Deepfake Text Detection on Social Media: A Comprehensive Review of FastText and Deep Learning Approaches

M. Shravani

PG scholor, M. TECH (CSE),

JNTUH University College of Engineering, Jagityal (Autonomous), Nachupally (Kondagattu), Tg

Dr. B. Sateesh Kumar

Prof, Head of The Department Of CSE

JNTUH University College of Engineering

Jagtial (Autonomous), Nachupally (Kondagattu), TG

Abstract

The rapid growth of social media platforms has provided an effective medium for communication, information sharing, and public discourse. However, it has also enabled the widespread dissemination of misinformation and synthetic content, such as deepfake text generated by advanced language models. Detecting machine-generated tweets is crucial to preserving the authenticity of online interactions and safeguarding against malicious manipulation. This review paper explores existing approaches to deepfake detection on social media with a particular emphasis on deep learning techniques and FastText embeddings. We present a comprehensive survey of traditional machine learning methods, recent advancements in natural language processing (NLP), and hybrid deep learning architectures that enhance detection accuracy. Further, we analyze the strengths, limitations, and challenges associated with semantic representation methods, contextual embeddings, and classification models. By consolidating current research, this paper highlights emerging trends and future directions in combating deepfake text, aiming to contribute to the development of more robust and scalable detection frameworks for real-world social media environments.

Introduction

The proliferation of social media platforms has revolutionized the way people communicate, share information, and engage in political, cultural, and social discourse. However, alongside its benefits, social media has become a fertile ground for the dissemination of misinformation, disinformation, and synthetic content. Among these, **deepfake text**—synthetic text generated by advanced language models—poses a growing threat to the integrity of online communication and public trust. Unlike manipulated images and videos that are often visually detectable, text-based deepfakes are subtle, contextually coherent, and increasingly indistinguishable from human-authored content, making their detection far more challenging (Westerlund, 2019; Liu et al., 2021).

The dangers of such machine-generated content extend beyond casual misinformation. Studies show that false news spreads more rapidly and widely than truthful information on social media (Vosoughi, Roy, & Aral, 2018), and organized disinformation campaigns are frequently amplified by bots and automated accounts (Bradshaw, Bailey, & Howard, 2021; Grimme et al., 2017). This has fueled concerns over political manipulation, erosion of trust in digital ecosystems, and large-scale social engineering attacks (Ternovski, Kalla, & Aronow, 2021). Consequently, the detection of deepfake tweets has emerged as a critical research area in natural language processing (NLP), data mining, and computational social science.

Early efforts at detecting social bots and machine-generated content relied on handcrafted features and rule-based heuristics (Siddiqui et al., 2017). However, the rapid advancement of generative language models such as GPT-2 and GPT-3 (Lee & Hsiang, 2020; Dale, 2021) has rendered these methods insufficient. Modern approaches increasingly leverage **deep learning architectures**, contextual embeddings, and hybrid frameworks to improve detection robustness (Zellers et al., 2019; Gehrmann, Strobelt, & Rush, 2019). Among embedding techniques, **FastText** has emerged as an efficient and scalable word representation model that captures semantic and subword-level information, making it particularly valuable for classifying noisy, short, and linguistically diverse social media texts (Fagni et al., 2021).

This review paper surveys the evolving landscape of deepfake text detection on social media, with a focus on deep learning and FastText-based approaches for identifying machine-generated tweets. It explores traditional methods, recent advances, benchmark datasets, and state-of-the-art frameworks while also discussing current limitations and future directions. By synthesizing insights from existing literature, this work aims to provide a comprehensive understanding of the challenges and opportunities in building effective detection systems capable of addressing the rising tide of AI-generated disinformation.

