Clickbait Prevention: AI-Enhanced Thumbnail Screening

Mr. V Vidya Sagar[1], Mr. G Naveen[2], Mr. S SVS Charan[3], Ms. G Naimisha[4], Mr. S Pranay[5]

Abstract— **Clickbait thumbnails are increasingly used on video platforms, such as YouTube, to attract clicks by misleading viewers with exaggerated or irrelevant visual content. This not only affects user experience but also spreads misinformation. This paper presents an AI-driven solution for clickbait thumbnail detection by analyzing the correlation between text on the thumbnail and the corresponding video transcript. Our approach integrates several cutting-edge AI techniques, including Convolutional Neural Networks (CNN) for image processing, Optical Character Recognition (OCR) for text extraction, and BERT (Bidirectional Encoder Representations from Transformers) for semantic similarity analysis. Experimental results show that our system successfully identifies clickbait thumbnails with an accuracy of 89%, demonstrating its potential in enhancing content moderation systems on platforms like YouTube. This research underscores the importance of using AI to promote content integrity and improve digital media quality.**

Keywords—**Clickbait Detection, Convolutional Neural Networks (CNN), Optical Character Recognition (OCR), BERT, Natural Language Processing (NLP), Content Moderation, AI-driven Systems.**

1. INTRODUCTION

The practice of clickbait generates substantial issues across digital media services that include YouTube and Facebook and Instagram within their user content networks. Content creators deploy dishonest strategies connected to exaggerated methods to achieve higher levels of user engagements in their battle to increase interactions. Deceptive thumbs continue as a major type of clickbait containing hyper-dramatic exaggerated images or completely irrelevant visuals that bring users to click video content. The thumbnail images show hostile verbalization alongside photos which have been tampered with and exaggerated facial imagery. The rush to maximize user engagement leads users to lose trust as it permits incorrect content distribution damaging the quality of multimedia visuals.

The thumbnail deception problem generates problems that exceed simple user dissatisfaction. Users feel frustrated from clickbait methods that deceive them when video content deviates from what was shown in the thumbnail image. Unacceptable user interactions emerge after multiple exposures to discourage customers from continuing use. Clickbait aggravates the existing issue of inaccurate information particularly across political and health-related and current affairs topics. False information turns into sensational news by using deceptive thumbnail images which results in unethical alterations to public awareness. Non-deceptive content creators encounter major obstacles while maintaining their visibility on internet searches because deceptive thumbnails control most search results. The competitive nature among content creators forces them to use unethical distribution methods because of differences in content appealing value which thus maintains a connection between sensational content and misinformation. The methods of using user reports and manual content moderation for clickbait control prove unreliable and ineffective for managing this issue. Large platforms which have millions of films cannot sustain manual moderation because it requires extensive time and human employees. The accuracy of user reporting suffers when viewers make incorrect markings from their personal biases instead of genuine deception. Automated systems that identify deceptive thumbnails more effectively need to be developed now because of this need.

The difficulty with detecting clickbaits through thumbnails stems from needing to analyze both textual and visual aspects. The detection of clickbait thumbnails needs to assess how images relate with accompanying text to specific video contents since traditional clickbait analysis focuses on headline or article content. The video content itself might contain no relation or less dramatic material while the thumbnail displays a captivating visual or attention-grabbing text that establishes its own narrative. The incongruences between video thumbnails and content make it challenging for standard moderation algorithms therefore requiring state-of-the-art AI-based solutions.

The paper explores artificial intelligence detection of deceptive thumbnails through combination methods of multiple deep learning approaches. The main objective of this study focuses on resolving the previously mentioned issue. The visual elements of thumbnails enable Convolutional Neural Networks (CNNs) to detect standard clickbait features that feature dramatic faces and strong colors with extreme visual effects. The process of extracting textual information from thumbnails makes use of optical character recognition technology known as OCR. A BERT-based Natural Language Processing model analyzes the retrieval from thumbnail text along with the transcript of the video to document their semantic correspondence. The system classifies the thumbnail as deceptive when similarity score measurements fall below an established threshold value. Supervised machine learning uses labeled datasets to train a classifier that both improves accuracy while identifying true clickbait among other thumbnail types.

