Volume 11, Issue 4, April – 2026 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr862

An Explainable Approach to Identifying Clickbait on YouTube: A Real-Time Framework for User Intervention

Ishika[1]; Hardik Mishra[2]; Mohammad Asim[3]

[1]CSE, SSCSE Sharda University, Greater Noida U.P, India
[2]CSE, SSCSE Sharda University, Greater Noida U.P, India
[3]CSE, SSCSE Sharda University, Greater Noida U.P, India

Publication Date: 2026/04/25

Abstract: The widespread use of deceptive engagement tactics, known as clickbait, has weakened trust in digital content, particularly on video-sharing platforms such as YouTube. Although automated systems exist to detect sensational headlines, they mostly operate as opaque models that provide no meaningful feedback to users, limiting their ability to develop long-term media awareness. This paper introduces a real-time framework that detects clickbait while simultaneously offering clear, human-readable explanations within the user's browsing environment. Integrated as a browser extension, the system highlights linguistic patterns such as curiosity gaps and exaggerated phrasing to support informed decision-making.

Keywords: *Clickbait Detection, Explainable Systems, Media Literacy, User Behavior Analysis, Browser Extensions, Deceptive Engagement, Human-Computer Interaction, Video-Sharing Platforms.*

How to Cite: Ishika; Hardik Mishra; Mohammad Asim (2026) An Explainable Approach to Identifying Clickbait on YouTube: A Real-Time Framework for User Intervention. *International Journal of Innovative Science and Research Technology,* 11(4), 1797-1804. https://doi.org/10.38124/ijisrt/26apr862

I. INTRODUCTION

Digital media has expanded rapidly and platforms like YouTube have become a hub of information, entertainment, and education. This has been accompanied by a rise in content that aims to tap into cognitive biases and curiosity gaps, which is often known as "clickbait". This kind of content usually generates a discrepancy between what a headline actually promises and what is the actual informational value of the content which causes user dissatisfaction, decreased trust in the platform.

Automated detection systems have been developed to deal with this issue, but most of them simply provide a binary answer: clickbait or non- clickbait without explanation of the rationale. Although these systems can be very accurate in their classification, users are not informed of what linguistic or semantic signals define deceptive content. It implies that the users will remain passive consumers of the algorithmic warning rather than become active learners who will be able to evaluate the quality of the media independently.

In order to overcome the shortcomings examined in this paper, a real-time system will be proposed that will enable the users to interact with the detection features and will provide instant feedback to them using their web browsers.The system will identify video title and description manipulation patterns which contain overstated information and fabricated time pressure and created gaps of interest and deliver results using simple to understand language. The approach allows users to have a reason as to why they should view certain content because it gives them contextual information that puts them in a position to make effective decisions when watching other content. This publication contributes to the progress of clickbait detection in four ways. The introduction of a hybrid scoring framework which includes four components enables the system to detect deceptive content through the combination of a fine-tuned BERT classifier and rule-based linguistic heuristics and semantic similarity assessment and named entity mismatch detection. The explainability layer produces human-friendly explanations which prioritize all detection outcomes to show users how the system arrived at its decisions. The framework functions as a lightweight Chrome browser extension which operates with a FastAPI inference server to perform instant classification without using post-engagement analytics. The evaluation of 100 YouTube videos which were manually selected and labeled with complete transcript access produced an AUC-ROC score of 0.8642 and a balanced accuracy of 0.8167 and a sensitivity of 0.9500, while bootstrap confidence intervals established the statistical reliability of these findings.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr862

II. LITERATURE SURVEY

Current academic research establishes a comprehensive definition of clickbait which refers to digital content that people make to draw viewers in so they will click on it to boost advertisement revenue [1][3]. YouTube shows this behavior due its monetization system that awards high click and watch time videos[2]. People use clickbait because of its financial advantages, and it also relies on psychological principles of Loewenstein's Information Gap Theory principle [5]. The theory explains how people develop curiosity when they believe that their current knowledge state contains unfilled knowledge gaps which they want to bridge. The cognitive tension in clickbait titles shows how this mechanism works because the titles use incomplete or teasing formats which include You Won't Believe What Happened Next to compel users into clicking to find out.

The current research also takes this theoretical insight a notch further by outlining differences between conventional and fraudulent clickbait. Whereas in the first case of traditional clickbait, exaggeration and hyperbole is used, in the second example of the deceptive clickbait, the headlines appear as authentic news but gives trivial information or misinformation. This change points to the increased maturity of clickbait approaches, in which the deception is more implicit and semantic in nature.

