Sentiment Analysis of Textual Data in Social Networks: A Survey of Existing Approaches

Vusal Shahbazov
Institute of Information Technology, Baku, Azerbaijan
v.shahbazov@iit.science.az

***Abstract*— The growth of social media platforms results in billions of user-generated messages daily, making automated text analysis critical. The global impact of disinformation, the evolution of cyber threats toward psychological tactics, and the growing political and commercial value of public opinion have made sentiment analysis an important and relevant area of research. Although existing reviews have examined sentiment analysis approaches in terms of methods and application areas, few have evaluated these approaches in the context of cyber threat detection. This paper examines sentiment analysis methods applied to social media text data, covering lexicon-based, machine learning, and deep learning approaches, including transformer-based architectures, as well as widely used datasets. The paper also discusses how sentiment analysis can be applied to the detection of threats that exploit human emotions, including phishing, disinformation, and social engineering. To support this, an empirical analysis of large language model performance is conducted, measuring their ability to detect emotionally manipulative content. The purpose of this paper is to provide readers with an objective understanding of sentiment analysis and its role as a defense against socially engineered cyber threats.**

***Keywords— sentiment analysis; emotional analysis; natural language processing; social networks; cyber threat detection*.**

I. INTRODUCTİON

Over the past decade, social media has transformed from a communication tool into the primary channel through which billions of people consume information daily. As of 2024, more than 5.4 billion people use social media platforms worldwide [1] generating a combined total of approximately 500 million posts, comments, and messages every day [2]. This volume of text is not neutral. It shapes opinions, influences decisions, and is increasingly being used for malicious purposes.

The exploitation of human psychological vulnerabilities predates their systematic study in the academic literature. While hacking a well-protected system requires technical skill, persistence, and luck, sending a convincing email requires none of these. Unlike technical systems, influencing people does not require exploiting software vulnerabilities; they are susceptible to psychological manipulation. This is the underlying logic of social engineering and phishing attacks. It is much easier to manipulate a person than to hack a secure system. The FBI's Internet Crime Complaint Center consistently ranks phishing among the most frequently reported cybercrimes [3], resulting in adjusted losses of over $2.7 billion in 2022 [4]. According to Verizon's 2023 Data Breach Investigations Report, the human element accounts for 74% of breaches, including social engineering such as phishing [5]. The effectiveness of these attacks is determined not by technical sophistication, but by emotional precision. The effectiveness of a phishing message does not depend on linguistic precision. Rather, it relies on its ability to elicit a targeted emotional response, such as urgency, fear, perceived authority, or anticipation, at a psychologically opportune moment. For example, messages alleging unauthorized account access are designed to exploit anxiety. Notifications of prizes or rewards target anticipation. Text is a means of emotional manipulation, and for this reason, the language within it has analytical value. The same logic applies to disinformation, coordinated campaigns that flood platforms with false narratives to sway public opinion [6]. Unlike phishing, which targets individuals, disinformation targets entire populations.

Sentiment analysis and emotion analysis are related, but not identical [7]. Sentiment analysis determines the polarity of a text as positive, negative, or neutral. This answers the question: is this opinion favorable or unfavorable? Emotion analysis goes deeper, identifying the specific emotional state being conveyed, such as joy, anger, fear, sadness, surprise, or disgust, in accordance with concepts such as Ekman's six basic emotions or Plutchik's wheel of emotions [8]. In the context of cybersecurity threats, this distinction has practical implications. A phishing email can convey a technically neutral or even positive sentiment, while simultaneously encoding strong emotional cues of urgency or authority designed to bypass rational judgment [9]. This field exists at a crossroads that is both technically challenging and practically relevant.

This article examines modern methods for identifying sentiment and emotional tone in text, focusing on approaches relevant to cybersecurity. It analyzes machine learning, deep learning, and transformer-based methods, discusses the datasets used for their training and evaluation, and provides an empirical comparison of the performance metrics presented in the reviewed literature. The goal is to determine the current state of the field, identify strengths, and identify remaining gaps.