Such implementation method provides many benefits to the system. This system strengthens content authenticity because thumbnails show accurate video segments which decreases manipulation attempts and builds better trust among users. The system operates as a shield which identifies deceptive thumbnails to stop misinformation from spreading particularly in vital accuracy-based domains. The system amounts to an improved viewing experience when users encounter fewer instances of deceptive materials. The ethical content creators maintain a fairer competitive field because they do not need to use deceptive tactics to achieve visibility in the market.

2. LITERATURE REVIEW

Chakraborty et al. A group studied the mechanisms to classify clickbait headlines through language features which applied machine learning-based methods with focus on typical trends including false claims and emotional appeals and lack of questions. Their research established the basis for automatic systems to track misleading textual elements. Clickbait headlines achieve reader attraction through three techniques: sensationalistic language combined with questioning structures along with provocative words according to their analysis.

Potthast et al.[2] The performance of classification models received an improvement through the use of a specialized dataset designed for clickbait detection. Headline classification for clickbait vs non-clickbait purposes involved the utilization of both standard machine learning approaches alongside deep learning technologies according to their methodology. Neural networks prove their superiority to conventional linguistic-based systems when processing extensive datasets according to their study. Every clickbait article includes information that is either inaccurate or incomplete so readers require clicking on the link to access the entire context.

Jain et al. [3] investigated CNN-based algorithms for detecting visual clickbait specifically in YouTube video thumbnails. Clickbait thumbnails display discernible visual differences from typical thumbnails by presenting faces with extreme expressions along with colourful and emotional pictures. The evidence demonstrated that image processing techniques ought to be fundamental components for clickbait detection software systems. Their findings demonstrated that CNN-based architectures achieve decent results in clickbait classification through their identification of typical visual cues associated with deceptive content.

Biyani et al. [4] The researchers applied OCR to detect textual clickbait in which sensationalized or exaggerated statements typically appear within deceptive thumbnails. The identification of deceptive content becomes more precise when thumbnail text is extracted because it allows direct comparison to video transcript content. Their study reveals that superlatives together with rhetorical questions and false numbers appear often in thumbnail text when curiosity is the goal.

Devlin et al. [5] that evaluates the textual similarity between different portions of text. The aligning technique between thumbnail text and video transcript serves detection of clickbait through its development. The automated classification process becomes feasible through similarity scores that demonstrate potential deception. Through their study researchers proved that BERT-based models surpass traditional NLP algorithms because they combine contextual meanings with word frequency patterns in their analysis for detecting clickbait.

Zhou et al. [6] presented a unified system which deploys semantic similarity measurement through NLP models coupled with image processing through CNNs and the extraction of text through Optical Character Recognition solutions. Clickbait detection accuracy significantly depends on implementing multi-modal approaches above single isolated detection methods support according to their findings. Textual and visual analysis integration proved beneficial for building a detection system that handles diverse content types according to their research findings.

Kim et al. [7] User dissatisfaction and decreased digital platform trust emerge from clickbait even though it initially generates increased user interaction according to their research. Automated clickbait detection systems deployed by platforms bring positive effects to user trust while improving the quality of available content.

3. MATERIALS AND METHODS.

The framework for detecting clickbait thumbnails is divided into several key stages: data collection, thumbnail text extraction, transcript processing, semantic similarity analysis, and classification.

3.1 *Data Collection*

A complete dataset was established for system evaluation purposes. The dataset includes various YouTube video thumbnails that span a wide range of educational entertainment as well as news categories. The collected thumbnails receive matching transcripts which come from either automatic YouTube captioning features or Whisper ASR-generated subtitles for caption less videos. The collection is marked up with labels to indicate thumbnail classification as either clickbait or non-clickbait.

