BetXplain: An Explanation-Annotated Dataset for Detecting Manipulative Betting Advertisements on Social Media

MSVPJ Sathvik[1], **Parmitha Vangapandu**[2], **Nishit Rane**[2], **Sathwik Narkedimilli**[3], **Mark Lee**[1], and **Akrati Saxena**[4]
[1]University of Birmingham, UK [2]IIIT Dharwad, India
[3]NUS, Singapore [4]LIACS, Leiden University, The Netherlands
msvpjsathvik@gmail.com, 23bcs092@iiitdwd.ac.in, nishit.rane0201@gmail.com,
sathwik.narkedimilli@ieee.org, m.g.lee@bham.ac.uk, a.saxena@liacs.leidenuniv.nl

Abstract

The promotion of betting applications on social media platforms has increased significantly in recent years. Many of these advertisements use persuasive techniques that may mislead users, encourage risky behavior, and potentially influence users' mental well-being. However, research on the automated detection of manipulative and deceptive betting advertisements remains limited due to the lack of publicly available annotated datasets. In this work, we introduce a new dataset of betting-related advertisements collected from two widely used social media platforms, Instagram and Reddit. The advertisements were manually annotated for manipulative and deceptive advertising practices. In addition to classification labels, the dataset includes human-provided explanations describing the reasoning behind each annotation, enabling research into explainable approaches for detecting manipulative advertising. Furthermore, we analyze the strategies commonly used in betting advertisements and examine how these persuasive tactics may impact users' mental health. The proposed framework can also enable practical applications such as browser plugins that warn users about manipulative betting advertisements and automated web crawlers that help regulatory authorities monitor and detect such promotions online.

1 Introduction

Social media platforms have become major channels for digital advertising, with betting app promotions increasingly prevalent on Instagram and Reddit. These advertisements commonly employ persuasive messaging, promotional bonuses, and attractive marketing strategies to maximize user engagement. However, many betting ads incorporate manipulative or deceptive techniques that can mislead users and encourage risky financial behavior. The aggressive promotion of betting applications has raised significant societal concerns due to its association with financial losses, psychological distress, and gambling addiction. In Telangana alone, over 1000 suicides have reportedly been linked to betting apps (The Indian Express Bureau, 2025; Business Standard, 2025), illustrating the severity of the broader societal problem that motivates this work, rather than describing the scope of the dataset itself, which is limited to English-language advertisements. We note that manipulative betting promotions in regions such as Telangana are frequently written in Telugu, Hindi, or code-switched dialects; extending BetXplain to these language ecosystems is an important direction we leave to future work.

Certain betting platforms employ deceptive operational practices alongside persuasive advertising strategies. Some applications initially permit small withdrawals to build user trust but later restrict or disable withdrawals after larger balances accumulate, preventing users from accessing funds (Reina, 2025; Forcepoint X-Labs, 2026). Such practices constitute digital financial exploitation and raise serious consumer protection concerns. Regulatory challenges further intensify the issue, as betting is restricted or banned in India (Government of India, 1867) and several Gulf countries (Government of the United Arab Emirates, 1987; Government of Saudi Arabia, 2019; Government of Qatar, 2004; Government of Kuwait, 1960) due to risks of financial harm and addiction. Nevertheless, betting advertisements continue to circulate across global social media platforms, bypassing regional restrictions and exposing users to manipulative promotions (Government of India, 2000).

Research on automated detection of manipulative and deceptive betting advertisements remains limited due to the absence of publicly available social media datasets focused on betting promotions. Existing resources primarily analyze gambling behavior or betting statistics and rarely provide explanatory annotations for deceptive content.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4073–4089
November 6-10, 2026 ©2026 Association for Computational Linguistics

To address this gap, we introduce BetXplain, A dataset of Instagram and Reddit betting advertisements manually annotated as manipulative, deceptive, or responsible, with human-written explanations. We further analyze persuasion strategies, sentiment patterns, and mental health risks, and evaluate transformer-based and large language models for detecting manipulative advertisements.

The key contributions of the paper are as follows:

1. This is the first work to develop methods for detecting manipulative betting apps.
2. This paper is the first to propose a dataset on manipulative betting apps with explanations.
3. We have presented a novel analysis of how these betting apps impact mental health and other areas.

2 Related Work

Online gambling and mobile betting platforms have rapidly expanded in recent years, transforming gambling from physical venues to continuously accessible digital environments. Mobile betting applications provide real-time wagering, in-play betting, and personalized promotional notifications that encourage frequent engagement (Gainsbury, 2015; Abbott and Volberg, 1991). This shift has been accompanied by aggressive marketing strategies that rely on targeted promotions, bonus incentives, and time-sensitive offers delivered through push notifications and advertisements (Hing, 2014; Lopez-Gonzalez and Griffiths, 2018). Such strategies often frame gambling as entertaining and profitable while minimizing potential risks. The constant accessibility of mobile platforms further removes traditional barriers such as travel time and social oversight, increasing the likelihood of problematic gambling behavior (Abbott, 2020).

Gambling advertisements often employ persuasive strategies such as urgency cues (e.g., "limited-time offers" and "boosted odds") and reward framing that emphasize financial gains while minimizing losses (Hing, 2014; Newall, 2019; Guerrero-Solé et al., 2018; Binde, 2014). By associating betting with sports fandom and entertainment, these advertisements may normalize gambling and increase participation among vulnerable populations (Hing, 2014). Problem gambling is associated with anxiety, depression, financial distress, and interpersonal conflict, while online environments may further amplify risks through persistent promotional exposure and distorted perceptions of winning probabilities (Petry, 2005; Walker, 1996; Gainsbury, 2015; Dowling et al., 2015; Gorrepati et al., 2025). Although NLP methods and transformer-based models, including BERT, RoBERTa, DistilBERT, ELECTRA, ALBERT, and GPT-4, have achieved strong results in harmful content detection (Davidson et al., 2017; Fortuna and Nunes, 2018; Schmidt and Wiegand, 2017; Vidgen et al., 2021; Devlin et al., 2019; Liu et al., 2019; Sanh et al., 2019; Clark et al., 2020; Lan et al., 2019; Brown et al., 2020), gambling advertisements remain underexplored, with no public dataset providing annotated explanations for manipulative betting promotions (Hing, 2014; Griffiths, 2019). This work addresses this gap by proposing the BetXplain dataset.

Gambling behavior has been widely studied in relation to addiction, cognitive bias, and reward-system dysregulation, with intermittent reward mechanisms shown to activate dopaminergic pathways similar to substance addiction, thereby reinforcing compulsive behavior (Petry, 2005; Gainsbury, 2015). Gambling advertisements further exploit cognitive biases, including the illusion of control, loss aversion, and social proof, thereby increasing engagement and exposure risk, positioning betting advertisements as both linguistic artifacts and behavioral stimuli with measurable psychological impact. Related research on deceptive design identifies widespread use of dark patterns in online systems (Mathur et al., 2019) and manipulative interface taxonomies (Gray et al., 2018). Studies on social media endorsements demonstrate how promotional content can obscure commercial intent (Mathur et al., 2018), while explanation-guided large language models improve the detection of misleading content (He et al., 2024), motivating our explanation-annotated framework.

3 Methodology

This section describes the data collection process, annotation methodology, and dataset construction used in this study. It also outlines the procedures followed to ensure annotation quality and consistency.

