

The Labeling Dilemma: Study of Clickbait Datasets and their Methodologies

Avinash Shrivastava, Anamika Gupta, Aayush Arora, Anjali Tomar *

Abstract. In today's digital ecosystem, there is a dominance of attention driven content platforms that promote sensationalism over informational quality. These platforms use various means to manipulate users. Clickbait is among them. It often uses misleading or exaggerated headlines to lure people to click on a link. This leaves them in discontent as the promises are never met. The aim is to gain user engagement by either routing them to a page with lots of advertisements that, in turn, boost their revenue or simply spreading misinformation. This necessitate the development of automatic clickbait detection models. This article serves as a systematic review of the work done in this domain, focusing on two important areas: existing datasets with their labeling techniques, and the evolution of various clickbait detection models from ML to DL to new pre-trained language models such as BERT and RoBERTa. This paper aims to serve academic researchers and industry professionals seeking an overview of clickbait detection methods with particular emphasis on ground truth datasets generation and their labeling strategies.

Keywords: clickbait detection, datasets, labeling methodologies, machine learning, natural language processing

1. Introduction

Clickbait is a form of Internet content, either text or visuals, that is specifically designed to lure users to click that link and view, read, stream, or listen to the linked piece of online content, typically being deceptive, sensationalized, or otherwise potentially confusing. The aim is to get more engagement so as to generate more revenue. It is similar to "agro" (derived from "aggravation", a Korean term), which is used to refer to content that deliberately employs eye-catching (provocative, peculiar)

*Department of Computer Science, Shaheed Sukhdev College of Business Studies, University of Delhi, Delhi, India

language or images to capture audience attention. Unlike clickbait, "agro" is not necessarily void of completeness, fairness, accuracy, or deception. ([13]) Clickbait exploits tricks like the curiosity gap and emotional hooks to lure audiences to click and dive deeper and deeper. This not only manipulates user behavior but also undermines trust in online content. This has paved the way for a new, yet significant, domain of research, clickbait detection.

Impact of clickbait content

As discussed, clickbait manipulates the user's understanding by cognitive and emotional triggers. It may lure them to click on a link that may redirect them to a malicious page. Thus, it poses a risk to identity, privacy, and even financial security. The quality of information retrieval has also been impacted by clickbait content. Since clickbait contents are often optimized for drawing attention rather than semantic accuracy, they distort the keyword distributions and contextual signals used by search engines for recommendations. The user's intent is often misaligned with the retrieved results. Thus, a sustainable clickbait flagging system can reduce the chances of spread of misinformation and low credibility content and boost trust, saves information retrieval time and prevents cyber fraud.

Motivation

The motivation for this review arises from the realization that building high-quality, labeled clickbait detection datasets is still a great challenge. This is because of the time-consuming annotation process, the lack of publicly available exhaustive datasets, and the continuous change in clickbait strategies across platforms; despite the crucial need for such datasets in the training and enhancement of detection models, the literature surprisingly lacks a systematic way to explore the approaches to dataset construction, labeling methods, and the effects on model performance, emphasizing the need for an exclusive review on current practices and to direct future research in dataset construction towards elevating the strength of clickbait detection systems.

Research Questions

The structure of this review paper is based on the following questions.

- RQ1: What are some publicly available datasets used for clickbait detection research?
- RQ2: What are the various labeling and annotations schemes employed for labeling the clickbaits?
- RQ3: How have the different models evolved for the detection of clickbait?

- RQ4: What are the various performance metrics used for evaluating the models?
- RQ5: What are the challenges associated with this area, and what is the potential for future development?

Organisation of the paper

Section 2 addresses the first research question (RQ1) by listing the available clickbait datasets. Next, the labeling techniques and annotation schemes are elaborated upon in section 3 (RQ2). Further, the models employed are discussed in section 4 (RQ3) and performance metrics are mentioned in section 5.1 (RQ4). Challenges in this area and future scope are discussed in sections 6 and 7 respectively (RQ5).

2. Datasets and Data Collection

This paper covers more than 50 different studies from 2020 to 2025 that directly address clickbait detection. For papers published between 2020 and 2023, only those with more than 4 citations are included, and for those between 2024 and 2025, we have included all relevant papers published up to July 2025. Papers that do not employ machine learning or deep learning models for clickbait detection or have fewer than 4 citations are not included. We have thoroughly examined the datasets researchers have used to build models. For simplicity, we have categorized the datasets into three groups based on how they are acquired, i.e., the source: 1) downloaded, 2) web-scraped, and 3) A Mixture of the above, i.e., partially downloaded and partially web scraped. Table 1 summarizes the datasets obtained entirely via direct download from repositories such as Kaggle, GitHub, and official challenge websites. Table 2 lists datasets collected exclusively through web scraping from social media platforms and news aggregators. Table 3 details datasets built using a combination of direct downloads and web-scraping techniques, often leveraging APIs or custom crawlers, to augment manually provided annotations.