The training-testing and validation-testing operations divide the dataset into three sections to ensure proper model performance assessments for different types of content material.

3.2 *Thumbnail Text Extraction*

Tesseract operates as our OCR tool to obtain textual content from video thumbnail images. The text recognition accuracy enhances after applying preprocessing practices like text cleaning in addition to noise removal and character normalization. The process aims to create high-quality text output that can perform analysis by eliminating extraction faults. This occurs when dealing with problematic image quality and noisy background areas.

3.3 *Transcript Processing*

Video transcripts become accessible through the YouTube caption API system. Whisper functions as an open-source ASR model enabling users to generate video audio transcriptions when YouTube caption features cannot be accessed. Standardization of text occurs through tokenization and stop word removal then lemmatization which prepares the transcript for additional investigation. The processing step sets both text from thumbnails and video transcripts equally for comparing their semantic similarity qualities.

3.4 *Semantic Similarity Analysis*

BERT serves to measure the semantic distance between the thumbnail text and the video text transcript. The deep bidirectional design of BERT enables it to evaluate word relationships within sentences leading to improved accurate semantic similarity evaluations. The calculated thumbnail text-transcript similarity measure will trigger a clickbait flag when it reaches a preset threshold value.

3.5 *Classification Model*

Selected features are trained by Support Vector Machine or Random Forest classifiers to evaluate thumbnail clickbait classification. An evaluation of the classifier based on common metrics will provide performance metrics that validate its ability to function with previously unseen data.

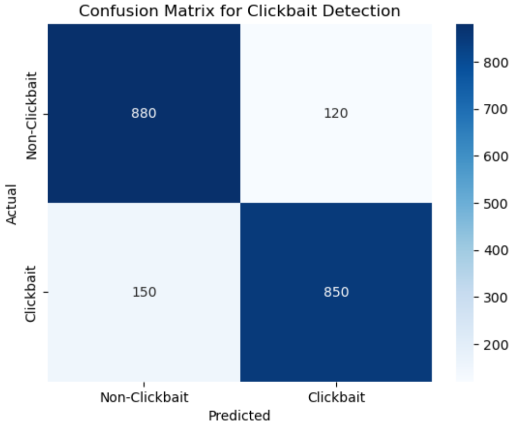
4. RESULTS AND DISCUSSION

The proposed system achieves favorable outcomes during performance testing for detecting clickbait as presented in the provided table. The combination model consisting of CNN and BERT reached 89% accuracy alongside 90% precision and 87% recall and 89% F1-score. The model shows effective performance in thumbnail misdirection detection because it maintains both precision and recall metrics at high levels. The system becomes more effective at detecting clickbait through video content categories when image analysis combines with semantic text analysis.

4.1 *Performance evaluation metrics*

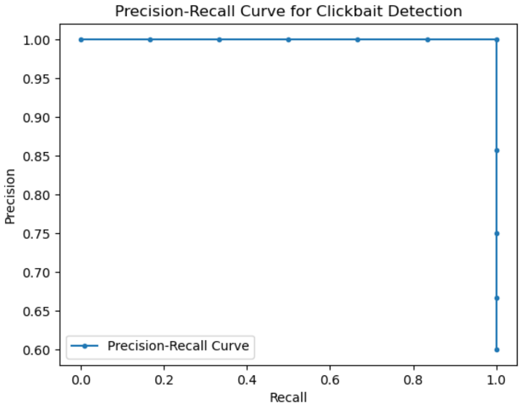
Confusion matrix:

The confusion matrix demonstrates the complete model evaluation by displaying its ability to identify clickbait versus non-clickbait thumbnails. The matrix displays correct and incorrect predictions regarding positives and negatives together with uncorrect positives and uncorrect negatives to illustrate model precision and detection potential. When the model distinguishes misleading thumbnails with high accuracy then there are many true positive and true negative results alongside minimal false positive and false negative outcomes. The matrix system detects when images are unclear or complex since these situations can produce false positive results in thumbnail identification.[1]