Literature Survey

The emergence of deepfake technologies has significantly transformed the landscape of online communication, with synthetic text generation becoming one of the most concerning developments. Early studies have highlighted the potential of deepfake technologies to manipulate information and disrupt trust in digital ecosystems (Westerlund, 2019). With the advent of large-scale generative models such as GPT-2 and GPT-3, researchers have observed that machine-generated text can closely mimic human language, making it difficult to detect and highly effective in spreading disinformation across platforms (Liu et al., 2021; Dale, 2021). Popular accounts of GPT-3, including experiments in which bots posted undetected on social platforms, underscore the risks associated with unregulated deployment of such systems (Heaven, 2020).

The threat of synthetic text becomes more severe in the context of social media, where misinformation spreads faster than verified information and often reaches a wider audience (Vosoughi, Roy, & Aral, 2018). Studies also reveal how organized disinformation campaigns rely heavily on social bots and semi-automated accounts to amplify misleading narratives at scale (Bradshaw, Bailey, & Howard, 2021; Grimme et al., 2017). Such campaigns not only create noise in digital communication but also influence political discourse,

manipulate public opinion, and erode trust in online communities (Ternovski, Kalla, & Aronow, 2021). The industrialized nature of online manipulation highlights the urgent need for effective detection strategies tailored to the dynamics of platforms like Twitter.

Early detection efforts relied on heuristic features such as posting frequency, n-gram distributions, and syntactic anomalies. While effective against simple bots, these methods were insufficient against modern generators. Statistical approaches such as GLTR (Gehrmann, Strobelt, & Rush, 2019) provided visual and statistical indicators to differentiate human from machine-generated text, but the improvement in generative models quickly reduced their effectiveness. More recent works employ deep learning classifiers and transformer-based architectures, such as Grover (Zellers et al., 2019), which demonstrated strong performance in distinguishing neural fake news but also highlighted the fragility of detectors when applied to new generators. The literature thus points to an arms race between increasingly sophisticated generation models and the detection strategies designed to counter them.

Among embedding techniques, FastText has gained considerable attention due to its ability to handle noisy, short, and linguistically diverse content typical of social media. Unlike traditional word embeddings, FastText incorporates subword information through character n-grams, which enhances its robustness against out-of-vocabulary tokens and informal text variations. Research such as TweepFake (Fagni et al., 2021) demonstrates the effectiveness of FastText features combined with deep classifiers in detecting machine-generated tweets. This is particularly relevant given the brevity and variability of tweets, where context is limited and standard embeddings often underperform. The efficiency of FastText also makes it practical for large-scale applications where computational resources are constrained.

Benchmark datasets play a critical role in advancing research in this area. TweepFake, for instance, provides a collection of human and machine-generated tweets, enabling researchers to train and evaluate detection models under realistic conditions (Fagni et al., 2021). Similarly, TuringBench (Uchendu et al., 2021) offers a benchmark for distinguishing human-authored text from outputs of multiple neural generators. Despite their utility, these datasets face limitations such as temporal drift, limited generator diversity, and label noise, which affect the generalization capability of detection models. Continuous dataset updates and inclusion of diverse generator families are emphasized across the literature as essential for building robust detection frameworks.

Beyond text-based features, some studies incorporate social network signals and metadata to improve classification accuracy. For example, posting patterns, account profiles, and retweet networks provide contextual cues that complement text-based embeddings and enhance robustness against adversarial attacks (Siddiqui et al., 2017; Grimme et al., 2017). However, reliance on social features introduces ethical and privacy concerns, particularly when applied at scale. Human evaluation studies also suggest that people often struggle to discern machine-generated content even when warnings are provided, indicating that automated systems remain central to the defense against deepfakes (Ternovski, Kalla, & Aronow, 2021).

Overall, the literature highlights both progress and persisting gaps in deepfake text detection research. While advances in deep learning and embedding techniques, particularly FastText, have enhanced detection accuracy on social media, the rapid evolution of generative models continues to challenge existing approaches. Research consistently underscores the need for generator-agnostic, adversarially robust methods, continual dataset updates, and hybrid detection frameworks that combine linguistic, social, and behavioral cues. At the same time, explainability, scalability, and policy integration emerge as critical areas for future exploration, ensuring that technological solutions are effective, transparent, and aligned with broader social and ethical considerations.