From a linguistic standpoint, a number of lexical and structural patterns have been found to be recurrent in clickbait. These would be excessive punctuation, capitalization, emotion-provoking words and the utilization of trigger words like Shocking, OMG and Unbelievable [4] [6]. Also, the forward-referencing features like use of cataphora, like this, he, etc., without a direct clarification, prompt the user to receive the context clarification by clicking [3]. Other prevalent designs encompass listicles and high-arousal phrasing that escalate curiosity on the part of the users.

Textual analysis alone is not sufficient in the case of a site such as YouTube, where the content is multimodal in nature. Scholars have therefore promoted fusion-based methods where differences between thumbnail images and the real video images are good predictors of deceptive behavior [7]. Detection pipelines are used along with computer vision to consider visual consistency by recognizing objects, analyzing facial expressions, or comparing keyframes and thumbnails [7][8]. Audio-based analysis of transcript-titles has also been investigated to reveal discrepancies between the promises and delivery of a video [8]. On the methodological side, the detection of clickbait has experienced significant development together with the improvement of machine learning. Previous methods used traditional classifiers like Naive Bayes and the Random Forests [9][10], which yielded competitive performance on curated benchmark data.

The steadily growing amount of large-scale data has propelled the deep learning techniques to become the most significant solution of big data processing. Convolutional Neural Networks (CNNs) can be used to extract features (both textual and image) automatically in a machine [6] whereas Long Short-Term Memory (LSTM) networks can be used to learn the sequence of words in a textual context [11]. The current research focus on BERT-based transformer models shows these models achieve better results because they can analyze both ways of understanding contextual information [2][12]. State-Of-The-Art architectures integrate both the semantic and syntactic information so that they can detect it better. An example is the hybrid promoting BERT representation semantics with Graph Attention Networks (GAT) to embed the structural pattern in the phrasing of a clickbait. These techniques enable the models to transfer to new topics whilst being sensitive to patterns of linguistic manipulation.

Other than content-based ones, scholars researched content-agnostic and post- click methods. The Online Video Clickbait Protector (OVCP) system uses the comments and interaction networks left after viewing a video to identify a clickbait based on patterns it detects in its audience response. In addition to that, other metadata capabilities in the form of like-to-dislike ratios, views, and the time of publication are added to introduce more detailed content credibility representation. An ensemble learning method along with SMOTE-based oversampling has been widely used to overcome another limit, namely class imbalance (i.e. there are a few instances of clickbait).

Various systems despite their improvements have a major drawback because they use post-engagement metadata which is not made immediately the user has interacted with the content [11]. This response method does not avert exposure to false information in the first place. Therefore, modern studies have evolved to real-time detection models which exploit pre-click characteristics, e.g., titles, thumbnails and transcripts at the time of upload [3][8]. The devices such as the Clickbait Prevention and Detection Model (CPDM) prove that it is possible to achieve a high level of detection (approximately 95 percent) without had-to-use post-clicks information. In operationalizing these developments, researchers recommend the implementation of detection systems in the form of extensions (utilized by web browsers) or platform-built-in components, which will allow simplified yet real-time sifting regarding clickbait content, when interacting with the user [4]. The role of this such systems is to operate like ad-blockers, practicing prointercession, as opposed to post-factum analysis.

Despite these advances, several important gaps remain unaddressed. Many current models rely on post-engagement signals, constraining their utility for pre-click intervention. Most systems also lack interpretability, returning only numeric scores or binary labels without offering users any rationale for the flagging decision. Deployment in real-world environments remains limited, with most research validated only through offline experiments. Moreover, even the advanced types of deceptive content, such as texts that use linguistic ambiguity, abbreviations, or homophones, are hard to detect. All these restrictions bring to the fore the importance of the detection systems that would be both robust, explainable, and well-integrated into the user-facing interfaces.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr862