II. LİTERATURE REVIEW

Over the past ten years, research in sentiment and emotion analysis has expanded significantly, driven by the growing volume of user-generated content on social media and the increasing use of text manipulation in online environments. This section reviews existing work in four areas: review articles, classical machine learning approaches, deep learning methods,

DOI: 10.25045/forensic.gov.2026.79

and transformer-based models. Fig. 1 illustrates the annual distribution of publications retrieved from Scopus between 2018 and 2026.

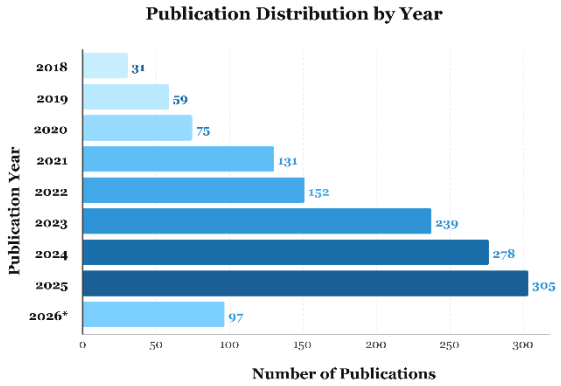

Figure 1. Annual publication distribution of sentiment and emotion analysis studies by year (Scopus, 2018–2026)

Review Studies

Nandwani and Verma (2021) presented a fundamental review covering both sentiment analysis and emotion detection in text. The authors compare approaches based on lexicon, machine learning, deep learning, and transfer learning. They also distinguish between multidimensional emotion models and categorical ones, such as Ekman's six basic emotions and Plutchik's wheel. The reported accuracy in the reviewed studies ranges from 80 to 90% for machine learning-based methods, with LSTM and CNN variants showing competitive results. Among the limitations identified by the authors: most datasets contain only English-language data; informal language degrades lexicon performance; sarcasm and irony cause misclassification. However, the review does not consider hostile or misleading texts, such as phishing, manipulation, and disinformation [7].

Rodriguez-Ibáñez et al. (2023) provides one of the most comprehensive reviews of sentiment analysis on social media. This work covers 2,306 articles from 2008 to 2021 in six domains: marketing, politics, economics, healthcare, emergencies, and general affairs. The review covers the entire methodological spectrum, from lexical-based and classical machine learning approaches to deep learning and transformer-based models. The work includes a direct reproducibility comparison at SemEval 2017, where fine-tuned models such as BERTweet significantly outperform GPT-3 and GPT-J in a zero-training setting. Additionally, the work shows that large linear machine learning models still require domain adaptation to be competitive. Key limitations: the datasets are largely static and domain-specific, making them ill-suited for evolving language, such as that of deceptive or manipulative texts. Comparison of research results is difficult due to inconsistencies in metrics and datasets. Most critically, the field of cybersecurity is absent from the review's taxonomy. This gap directly motivates the present study [10].

Despite the broad scope of these reviews, few studies address sentiment and emotion analysis in the context of cyberthreats. The reviews reviewed focus on areas such as marketing, politics, and healthcare. Hostile text forms such as phishing, social engineering, and disinformation have not been systematically addressed. This lack of research limits the applicability of existing findings to cybersecurity. A notable exception is recent work applying sentiment and emotion analysis directly to phishing victim data at scale [11], which demonstrates the practical value of NLP-based emotional profiling in real cybersecurity contexts.

Machine learning Approaches

Classical machine learning methods represent the earliest and most widely used basic approach to sentiment and emotion classification in the cybersecurity context. These approaches, such as SVM, Naive Bayes, logistic regression, and random forest, are typically combined with TF-IDF or bag-of-words feature representations and remain competitive in resource-constrained settings or where interpretability is critical. Table 1 summarises the reviewed studies, presenting the methods applied, evaluation metrics reported, and the cybersecurity domain addressed.