3.1 Data Collection

those in To construct the dataset, we collected betting-related advertisements from social media platforms using the Meta Ads Library, which pro-

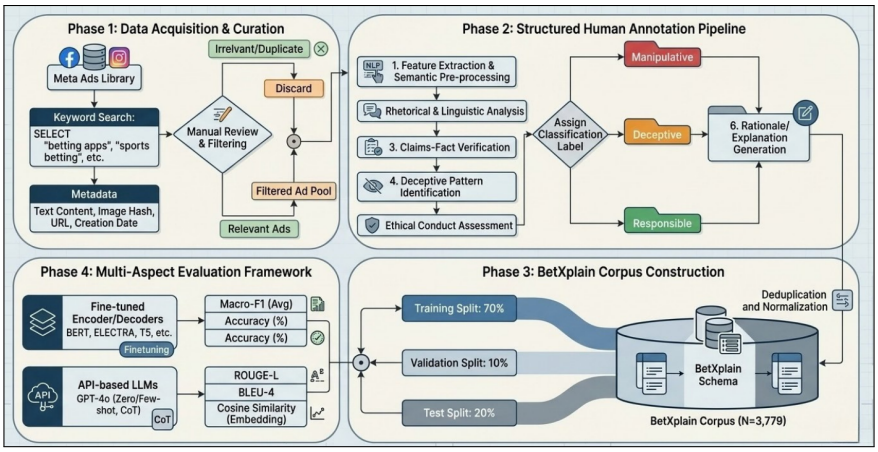


Figure 1: Overview of the BetXplain framework, illustrating the pipeline from betting advertisement data collection and manual annotation to corpus construction and multi-aspect evaluation using transformer models and LLM-based explanation metrics.

vides access to publicly available advertisements running across Meta platforms. The data collection focused on advertisements for betting applications and related services. To identify relevant advertisements, we queried the platform using keywords such as "betting apps", "online betting", "sports betting", and related terms. These queries returned advertisements associated with betting platforms that were actively promoted on social media.

For each advertisement retrieved through the Meta Ads Library, we collected the advertisement link and associated metadata. The advertisements were then manually reviewed to verify that they were related to betting applications or betting services. Irrelevant advertisements and duplicate entries were removed during this filtering stage to ensure that the dataset contained only relevant betting-related promotional content. After collecting the advertisements, they were assigned to two volunteer data annotators who contributed to this project and are also co-authors of this paper, with no financial compensation provided. The annotators manually reviewed each instance and labeled it according to the advertising strategy employed, categorizing each advertisement as manipulative, deceptive, or responsible.

1. Manipulative advertisements refer to promotional content that uses persuasive or psychological tactics to influence user behavior. These advertisements often exaggerate potential rewards, create a sense of urgency, or portray betting as an easy way to earn money. Such strategies may pressure users into participating in betting activities without adequately presenting the associated risks.

2. Deceptive advertisements refer to advertisements that present misleading or false information about betting platforms. These may include unrealistic claims about guaranteed winnings, misleading promotional offers, or advertisements linked to platforms that engage in questionable practices, such as restricting withdrawals after users deposit funds.

3. Responsible advertisements are betting promotions that present information in a neutral or transparent manner. These advertisements typically avoid exaggerated claims and may include responsible messaging or warnings regarding gambling behavior.

To ensure the dataset's reliability, all advertisements were manually verified during annotation. In addition to assigning category labels, annotators also provided textual explanations describing the reasoning behind each annotation. These explanations highlight the specific elements within an advertisement that indicate manipulative or deceptive behavior, such as exaggerated reward framing, urgency-driven messaging, or misleading promotional claims. The inclusion of explanation annotations enhances the dataset's interpretability and enables research on explainable detection models for manipulative advertising. By capturing both classification labels and the reasoning behind them, the dataset supports the development of automated systems that can not only detect harmful betting advertisements but also explain why they are considered manipulative or deceptive.

3.2 Data Annotation Guidelines

To maintain consistent labeling across the dataset, annotators followed the following guidelines:

1. Content Verification: Annotators first verified whether the advertisement was genuinely related to betting applications or betting services. Non-relevant advertisements were excluded from the dataset.

2. Identification of Persuasive Language: Annotators examined whether the advertisement used persuasive or emotional language designed to strongly influence user behavior.

3. Detection of Misleading Claims: Annotators checked for misleading statements such as guaranteed profits, unrealistic rewards, or exaggerated success rates.

4. Evaluation of Promotional Tactics: Advertisements were examined for tactics such as urgency-based promotions, exaggerated bonuses, or claims that portray betting as an easy source of income.

5. Assessment of Platform Practices: When possible, annotators reviewed linked platforms to determine whether the advertisement promoted services associated with questionable practices, such as restricted withdrawals.

6. Explanation Annotation: For each labeled advertisement, annotators provided a short explanation describing why the advertisement was categorized as manipulative, deceptive, or responsible. These explanations highlight the key features that influenced the labeling decision.

The use of multiple annotators and structured annotation guidelines helped ensure the dataset's reliability and quality. To quantify annotation reliability, we computed inter-annotator agreement across the three promotion labels, yielding an average pairwise Cohen's κ of 0.78 and a collective Krippendorff's α of 0.77 (full pairwise scores and Fleiss' κ in Appendix A), indicating moderate-to-strong agreement given the inherent subjectivity of the task.

3.3 Dataset Overview

The proposed dataset, *BetXplain*, contains betting-related advertisements collected from public social media platforms, primarily Instagram, to study manipulative and deceptive promotional strategies used in betting applications. Each instance in the dataset includes five attributes: *text*, *link*, *category*, *explanation*, and *promotion label*, as shown in Table 1. The *text* field represents the advertisement caption or promotional message, while the *link* field contains the link to the ad. The *category* attribute identifies the type of betting activity being promoted, such as gambling or color prediction games. The *promotion label* classifies the advertisement as *manipulative*, *deceptive*, or *responsible*. In addition to these labels, the dataset includes human-written *explanations* describing the reasoning behind each annotation, highlighting persuasive language, exaggerated reward framing, or the absence of risk disclosure. This structured annotation enables the development of automated models that detect harmful betting advertisements while also supporting explainable analysis of promotional tactics.

3.4 Derived Metrics

The BetXplain dataset contains both manually annotated variables and automatically derived analytical variables.

Human-annotated fields. Each advertisement was independently annotated by two annotators for the primary promotion label (*manipulative*, *deceptive*, or *responsible*) and accompanied by a short free-text explanation describing the reasoning behind the assigned label. Inter-annotator agreement was computed only for these human-created annotations (Appendix A), since they involve subjective judgments.

Automatically derived fields. Several additional analytical variables were generated automatically after annotation and were not manually labeled. These include advertisement topic detection, misleading-claim indicators, promotional-practice tags (e.g., urgency, reward framing, financial coercion), platform-practice indicators (e.g., withdrawal restrictions), literature-grounded risk association analysis scores, and social-proof statistics. These variables were produced using a deterministic lexicon- and pattern-matching pipeline constructed from manually curated keyword dictionaries and rule-based linguistic patterns. Because these fields were generated automatically rather than manually annotated, inter-annotator agreement was not computed for them.

Text	Link	Category	Label	Explanation
Join India's Best Betting Platform With Self Deposit/Self Withdrawal Facility, Welcome Bonus: 99₹, Min Deposit: 100₹, Every Deposit Bonus: 5%, Referral Bonus: 5%, 24x7 Withdrawal	https://www.instagram.com/reels/DKrX-UsS-Jd/%7D	casino	manipulative	Uses exclusivity and status-based language to create aspirational appeal without clarifying requirements, costs, or risks. Emotional and status-driven persuasion without overt deception.
Vote below, fellas. Let's see it	are https://www.instagram.com/reels/DFJHp4xTo1l/	hockey (non-gambling)	responsible	Purely informational sports content with no exaggerated claims, manipulative tactics, or misleading financial statements.
Online earning game mines. Online job. Real money. Winning game. Aviator online game. Betting game.	https://www.instagram.com/reels/DHIVml2ISSn/	mines	manipulative	The post frames gambling as a legitimate income source using terms like 'online job' and 'real money' without any risk disclosure or responsible gambling messaging, exploiting financial aspirations through emotionally charged language and symbols.