Table: Comparison of Existing Works on Deepfake Text Detection

Paper / Author	Methodology	Dataset Used	Strengths	Limitations
Siddiqui et al. (2017) [2]	Rule-based + heuristic bot detection	Social media accounts	Simple, interpretable features	Ineffective against advanced text generators
Westerlund (2019) [3]	Review of deepfake technologies	–	Broad overview of risks and applications	No technical detection framework
Vosoughi et al. (2018) [5]	Empirical analysis of misinformation spread	Twitter (true vs false news)	Shows how false news spreads faster than truth	Does not focus on automated detection
Bradshaw et al. (2021) [6]	Global inventory of disinformation	Multiplatform analysis	Highlights industrialized manipulation	Lacks algorithmic detection methods
Gehrmann et al. (2019) [14]	GLTR – statistical analysis of token likelihood	Machine vs human text samples	Visualization and interpretable detection	Performance drops with advanced models
Zellers et al. (2019) [9][16]	Grover – transformer-based fake news detector	News articles dataset	High accuracy on known generators	Poor generalization to unseen models
Adelani et al. (2020) [15]	Neural models generating fake reviews + detection	Online review datasets	Tests human + machine detection	Domain-specific, not social media tweets
Uchendu et al. (2021) [18]	TuringBench – benchmark for neural text detection	Multi-generator datasets	Standardized evaluation framework	Dataset limited in diversity and scale
Fagni et al. (2021) [19]	TweepFake – FastText + classifiers	Human & machine-generated	Focused on Twitter, scalable embedding	Needs integration with deep contextual

		tweets		models
Stiff & Johansson (2022) [20]	ML-based disinformation detection	Multisource datasets	Explores disinformation beyond text	General, not optimized for short tweets

Discussion

The literature clearly indicates that the rapid evolution of generative language models has created a dynamic arms race between text generation and detection. While early approaches based on handcrafted features and statistical heuristics offered partial success, they have quickly become obsolete against models such as GPT-2 and GPT-3 that produce highly coherent and contextually rich text. Deep learning approaches, particularly those leveraging embeddings, have shown greater promise, yet their effectiveness is still tied to the diversity of training data and their ability to generalize to unseen generators. This raises an important challenge of adaptability: detection systems must not only recognize known patterns but also anticipate future variations of synthetic text.

A recurring theme across the reviewed works is the importance of embeddings in handling noisy and short social media text. FastText, with its subword-aware representation, has demonstrated significant potential in capturing the nuances of tweets where traditional embeddings or transformers often struggle due to limited context. Its scalability and robustness to vocabulary variations make it especially well-suited for real-world deployment. However, FastText alone may not be sufficient against adversarial or context-rich synthetic content, suggesting that hybrid models combining FastText with transformer-based contextual embeddings could provide a more balanced and resilient approach.

Another dimension highlighted in the literature is the role of social and behavioral features. While content-based detection remains the central focus, metadata such as posting frequency, network patterns, and retweet cascades can enhance detection accuracy by identifying coordinated or bot-like behavior. Integrating such multimodal signals, however, raises ethical and privacy considerations that require careful handling, especially when applied at scale across social platforms. Moreover, the sustainability of detection efforts depends heavily on continuously updated benchmark datasets such as TweepFake and TuringBench, which must evolve alongside emerging generators to avoid obsolescence.

Despite progress, several gaps remain unresolved. Generalization across unseen generators is still weak, explainability of detection models is limited, and adversarial robustness remains a pressing issue. Additionally, the regulatory and ethical dimensions of deploying automated detectors are still underdeveloped, with existing policies showing inconsistencies across jurisdictions. These challenges suggest that technical advances must be complemented by policy frameworks and human-in-the-loop systems to ensure reliability, transparency, and trustworthiness.