Table 1 Literature Survey

Reference	Data / Features	Classification	Method	Strengths (+) and Limitations (−)
Winarto et al.	31,987 YouTube titles; BoW, TF-IDF, word tokenization	Binary (clickbait vs. non-clickbait)	Naive Bayes (Bernoulli), SVM (linear kernel), LSTM	+ 98.53% accuracy (Kernel TF-IDF SVM); outperformed previous benchmarks. − LSTM performed poorly due to dataset simplicity.
Mowar et al.	BollyBAIT (1,000 titles) and MVD datasets; keyframes, audio, titles, thumbnails	Real-time prevention and detection	Stacking classifier (KNN, SVM, XGBoost, NB, LR, MLP) with RF meta-classifier; BERT-Base; ResNet-50	+ Multilingual (200+ languages); works without meta-features. − High architectural complexity due to ensemble stacking.
Varshney & Vishwakarma	987 videos; keyword cues like "Shocking", "OMG"	YouTube clickbait detection	Unified textual analysis approach	+ 98.89% accuracy. − Easily bypassed by altering keywords/styles.
Vadde et al.	Titles, comments, views, likes, dislikes	Binary classification	SVM, LSTM, Random Forest	+ Multi-modal using metadata & engagement features. − Lower accuracy (SVM 96.76%, LSTM 93.79%) than recent 2023 studies.
Khater et al.	21,982 posts (2,495 train; 19,487 validation); textual data	Clickbait detection	SVM, Logistic Regression	+ 79.4% validation accuracy. − No major performance difference between models.
Shang et al.	Network features, metadata (views), comments (Doc2Vec)	Content-agnostic online video protection	OVCP using Adaboost	+ No need to process heavy video files. − Does not use visual descriptors like thumbnails.
Gothankar et al.	8,219 labeled videos; titles, descriptions, comments; Word2Vec, BERT, DistilBERT	Binary classification	Logistic Regression, RF, MLP	+ BERT improved contextual representation; accuracy increased with feature density. − Longer training time for BERT.
Chakraborty et al.	15,000 headlines; 14 linguistic + N-gram features	News & social media clickbait detection	SVM, Decision Tree, RF	+ 89% accuracy (SVM); browser extension developed. − Text-focused; ignores visual/video aspects.
Zannettou et al.	206k video metadata; CNN (thumbnails), sent2vec (headlines)	Semi-supervised YouTube detection	VAEs with mixture models & gating network	+ Works with limited labeled data; multi-modal fusion. − High model complexity (thumbnail deconvolution layers).
Agrawal	2,388 headlines (Reddit, Facebook, Twitter); Word2Vec	Social media clickbait detection	CNN	+ No manual feature engineering; cross-platform generalization. − Small dataset compared to recent studies.
Potthast et al.	2,992 tweets; 215 linguistic & readability features	Automatic clickbait filtering	RF, Logistic Regression, NB	+ First public clickbait corpus; detailed stylistic analysis. − Moderate ROC-AUC (0.79) vs. modern DL methods.
Scott (2023)	UK tabloid headlines and Google Discover feed; analyzed via lexical, semantic, and part-of-speech choices.	Qualitative: "Classic" (formulaic) vs "Deceptive" (traditional news style) clickbait. Linguistic analysis of "Lies" vs "Misleading".	Relevance-theoretic pragmatics analysis that is concerned with explicature and intentionality.	+ Gives a thorough linguistic basis to classify deceptive clickbait as deliberate lies; develops definitions to explainable AI. −Theoretical and qualitative methodology.
Shang et al.	YouTube dataset (500 train/125 test); features from topological/semantic user comment networks, Doc2vec	Binary: Clickbait vs. Non-clickbait videos.	Online Video Clickbait Protector (OVCP): Content-agnostic scheme using Random Walk	+Robust against sophisticated creators by identifying "frustration signals" in audience feedback; content-agnostic. −Requires significant "post-click" user comments to function, limiting its

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr862

	linguistic features, and metadata (likes, views).		on directed graphs, stacked autoencoders	effectiveness for brand-new uploads.
Liu et al.	9,708 WeChat article titles; Semantic (BERT/Bi-LSTM), Syntactic (RGAT/dependencies)	Ternary: Non-clickbait, Malicious-clickbait, and General-clickbait.	MFWCD Framework: Integrates semantic and syntactic info using BERT/Bi-LSTM	+ Effectively captures recurring local syntactic patterns regardless of topic; includes a lightweight model for speed. − Performance decreases with deeper RGAT layers due to over-smoothing.
Winarto et al.	31,987 English YouTube titles; features extracted via Bag of Words (BoW), TF-IDF, and word tokenization.	Binary: Clickbait or Non-clickbait YouTube titles.	Comparison between Kernel TF-IDF Support Vector Machine (SVM), Bernoulli Naïve-Bayes as well as LSTM.	+ Achieved a state-of-the-art accuracy of 98.53% with the Kernel SVM model; identifies specific "baity" lexical markers. − Focuses solely on textual titles, ignoring visual disparity or user engagement statistics.
Sardana et al.	Webis-CBC-16 dataset (2,992 Twitter tweets)	Binary: Clickbait vs. Non-clickbait news.	Random Forest, NB, LR, DT, and SVM.	+Handles data skewness effectively using SMOTE (oversampling); Random Forest.− Multi-layered ensemble techniques increase computational complexity.

