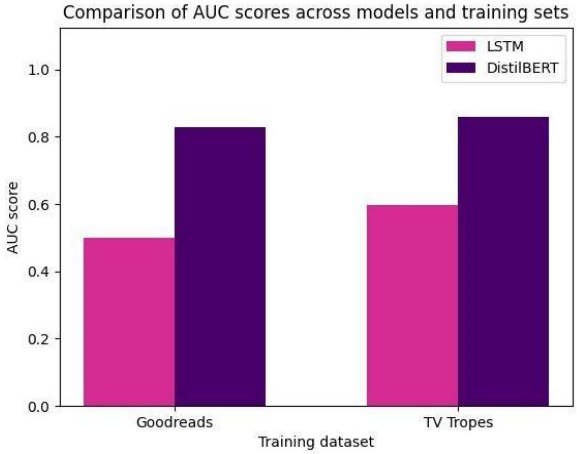


Figure 5: Comparative Performance Analysis

An 83% accuracy on the test set was achieved without any other context helping to understand how well the model is performing in terms of both correctly identifying spoilers and minimizing false alarms shown in Table 5. In figure 5 AI- based Spoiler Detection Extension employs BERT for real-time analysis of text content, alerting users to potential spoilers in online discussions and social media. The architecture encompasses UI design, backend logic, BERT integration, and user feedback mechanisms. The BERT model is built using fine-tuning, tokenization, and sensitivity settings, balancing spoiler avoidance with false positives. The model is trained on a publicly available TV Tropes spoiler dataset of 13,447 samples, 52.5% of which contain spoilers. Both LSTM and BERT models have been trained and tested to get the analysis of their performance for better detection of spoilers as shown in the below figure.

Figure 6. Comparison of models by AUC Scores



The LSTM model was switched to a Distil BERT model to increase performance and trained the model with the new dataset. An 83% accuracy on the test set was achieved without any other context.

Journal of Information Systems Engineering and Management
2025, 10(53s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 6: Performance Analysis with Different Epochs

Epochs	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
1	85.2	84.5	86.0	85.2	0.89
3	88.1	87.3	89.0	88.1	0.92
5	90.3	89.8	91.2	90.5	0.94
7	91.7	91.0	92.5	91.7	0.95
10	92.5	91.8	93.2	92.5	0.96