2.1. Prominent Datasets

In clickbait detection research, the Webis Clickbait Corpus 2017 is one of the most widely used datasets, consisting of 38,517 tweets from 27 top US news outlets over a 150-day window, each rated on a four-point scale by five crowdworkers for both regression and classification tasks [44]. Another popular Kaggle dataset offers 32,000 headlines from BuzzFeed and Upworthy with binary clickbait annotations that make comparative testing easier [56]. Paranjape’s GitHub repository provides 32,519 entries with metadata and reproducible labels for standardized experimentation across models [27]. For multimodal analysis, the Amitness Clickbait corpus has 31,987 YouTube video titles, collected through the YouTube Data API, providing support for studies

Table 1. Summary of Downloaded Clickbait Detection Datasets

Reference	Size	Source & Access
[41], [17], [1], [20]	~20K–80K tweets	Clickbait Challenge datasets (Bauhaus University)
[51]	32K + 38.5K posts	NYT, BuzzFeed (Kaggle); Twitter (Zenodo)
[53]	1K samples	Kaggle – Aman Rai dataset
[56]	100K headlines	BuzzFeed (Clickbait); Reuters, NYT (Genuine)
[30]	32K items	GitHub – Paranjape dataset
[2],[10]	32K headlines	Kaggle (BuzzFeed, Upworthy)
[58]	~81K headlines	DL-Clickbait, CND, SC-Clickbait (EN+ZH)
[27]	7.8K (CLDI), 10.3K (WCC)	Instagram (Harvard Dataverse), Twitter
[7]	10.5K tweets	Twitter (Arabic)
[59]	~32K titles	GitHub – YouTube titles

bridging headline detection and video content [35].

2.2. Datasets Characteristics and Limitations

In the current fast-paced digital world, datasets for clickbait detection show enormous variability in scale, language, and modality ([26]). It is worth noting from the above tables that numerous sets now consist of over 100,000+ examples across platforms ranging from Twitter and YouTube to local news websites, in languages such as Arabic and Telugu, and thus provide extensive topical and cultural coverage ([35, 7]). Annotation schemes vary from fine-grained multi-point scales that reflect the level of clickbait intensity to binary labels that are appropriate for large-scale web-scraped corpora ([10, 44]). Yet, these corpora tend to be plagued by annotation inconsistency, as the judgments of crowdworkers may differ by region and platform, resulting in noisy labels that prevent model generalization ([44]).

With the passage of time, metadata on items within these datasets changes, thereby rendering these datasets more susceptible to inconsistencies and shift in the feature distributions ([26]). Attributes like engagement counts, timestamps, geolocation tags, and user-provided fields over time can change or go stale, and repeated updates or versioning techniques have to be followed to ensure dataset validity and model performance.

Table 2. Summary of Web-Scraped Clickbait Datasets

Reference	Size	Source
[3] (2020)	32K headlines	BuzzFeed, Upworthy, Wiki News, NYT, Hindu, Guardian
[62] (2024)	38.5K text, 18.4K images	Reddit, Twitter, Facebook, BuzzFeed, Upworthy, Clickbait Challenge
[5] (2021)	54.8K items	Arabic news (Twitter)
[15] (2022)	32K headlines	Popular websites (unspecified)
[35] (2021)	112.7K headlines	Telugu websites: Prajasakti, Guite, Manatelangana
[44] (2018)	38.5K tweets	Twitter
[19] (2021)	6.6K training, 3.2K test	12 Indonesian news sites (via William and Sari, 2020)
[34] (2020)	32K headlines	Clickbait: BuzzFeed, ViralNova, etc.; Non-clickbait: WikiNews
[43] (2020)	2.4K samples	Reddit, Twitter, Facebook

3. Labeling Methodologies and Annotation Schemes

There is a range of labeling methodologies and annotation schemes that are used in clickbait detection research to create ground-truth data. We have focused on this topic because of its high relevance in training a supervised learning model.

Crowdsourced Multi-point Ratings:

This method employs crowdworkers to score every headline on a multi-point scale (traditionally four levels: 0.0, 0.33, 0.66, 1.0), capturing the level of "clickbaitiness." For instance, the Webis Clickbait Corpus 2017 instructed five Amazon Mechanical Turk annotators to assign these values to 38,517 tweets, allowing for regression and classification modeling of clickbait intensity ([44]). This approach is suitable when clickbait is viewed as a continuous phenomenon rather than a binary property. The training of regression models, ranking systems, or threshold-based classifiers benefits from this approach to labeling technique. When budgets are limited or when fast scale labeling is required, this technique is less appropriate as it imposes cognitive load on the annotators.

Binary Crowdsourced Labels: Easier to administer than multi-point scales, binary labeling assignments prompt crowdworkers to indicate whether each item is clickbait (1) or not (0). The majority voting is commonly used, where most annotators' label is the final annotation. This system is employed in a number of Kaggle datasets with 32,000 headlines to enable easy classification tasks ([10]). This approach is efficient, cost effective, and easy to administer, making it ideal for benchmarking models. On the other hand, it oversimplifies ambiguous cases and can discard useful signals, particularly in the datasets where headlines may be at the border of informative and deceptive.

Table 3. Partially Downloaded and Partially Web-Scraped Clickbait Datasets

Reference	Size	Source and Method
[33] (2023)	49.9K (train), 350 (test)	Self-collected news articles; fake headlines generated algorithmically (scraping method not specified)
[16] (2023)	11.5K headlines	9.5K downloaded (Clickbait Challenge), 2K scraped from Reddit
[24] (2021)	8.2K videos	YouTube API (scraped) + GitHub repo (downloaded video IDs)
[57] (2023)	22K (Dataset1), 32K (Dataset2)	Dataset1: Zenodo (downloaded); Dataset2: Clickbait sites like Scoop-whoop, ViralNova, NYT (scraped)
[29] (2023)	15K labeled, 1M unlabeled	Labeled: downloaded; Unlabeled: web-scraped (Mahtab et al., 2023)

Heuristic Rule-Based Annotation: Heuristic labeling uses pre-defined rules—e.g., low semantic similarity between headline and content based on TF-IDF cosine similarity thresholds—to label items automatically as clickbait. These rules are able to generate large-scale annotations without any human effort, though they will likely miss subtle examples of misleading headlines.