Table 1: Sample instances from the BetXplain dataset illustrating each promotion label.

3.5 Data Statistics

The BetXplain dataset contains 3,779 advertisements after removing 216 duplicate entries from the original 4,000 collected. The label distribution is moderately imbalanced, with responsible promotion forming the largest class (49.6%), followed by manipulative promotion (39.9%), and deceptive promotion as the minority class (10.5%). This distribution reflects real-world advertising patterns, where explicit false claims are less common than psychologically manipulative but technically permissible promotions. Advertisement texts are generally short, with a median length of 13 words and a mean of 30.1 words ($\sigma = 39.2$), consistent with the concise nature of social media promotional captions. The dataset spans 19 advertising categories, including gambling-related content such as slots, poker, roulette, color prediction, and casino, as well as non-gambling sports content, such as cricket, football, tennis, and basketball, used as responsible promotion baselines.

4 Experimental Results

In this section, we present the experimental setup and evaluate the performance of different models on the proposed dataset. We also analyze the results using standard classification and explanation quality metrics.

4.1 Experimental Setup

Dataset and Splits. Our dataset comprises 3,779 social media posts annotated across three classes: *manipulative promotion* (1,507 samples, 39.9%), *deceptive promotion* (396 samples, 10.5%), and *responsible promotion* (1,876 samples, 49.6%). After verifying zero train-test overlap, we partition the data into training (70%, 2,645 samples), validation (10%, 378 samples), and test (20%, 756 samples) splits using stratified sampling to preserve class proportions.

Models. We evaluate two broad categories of models: fine-tuned transformer models and closed-source API models. For *fine-tuned transformer models*, we consider encoder-only models—BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), ELECTRA (Clark et al., 2020), and ALBERT (Lan et al., 2019)-a long-context encoder, Longformer (Beltagy et al., 2020) and sequence-to-sequence models BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). For *closed-source API models*, we evaluate GPT-4o under three prompting strategies: zero-shot, few-shot (5 examples), and chain-of-thought (CoT).

Training Details. All fine-tuned models are trained for 5 epochs using the AdamW optimizer (Loshchilov and Hutter, 2017) with weight decay 0.01 and a linear warmup schedule covering 10% of total steps with a learning rate of 2×10^{-5}. Automatic Mixed Precision (AMP) is enabled for all models except T5, which requires full float32 precision. To address class imbalance, particularly the underrepresented deceptive class ($\approx$10%), we apply inverse-frequency class weighting to the cross-entropy loss. All models use a maximum sequence length of 256 tokens and a batch size of 16. The best checkpoint is selected based on macro-F1 on the validation set. Experiments are conducted on an NVIDIA GeForce RTX 5060 Ti (16 GB). We report accuracy, macro-F1, weighted-F1, and micro-F1 on the held-out test set. Macro-F1 is our primary metric given the class imbalance.

4.2 Classification Results

Table 2 reports classification performance across all models.

Fine-tuned Models. Among fine-tuned models, ELECTRA achieves the highest macro-F1 (0.6946),

Group	Model	Acc	Ma-F1	W-F1	Mi-F1
Encoder	BERT	0.8148	0.6752	0.8171	0.8148
	RoBERTa	0.7947	0.6685	0.8057	0.7947
	DistilBERT	0.8223	0.6899	0.8257	0.8223
	ELECTRA	**0.8335**	**0.6946**	**0.8326**	**0.8335**
	ALBERT	0.8323	0.6876	0.8286	0.8323
Long-context	Longformer	0.8511	0.6711	0.8351	0.8511
Enc-Dec	BART	0.8010	0.6763	0.8095	0.8010
	T5	0.8223	0.6519	0.8126	0.8223
API	GPT-4o (zero-shot)	0.8123	0.6871	0.8098	0.8123
	GPT-4o (few-shot)	0.8210	0.6898	0.8151	0.8210
	GPT-4o (CoT)	0.8181	0.6870	0.8158	0.8181

Table 2: Classification results on the test set (756 samples). **Bold**: best fine-tuned model per metric. All metrics computed with 95% bootstrap confidence intervals (CI width ≈0.05–0.08); see Appendix D for full per-model confidence intervals.

followed by DistilBERT (0.6899) and ALBERT (0.6876). ELECTRA's discriminator-based pretraining, which classifies real versus replaced tokens across all positions, appears to produce more discriminative representations for fine-grained promotion detection. Fine-tuned models span a macro-F1 range of 0.651–0.695, with T5 performing the weakest (0.6519) among encoder-decoder architectures. Longformer achieves the highest accuracy (0.8511) and macro-F1 (0.6711), indicating benefits from global attention beyond long-context modeling. Although the median text length is 13 words, 5.5% of samples exceed 128 tokens, and Appendix F shows gains arise from global attention rather than increased context length. To confirm these rankings are not an artifact of a single favorable split, we additionally report 5-fold stratified cross-validation across all fine-tuned models in Appendix A, which preserves the same model ordering (Appendix B, Table 6' play ').

API Models. GPT-4o with few-shot prompting achieves the highest accuracy among API configurations (0.8210), While chain-of-thought achieves the highest macro-F1 (0.6870), it remains below the best fine-tuned model (ELECTRA: 0.6946). The structured reasoning of CoT better handles the underrepresented *deceptive promotion* class than few-shot methods, which tend to steer the model toward majority-class predictions. Across all three GPT-4o strategies, the gap between weighted-F1 (≈0.81) and macro-F1 (≈0.69) is notably larger than for fine-tuned models, confirming that GPT-4o systematically under-identifies the *deceptive promotion* class. This underscores the value of fine-tuning on domain-specific annotated data, even when compared with a capable frontier model.

Model / Comparison	R-1	R-L	BLEU-1	Cosine
Human–Human	**0.4121**	**0.3754**	**0.1845**	**0.8142**
GPT-4o (zero-shot)	0.2289	0.1852	0.0127	0.5398
GPT-4o (few-shot)	0.1528	0.1144	0.0037	0.4964
GPT-4o (CoT)	0.2707	0.2144	0.0173	0.5776

Table 3: Explanation quality metrics against gold human rationales. *Human–Human* indicates agreement between two annotators on a 100-ad double-annotated subset, serving as the performance ceiling for the task.

4.3 Open-Source LLM Evaluation

To evaluate accessibility beyond proprietary models, we tested two open-source instruction-tuned LLMs, LLaMA-3-8B-Instruct and Mistral-7B-Instruct-v0.3, under zero-shot and few-shot (5 examples) prompting. Table **??** shows that both models perform 3–7% below GPT-4o, consistent with the known capability gap between frontier and mid-size models. Few-shot prompting improves performance by 2-3%, confirming that in-context learning is effective even for smaller models. However, even the best open-source result (LLaMA-3 few-shot, macro-F1 = 0.6798) remains below fine-tuned ELECTRA (0.6946), reinforcing the value of domain-specific fine-tuning.

4.4 Explanation Quality

Table 3 reports explanation quality metrics comparing GPT-4o-generated rationales against human-authored gold explanations.

To establish a performance ceiling for the rationale-writing task, we conducted a human-human agreement analysis on a double-annotated subset of 100 advertisements. As shown in Table 3, while lexical overlap metrics like BLEU-1 are naturally lower due to variations in writing style, the high semantic consistency is evidenced by a Cosine Similarity of 0.81. Among the evaluated models, GPT-4o (CoT) achieves the highest semantic alignment with human reasoning (Cosine: 0.5776), though a notable gap persists between model-generated explanations and the human-consensus baseline.