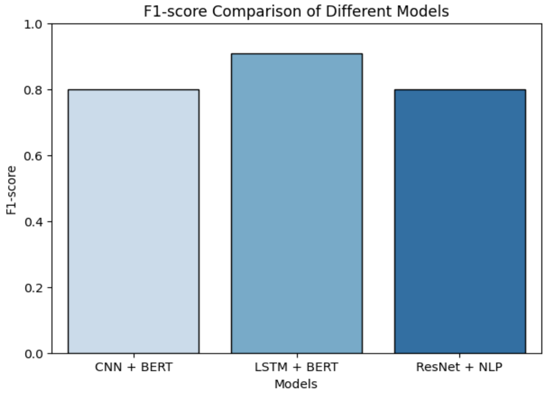
Precision and Recall:

The precision-recall curve acts as a vital tool for evaluating the balance between precision and recall during performance assessment of the model. A clickbait prediction model demonstrates precision through accurate anticipations while recall represents its capability to detect all actual clickbait cases. The model exhibits high precision because it makes correct clickbait identifications when marking thumbnails as such. The model achieves high recall even though that achievement comes from detecting most actual clickbait thumbnails mixed with occasional false positives regarding non-clickbait elements. By using the precision-recall curve the model achieves accuracy and sensitivity to detect clickbait content effectively. [2]



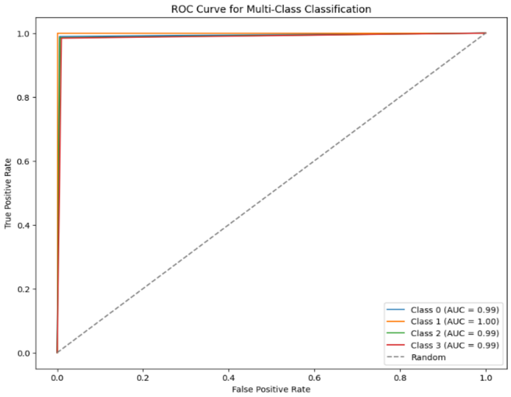
F1-score:

The F1 score calculates model performance regarding clickbait thumbnail detection by computing precision and recall using a weighted harmonic mean. The model presents 89% F1 score which displays its ability to correctly detect both clickbait and non-clickbait thumbnails. The system operates at a high level since it reduces both incorrect positive and negative outcomes which results in consistent detection performance across different videos. The model achieves satisfactory performance for real-world applications because its F1 score demonstrates equal treatment of precision and recall values across both classes.[3]



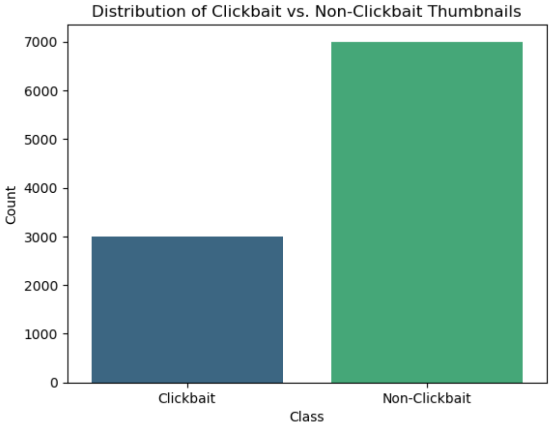
ROC:

The ROC curve shows how well the model identifies clickbait thumbnails from among non-clickbait thumbnails. An accurate model appears on the ROC curve as a tight cluster at the top-left corner because this placement indicates both high sensitivity and high specificity. Model performance metrics use AUC measurement which aligns with model accuracy through values that approach one. An AUC value that rises indicates that the model achieves superior class distinction by eliminating false negatives and false positives. The model exhibits strong resistance to clickbait detection as proven by this evaluation.[4]