Weak Supervision via Labeling Functions: Weak supervision models integrate several noisy labeling functions together to produce training labels. In the case of clickbait thumbnails, functions may evaluate title subjectivity, channel metrics, and thumbnail-content incongruity. The function outputs are combined (e.g., using a generative model) to produce final weak labels, which are subsequently optimized using co-teaching in order to reduce noise ([60]). In co-teaching, two or more models, trained in parallel on different views or subsets of data, and iteratively teach one another by identifying and correcting each other’s mistaken predictions. This technique is less effective when the models are highly correlated, and as a result, they provide little unique information, limiting both the reduction of noise and the label quality.

Expert Linguist Annotations: In specialized or multilingual datasets such as Bangla or Arabic corpora, trained linguists put labels to ensure high-quality annotations. Although expensive and less scalable, expert annotations attain higher consistency and cultural sensitivity, especially for languages with smaller crowdworker pools.

Synthetic Label Generation via Data Augmentation: Certain works utilize generative models to generate synthetic clickbait or non-clickbait examples. Synthetic samples inherit their style labels from the generation, enriching sparse labeled data. A style discriminator subsequently checks label fidelity, ensuring generated examples preserve the desired clickbait attributes ([54]).

Binary Labels from Engagement Metrics: Various user engagement meta-data—such as low dwell time (the short duration users spend on a page before leaving, indicating unmet expectations), click-through rate (the proportion of users who click

a headline or link after seeing it, reflecting initial curiosity), and comment-to-view ratio (the number of comments relative to total views, capturing the level of user reaction)—can be used to infer binary labels for clickbait detection. Content with abnormal engagement patterns signalling user disappointment or swift abandonment is marked as clickbait based on inferring deceit from behavioral cues ([62]). This technique allows us to categorize content efficiently, but it depends heavily on the accuracy of the engagement metrics and the threshold as they may be affected by topic popularity, recommendation algorithms or other contextual factors.

Automatic Headline–Body Incongruence Detection: This scheme builds clickbait labels by combining headlines with incongruous article bodies. Corpora such as NELA employ “unrelated” body content to create negative examples, and subsequently tag the headline-body pair as clickbait. This approach circumvents manual annotation through the use of artificially caused incongruence ([61]).

The landscape of clickbait detection has undergone a significant transformation, evolving from traditional machine learning approaches reliant on explicit feature engineering to sophisticated deep learning models and, more recently, to powerful pre-trained and large language models. This evolution reflects a continuous effort to enhance accuracy, reduce manual effort, and adapt to the increasingly complex nature of deceptive headlines.

4. Models

Early Approaches: Reliance on Hand-Crafted Features

Initial endeavors in clickbait detection primarily utilized conventional machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), Logistic Regression, Decision Trees, Random Forest, Gradient Boosted Decision Trees (GBDT), and k-Nearest Neighbors (k-NN) ([14, 55, 52, 57, 6]). These methods heavily depended on meticulous hand-crafted feature engineering derived from various textual properties ([14, 10, 52, 50, 6]).

Key features extracted often included lexical characteristics, such as word frequencies, presence of specific punctuation (e.g., “!” or “?”), numbers, capitalization, and the use of “5W1H” words (who, what, where, when, why, how) ([40, 10, 52, 34, 59]). Researchers also focused on syntactic features, analyzing the sentence structure, part-of-speech (POS) tags (e.g., nouns, verbs, adjectives, adverbs, pronouns), grammatical patterns, and readability scores (such as formality, complexity, and average sentence/word length) ([40, 58, 16, 23, 6]). Beyond these, sentiment analysis and features related to emotional tone were sometimes incorporated ([56, 34, 37, 50]). Furthermore, the similarity between a headline and its corresponding article content was a crucial indicator, with lower similarity often signaling clickbait ([51, 5, 52, 39, 59]). These traditional models demonstrated competitive performance, with reported accuracies reaching high levels, for instance, up to 98% with SVM or GBDT ([10, 52]) and 91.16% with Random Forest ([56, 52]).

Advancements with Deep Learning Models

The inherent challenge of manual feature engineering and the desire to automatically capture more complex textual nuances led to the adoption of deep learning models ([40, 16, 58, 50, 59]). Models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs) became prominent ([40, 30, 16, 15, 27]). These architectures typically integrate word embeddings—such as Word2Vec (which learns word representations based on local context prediction), GloVe (which captures global word co-occurrence statistics from the corpus), and FastText (which represents words using subword character n-grams to handle rare and morphologically rich terms)—to encode words as dense vectors, enabling models to learn semantic relationships without explicit feature definition ([40, 51, 63, 15, 9]). Furthermore, sentence embeddings (such as Skip-Thought or SBERT) have been employed to capture the meaning of entire sentences, proving particularly effective in morphologically rich languages like Urdu ([40, 16, 35]).