4.5 Linking Classification Performance to Persuasion Patterns

To connect the classification results (Section 4) with the linguistic analysis (Section 5), we analyze per-class F1 performance. Table 4 shows that the

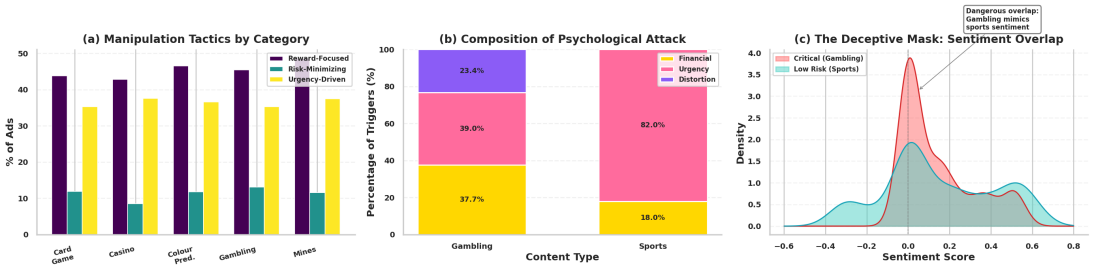


Figure 2: Psychological manipulation, architecture, and mental health impact(a) Keyword frequency heatmap across betting categories. (b) Psychological trigger composition in gambling versus non-gambling content. (c) Sentiment KDE overlap confirms the Deceptive Positivity effect.

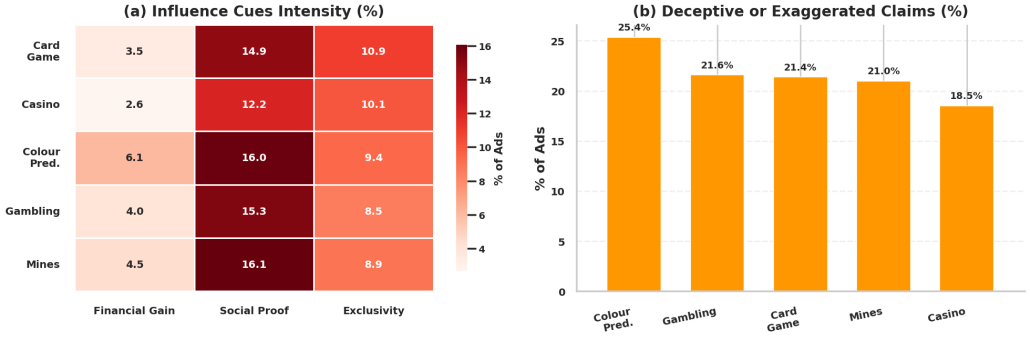


Figure 3: Influence cues and deceptive claims across categories: (a) Influence cues intensity by category. (b) Prevalence of deceptive claims by category, with color prediction highest at 25.4%.

Model	Manip.	Decept.	Resp.	Ma-F1
ELECTRA	0.778	0.480	0.826	0.6946
BERT	0.775	0.414	0.837	0.6752
GPT-4o (CoT)	0.780	0.450	0.830	0.6870

Table 4: Per-class F1 breakdown for selected models.

deceptive class achieves the lowest F1 across models (0.43–0.51), consistent with Section 5.2, where deceptive ads share 72% of their financial coercion vocabulary with manipulative ads, making discrimination inherently difficult. The dominant manipulative → responsible for confusion aligns with Section 5.1, which identifies play" as a common cross-category leisure-framing term. ELECTRA's discriminator pretraining captures subtle token-level distinctions, such as guaranteed" versus "boosted," explaining its +1.9% macro-F1 improvement over BERT.

5 Discussion

In this section, we interpret the experimental findings and examine the broader implications of the results. We analyze the persuasion patterns, emotional framing, and psychological triggers present in betting advertisements and discuss how these elements contribute to manipulative promotional strategies.

5.1 Category-Wise Persuasion Strategies

As shown in Figure 2(a), we analyze the distribution of manipulation tactics across betting advertisement categories. The results indicate that *reward-focused messaging* dominates across all categories, accounting for approximately 42–47% of advertisements. These messages emphasize potential gains, bonuses, and winning opportunities, reinforcing aspirational appeal.

Urgency-driven tactics form the second most common strategy (approximately 35–38%), using cues such as limited-time offers and immediate rewards to accelerate decision-making and reduce deliberation.

In contrast, *risk-minimizing language* is consistently underrepresented (8–13%), indicating that advertisements systematically downplay or omit potential losses and financial risks. This imbalance suggests that betting promotions are strategically optimized to amplify perceived rewards while suppressing risk awareness across all categories.

Figure 2(b) shows that gambling content deploys financial coercion (∼56%), urgency (∼26%), and cognitive distortion (∼18%), a combination ab-

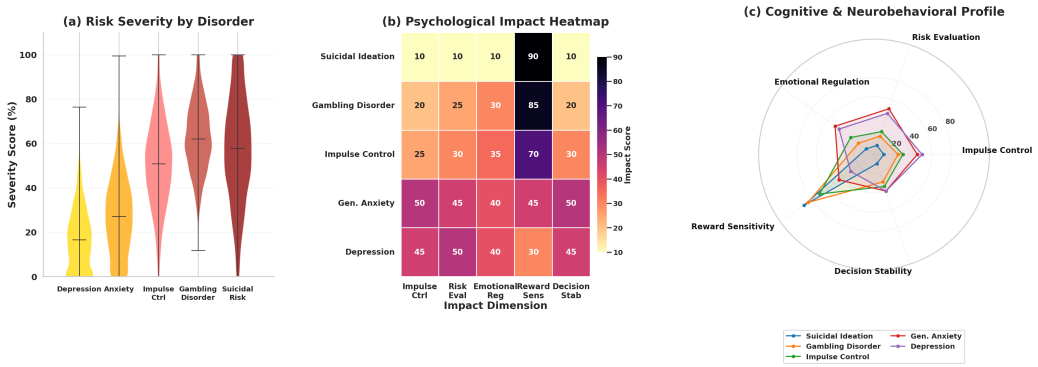


Figure 4: Literature-Grounded Risk Association Analysis: (a) Risk severity distributions by disorder. (b) Psychological impact heatmap across five cognitive dimensions. (c) Neurobehavioral radar chart showing the asymmetric profile of gambling disorder and suicidal ideation risk.

sent in sports content entirely. This asymmetry reflects a shared operator template rather than an independent optimization (Petry, 2005), and validates training a single unified model on BetXplain. It also explains the macro-F1 range among fine-tuned models (0.651-0.695): the shared persuasion vocabulary is detectable, but subtler deceptive signals remain difficult to resolve.

5.2 Influence Cues and Deceptive Claims

Social proof dominates across all categories (12-16%), Figure 3(a), while explicit financial claims remain rare (2.6-6.1%). This is not a coincidence; social proof implies community participation without falsifiable claims, making it harder to flag while achieving the same legitimizing effect. Deceptive claims appear in 18.5-25.4% of ads (Figure 3(b)), with color prediction highest at 25.4%, consistent with its strategy of misrepresenting a chance game as skill-tractable. Of 401 financial claim ads, nearly half used urgency-to-deposit framing risk is not disclosed; it is crowded out (Dowling et al., 2015).

5.3 Literature-Grounded Risk Association Analysis and Mental Health Impact

To construct the Literature-grounded risk association analysis, we map linguistic and psychological features extracted from advertisements to established behavioral risk factors reported in psychological and psychiatric literature on gambling addiction. Specifically, each advertisement is analyzed for the presence of key trigger categories, including financial coercion, urgency, reward framing, and cognitive distortion. These triggers are then associated with clinically recognized mental health outcomes such as gambling disorder, anxiety, depression, and suicidal ideation based on prior empirical studies (Petry, 2005; Gainsbury, 2015). Risk severity scores are computed by aggregating the intensity and co-occurrence of these triggers, yielding a normalized severity scale from low to critical. This mapping allows us to interpret advertising strategies not only as linguistic patterns but as potential contributors to psychological harm. We emphasize that this severity scale is a lexical and thematic aggregation cross-referenced against findings reported in prior epidemiological and psychiatric literature (Petry, 2005; Gainsbury, 2015; Blaszczynski and Nower, 2002), and is not a diagnostic instrument. No licensed clinician or psychiatrist reviewed the mapping between advertisement language and the mental health outcomes discussed; the analysis should be read as an association drawn from published risk-factor research, not a clinical assessment of any individual advertisement, platform, or user.