Conclusion

The detection of deepfake text on social media has emerged as a critical research challenge in the era of advanced language models. This review highlights that while deep learning and embedding-based approaches have achieved substantial progress, the problem remains far from solved due to the constantly evolving sophistication of text generators. FastText, with its subword-level representation, stands out as a particularly effective technique for dealing with short and noisy social media text, making it a valuable component in modern detection pipelines. However, to achieve robust and scalable detection, FastText must be integrated with deep contextual models, adversarially robust architectures, and multimodal signals that extend beyond textual features.

Looking ahead, the development of dynamic, generator-agnostic benchmarks, hybrid detection frameworks, and explainable AI systems will be essential for strengthening defenses against machine-generated disinformation. Moreover, detection technologies must be aligned with ethical guidelines and supported by policy interventions to address the societal risks posed by deepfake content. Ultimately, the fight against synthetic disinformation requires a multidisciplinary effort that combines technical innovation, social awareness, and regulatory oversight to preserve the integrity of digital communication in an increasingly AI-driven world.

References:

[1] Jai Prakash Verma, Smita Agrawal, Bankim Patel, and Atul Patel. Bigdata analytics: Challenges and applications for text, audio, video, andsocial media data. International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), 5(1):41–51, 2016.

[2] Husna Siddiqui, Elizabeth Healy, and Aspen Olmsted. Bot or not. In2017 12th international conference for internet technology and secured transactions (ICITST), pages 462–463. IEEE, 2017.

[3] Mika Westerlund. The emergence of deepfake technology: A review. Technology innovation management review, 9(11), 2019.

[4] John Ternovski, Joshua Kalla, and Peter M Aronow. Deepfake warningsfor political videos increase disbelief but do not improve discernment:Evidence from two experiments. 2021.

[5] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. science, 359(6380):1146–1151, 2018.

[6] Samantha Bradshaw, Hannah Bailey, and Philip N Howard. Industrialized disinformation: 2020 global inventory of organized social media manipulation. Computational Propaganda Project at the Oxford Internet Institute,2021.

[7] Christian Grimme, Mike Preuss, Lena Adam, and Heike Trautmann.Social bots: Human-like by means of human control? Big data, 5(4):279–293, 2017.

[8] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian,Zhilin Yang, and Jie Tang. Gpt understands, too. arXiv preprintarXiv:2103.10385, 2021.

[9] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, AliFarhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. Advances in neural information processing systems, 32, 2019.

[10] Logan Beckman. The inconsistent application of internet regulations and suggestions for the future. Nova L. Rev., 46:277, 2021.

[11] Jieh-Sheng Lee and Jieh Hsiang. Patent claim generation by fine-tuningopenai gpt-2. World Patent Information, 62:101983, 2020.

[12] Robert Dale. Gpt-3: What's it good for? Natural Language Engineering,27(1):113–118, 2021.

[13] Will Douglas Heaven. A gpt-3 bot posted comments on reddit for aweek and no one noticed. MIT Technology Review. Retrieved November,24:2020, 2020.

[14] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr:Statistical detection and visualization of generated text. arXiv preprintarXiv:1906.04043, 2019.

[15] David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H Nguyen,Junichi Yamagishi, and Isao Echizen. Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection. In Advanced Information Networking and Ap-plications: Proceedings of the 34th International Conference on Advanced Information Networking and Applications (AINA-2020), pages 1341–1354. Springer, 2020.

[16] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, AliFarhadi, Franziska Roesner, and Yejin Choi. Grover-a state-of-the-artdefense against neural fake news, 2019.

[17] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong,and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858, 2019.
[18] Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee.Turingbench: A benchmark environment for turing test in the age of neural text generation. arXiv preprint arXiv:2109.13296, 2021.

[19] Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella,and Maurizio Tesconi. Tweepfake: About detecting deepfake tweets. Plosone, 16(5):e0251415, 2021.

[20] Harald Stiff and Fredrik Johansson. Detecting computer-generated disinformation. International Journal of Data Science and Analytics,13(4):363–383, 2022.