Deep learning models have shown impressive results, with LSTMs (Long Short-Term Memory, which is a type of neural network that can remember important information from earlier words in a sequence) achieving up to 97% accuracy for Turkish clickbait detection ([51]) and 95.031% for English datasets ([14, 57]). For Urdu news headlines, Bi-LSTM models with sentence embeddings reached 88% accuracy ([40]). Specialized deep learning architectures have also emerged, such as Parallel Neural Networks (PNNs) which ensemble CNNs to learn relationships between post text, headlines, and articles ([1]). Another notable approach, MuSeM (Mutual Attentive Semantic Matching), detects incongruence by modeling relationships between original and synthetically generated headlines through inter-mutual attention ([39]). Similarly, the LSACD (Lure and Similarity for Adaptive Clickbait Detection) model combines a "luring degree detection module" and a "similarity computing module" with an adaptive prediction mechanism, aiming to mimic human judgment more closely ([63]). Despite their strengths, these models can be vulnerable to adversarial attacks, where subtle text alterations can mislead the classifier ([8]).

The Rise of Pre-Trained and Large Language Models

The latest frontier in clickbait detection involves the application of Pre-trained Language Models (PLMs) and Large Language Models (LLMs), including BERT, RoBERTa, ALBERT, ELECTRA, XLNet, GPT-2, and ChatGPT ([8, 40, 25, 42, 52]). These models use vast amounts of pre-training data to gain a deep understanding of language. These models can then be tailored for specific downstream tasks like clickbait detection ([25, 58]). PLMs have demonstrated powerful capabilities, especially in multilingual contexts, and can often outperform models trained from scratch ([40, 58, 35]). For instance, BERT, RoBERTa, ALBERT, and ELECTRA models trained on Telugu corpora outperformed existing pre-trained embeddings in binary clickbait detection tasks ([35]).

However, despite their advanced capabilities, LLMs like ChatGPT currently present certain limitations. They have shown "unsatisfying results" in few-shot and zero-shot scenarios when relying solely on headlines for clickbait detection, particularly when compared to fine-tuned PLMs ([58]). Furthermore, LLMs can sometimes "hallucinate answers" or be "far from perfect" in direct applications ([25]). A critical vulnerability across these advanced models is their susceptibility to adversarial attacks, where imperceptible character-level or word-level perturbations can cause misclassification ([8]). This ongoing challenge necessitates further research into developing more robust and generalizable detection systems that can withstand evolving clickbait tactics and adversarial manipulations.

5. Comparative Analysis and Benchmarking

5.1. Performance Metrics

Clickbait detection models are typically evaluated using standard classification metrics, most notably Accuracy, Precision, Recall and F1-score. Accuracy means the number of correctly predicted outputs divided by the total number of data samples. It shows how correct the model is overall, but sometimes it is not very reliable when the dataset is unbalanced. Precision is the ratio of correctly predicted positive samples to the total predicted positive samples. It tells how many predicted clickbait news are actually clickbait. Recall means the ratio of correctly predicted positive samples to all actual positive samples, and it shows how well the model can find clickbait news. F1-score is the harmonic mean of precision and recall, and it is used to balance both values, especially when the dataset is uneven.([40, 25, 55, 42, 52]).

These are often derived from a confusion matrix. It is a table that shows the performance of classification model. It contains True Positive(TP), False Positive(FP), True Negative(TN) and False Negative(FN).

Additional metrics include Mean Squared Error(MSE) and Normalised Mean Squared Error(NMSE). MSE is the average of the squared difference between predicted values and actual values. It is mainly used to measure the amount of error in prediction. NMSE is a modified version of MSE where the error is normalized using the range or variance of the output values, so that results can be compared easily between datasets ([44]).

The Receiver Operating Characteristic (ROC) Curve, Area Under ROC Curve (AUC), and Matthews Correlation Coefficient (MCC) are also used to evaluate model performance. ([40, 42, 5]). ROC shows the relation between True Positive Rate (TPR) and False Positive Rate (FPR) at different threshold values and AUC gives a single value between 0 and 1, where higher value means better classification performance. A higher AUC shows that the model can better separate clickbait and non-clickbait data. MCC is useful when the dataset is imbalanced. It considers all four values of confusion matrix and gives a result between -1 and +1. A value close to +1 means good prediction, while 0 means random prediction. Because of this reason, MCC is

considered a reliable metric for classification problems.

For text generation tasks, Bilingual Evaluation Understudy score(BLEU Score) is utilized ([42]). It checks the similarity between generated text and reference text using n-gram matching. It helps in measuring how close the generated text is to human-written text.

In adversarial learning-based studies, some additional metrics like Attack Success Rate(ASR) and Average Perturbed Rate(APR) are used. Attack Success Rate (ASR) shows the percentage of adversarial samples that successfully fool the model. Average Perturbed Rate (APR) represents the average amount of changes done in the input text during adversarial attack experiments([8]).

For unsupervised approaches such as topic modeling, semantic similarity measures are used to determine how similar two texts are in meaning. Topic modeling is a technique that automatically identifies common topics or themes present in a collection of documents without using labeled data. It helps in understanding whether the headline and the content belong to the same topic or not. Semantic similarity refers to the degree to which two texts convey the same meaning. Common methods include cosine similarity, which measures the angle between two text vectors([20, 21]) and Word Mover's Distance(WMD), which calculates the minimum effort required to transform one text into another using word embeddings. These methods help in evaluating how closely a headline matches the actual content, which is useful for detecting clickbait ([20]).

These methods mainly use word embeddings, which are vector form of words that show their meaning and relation with other words. Word embeddings help the model to understand how words are related even if they are different. Along with this, textual coherence is also checked, which means the logical flow and consistency of the text. If a headline is well connected with the content, then it is considered meaningful, but clickbait headlines usually lack this flow.([35, 20]).