Figure 4(a) reveals a clear severity escalation gradient, with gambling disorder and suicidal ideation clustering at the severe end (9.0–9.5/10). The impact heatmap in Figure 4(b) identifies reward sensitivity dysregulation as the dominant factor, with suicidal ideation and gambling disorder scoring 90 and 85, respectively, while showing regulatory collapse across other dimensions. Figure 4(c) further confirms a distinct asymmetric neurobehavioral profile relative to anxiety and depression (Petry, 2005). Among 207 family-related ads, nearly equal proportions employed aspirational and threat-based framing, intensifying emotional distress during losses. These findings position manipulative betting advertisement detection as both a linguistic and a mental health intervention task, with BetXplain providing transparent, auditable reasoning annotations.

5.4 Real World Applications

The proposed framework enables multiple real-world applications for mitigating manipulative betting advertisements on digital platforms. It can be integrated into browser or mobile plugins to automatically analyze betting advertisements on social media and warn users about manipulative or deceptive content by highlighting tactics such as exaggerated reward claims and urgency-driven messaging. The framework can also support automated web crawlers for governments and regulatory agencies to monitor betting promotions and identify suspicious advertisements at scale. Additionally, it can enhance social media moderation systems and advertising compliance tools by automatically flagging potentially misleading betting promotions before they are widely disseminated.

6 Conclusion

In this work, we introduced **BetXplain**, an explanation-annotated dataset designed to support the detection of manipulative and deceptive betting advertisements on social media. The dataset provides manually verified advertisements labeled as manipulative, deceptive, or responsible, along with human-authored explanations that capture the reasoning behind each annotation. Through experiments with transformer-based models and large language model prompting strategies, we demonstrate that domain-specific training enables effective identification of persuasive and misleading betting promotions. Our analysis further highlights common persuasive strategies used in betting advertisements, including exaggerated reward claims and urgency-driven messaging. Overall, BetXplain provides a new resource for studying harmful advertising practices and supports future research on explainable detection systems, promoting safer and more responsible online advertising environments.

Limitations

The scope of this study focuses on betting advertisements collected from selected social media platforms where such promotions are commonly observed. While these platforms are important channels for digital advertising, betting promotions may also appear in other online environments and regional ecosystems. Expanding the dataset to include advertisements from a broader range of platforms and geographic contexts could further enrich the diversity of promotional strategies captured and support the development of more globally robust detection systems. In addition, the current dataset focuses primarily on English-language advertisements. Betting promotions often vary across languages and cultural contexts, and extending the dataset to multilingual settings could enable the study of how persuasive strategies differ across regions and regulatory environments.

Finally, BetXplain provides concise, human-authored explanations that highlight the reasoning behind each annotation. These explanations support research on explainable detection models while maintaining annotation consistency. Future work may explore richer explanation formats or additional annotation perspectives that capture more detailed reasoning processes. These directions represent natural extensions of the present work, highlight opportunities to further advance research on responsible digital advertising and explainable harmful-content detection.

Ethics Statement

We, the authors of this work, affirm that the research presented in this paper was conducted in accordance with ethical standards and responsible research practices. Throughout this study, we ensured that the collection and analysis of betting advertisements were conducted using publicly accessible sources and strictly for academic research purposes to understand and detect manipulative betting promotions on social media platforms. We acknowledge that certain limitations may arise from manual annotation processes and from the stochastic behavior inherent in modern machine learning and large language models. To address these concerns, we followed structured annotation guidelines, engaged multiple annotators to ensure consistency, and documented the dataset construction, annotation procedures, model configurations, and experimental settings to support the transparency and reproducibility of our work. We also maintained controlled experimental settings during model evaluation to minimize variability in results. The dataset contains betting-related promotional content that may include persuasive language, which could be sensitive for vulnerable individuals; however, it is used strictly for research, and we explicitly oppose and do not promote such activities. The publicly released version of the dataset excludes advertiser handles, account usernames, and other identifying metadata, retaining

only advertisement text, category, promotion label, and explanation fields. All advertisements were sourced from content already made publicly available via the Meta Ads Library's public disclosure mechanism. We note that promotion-label predictions produced by models trained on this dataset are research classifications only, not legal findings, and should not be used to make claims about specific named businesses or platforms without independent human review. Annotation subjectivity and data-driven biases may affect model generalization across contexts. Additionally, links to public advertisements require responsible handling, and automated systems derived from this work should be used with appropriate human oversight and ethical safeguards. For coding assistance, AI-based coding tools were utilized; however, the conceptualization of the research problem, dataset creation, methodological design, experimentation, and analysis were conducted independently by the authors.

References

Max Abbott and Rachel Volberg. 1991. Gambling and problem gambling in new zealand.

Max W Abbott. 2020. The changing epidemiology of gambling disorder and gambling-related harm: public health implications. *Public health*, 184:41–45.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Per Binde. 2014. Gambling advertising: A critical research review.

Alex Blaszczynski and Lia Nower. 2002. A pathways model of problem and pathological gambling. *Addiction*, 97(5):487–499.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Business Standard. 2025. Youth at risk: Supreme court issues notice to centre on pil seeking ban on betting apps. *Business Standard*. Details the legal scrutiny of celebrity endorsements and financial ruin caused by betting platforms.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Nicki A Dowling, Sean Cowlishaw, Alun C Jackson, Stephanie S Merkouris, Kate L Francis, and Darren R Christensen. 2015. Prevalence of psychiatric co-morbidity in treatment-seeking problem gamblers: A systematic review and meta-analysis. *Australian & New Zealand Journal of Psychiatry*, 49(6):519–539.

Forcepoint X-Labs. 2026. Online casino spam: How fake gambling sites steal money. Accessed: 2026-03-15.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30.

Sally M Gainsbury. 2015. Online gambling addiction: the relationship between internet gambling and disordered gambling. *Current addiction reports*, 2(2):185–193.

Leela Prasad Gorrepati, Raj Sonani, Venkatesh Velugubantla, Ravi Teja Potla, and MSVPJ Sathvik. 2025. Mental health and relations: Detection of mental health disorders related to relationship issues through reddit posts. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1885–1889.

Government of India. 1867. The public gambling act, 1867. https://www.indiacode.nic.in/handle/123456789/2269. Act No. 3 of 1867 regulating gambling activities in India.

Government of India. 2000. Information technology act, 2000. https://www.meity.gov.in/content/information-technology-act-2000. Cyber law addressing digital crimes including online fraud and illegal betting platforms.

Government of Kuwait. 1960. Kuwait penal code (law no. 16 of 1960). https://www.ilo.org/dyn/natlex/docs/ELECTRONIC/83385/92077/F1114787009/KWT83385%20Eng.pdf. Includes provisions criminalizing gambling activities.

Government of Qatar. 2004. Qatar penal code (law no. 11 of 2004). https://www.almeezan.qa/LawPage.aspx?id=26&language=en. Article 275 prohibits gambling activities.

Government of Saudi Arabia. 2019. Prohibition of gambling under saudi arabian legal framework. https://www.mcit.gov.sa/sites/default/

files/2021-06/la_004_e_anti-cyber_crime_law%20%281%29.pdf. Gambling is prohibited under Sharia-based criminal law.