TABLE I. OVERVİEW OF MACHİNE LEARNİNG APPROACHES APPLİED TO SENTİMENT ANALYSİS

Reference	Methods	Metrics	Domain
[12]	SVC, LR, Multinomial NB, RF, TF-IDF	SVC Acc: 88.0% LR: 86.9% NB Acc: 80.1% RF Acc: 79.0%	Cybersecurity / Privacy / Social Media
[13]	SVM + lexicon-based SA	Acc: 98.6% F1: 98.8%	Cybersecurity / Troll Detection
[14]	Extra Tree +sentiment features	Acc: 95.38% F1: 95%	Cybersecurity / Cyberbullying
[15]	LR, DT, NB, SVC, XGBoost	XGBoost (Best) Acc: 90%	Political / Conflict Analysis

The results of these studies demonstrate a robust performance ceiling for classical machine learning. Classical machine learning approaches have two practical advantages. They are computationally lightweight and provide interpretability. This is crucial in certain tasks where model transparency is required. However, these methods have a fundamental limitation. They rely on fixed, shallow feature representations, such as bag-of-words or TF-IDF, which reflect little contextual information. As a result, they struggle to generalize to new phrases. For new phrases not included in the training set, they will perform poorly.

Deep Learning Approaches

Deep learning models, primarily LSTM, BiLSTM, GRU, CNN, and RNN, have supplanted classical machine learning methods for sentiment analysis and threat detection in text, particularly in cases involving sequential context, long-term dependencies, and informal language patterns. Unlike the bag-of-words representations used in machine learning, deep learning architectures learn feature representations directly from data, making them more suitable for the highly emotionally charged language typical of phishing emails, fake news, and

social engineering attacks. Their main drawback remains interpretability and the need for large, labeled datasets, which are extremely rare in the cybersecurity context. Table 2 summarises the reviewed deep learning studies, presenting the architectures applied, evaluation metrics reported, and the cybersecurity domain addressed.

TABLE II. OVERVİEW OF DEEP LEARNİNG APPROACHES APPLİED TO SENTİMENT ANALYSİS

Reference	Methods	Metrics	Domain
[16]	GRU, LSTM, RNN with 39 hand-crafted features	GRU (best) Acc: 86.12%, Precision: 84% Recall: 83% F1: 83%	Fake News Disinformation
[17]	BiLSTM	Acc: 96.89%, F1: 97.81%	Fake News / Disinformation
[18]	LSTM, BiLSTM, GRU, BiGRU, CNN	BiGRU (best) Sad: F1=78.9% Angry: F1=74.8% Happy: F1=70.4%	Emotion Detection
[19]	LSTM	Acc: 98.75%	Medical
[20]	BiLSTM + GloVe	Acc: 96.63%	Social Media / NLP
[21]	LSTM + Word2Vec	Acc: 88.7% Precision: 85% Recall: 90% F1: 87%	Social Media / Public Opinion

Deep learning architectures offer a significant advantage over classical machine learning. They allow for the capture of richer contextual information in text, enabling better processing of language and emotional nuances. However, limitations are particularly relevant in the context of cybersecurity. Deep learning models require significantly larger labeled datasets to learn stable representations. Additionally, these models require more computing resources.

Transformer-Based Approaches

Transformer-based models, particularly BERT and its variants such as RoBERTa and DistilBERT, represent the state-of-the-art in threat detection based on natural language processing. Unlike LSTM or CNN architectures, Transformers employ self-attention mechanisms that capture contextual relationships throughout a text sequence, making them effective at detecting hidden emotions, sarcasm, and deceptive wording, which are central to phishing and disinformation. Pre-trained on large corpora and refined on domain-specific data, these models demonstrate high performance on relatively small labeled datasets. Recently, large language models such as GPT have been evaluated in zero- and small-example settings, offering the potential for threat detection without task-specific training. Table 3 summarises the reviewed transformer-based studies, presenting the architectures applied, evaluation metrics reported, and the cybersecurity domain addressed.

Transformer-based models offer two key advantages over previous approaches. First, their self-attention mechanism makes them globally context-aware. Transformers can associate any token with any other, enabling them to detect subtle cues such as sarcasm and deceptive phrasing. Second, pre-trained versions of Transformers consistently perform better, even without training. However, these advantages come at a cost. Training and fine-tuning Transformer-based models require significant computational resources. This limitation makes them impractical in resource-constrained settings.