Apart from this, path similarity is also used, which is taken from lexical databases like WordNet. It measures how closely two words are related based on the distance between them in a semantic structure. Lexical databases store words according to their meanings and relationships. This helps in comparing texts and finding similarity between headline and content, which is useful in clickbait detection tasks ([17]).

6. Challenges

The detection of clickbait is achieved efficiently by overcoming some major challenges, mainly concerning the availability and quality of the datasets and the constraints in the model development and performance.

Dataset Availability and Labeling Challenges

Obtaining appropriate datasets is one of the main challenges of clickbait detection research. While a few works have used available publicly sources, e.g., an open-source collection of 31,987 English YouTube video titles ([59]), or a bigger corpus of 32,000 news headings ([44]), there is still a common lack of available large-scale, publicly available, benchmark datasets on various languages and topics ([40, 5, 35,

29]). This has often forced researchers to do the time-consuming task of developing their own datasets. For example, the initial Arabic clickbait headline news corpus was carefully built from 54,893 Arabic news articles taken from Twitter, with features being carefully chosen and labels added by media professionals ([5]). In a similar vein, a Telugu dataset of 112,657 headlines was labeled by five native speakers by hand ([35]). For Romanian, the SciTechBaitRO corpus was created, having nearly 11,000 instances specifically for science and technology news ([23]).

The process of manual labeling itself poses significant challenges. It tends to be subjective in nature, making it hard and time-consuming to ensure consistent annotation. For instance, in developing a youtube dataset, among more than 10000 sampled records, only 1881 were used in the final dataset, annotated by three annotators, it emphasizes the effort needed to acquire high-quality labeled data ([11]). In addition, the consistency of labeling can be affected using automated tools, such as the case of ChatGPT, which can have inconsistencies in its responses due to insufficient human judgment and context awareness ([23]). To reduce reliance on large labeled datasets, some studies have explored unsupervised approaches such as topic modeling using models like Latent Dirichlet Allocation (LDA), which identifies hidden topics by grouping words that frequently occur together, and Latent Semantic Analysis (LSA), which uncovers underlying semantic patterns by analyzing word-document relationships, along with BERTopic, which can detect clickbait without having to pre-label examples ([21, 22]).

Challenges in Model Development and Performance

The development of clickbait detection models has transitioned from basic machine learning to more complex deep learning and pre-trained language models, each with its own challenges. The initial methods were based mainly on machine learning algorithms like Naïve-Bayes, Support Vector Machines (SVM), and Random Forest ([10, 52, 59]). These relied highly on labor-intensive hand-crafted feature engineering ([37]). These attributes were obtained from different text properties, such as lexical patterns, grammatical composition, sentiment orientation, readability, and existence of certain punctuation marks or digits, typically employing methods such as Bag-of-Words or TF-IDF ([17, 63, 10, 35, 34]).

Whereas deep learning models, such as CNNs, RNNs, and LSTM networks, incorporate word embeddings (such as Word2Vec and GloVe) to take advantage of more subtle semantic connections, thereby minimizing the amount of manual feature engineering required ([40, 63, 35, 9]), they present new weaknesses. These sophisticated models have vulnerabilities to adversarial attacks, wherein subtle, sometimes imperceptible, modifications to text can deceive the model into producing incorrect predictions ([8]).

The recent developments include Pre-trained Language Models (PLMs) and Large Language Models (LLMs) like BERT, RoBERTa, ALBERT, ELECTRA, GPT-2, and ChatGPT ([25, 58, 42, 35]). Although they are mighty enough, particularly in the multilingual setting, LLMs today have "unsatisfying results" in few-shot and zero-shot cases while solely depending on headlines for detecting clickbait, especially when compared to fine-tuned PLMs ([58]). Additionally, LLMs such as ChatGPT have been reported to at times "hallucinate answers" or be "far from perfect" ([25]), showing

lapses in direct use for this task.

Generalizability is a key ongoing concern. It is still difficult to train one algorithm that works best on various news domains because they have distinct textual architectures ([4]). Additionally, whereas work on the English language has been quite extensive, there remains less for other languages, especially low-resource languages ([40, 5, 35, 29]). Detection of clickbait in the early stages is also quite challenging because there is limited information during the initial phase of content spread ([50]). All these factors combined together increase the complexity of developing solid and universally workable clickbait detection systems.

7. Future Scope

Despite significant progress in clickbait detection, the dynamic nature of deceptive tactics necessitates continuous innovation and refinement of detection models ([26, 55, 2, 15, 6]). A primary future direction involves developing more robust detection systems that are resilient to adversarial attacks, where subtle manipulations can currently mislead classifiers ([8, 36]). Incorporating adversarial training methodologies could build frameworks demonstrating greater resilience and improved precision of classification algorithms ([8]). Furthermore, balancing high accuracy with model interpretability is crucial ([2]), particularly for deep learning approaches, to enhance comprehensibility and broader applicability across diverse fields and datasets ([2]).

Current models often rely primarily on textual features ([34]), yet clickbait frequently incorporates other elements ([2, 42]). Future efforts should focus on integrating multimodal features, including analyses of images, videos, URLs, and user comments alongside text, to provide a more comprehensive understanding of deceptive content ([51, 25, 2, 42, 21]). The creation of larger, more diverse, and more meticulously annotated datasets across various domains and platforms is crucial for training and validating these advanced models ([25, 5, 10, 21, 29]). A significant gap remains in multilingual clickbait detection, especially for low-resource languages like Telugu, Romanian, Arabic, and Urdu ([40, 26, 42, 7, 23]), indicating a need for systems capable of processing content in various linguistic contexts ([40, 42, 35, 23]). This specifically includes collecting more Arabic content to enhance analysis ([5, 11, 7]).