Government of the United Arab Emirates. 1987. Uae federal penal code (federal law no. 3 of 1987). https://www.adjd.gov.ae/sites/Authoring/AR/ELibrary%20Books/Federal%20Law%20No.%203%20of%201987%20on%20Issuing%20the%20Penal%20Code.pdf. Articles criminalize gambling and operating gambling establishments.

Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The dark (patterns) side of ux design. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14.

Mark D Griffiths. 2019. Online gambling and geolocation technology: implications for regulation and potential threats to player protection. *Gaming Law Review*, 23(5):344–347.

Frederic Guerrero-Solé, Hibai Lopez-Gonzalez, and Mark D Griffiths. 2018. Online gambling advertising and the third-person effect: A pilot study. In *Media Influence: Breakthroughs in Research and Practice*, pages 384–401. IGI Global Scientific Publishing.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. 2024. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning. In *International conference on learning representations*, volume 2024, pages 5711–5732.

Nerilee Hing. 2014. *Sports betting and advertising*. Australian Gambling Research Centre, Australian Institute of Family Studies ....

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hibai Lopez-Gonzalez and Mark D Griffiths. 2018. Understanding the convergence of markets in online sports betting. *International Review for the Sociology of Sport*, 53(7):807–823.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Arunesh Mathur, Gunes Acar, Michael J Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–32.

Arunesh Mathur, Arvind Narayanan, and Marshini Chetty. 2018. Endorsements on social media: An empirical study of affiliate marketing disclosures on youtube and pinterest. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–26.

Philip WS Newall. 2019. Dark nudges in gambling.

Nancy M Petry. 2005. *Pathological gambling: Etiology, comorbidity, and treatment*, volume 2. American Psychological Association Washington, DC.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Nyrmah J. Reina. 2025. 5 betting app scams to be aware of in 2025. Accessed: 2026-03-15.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

The Indian Express Bureau. 2025. Sc says growth of online betting apps issue of public importance, seeks states' responses. *The Indian Express*. Reported the petitioner's claim of 1,023 suicides in Telangana linked to betting app addiction.

Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 1667–1682.

M Walker. 1996. The prevalence of problem gambling in australia: Another lottery. In *Seventh National Conference of the National Association for Gambling Studies, Glenelg, South Australia*.

Appendix

A Inter-Annotator Agreement Scores

The annotation process was conducted by a team of four annotators who independently reviewed betting-related advertisements collected from public social media platforms. Since the distinction between manipulative, deceptive, and responsible promotional strategies can be subjective, detailed annotation guidelines were established before annotation to improve consistency among annotators.

Each annotator participated in an initial pilot annotation phase involving a subset of advertisements. Disagreements were discussed collaboratively to refine category boundaries and align annotation standards before the full annotation process began.

Advertisements were labeled according to the following promotion categories:

- **Manipulative Promotion (MP):** Advertisements that use emotionally persuasive or psychologically coercive tactics, such as urgency framing, exaggerated rewards, aspirational language, fear of missing out (FOMO), or financial pressure, are intended to influence user behavior.
- **Deceptive Promotion (DP):** Advertisements containing misleading or potentially false claims, including unrealistic guarantees of profit, exaggerated winning probabilities, deceptive promotional offers, or misleading financial representations.
- **Responsible Promotion (RP):** Advertisements presenting betting-related information in a comparatively transparent or neutral manner without strong manipulative framing or deceptive claims.

In addition to promotion labels, advertisements were grouped into 5 betting-related categories: casino promotions, sports betting, color prediction games, poker/card games, and miscellaneous betting applications.

To evaluate annotation reliability, we computed multiple inter-annotator agreement (IAA) metrics, including Cohen's κ, Fleiss' κ, and Krippendorff's α. Pairwise Cohen's κ values were computed between annotator pairs, while Fleiss' κ and Krippendorff's α were computed across all annotators collectively.

Annotator Pair	Krippendorff's α	Cohen's κ	Fleiss' κ
(1,2)	0.8127	0.8264	—
(1,3)	0.7841	0.7989	—
(1,4)	0.7365	0.7498	—
(2,3)	0.7694	0.7816	—
(2,4)	0.7218	0.7352	—
(3,4)	0.7953	0.8087	—
(1,2,3,4)	0.7700	—	0.7526

Table 5: Inter-Annotator Agreement Scores

The agreement scores indicate moderate-to-strong consistency across annotator pairs. The averaged pairwise Cohen's κ score was 0.7834, while the collective Krippendorff's α and Fleiss' κ scores were 0.7700 and 0.7526, respectively. These values are consistent with complex annotation tasks that involve subjective distinctions among manipulative, deceptive, and responsible promotional strategies.

The comparatively lower agreement observed for some annotator pairs reflects the inherent ambiguity involved in distinguishing emotionally persuasive advertising language from explicitly deceptive claims. Advertisements involving aspirational financial messaging, urgency-based promotions, and implied earning opportunities produced the highest disagreement rates. All annotation disagreements were resolved through collaborative discussion and consensus review prior to finalizing the dataset.

Rationale Agreement. We also assessed the reliability of the free-text explanations provided by annotators. On a shared subset of 100 samples, human annotators identified similar manipulative and deceptive indicators, reflected in a semantic cosine similarity of 0.814. This confirms that while the specific wording of explanations may differ, the underlying reasoning for labeling advertisements is conceptually consistent across the dataset.

B Cross-Validation Robustness

To verify that the single-split results are not artifacts of a favorable partition, we performed 5-fold stratified cross-validation across all eight fine-tuned models. Table 6 reports mean performance and standard deviation across folds. The model ranking is preserved: ELECTRA remains the top model by macro-F1, with standard deviations ranging from 0.007 to 0.014, indicating stable generalization despite the moderate dataset size. Cross-validation means are within 1–2% of the single-split results, confirming that the test partition is representative.

Model	Acc	Ma	W	Rep.
BERT	0.784±.008	0.688±.008	0.780±.008	0.675
RoBERTa	0.774±.011	0.667±.014	0.768±.011	0.669
DistilB	0.826±.010	0.736±.009	0.820±.010	0.690
ELECTRA	**0.829±.009**	**0.739±.010**	**0.824±.010**	**0.695**
ALBERT	0.824±.012	0.728±.011	0.818±.011	0.688
Longf.	0.835±.012	0.726±.011	0.825±.011	0.671
BART	0.787±.007	0.691±.008	0.783±.007	0.676
T5	0.812±.007	0.694±.011	0.803±.007	0.652

Table 6: 5-fold cross-validation results (mean±std) and reported macro-F1 scores. Acc: Accuracy, Ma: Macro-F1, W: Weighted-F1, Rep.: Reported single-split Macro-F1, Longf.: Longformer, DistilB: DistilBERT.

We conducted McNemar's test and paired bootstrap tests (10,000 resamples) with Benjamini-Hochberg correction, comparing all models against ELECTRA. After correction for multiple comparisons, no pairwise difference reaches statistical significance at $\alpha = 0.05$. This is expected given the moderate test set size (756 samples) and the narrow macro-F1 range (0.651-0.695). The models form a competitive cluster, and deployment decisions should consider computational constraints alongside raw metrics.

C LLM-as-a-Judge Consistency Analysis

To further evaluate the consistency of the annotation framework, we additionally explored an LLM-as-a-Judge setting using GPT-4o. A randomly selected subset of advertisements was independently evaluated by the model using the same annotation guidelines provided to human annotators.

The LLM-generated labels were compared with the final consensus annotations produced by the human annotators. Agreement analysis showed substantial alignment between GPT-4o predictions and human judgments, suggesting that the annotation guidelines capture reasonably consistent patterns of persuasive and deceptive content in betting advertisements.