III. METHODOLOGY

This section provides detail of the architecture and implementation of the proposed real-time clickbait detection. The pipeline used in the system is a multi-component system comprising of a deep learning-based BERT classifier and a rule based linguistic analyzer, “semantic similarity” module and a named entity mismatch module Everything is packaged as a Chrome browser extension that is supported by a FastAPI inference server.

➢ *System Architecture*

The scheme complies with three level client server architecture. The frontend, which is a Chrome browser extension, is made using HTML5, CSS3 and JavaScript (Manifest V3).The second layer is a FastAPI backend server written in Python and it contains all inference logic. The third one is the machine learning models and the NLP models that were loaded on server start up. Upon visiting the page of a video on the version of YouTube, the extension extracts the video URL and pushes it asynchronously to the predict endpoint. The server runs a pipeline of analysis and provides a structured JSON to the information about the clickbait category, the level of the risk, the confidence scores, and the ability to explain it to human beings. This is made into a color coded overlay on the browser by the extension and this is shown in figure 1.

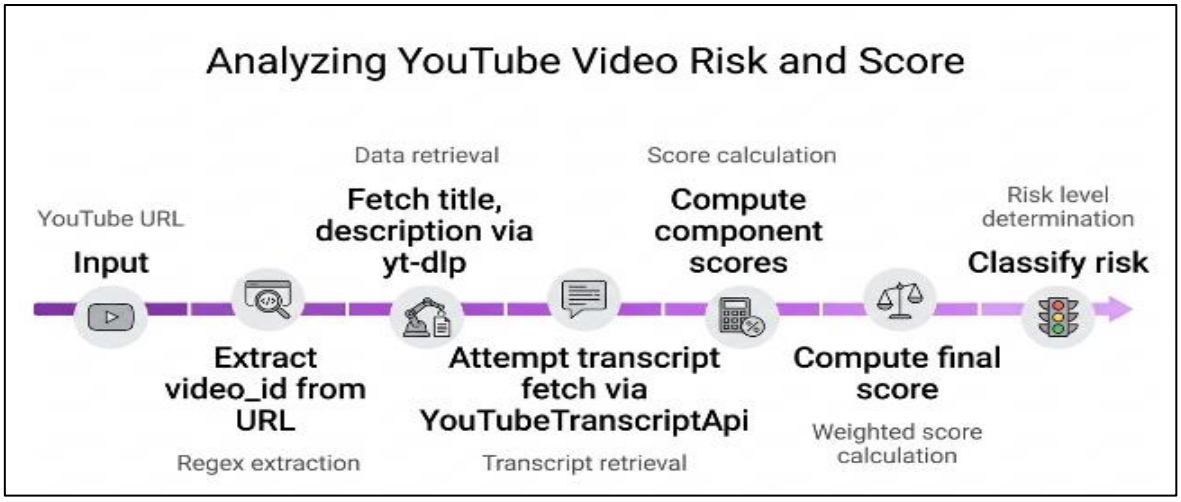

Fig 1 System Architecture

➢ *Input Processing and Video Metadata Extraction*

When it is passed a URL to a video on YouTube, the backend processes the URL using regular expression matching on three recommended formats of URL variants: standard watch URLs (v=), shortened URLs and YouTube Shorts URLs (shorts/). The metadata of the video, namely the title and description are retrieved with the help of the yt-dlp library, which uses the internal API of YouTube to access the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr862

metadata (no authentication is required, however). The title is the key input in all the classification and linguistic analysis processes even when transcript is not available.

➢ *Transcript Retrieval*

The system uses YouTube Transcript Api library to extract video transcript which it needs for comparing title with actual video content. The system implements a two-pass retrieval method which first attempts to fetch English captions with these language codes: (en and en-US and en-GB and en-AU and en-CA) and this fails to find any results the system proceeds to search all available transcripts which exist in every language. The comparison corpus consists of complete transcript text combined with video description which forms a single entity. The system will apply a conservative weighting method which depends on AI classifier outputs and rule-based scores when no transcript exists.