The evolution of clickbait detection will continue with the integration of sophisticated deep learning architectures ([14, 2, 48, 21, 6]), such as Hierarchical Attention Networks (HAN) which model documents in a hierarchical way by first focusing on important words within sentences and then on important sentences within a document, using attention mechanisms to highlight the most relevant information ([23]), and the exploration of hybrid models that combine different learning paradigms ([63]). Refined word and sentence embeddings, alongside novel attention mechanisms, will be vital for capturing intricate semantic and syntactic nuances ([40, 51, 42, 15, 43]). Crucially, as Large Language Models (LLMs) like GPT continue to advance, their potential for clickbait detection will be further leveraged ([40, 25, 58, 42, 16]), although efforts must focus on mitigating issues like "hallucinations" and improving performance in few-shot and zero-shot scenarios compared to fine-tuned PLMs ([25, 58]).

This includes developing more effective prompt engineering strategies for LLMs ([23]) and exploring domain-adaptive PLMs ([29]).

Research will delve into additional temporal, structural, and pragmatic features to enhance detection accuracy ([45, 5, 10, 35, 43]). This involves moving beyond hand-crafted features to automated approaches ([35]). Refining existing Natural Language Processing (NLP) techniques, such as TF-IDF and sentiment analysis, remains a key area for improving feature extraction ([56, 42, 52, 34, 59]). Extracting more semantic and linguistic features, including idioms and metaphors, remains a key area for improvement ([40, 45, 63, 35, 11]). Beyond simple binary classification, future models could output a "clickbait strength" or probability score, rather than a definitive yes/no, to better reflect the subjective nature of clickbait ([25, 42, 35, 44, 1]). This also involves investigating end-to-end models where processes like synthetic headline generation and semantic matching are combined ([42, 63]). Furthermore, utilizing topic modeling to derive features from various textual elements holds promise for enhancing predictive capabilities ([21, 22]).

Ultimately, the goal is to translate research into tangible solutions ([46, 10]). This includes developing real-time clickbait detection systems capable of classifying headlines as they are published, enabling immediate identification and filtering of deceptive content ([46, 52]). Creating user-friendly applications, browser extensions, or APIs that allow users to instantly check headlines, potentially incorporating feedback mechanisms, will foster a more informed online environment ([51, 46, 10, 34, 18]). The constant evolution of clickbait tactics demands that detection methods be adaptable and regularly updated to maintain their effectiveness over time ([45, 2, 50]). Moreover, combating the prevalence of clickbait should be a community initiative, with open-source contributions encouraging collective efforts ([34]).

8. Conclusion

This review takes a good look at the research out there on figuring out clickbait. It starts by talking about what clickbait actually is, why it is a problem, and why we really need to do something about it. Then, it dives into the main sets of data that researchers use for this, where they come from, and most importantly, how they were labeled – that is the special part of this paper.

The paper also checks out how the ways we detect clickbait have changed over time, looking at both methods that use labeled data and those that do not, along with how well they have performed. It also covers the tough bits in this research area and suggests where future studies could go.

Overall, the paper is hopeful, and aims to help people get a better grasp of what is happening in clickbait detection and excite them to dig deeper, pushing the research forward.

Acknowledgment

The authors acknowledge the support and resources provided by the Department of Computer Science, Shaheed Sukhdev College of Business Studies, University of Delhi.

References

- [1] Adelson P, Arora S, Hara J (2016) Clickbait Detection Using Parallel Neural Networks. *Clickbait Challenge*.
- [2] Adrian FHN, Gunawan AAS, Handradika NN, et al. (2024) Clickbait Detection on Online News Headlines Using Naive Bayes and LSTM. *2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*.
- [3] Ahmad I, Alqarni MA, Almazroi AA, et al. (2020) Experimental Evaluation of Clickbait Detection Using Machine Learning Models. *Intelligent Automation & Soft Computing* 26(6).
- [4] Ahmad I, Yousaf M, Yousaf S, et al. (2020) Fake News Detection Using Machine Learning Ensemble Methods. *Complexity* 2020.
- [5] Al-Sarem M, Saeed F, Al-Mekhlafi ZG, et al. (2021) An Improved Multiple Features and Machine Learning-Based Approach for Detecting Clickbait News on Social Networks. *Applied Sciences* 11(9487).
- [6] Alanazi W, Alshaibani AB, Mohammed BA, et al. (2025) Machine Learning Models to Identify and Classify Clickbait Headlines Accurately. *Journal of Applied Artificial Intelligence* 6(1): 1-17.
- [7] Alhanaya R, Algarawi D, Alharbi B, et al. (2024) Mushakkal: Detecting Arabic Clickbait Using CNN with Various Optimizers. *Journal of Information Technology Management* 16(4): 64-78.
- [8] Bajaj A, Vishwakarma DK (2024) Non-Alpha-Num: A Novel Architecture for Generating Adversarial Examples for Bypassing NLP-Based Clickbait Detection Mechanisms. *International Journal of Information Security* 23: 2711-2737.
- [9] Bhoj N, Dwivedi AR, Tripathi A, et al. (2021) LSTM Powered Identification of Clickbait Content on Entertainment and News Websites. *13th International Conference on Computational Intelligence and Communication Networks*, pp. 30-34.
- [10] Bronakowski M, Al-khassaweneh M, Bataineh AA (2023) Automatic Detection of Clickbait Headlines Using Semantic Analysis and Machine Learning Techniques. *Applied Sciences* 13(4): 2456.