TABLE III. OVERVİEW OF TRANSFORMER-BASED APPROACHES APPLİED TO SENTİMENT ANALYSİS

Reference	Methods	Metrics	Domain
[22]	8 emotions + BERT	Acc: 91% Precision: 90.5% Recall: 81.3% F1: 91%	Political Security Threat
[23]	SiEBERT+ DistilRoBERT Ekman's 6 emotions	Sentiment model Acc: 75% Emotion model Acc: 39%	Sentiment and Emotion Analysis Media Studies
[24]	RoBERTa + SenWave , 10 sentiment labels	F1: 59.1%	News Media COVID-19
[25]	DeBERTa + GRU	Acc: 96.87% Precision: 97% Recall: 97% F1: 97%	Sentiment Analysis Social Media / Twitter
[26]	DistilBERT + LSTM + CNN	IMDB: Acc: 90%, Precision: 90%, Recall: 90%, F1: 90% SST-2: Acc: 88%, Precision: 88%, Recall: 88%, F1: 88%	Movie reviews / General sentiment

Explainability and Interpretability

Despite the models' high prediction accuracy, they largely operate as black boxes. They provide little information about the reasons for their classification. In a security context, this is a significant limitation. Knowing that a model has labeled a message as emotionally manipulative is much less useful without understanding the characteristics that led to this conclusion.

Explainable artificial intelligence (XAI) methods, such as LIME and SHAP, were developed precisely to address this problem. These methods have been shown to be applicable to both natural language processing (NLP) and cybersecurity classification tasks. LIME has been used to explain predictions based on tweet data and malware samples [27]. SHAP, LIME, and decision trees have been applied to emotion-based security classifiers, demonstrating that XAI can reveal how adversarial manipulation biases model behavior in emotion recognition systems [28]. Despite these advances, applying XAI to text-based emotion classifiers specifically designed for cyberthreat detection remains an open problem.

Datasets

Dataset selection is a critical and often underappreciated factor in sentiment and emotion analysis research. Datasets vary significantly in size, annotation scheme, language, domain, and label granularity, making direct comparisons between studies unreliable. In the context of cybersecurity, this problem is compounded by the scarcity of labeled corpora specific to specific attacks. Most of the reviewed studies rely on generic social media datasets that roughly, but not directly, represent the language of phishing, manipulation, or disinformation campaigns. This section examines the most frequently used

datasets in the reviewed literature, outlining their key characteristics. Table 4 summarises the top datasets identified across all reviewed papers, ordered by frequency of use.

TABLE IV. DATASETS USED İN SENTİMENT AND EMOTİON ANALYSİS RESEARCH RELEVANT TO CYBERSECURİTY AND THREAT DETECTİON

Dataset	Size	Labels	Domain
Twitter / X datasets	10k–2.5M	Positive, Negative, Neutral	cyberbullying, disinformation, phishing
VADER Lexicon	~7,500 lexicon entries	Sentiment polarity + intensity score	Social media text
SemEval (2014–2018)	5,936–62,000	3–4 classes (sentiment, emotion)	Twitter, restaurant and laptop reviews
GoEmotions	58,000 Reddit comments	27 emotions + neutral	Social media
COVID-19 Tweets	50k–17M	Sentiment, emotion, fake news	Health disinformation
ISEAR	~7,500 sentences	7 emotions (Ekman-based)	Self-reported emotional experiences
IMDB Reviews	50,000 reviews	Binary (positive, negative)	Movie sentiment

A common limitation of all the datasets examined is their clear bias toward English-language content. Most widely used corpora were created primarily from English-language sources. This creates a significant gap in multilingual threat detection. Phishing, disinformation, and social engineering threats can be conducted in multiple languages, and applying these models will not yield satisfactory results. The lack of labeled data in other languages for sentiment and emotion analysis in cybersecurity scenarios remains an unsolved problem in this field.