➢ *Multi-Component Scoring Pipeline*

The essence of the methodology is that it has four-component scoring pipelines that is intended to generate the final weighted clickbait risk score in a cumulative way. Weights were determined based on ablation experiments (Table 2). Every element reflects a different aspect of clickbait behavior as explained below and summarized in Table 2.

Table 2 Ablation Study — Performance Comparison (CB = Clickbait Class)

Component	No Transcript	With Transcript	Role
BERT classifier	0.60	0.40	Contextual understanding
Rule-based heuristics	0.40	0.20	Trigger phrase detection
Semantic similarity	—	0.25	Title-content alignment
Entity mismatch	—	0.15	Misleading entity refs.

- BERT-Based AI Classifier: This is a fine-tuned BERT transformer loaded from Hugging Face. The bi-directional attention mechanism of BERT allows the model to identify deep contextual connections between brief title texts and is therefore well able to detect soft sensationalist language styles that cannot be identified via a keyword-based method [14].The model provides a label (CLICKBAIT vs NOT_CLICKBAIT) and a normalised score to 0-100 as the AI score.

- Rule-Based Linguistic Scorer: A linguistic scorer of titles is a rule-based scorer that matches the title to two sets of 14 known clickbait trigger words (e.g., shocks, you will not believe, leaked) and 15 emotional intensifiers (e.g., mind-blowing, terrifying). Other heuristic fines are the extraneous use of all-caps exclamation marks (+2 /mark) and all-caps question marks (+2 /mark), all-caps title (+10), all-caps listicle designs (+5). The aggregate rule score will not exceed 40 in order to avoid being too overwhelming on the final score.

- Semantic Similarity Analysis: In order to recognize title-content mismatch, both the video title and combined description-transcript corpus are encoded into dense vectors using a sentence embedding model (all-MiniLM-L6-v2 of R SentenceTransformers). The cosine similarity between the two embeddings is calculated and percentage of similarity is obtained on a scale of 0-100. The low score in similarity means that the title promises or frames things that are not represented substantively in the rest of the video was content, which is one of the generated characteristics of advanced clickbait that cannot be detected by merely looking at words like words.

- Named Entity Mismatch: Named entities (persons, organizations, geopolitical entities, products and events) are identified in the video title with the help of spaCy en_core web sm model. All of the identified entities are identified in the transcript text. The entity mismatch score is the ratio of title entities that are absent in the transcript. This is known as a particular clickbait element where well-known names or incidents are included in headings only to generate a good number of clicks but is not addressed in the actual video.

➢ *Final Score Computation and Risk Classification:*

The four component scores are added using a dynamic weighting scheme which varies with the availability of transcripts. In case of an available transcript:

$$\text{Final Score} = 0.40 \times \text{AI Score} + 0.20 \times \text{Rule Score} + 0.25 \times (100 - \text{Similarity}) + 0.15 \times \text{Entity Mismatch Score}$$

- *In Cases where the Transcript is Not Available:*

$$\text{Final Score} = 0.60 \times \text{AI Score} + 0.40 \times \text{Rule Score}$$

This design guarantees that deep semantic cues (similarity and entity mismatch) have more collective power (40%) compared to the rule-based score when adequate content exists but will provide a fallback when transcripts are unavailable. The final score maps to one of four risk levels: Low (< 45), Medium (45–59), High (60–74), and Very High (≥ 75). A binary clickbait flag will become true when the final score lies above 55.

- *The Flow is Structured as:*

✓ Input: YouTube URL
✓ Intermediate: title, description, transcript
✓ Output: {score, label, risk level, explanation}

➢ *Explainability Layer*

One of the fundamental design goals is to go beyond qualitative binary categorization and provide clear and actionable explanations to the users. The generating

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr862

explanation method combines the output of each of the scoring parts in a list of human readable justifications in a ranked order. The possible explanation strings can be a find of certain clickbait trigger phrases that exist in the title; a report of an emotionally charged vocabulary; a note when the AI model scores high sensationalism (> 60); a content-mismatch warning when the similarity score is below 30. This set of explanations is sent over the API response and using the extension displayed as an expandable tooltip, which makes sure the user knows not only that a specific title was flagged, but also what signals were used to get to this classification.

➢ *Browser Extension Integration*

The Chrome (Manifest V3) extension is a content script that is inserted into the pages of YouTube videos. Once loaded or visited by the user, the video the script is loaded on, the script will extract the current URL and send it asynchronously to the FastAPI backend. Upon receiving the JSON, the extension presents a risk badge in a color-code fashion (green (Low), amber (Medium), orange (High), and red (Very High)), and the title of the video. The entire description list and singular component scores appear on an expandable space whereby one can read further about the rationale of detection without the need to exit the YouTube interface.