- [11] Bsoul MA, Qusef A, Abu-Soud S (2022) Building an Optimal Dataset for Arabic Fake News Detection. *Proceedings of the 13th International Conference on Ambient Systems, Networks and Technologies (ANT 2022) and the 5th International Conference on Emerging Data and Industry 4.0 (EDI40 2022)*, pp. 665–672.
- [12] Choi B, Cho H, Lee J, et al. (2022) Detecting Agro-Korean Trolling and Click-baiting Behaviour in Online Environments. *2022 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*.
- [13] Choi EB, Kim J, Jeong D, et al. (2022) Detecting agro: Korean trolling and clickbaiting behaviour in online environments. *Journal of Information Science* 50(1): 3-16.
- [14] Chowanda A, Nadia, Kolbe LMM (2023) Identifying Clickbait in Online News Using Deep Learning. *Bulletin of Electrical Engineering and Informatics* 12(3): 1755-1761.
- [15] Churi M, Patil D (2022) Clickbait Detection Using Machine Learning Algorithm. *International Journal of Research and Analytical Reviews (IJRAR)* 9(4): 616–620.
- [16] Cornn K (2023) Clickbait Article Detection Using Deep Learning: These Results Will Shock You!. GitHub repository.
- [17] Daoud DM, El-Seoud MSA (2019) An Effective Approach for Clickbait Detection Based on Supervised Machine Learning Technique. *International Journal of Online and Biomedical Engineering (iJOE)* 15(3).
- [18] Dubey R, Dubey RK, Mishra S, et al. (2024) Identifying False News Using Enhanced Passive Aggressive Classifier Methods. *2024 2nd International Conference on Disruptive Technologies (ICDT)*, pp. 101–106.
- [19] Fakhruzzaman MN, Jannah SZ, Ningrum RA, et al. (2021) Clickbait Headline Detection in Indonesian News Sites using Multilingual Bidirectional Encoder Representations from Transformers (M-BERT). *arXiv preprint arXiv:2102.01497*.
- [20] Ghosh S (2019) *Is it a click bait? Let's predict using Machine Learning*. M.Tech. Dissertation, Birla Institute of Technology & Science, Pilani (Rajasthan).
- [21] Ghosh S, Sen D, Ghosh S (2024) Enhancing Clickbait Detection with Cross-Modal Topic Modeling in Social Networks. *2024 7th International Conference on Information and Computer Technologies (ICICT)*, pp. 289–294.
- [22] Ghosh S, Jhalani S, Oshiro CJ (2024) Topic Modeling-Driven Feature Engineering to Enhance Clickbait Detection in Social Networks. *2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pp. 1–8.

- [23] Ginga R, Uban AS (2024) SciTechBaitRO: ClickBait Detection for Romanian Science and Technology News. *Proceedings of the Third Workshop on NLP for Positive Impact*, pp. 188–201.
- [24] Gothankar R (2021) Clickbait Detection in YouTube Videos. Master's Project, San Jose State University.
- [25] Hu D, Anderies, Chowanda A (2024) YouTube Videos Clickbait Classification Utilizing Text Summarization and Similarity Score via LLM. *2024 6th International Conference on Cybernetics and Intelligent Systems (ICORIS)*, pp. 1–6.
- [26] Jacobo-Morales D, Marino-Jiménez M (2024) Clickbait: Research, challenges and opportunities – A systematic literature review. *Online Journal of Communication and Media Technologies* 14(4): e202458.
- [27] Jain M, Mowar P, Goel R, et al. (2021) Clickbait in Social Media: Detection and Analysis of the Bait. *arXiv preprint*.
- [28] Jha S, Choudhary M, Saxena D, et al. (2021) A Review of Fake News Detection Methods using Machine Learning. *2021 2nd International Conference for Emerging Technology (INCET)*.
- [29] Joy SS, Aishi TD, Nodi NT, et al. (2023) BanglaClickBERT: Bangla Clickbait Detection from News Headlines using Domain Adaptive BanglaBERT and MLP Techniques. *Proceedings of Recent Advances in Natural Language Processing*, pp. 744–754.
- [30] Kachawa F, Bhatt S, Siddique K, et al. (2024) Clickbait Detection Using Deep Learning Model. *Journal of Emerging Technologies and Innovative Research (JETIR)* 11(4).
- [31] Kamatchi CB (2025) Avoiding Clickbait in Social Media using Gradient Boosting Algorithm. *2025 International Conference on Inventive Computation Technologies (ICICT)*, pp. 76–765.
- [32] Kaushal V, Vemuri K (2025) Clickbait - Trust and Credibility of Digital News. *Perception Engineering Group, Cognitive Sciences Lab, IIIT Hyderabad*.
- [33] Kim J, Park S, On B (2023) A Pooled RNN-based Deep Learning Model based on Data Augmentation for Clickbait Detection. *Journal of KIIT* 21(4).
- [34] Lamichhane P, Shrestha K (2020) Implementation of Machine Learning Approach to Detect Clickbaits in Online News. *Preprint*.
- [35] Marreddy M, Reddy OS, Sireesha VL, et al. (2021) Clickbait Detection in Telugu: Overcoming NLP Challenges in Resource-Poor Languages using Benchmarked Techniques. *Technical Report, International Institute of Information Technology, Hyderabad*, (IIIT/TR/2021/-1).