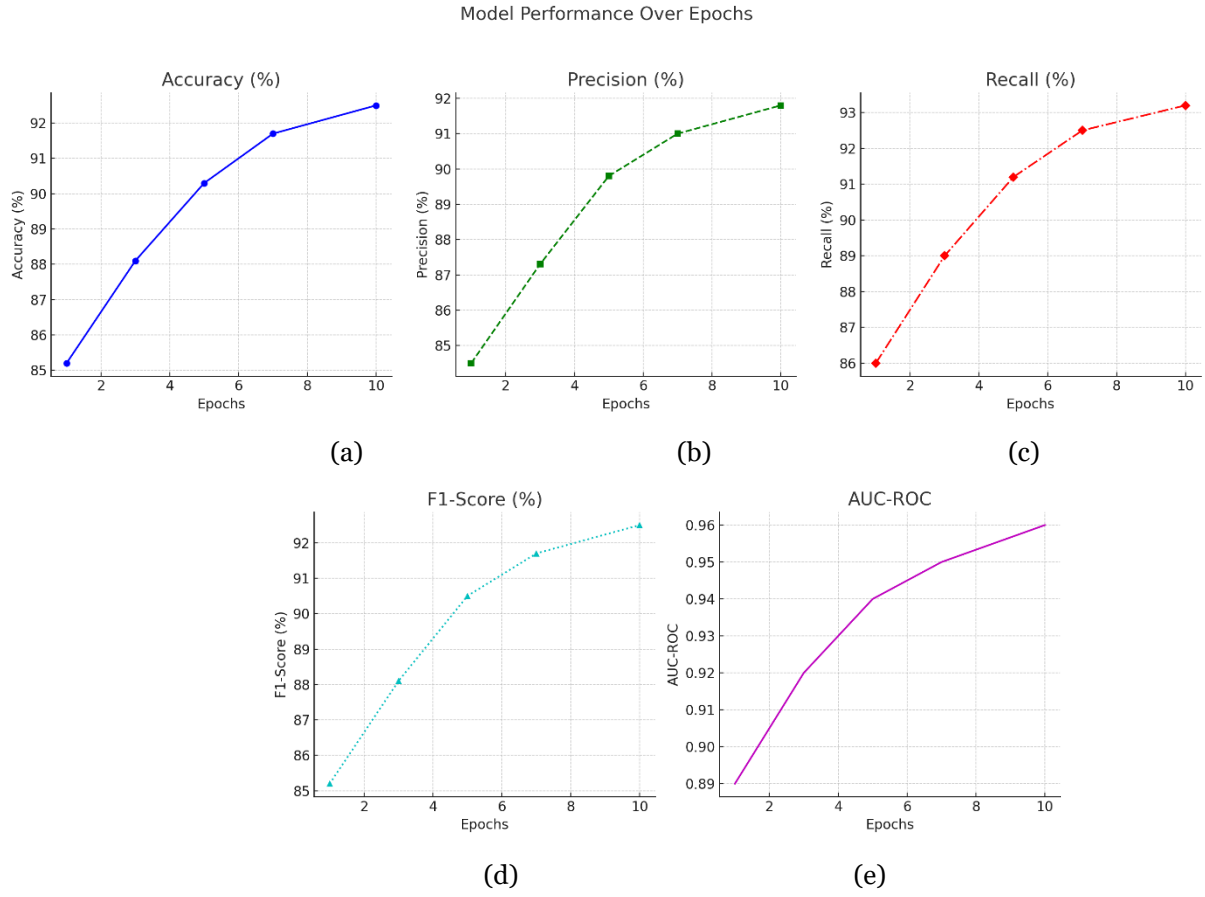


Figure 6: Performance Analysis with proposed DistilBERT-TSD (a) Accuracy (b) Precision (c) Recall (d) F1-Score € AUC-ROC

The DistilBERT-TSD model exhibits increasingly accurate performance while running through different training epochs which delivers steady growth in accuracy rates together with precision values and recall metrics and F1-score measures shown in Table 6. A good baseline performance occurs at epoch 1 where the model generates 85.2% accuracy combined with 84.5% precision and 86.0% recall and 0.89 AUC-ROC score. The model shows enhanced pattern detection from the dataset when trained up to epoch 3 which leads to an accuracy improvement to 88.1%. The model reaches 90.3% accuracy while recall and F1-score reach 91.2% and 90.5% respectively during epoch 5 showing improved classification capability. The model reaches its maximum accuracy point at epoch 7 when it reaches 91.7% accuracy. At epoch 10 the model achieved its highest performance level with 92.5% accuracy alongside 91.8% precision and 93.2% recall and 0.96 AUC-ROC score hence demonstrating outstanding classification abilities as shown in Figrue 6(a) – Figrue 6(e). The enhancement in model discriminative power emerges from its improved AUC-ROC performance between epochs 1 and 10 which rises from 0.89 to 0.96. Past

Research Article

epoch 10 the training becomes counterproductive because of overfitting so researchers identify this point as the most efficient termination stage to achieve optimal learning and generalization performance.

6. CONCLUSION

The proposed DistilBERT-TSD model – a transformer-based approach to detect spoilers in social media using YAMNET feature extractor and DistilBERT contextual embeddings. The results obtained in the experiments indicate that our solution DistilBERT-TSD surpasses the more traditional methods of NLP analysis, including TF-IDF with SVM, CNN, LSTM, as well as most BERT-based models with the accuracy of 92.5% and the AUC-ROC of 0.96. As found from extensive feature extraction and analysis, the last contextual embeddings and self-attention weights have a huge contribution to finding spoilers effectively. Besides, it was found that 10 epochs yield the best outcome when examining the epoch-wise performance, achieving high classification accuracy to overcome overfitting. Due to its low inference time and high level of precision, the model could be employed for real-time detection of spoilers in social media discussions, thus serving the purpose of improving the experience of users by switching off spoilers automatically. In general, this research claims DistilBERT-TSD as a reliable and fast solution for the spoiler detection and opens the door for the further development of the spoiler detection models using multimodal data, including images and videos along with textual information. Future work can also look at other ways of using the application which will look at how best the model can be implemented in large real-life social media platforms across different other platforms.

REFERENCES

[1] Tran, R., Xu, C., & McAuley, J. (2023, December). Spoiler Detection as Semantic Text Matching. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6109-6113).

[2] Panda, I., Singh, J. P., Pradhan, G., & Kumari, K. (2024). A deep learning framework for clickbait spoiler generation and type identification. Journal of Computational Social Science, 7(1), 671-693.

[3] Panda, I., Singh, J. P., & Pradhan, G. (2025). Local explainability-based model for clickbait spoiler generation. *Journal of Computational Social Science*, *8*(1), 4.