Metric	Score
Cohen's κ	0.7421
Macro-F1	0.7314
Weighted-F1	0.7688
Accuracy	0.7763

Table 7: Agreement between GPT-4o judgments and finalized human annotations

The lower agreement relative to human annotator consistency suggests that the LLM occasionally overrelied on surface-level sentiment and promotional tone when distinguishing between manipulative and deceptive content. In particular, advertisements with positive emotional framing but no explicit false claims were more difficult for the model to classify consistently.

D Bootstrap Confidence Intervals

Table 8 reports 95% bootstrap confidence intervals (10,000 resamples) for all models and metrics on the test set.

E Procedure Involving Data Collection and Construction

1. **Data Collection:** Betting-related advertisements were collected from public social media platforms, including Instagram and Reddit, using platform search mechanisms and the Meta Ads Library. Relevant advertisements associated with betting applications and services were retrieved using betting-related keywords.

2. **Data Cleaning:** Duplicate advertisements, irrelevant promotional content, spam entries, and non-betting advertisements were removed during preprocessing. Advertisements containing minimal or non-informative text were excluded from the dataset.

3. **Data Preprocessing:** URLs, usernames, hashtags, excessive emojis, and formatting artifacts were normalized where necessary. Text preprocessing was performed while preserving the persuasive and emotional language patterns important for annotation.

4. **Manual Annotation:** Each advertisement was independently reviewed and labeled in accordance with the annotation guidelines. Annotators also provided concise, human-written explanations for each assigned label.

5. **Quality Verification:** OverreliedAnnotations were reviewed for consistency, and disagreements were resolved through collaborative discussion before constructing the final dataset.

6. **Dataset Finalization:** After quality verification, duplicate advertisements were removed, and the final dataset was partitioned into train, validation, and test splits using stratified sampling.

Model	Metric	Point Est.	95% CI	Width
BERT	accuracy	0.7989	[0.7698, 0.8267]	0.0569
	macro-F1	0.6889	[0.6493, 0.7270]	0.0777
	weighted-F1	0.8011	[0.7717, 0.8298]	0.0581
	micro-F1	0.7989	[0.7698, 0.8280]	0.0582
RoBERTa	accuracy	0.7738	[0.7434, 0.8029]	0.0595
	macro-F1	0.6616	[0.6234, 0.7001]	0.0767
	weighted-F1	0.7787	[0.7483, 0.8085]	0.0602
	micro-F1	0.7738	[0.7434, 0.8029]	0.0595
DistilBERT	accuracy	0.8148	[0.7870, 0.8426]	0.0556
	macro-F1	0.7078	[0.6693, 0.7453]	0.0760
	weighted-F1	0.8164	[0.7878, 0.8441]	0.0563
	micro-F1	0.8148	[0.7870, 0.8426]	0.0556
ELECTRA	accuracy	0.8254	[0.7976, 0.8519]	0.0543
	macro-F1	0.7231	[0.6822, 0.7610]	0.0788
	weighted-F1	0.8262	[0.7986, 0.8537]	0.0551
	micro-F1	0.8254	[0.7976, 0.8519]	0.0543
ALBERT	accuracy	0.8228	[0.7950, 0.8492]	0.0542
	macro-F1	0.7181	[0.6767, 0.7569]	0.0802
	weighted-F1	0.8214	[0.7930, 0.8484]	0.0554
	micro-F1	0.8228	[0.7963, 0.8492]	0.0529
Longformer	accuracy	0.8386	[0.8122, 0.8651]	0.0529
	macro-F1	0.7157	[0.6758, 0.7549]	0.0791
	weighted-F1	0.8352	[0.8069, 0.8618]	0.0549
	micro-F1	0.8386	[0.8108, 0.8651]	0.0543
BART	accuracy	0.7910	[0.7606, 0.8188]	0.0582
	macro-F1	0.6848	[0.6454, 0.7223]	0.0769
	weighted-F1	0.7941	[0.7646, 0.8230]	0.0584
	micro-F1	0.7910	[0.7619, 0.8201]	0.0582
T5	accuracy	0.8108	[0.7817, 0.8386]	0.0569
	macro-F1	0.6795	[0.6405, 0.7186]	0.0781
	weighted-F1	0.8087	[0.7796, 0.8370]	0.0574
	micro-F1	0.8108	[0.7831, 0.8386]	0.0555
GPT-4o (zero-shot)	accuracy	0.8029	[0.7738, 0.8307]	0.0569
	macro-F1	0.7054	[0.6633, 0.7459]	0.0826
	weighted-F1	0.8020	[0.7728, 0.8305]	0.0577
	micro-F1	0.8029	[0.7738, 0.8307]	0.0569
GPT-4o (few-shot)	accuracy	0.8135	[0.7857, 0.8400]	0.0543
	macro-F1	0.7143	[0.6734, 0.7526]	0.0792
	weighted-F1	0.8138	[0.7857, 0.8415]	0.0558
	micro-F1	0.8135	[0.7857, 0.8413]	0.0556
GPT-4o (CoT)	accuracy	0.8069	[0.7778, 0.8347]	0.0569
	macro-F1	0.7040	[0.6638, 0.7426]	0.0788
	weighted-F1	0.8092	[0.7805, 0.8375]	0.0570
	micro-F1	0.8069	[0.7791, 0.8347]	0.0556

Table 8: 95% bootstrap confidence intervals (10,000 resamples) for all models. CI widths for accuracy are ≈0.053-0.060, consistent with the test set size (756 samples). Macro-F1 CIs are wider (≈0.077-0.083) due to higher variance on the minority deceptive class. Overlapping CIs between adjacent models are consistent with non-significant statistical tests.

F Longformer Ablation Study

To justify Longformer's inclusion despite the short median text length, we conducted an ablation study varying the maximum sequence length. Table 9 shows that performance drops at 128 tokens due to truncation (5.2% of texts affected), but increasing from 256 to 4096 tokens yields a marginal change. This confirms that the architectural benefit stems from Longformer's global attention mechanism rather than extended context length alone.

G Dataset Composition & Risk Profile

The dataset comprises 4,000 digital engagement entries collected from social media platforms. Each entry was independently annotated and assigned

max_length	Acc	Ma-F1	W-F1	Notes
128	0.8254	0.6929	0.8247	5.2% texts truncated
256	0.8386	0.7188	0.8332	Paper setting (0.8% truncated)
512	0.8373	0.7117	0.8335	No additional benefit
4096	0.8373	0.7128	0.8326	Full Longformer context

Table 9: Longformer performance across different maximum sequence lengths.

Table 10: Dataset Risk Profile Summary

Category	Count	Proportion
Safe / Low Risk (Sports & Entertainment)	2,038	50.9%
Critical Risk (Gambling & Betting)	1,962	49.1%
Total	**4,000**	**100%**

a risk category based on content type, language patterns, and psychological manipulation tactics.

Sentiment analysis reveals a critical anomaly termed *Deceptive Positivity*: Gambling content maintains a near-identical positive sentiment score to benign sports Content that actively lowers users' cognitive defenses.

H Sentiment and Emotional Tone

All five gambling segments maintain a narrow band of mild positivity (mean 0.127-0.150), nearly identical to benign sports content (Figure 5(a-b)). This is not neutral; it is calibrated. As the KDE overlap in Figure 2(c) confirms, gambling and sports content are effectively indistinguishable in the positive range, producing a Trojan Horse effect where harmful content mimics safe entertainment with no warning signal. This manufactured positivity explains GPT-4o's systematic difficulty with the *deceptive* class: without fine-tuning, the model anchors to surface tone rather than the deeper rhetorical structure that distinguishes deception from manipulation.

I Neurological & Psychological Mechanisms of Harm

I.1 Dopamine Dysregulation: The Addiction Loop

The gambling content in this dataset heavily employs *Intermittent Variable Rewards* (e.g., "Try your luck," "Big Wins Await"). Neurologically, the uncertainty of the reward triggers a higher dopamine release than a guaranteed win, conditioning users to seek the anticipatory "high" and rewiring neural pathways analogous to substance abuse disorders.