- [36] McCutcheon A, Brogly C (2024) Interrupted time series analysis of clickbait on worldwide news websites, 2016-2023. *arXiv preprint arXiv:2408.06376*.
- [37] Meel P, Vishwakarma DK (2020) Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications* 153: 112986.
- [38] Meel P, Vishwakarma D (2021) Machine Learned Classifiers for Trustworthiness Assessment of Web Information Contents. *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp. 29–35.
- [39] Mishra R, Yadav P, Calizzano R, et al. (2020) MuSeM: Detecting Incongruent News Headlines using Mutual Attentive Semantic Matching. *arXiv preprint arXiv:2010.03617*.
- [40] Muqadas A, Khan HU, Ramzan M, et al. (2025) Deep learning and sentence embeddings for detection of clickbait news from online content. *Scientific Reports* 15(13251).
- [41] Omidvar A, Jiang H, An A (2018) Using Neural Network for Identifying Clickbaits in Online News Media. *arXiv preprint arXiv:1806.07713*.
- [42] Panda I, Singh JP, Pradhan G, et al. (2024) A deep learning framework for clickbait spoiler generation and type identification. *Journal of Computational Social Science* 7: 671–693.
- [43] Patil S, Chauhan H, Koul M, et al. (2021) Detecting and Categorization of Click Baits. *International Journal of Engineering Research & Technology (IJERT)* 9(3): 449–454.
- [44] Potthast M, Gollub T, Hagen M, et al. (2018) The Clickbait Challenge 2017: Towards a Regression Model for Clickbait Strength. *Proceedings of the 2018 World Wide Web Conference*, pp. 759–766.
- [45] Pujahari A, Sisodia DS (2021) Clickbait detection using multiple categorisation techniques. *Journal of Information Science* 47(1): 118–128.
- [46] Qureshi KA, Malick RAS, Sabih M, et al. (2021) Complex Network and Source Inspired COVID-19 Fake News Classification on Twitter. *IEEE Access* 9: 139636–139656.
- [47] Radhi AD, Al Naffakh HA, Fuqdan A, et al. (2024) A comprehensive review of machine learning-based models for fake news detection. *BIO Web of Conferences* 97: 00123.
- [48] Raju GNV, Nyalakanti N, Kambampati P, et al. (2023) Clickbait Post Detection using NLP for Sustainable Content. *E3S Web of Conferences (ICMPC)* 430.
- [49] Rao KP (2022) Fake News Detection using Machine Learning Algorithm. *International Journal of Computer Trends and Technology* 70(3): 16–18.

- [50] Rastogi S, Bansal D (2023) A review on fake news detection 3T's: typology, time of detection, taxonomies. *International Journal of Information Security* 22: 177–212.
- [51] Razaque A, Alotaibi B, Alotaibi M, et al. (2022) Clickbait Detection Using Deep Recurrent Neural Network. *Applied Sciences* 12(504).
- [52] Sardana N, Varshney D, Luthra S (2025) Enhanced Clickbait Detection through Ensemble Machine Learning Techniques. *Procedia Computer Science* 258: 599–608.
- [53] Sekharan CN, Vuppala PS (2023) Fine-Tuned Large Language Models for Improved Clickbait Title Detection. *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*.
- [54] Shu K, Wang S, Le T, et al. (2018) Deep Headline Generation for Clickbait Detection. *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 467–476.
- [55] Singla D, Mapari S, Gupta D, et al. (2025) From Rumors to Reality: Exploring Machine Learning Approaches for Fake News Detection. *2025 International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI)*, pp. 297–301.
- [56] Sisodia DS (2019) Ensemble Learning Approach for Clickbait Detection Using Article Headline Features. *Informing Science: the International Journal of an Emerging Transdiscipline* 22: 31–44.
- [57] Supriya S, Singh JP, Kumar G (2023) Identification of clickbait news articles using SBERT and correlation matrix. *Social Network Analysis and Mining*.
- [58] Wang H, Zhu Y, Wang Y, et al. (2023) Clickbait Detection via Large Language Models. *arXiv preprint*.
- [59] Winarto TS, Wijaya K, Faqih MA, et al. (2023) Tackling Clickbait with Machine Learning: A Comparative Study of Binary Classification Models for YouTube Title. *Procedia Computer Science* 227: 282–290.
- [60] Xie T, Le T, Lee D (2021) Detecting Clickbait Thumbnails with Weak Supervision and Co-teaching. *Machine Learning and Knowledge Discovery in Databases. Research Track (ECML PKDD 2021)*, pp. 549–565.
- [61] Yoon S, Park K, Shin J, et al. (2018) Detecting Incongruity Between News Headline and Body Text via a Deep Hierarchical Encoder. *arXiv preprint arXiv:1811.07066*.
- [62] Yu J, Wang S, Yin H, et al. (2024) Multimodal Clickbait Detection by Deconfounding Biases Using Causal Representation Inference. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10300–10317.

- [63] Zheng J, Yu K, Wu X (2021) A deep model based on lure and similarity for adaptive clickbait detection. *Knowledge-Based Systems* 214: 106714.
- [64] Zuhroh NA, Rakhmawati NA (2020) Clickbait detection: A literature review of the methods used. *Register: Jurnal Ilmiah Teknologi Sistem Informasi* 6(1): 1–10.

Received 18.11.2025, Accepted 30.03.2026