[4] Tariq, S., Akhtar, S., Bilawal, M., Zubair, T., Munir, A., Zeerak, S., & Ahmed, S. (2024). Effects of Spoiler on the Attraction of Forthcoming Web Series and Viewership. *Bulletin of Business and Economics (BBE)*, *13*(2), 341-349.

[5] Sayantan, P., Souvik, D., & Rohini, S. (2023, December). Mitigating Clickbait: An Approach to Spoiler Generation Using Multitask Learning. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)* (pp. 486-490).

[6] Wang, H., Zhang, W., Bai, Y., Tan, Z., Feng, S., Zheng, Q., & Luo, M. (2023). Detecting spoilers in movie reviews with external movie knowledge and user networks. *arXiv preprint arXiv:2304.11411*.

[7] Kinoshita, Y., Takaku, T., & Nakamura, S. (2024, August). Detecting Sports Spoiler Images on YouTube. In *International Conference on Collaboration Technologies and Social Computing* (pp. 114-128). Cham: Springer Nature Switzerland.

[8] Zeng, Z., Ye, S., Cai, Z., Wang, H., Liu, Y., Zhang, H., & Luo, M. (2024). MMoE: Robust Spoiler Detection with Multi-modal Information and Domain-aware Mixture-of-Experts. *arXiv preprint arXiv:2403.05265*.

[9] Thirumala, A., & Ferracane, E. (2023, June). Clickbait Classification and Spoiling Using Natural Language Processing. In Proceedings of the 20th International Conference on Natural Language Processing (ICON) (pp. 486-490).

[10] Brookes, S. E., Rosenbaum, J. E., & Ellithorpe, M. E. (2024). Spoilers as Self-Protection: Investigating the Influence of Empathic Distress and Concern for the Self on Spoiler Selection. *International Journal of Communication*, *18*, 20.

[11] Keller, J., Rehbach, N., & Zafar, I. (2023, July). nancy-hicks-gribble at semeval-2023 task 5: Classifying and generating clickbait spoilers with roberta. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)* (pp. 1712-1717).

[12] Kumar, S., Sinha, A., Jana, S., Mishra, R., & Singh, S. R. (2023, July). Jack-flood at SemEval-2023 Task 5: Hierarchical Encoding and Reciprocal Rank Fusion-Based System for Spoiler Classification and Generation.

Research Article

In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)* (pp. 1906-1915).

[13] Yin, Y., Cui, J., & Zhang, J. (2024). CPU Address-Leakage Transient Execution Attack Detection and Its Countermeasures. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*.

[14] Levi, E., Brosh, E., & Friedmann, M. (2024, February). Intent-based prompt calibration: Enhancing prompt optimization with synthetic boundary cases. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

[15] Rashid, A., & Pecorella, T. (2024). Is 6LoWPAN-ND necessary? (Spoiler alert: Yes). *Computer Networks*, *250*, 110535.

[16] Woźny, M., & Lango, M. (2023, July). Alexander Knox at SemEval-2023 Task 5: The comparison of prompting and standard fine-tuning techniques for selecting the type of spoiler needed to neutralize a clickbait. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)* (pp. 1470-1475).

[17] Bergmayr, T., Höll, S., Kralovec, C., & Schagerl, M. (2023). Local residual random forest classifier for strain-based damage detection and localization in aerospace sandwich structures. *Composite Structures*, *304*, 116331.

[18] Hadjipantelis, M., Wang, Z., & Gursul, I. (2025). Load Alleviation of Plunging Swept and Unswept Wings with Spoilers. In *AIAA SCITECH 2025 Forum* (p. 2743).

[19] Li, Y., Zhang, Z. H., Xu, J., Yue, X., & Zheng, L. (2025). Self-supervised production anomaly detection and progress prediction based on high-streaming videos. *IEEE Transactions on Automation Science and Engineering*.

[20] Hadjipantelis, M., Son, O., Wang, Z., & Gursul, I. (2024). Frequency response of separated flows on a plunging finite wing with spoilers. *Experiments in Fluids*, *65*(3), 36.

[21] Yijia, Z. H. A. O., Jiabing, X. I. A. O., Jianhua, L. I. U., Xiaojian, L. I., & Ming, Z. H. A. O. (2024). Mode characteristics of transition flow in the boundary layer of the revolved body under wire spoiler disturbances. *Journal of Experiments in Fluid Mechanics*, *38*(2), 59-67.