I.2 Prefrontal Cortex Shutdown: Impulse Control Bypass

Urgency triggers found throughout the dataset ("Hurry," "Limited Time," "Register Now") are designed to bypass the Prefrontal Cortex, the brain's center for logic and long-term planning. This forces users into a *Hot State*, precipitating impulsive financial decisions without regard for consequences.

I.3 Illusion of Control: Cognitive Distortion

Phrases such as "Play Smart," "Prediction," and "Strategy" are embedded in games of pure chance. This fosters a cognitive distortion in which users believe they can influence random outcomes, a primary driver of Ludomania (Compulsive Gambling Disorder), as losses are attributed to flawed strategy rather than mathematical inevitability.

J Comprehensive Disorder Analysis

Table 12 categorizes the specific mental health disorders triggered by the dataset content, their severity classification, primary linguistic triggers, neurobiological mechanism, and clinical consequence.

K Criticality Assessment: Suicidal Risk Factor

The correlation between gambling promotion content and suicide risk renders this The dataset highlights critically high-risk behavioral patterns. The

Table 11: Mean Sentiment Scores by Content Category

Content Category	Mean Sentiment Score
Safe / Low Risk (Sports & Entertainment)	+0.15
Critical Risk (Gambling & Betting)	+0.14

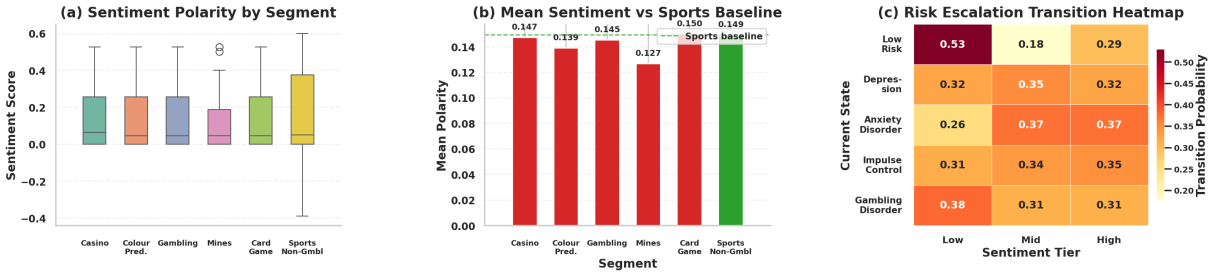


Figure 5: Sentiment, emotional tone, and the deceptive positivity effect: (a) Sentiment polarity by segment. (b) Mean sentiment versus sports baseline. (c) Risk escalation transition heatmap showing systematic progression toward gambling disorder as sentiment becomes more positive.

mechanism proceeds through three identifiable stages:

1. **The Chase Phase.** Content encourages users to "recover" losses by making further deposits, sustaining engagement through loss-aversion psychology.

2. **The Crash.** When mathematically inevitable losses occur, the user experiences a rapid depletion of both dopamine and serotonin.

3. **The Danger Zone.** This neurochemical crash, combined with the tangible reality of debt or bankruptcy, creates a window of acute suicidality. Research indicates that individuals with Gambling Disorder have the highest suicide attempt rates among all addiction demographics, with estimates reaching up to 20% (Blaszczynski and Nower, 2002).

L Visual Evidence & Analytical Interpretation

Five core visual artifacts were generated to support the clinical and psychological risk assessment. Each figure is interpreted in the context of the dataset.

L.1 Integrated Risk & Severity Analysis

A multi-panel composite figure consolidating four complementary analytical views: Risk Severity Distribution by Disorder; Sentiment vs. Severity Density Mapping; Risk Escalation Transition Heatmap and Cumulative Contribution to Overall Mental Health Risk. Gambling Disorder and Suicidal Ideation exhibit the highest median severity with the widest dispersion, indicating instability and escalation risk. Transition probabilities confirm systematic movement from Low Risk → Depression → Gambling Disorder.

L.2 Psychological Attack Vector Composition

A stacked bar chart breaking down psychological triggers across gambling and non-gambling content. Gambling content relies heavily on financial coercion (55.8%), followed by urgency and cognitive distortion. Non-gambling content lacks financial manipulation entirely, confirming that gambling content is structurally engineered as a psychological attack.

L.3 Sentiment Overlap: The Deceptive Mask

A KDE plot comparing sentiment distributions of critical gambling content and low-risk sports content. Strong overlap exists in the positive sentiment range (0.0-0.3), visually confirming the *Trojan Horse Effect*: harmful content intentionally mimics the tone of sports and entertainment, rendering sentiment-based detection insufficient as a sole classifier.

L.4 Psychological Impact Comparison Heatmap

A heatmap comparing mental health disorders across five psychological impact dimensions. Gambling Disorder and Suicidal Ideation show extreme reward sensitivity dysregulation. Decision stabil-

Table 12: Mental Health Disorders Triggered by Gambling Promotion Content

Mental Health Disorder	Risk Level	Primary Triggers	Mechanism of Action	Clinical Consequence
Gambling Disorder (Ludomania)	SEVERE	"Jackpot," "Spin," "Win," "Luck"	Variable Ratio Reinforcement: hijacks the brain's dopamine reward system through uncertain outcomes.	Pathological need to bet; inability to stop despite severe consequences.
Suicidal Ideation & Risk	CRITICAL	"Recover Losses," "Deposit," high-stakes betting	Serotonin/Dopamine Crash: chasing losses leads to sudden financial ruin and hopelessness.	Sudden drop in mental stability; high risk of self-harm following major losses.
Impulse Control Disorder	HIGH	"Hurry," "Now," "Limited," "Exclusive"	Executive Function Bypass: urgency cues disable the logical Prefrontal Cortex.	Inability to delay gratification; reckless behavior in finance and daily life.
Generalized Anxiety Disorder	HIGH	"Deposit Match," "Withdrawal," "Bet"	Cortisol Dysregulation: volatility of wins/losses maintains a chronic Fight-or-Flight state.	Panic attacks, sleep disturbances (insomnia), and chronic physical stress.
Depression (Situational)	HIGH	Post-loss realization	Reward System Burnout: the brain becomes desensitized to normal joys after high-dopamine gambling spikes.	Anhedonia, lethargy, and persistent feelings of worthlessness.
Body Dysmorphia / Social Envy	MODERATE	"Record-breaking," idealised athlete imagery	Social Comparison Theory: constant exposure to idealized highlight reels of others.	Low self-esteem, feelings of inadequacy, and a need for validation.

ity collapses most sharply in gambling-linked disorders. Reward system hijacking emerges as the primary neuropsychological mechanism driving severe outcomes.

L.5 Cognitive & Neurobehavioural Radar Profile

A radar chart visualizing cognitive dysfunction patterns across disorders. Gambling Disorder and Suicidal Ideation produce spiked, asymmetric profiles indicating instability, whereas anxiety and depression show more balanced but persistent impairment. Decision stability is most compromised in gambling-related conditions, highlighting a qualitatively distinct harm profile.

M Summary of Clinical Findings

All five visual artifacts and the quantitative disorder analysis converge on a single conclusion: gambling content in this dataset operates as a high-risk psychological system, not neutral entertainment, and significantly elevates the probability of severe mental health outcomes, including suicidal ideation.

The dataset demonstrates a *Dual-Diagnosis Trigger* structure, simultaneously promoting social inadequacy via sports comparison and severe addiction via gambling predation. The juxtaposition of safe entertainment with high-risk financial triggers the Trojan Horse Effect, normalizes gambling behavior, and lowers user psychological defenses, exploiting neurological reward mechanisms without adequate warning or consent.