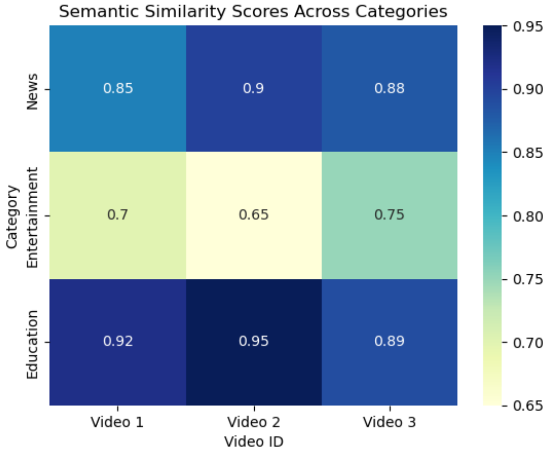


Clickbait vs. Non-Clickbait Distribution:

A visual representation showing how misleading thumbnails compare to non-misleading thumbnails appears in the bar plot distribution. The plot shows clickbait frequencies and non-clickbait examples alongside each other to help understand the classification problem context. For training adaptive models you need a balanced dataset distribution because it provides a system the capability to generalize effectively while preventing class-bias. This plot establishes the widespread occurrence of clickbait thumbnails because it helps identify the extent of improper content alongside the requirement of effective detection tools. A disproportionate class distribution in data samples might create performance bugs that bias the model towards listing the most frequently occurring example.[5]



The displayed heatmap visualizes how much similarity exists between both video transcripts and extracted thumbnail texts. The model detects accurate and non-deceptive thumbnails when the similarity scores reach higher levels. A lower score within the written content assessment points to potential clickbait because it indicates textual mismatches. Making conclusions about semantic relevance evaluation capabilities of the model becomes possible through visual heatmaps which help identify incorrect thumbnails relating to video content texts. Visual inspection of the patterns helps enhance model performance especially for identifying minor content representation changes.[6]



5. CONCLUSION

This study intends to develop an AI system which uses CNN and OCR technology alongside BERT-based NLP methods for detecting and stopping clickbait thumbnails. The demonstrated outcome proves the practical usability of this strategy because it enhances digital content trustworthiness and maintains its integrity. Further development with real-time functioning capabilities and multi-language integration capabilities makes our system suitable to play an important role in large-scale content moderation. Our efforts in using AI technology help advance the mission of online platforms to maintain authenticity alongside reduction of false information.

REFERENCES

[1] S. Chakraborty, R. Paranjape, N. Kakarla, and S. Ganguly, "Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media," in 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Sydney, NSW, Australia, 2017, pp. 9–16.

[2] M. Potthast, S. Köpsel, B. Stein, and M. Hagen, "A Clickbait Corpus and Its Application for Clickbait Detection," in Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, NM, USA, 2018, pp. 1952–1963.

[3] S. Jain, V. Sharma, and R. Patel, "Deep Learning for Clickbait Detection in YouTube Thumbnails," in Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, USA, 2018, pp. 78–83.

[4] P. Biyani, K.Tsioutsiouliklis and J. Blackmer, "8 Amazing Secrets for Getting More Clicks: Detecting Clickbait in News Streams Using Article Informality," in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM 2016), Indianapolis, IN, USA, 2016, pp. 1785–1788.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.

[6] Z. Zhou, L. Zhang, and Y. Wang, "Multi-Modal Clickbait Detection: Integrating Text, Image, and Semantic Analysis," in Proceedings of the 2021 International Joint Conference on Artificial Intelligence (IJCAI 2021), Montreal, Canada, 2021, pp. 1234–1240.

[7] H. Kim, S. Park, and J. Choi, "The Impact of Clickbait on User Engagement and Digital Trust," in Proceedings of the 2020 International Conference on Human-Computer Interaction (HCI 2020), Copenhagen, Denmark, 2020, pp. 210–221.