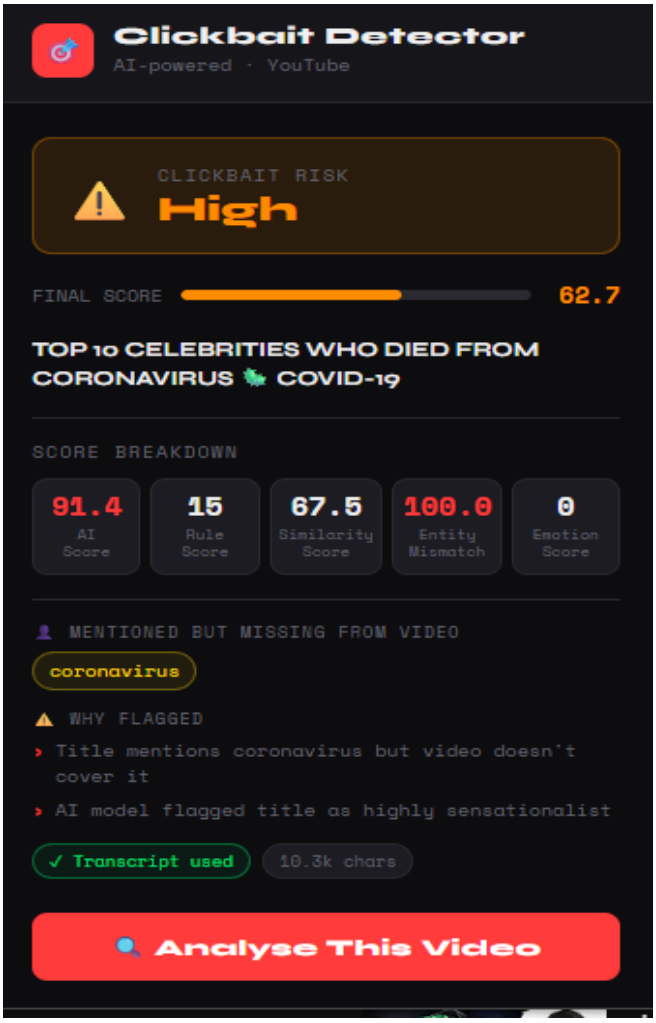

Fig 2 Browser Extension

IV. RESULTS

➢ *Dataset and Experimental Setup*

The evaluation dataset comprises 100 manually curated and labelled YouTube videos spanning 13 content categories, including education, health, science, technology, entertainment, and lifestyle. Of these, 40 videos were annotated as clickbait and 60 as non-clickbait. Although dataset is quite small, it was manually curated to ensure high quality annotations and balanced representation across categories. Transcripts were successfully retrieved for all videos, allowing the full-transcript weighting scheme (BERT×0.40 + Rule×0.20 + Similarity×0.25 + Entity×0.15) to be applied uniformly. Table 3 indicates confusion matrix that was performed over manually annotated clickbait and non-clickbait dataset of 100 youtube videos with their transcript.

Table 3 Confusion Matrix — Full Pipeline, THR = 0.38 (N = 100)

	Pred: Not CB	**Pred: Clickbait**
Actual: Not CB	41 (41.0%)	19 (19.0%)
Actual: CB	2 (2.0%)	38 (38.0%)

At the selected decision threshold of 0.38, the pipeline correctly identifies 38 of 40 clickbait videos while producing 19 false positives from the 60 non-clickbait samples, yielding a specificity of 0.6833. This precision-recall trade-off is an intentional design choice for an advisory tool: prioritising recall ensures that nearly all deceptive content is surfaced for user awareness, accepting a moderate rate of false alerts. Under the lower optimised threshold of 0.205, the system shifts further toward high recall, flagging a larger share of true positives at the expense of additional false positives; this configuration may suit users who prefer comprehensive coverage over conservative flagging.

➢ *Explainability and User-Centric Insights*

A distinguishing feature of the proposed system is its explainability layer, which accompanies every prediction with a ranked set of plain-language justifications. Rather than presenting a bare numeric score, the interface communicates specific signals such as the presence of clickbait trigger phrases, emotionally charged vocabulary, named entities mentioned in the title but absent from the video transcript, and low semantic alignment between the title and the content. This openness transforms what would have been a black-box classifier into a decision-support system: users have the opportunity to see directly the signals that triggered a flag, the persuasion strategies being used, and make more conscious decisions about whether to interact with the content. This design decision is supported by previous studies, which have indicated that human readable explanations are more effective in making users more trustful and possessing critical awareness when assessing flagged content. [15].

➢ *Pilot User Study*

The researchers measured the explainability layer through a pilot study that tested the browser extension with 50 users who watched YouTube videos. The recruiters used convenience sampling to select participants for their study.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr862

The age distribution showed that 66% of participants belonged to the 18 - 22 age group while 29% of participants belonged to the 22 - 30 age group and 5% of participants belonged to the age group of 30 and above. The sample size represents well the target user population as 76 percent of the participants indicated they use YouTube on a daily basis. Participants completed four five-point Likert ratings (1 = strongly disagree; 5 = strongly agree) which measured their perceived accuracy and explanation utility and trust and overall satisfaction. The system received high effectiveness ratings for clickbait detection because participants found that plain-language explanations (e.g., trigger-word alerts and entity-mismatch warnings) helped them understand each detection better.

When asked whether they would adopt the extension in everyday browsing, 61% answered Yes, 37% answered Maybe, and only 2% said No, yielding a cumulative positive intent rate of 98%. These findings provide early empirical support for the value of human-readable explanations: participants perceived them as trustworthy and practically effective, corroborating the core design premise that transparent, reasoned feedback enables more deliberate media-engagement decisions than unlabelled binary alerts.

V. DISCUSSION

The results indicate that the backbone of a BERT-based system, a rule-based heuristic, and transcript-sensitive semantic modules provide an effective and deployable system. The entire pipeline has an AUC-ROC of 0.8642 and a balanced accuracy of 0.8167, which means that it has a high level of discriminative ability between clickbait content and non-clickbait content. The sensitivity of 0.9500 is especially useful in an advisory context, where most of the misleading information is detected before the user can interact.

Positive Predictive Value (PPV = 0.6667) and high Negative Predictive Value (NPV = 0.9535) indicates strong reliability in identifying non-clickbait content. Also, Matthews Correlation Coefficient (MCC = 0.6267) and Cohen's Kappa ($\kappa = 0.5914$) show significant agreement with ground-truth labels. The performance of categories at the category level differs in domains with high accuracy being found in semantically differentiated categories like lifestyle and cooking as these domains exhibit clearer semantic differentiation whereas informal and ambiguous language in other categories like education reduces classification accuracy.

The main constraint of this research is the small sample size (n=100), which limits the statistical generalization of the obtained findings. Though bootstrap confidence intervals are more robust up to some extent, a bigger and a more diverse dataset is needed to confirm the performance of the model in a more comprehensive way. The moderate specificity (0.6833) suggests a bias towards false positives, which is a deliberate trade-off in advisory systems that favors recall. The next wave of work should be on threshold calibration and expansion of data to enhance accuracy without affecting detection sensitivity.

VI. CONCLUSION

This paper has introduced a real-time, explainable clickbait detection system for YouTube, delivered as a Chrome browser extension. The system integrates a fine-tuned BERT classifier with a rule-based linguistic scorer, semantic similarity analysis, and a named entity mismatch detector into a unified hybrid pipeline that produces both a calibrated risk score and a set of human-readable justifications for each prediction. Live evaluation across 100 annotated YouTube videos with full transcript retrieval demonstrates strong detection performance: AUC-ROC = 0.8642, balanced accuracy = 0.8167, sensitivity = 0.9500, and F1 = 0.7835 on the clickbait class. By surfacing the concrete linguistic and semantic signals behind each detection, the explainability layer transforms the system from a black-box classifier into a decision-support tool that encourages users to develop independent critical-media skills over time. This positions the system as a practical tool for real-time clickbait detection in real world browsing environments.

FUTURE WORK

Several directions need additional research. The evaluation needs expansion to a more extensive multilingual annotated corpus which will enhance reliability for each category while validating the complete transcription process across different content types. The research requires a longitudinal user study which will examine whether users develop permanent clickbait recognition abilities through recurrent interaction with the extension. The existing static rule-based lexicon can be enhanced using embedding-based soft matching techniques to capture paraphrased clickbait expressions. The system needs current multimodal state-of-the-art model benchmarks which will show its position in relation to existing research in the field. The project needs to extend its existing framework to new platforms which include social networks and news aggregators while developing thumbnail-based visual deception detection methods.

REFERENCES

[1]. K. Scott, Deceptive clickbait headlines: Relevance, intentions, and lies, Journal of Pragmatics, vol. 218, pp. 71–82, 2023, doi: 10.1016/j.pragma.2023.10.004.
[2]. Shang, Y. Zhang, Z. Wang, Y. Lai, and Z. Wang, "Online clickbait video detection system independent of video content," Knowledge-Based Systems, vol. 182, Art. no. 104851, 2019, doi: 10.1016/j.knosys.2019.07.022.
[3]. T. Liu, Y. Ke, L. Wang, X. Zhang, H. Zhou, and X. Wu, "Detecting clickbait on WeChat using a deep learning model based on semantic and syntactic features, Knowledge-Based Systems, vol. 245, 2022.
[4]. T. S. Y. Winarto, K. Wijaya, M. A. Faqih, S. Y. Prasetyo, and Y. Muliono, Tackling clickbait with machine learning: A comparative study of binary classification models on YouTube title, Procedia Computer Science, vol. 227, pp. 282–290, 2023, doi: 10.1016/j.procs.2023.10.526.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr862

[5]. N. Sardana, D. Varshney and S. Luthra, "Enhancing clickbait identification with ensemble machine learning methods, Procedia Computer Science, vol. 258, pp. 599–606, 2025, doi: 10.1016/j.procs.2025.04.294.
[6]. R. Gothankar, F. Di Troia and M. Stamp, Clickbait detection in YouTube videos, Dept. Comput. Sci., San Jose State Univ., 2021. [Online]. Available: https://arxiv.org/abs/2107.12791
[7]. P. Mowar, M. Jain, R. Goel and D. K. Vishwakarma, Clickbait on YouTube: Preventing, detecting, and analyzing clickbait using ensemble learning, Dept. Inf. Technol., Delhi Technological Univ., 2021. [Online]. Available: https://arxiv.org/abs/2112.08611
[8]. W. Navid, Z. A. Uzmi, and Z. A. Qazi, ThumbnailTruth: A multimodal LLM method to identify misleading thumbnails on YouTube, Dept. Comput. Sci., Lahore Univ. Manage. Sci., 2024. [Online]. Available: https://github.com/wajihanaveed/ThumbnailTruth
[9]. L. Nofar, T. Portal, A. Elbaz, A. Apartsin and Y. Aperstein, "An interpretable benchmark of clickbait detection and tactic attribution, Holon Inst. Technol., 2025. [Online].Available: https://github.com/LLM-HITCS25S/ClickbaitTacticsDetection
[10]. B. Gamage, A. Labib, A. Joomun, C. H. Lim and K. Wong, Baitradar: A multi-model clickbait detection algorithm using deep learning, in Proc. IEEE ICASSP, 2021. [Online]. Available: https://baitradar.bhanukagamage.com
[11]. Deng, Y. Zhu, Y. Wang, J. Qiang, Y. Yuan, Y. Li and R. Zhang, Prompt tuning in clickbait detection with text summarization, School Inf. Eng., Yangzhou Univ., 2024. [Online]. Available: https://arxiv.org/abs/2404.11206
[12]. R. A. Ginga and A. S. Uban, SciTechBaitRO: Clickbait detection in Romanian science and technology news, Univ. Bucharest, 2024. [Online]. Available: https://www.kaggle.com/datasets/andreeaginga/clickbait
[13]. S. Zannettou, S. Chatzis, K. Papadamou, and M. Sirivianos, The good, the bad and the bait: Detecting and characterizing clickbait on YouTube, in Proc. IEEE Security and Privacy Workshops (SPW), 2018, pp. 63–69, doi: 10.1109/SPW.2018.00018.
[14]. A. Waingankar and P. P. Anusha, Clickbait detection by using emotions to manipulate using machine learning approaches, Journal of Emerging Technologies and Innovative Research, 2024. [Online]. Available: http://www.jetir.org
[15]. Agrawal, Clickbait detection with deep learning, in Proc. 2nd Int. Conf. Next Generation computing technologies (NGCT) 2016, p. 268–272, doi: 10.1109/NGCT.2016.7877426.
[16]. M. Potthast, S. Koepsel, B. Stein and M. Hagen, Clickbait detection, in Advances in Information Retrieval (ECIR), 2016, pp. 810–817, doi: 10.1007/978-3-319-30671-1_72.