III. METHODOLOGY

Overview

This section describes the empirical methodology used to investigate whether emotional cues embedded in fake content have discriminatory power for detecting dangerous texts. The study is based on a three-step process: dataset preprocessing, emotional feature extraction, and text classification. Each step is described in detail in the sections below.

The empirical analysis is conducted on the WELFake dataset [29]. This dataset used for is for fake news detection. It is publicly available. Each entry contains title, full article text and a binary label (fake, real). The headline and article text are used as input for an emotional extractor. About 1.3% of messages were rejected because they were not in English. Repeating spaces and unnecessary characters were removed. The dataset contains 72,134 labelled news articles.

Emotion Feature Extraction

Emotion features are extracted using SamLowe/roberta-base-go_emotions, a RoBERTa-based model fine-tuned on the GoEmotions dataset [30]. GoEmotions is a large-scale, annotated dataset . The model was chosen for three reasons: it is based on a widely cited dataset; it produces fine-grained multiclass probability distributions across 28 categories; and it is publicly available via the HuggingFace Model Hub. This last point ensures full reproducibility. The model associates each input text with 28 emotions. The full set is presented in Table 5.

TABLE V. THE 28 AFFECT CATEGORİES İN THE GOEMOTİONS TAXONOMY

#	Emotion	#	Emotion	#	Emotion	#	Emotion
1	admiration	8	desire	15	grief	22	relief
2	amusement	9	disappointment	16	joy	23	remorse
3	anger	10	disapproval	17	love	24	sadness
4	annoyance	11	disgust	18	nervousness	25	surprise
5	approval	12	embarrassment	19	optimism	26	gratitude
6	caring	13	excitement	20	pride	27	curiosity
7	confusion	14	fear	21	realization	28	neutral

For each document d in the dataset, the model produces a probability score for each of the 28 affect categories. These scores constitute an emotion feature vector E(d), defined as follows:

$$E(d) = [e_1, e_2, e_3, ..., e_{27}, e_{28}] \quad (1)$$

$$\text{where } e_i = P(\text{emotion}_i \mid d),\ \sum_i e_i = 1,\ e_i \in [0, 1] \quad (2)$$

where e_i is the softmax-normalised probability score for the i-th emotion class given document d. For documents exceeding the 512-token limit of RoBERTa, only the first 512 tokens are used as input.

Classification

Three supervised classifiers are trained and evaluated. Table 6 shows applied machine learning classifiers

TABLE VI. SUPERVİSED CLASSİFİERS AND OPTİMİZED HYPERPARAMETERS USED İN THE CLASSİFİCATİON EXPERİMENTS

Classifier	Abbreviation	Best Parameters
Logistic Regression	LR	*C = 100. tol = 0.001*
Random Forest	RF	*n_estimators = 400 max_depth = None*
Support Vector Machine	SVM	*C = 1000.0 max_iter = 50,000*

Logistic regression (LR) is one of the most established algorithms. Instead of simply assigning a rigid label, it estimates the probability of a given observation belonging to each class, making it naturally interpretable and computationally lightweight. Despite its simplicity, it serves as a robust and transparent base model, especially when the underlying relationship between features and class labels is linear.

Random forest (RF) takes a fundamentally different approach: instead of training a single model, it constructs a large set of decision trees, each trained on a different random subset of data and features. The resulting predictions are combined into a final prediction. The result is a robust nonlinear classifier that handles high-dimensional features well.

Support vector machine (SVM) , implemented in this study as LinearSVC from scikit-learn, works by finding a decision boundary that separates two classes with the maximum possible

margin. The linear variant is particularly practical in high-dimensional spaces, where it remains computationally efficient without sacrificing discriminatory power.

IV. RESULTS

Features Discriminative Signatures

Fig. 2 presents the six emotion categories yielding the largest average differences in softmax scores between real and fake articles. A consistent directional pattern was observed across these categories with the highest discriminant scores: fake articles received higher classifier scores than real ones for most emotions, with the notable exception of approval.

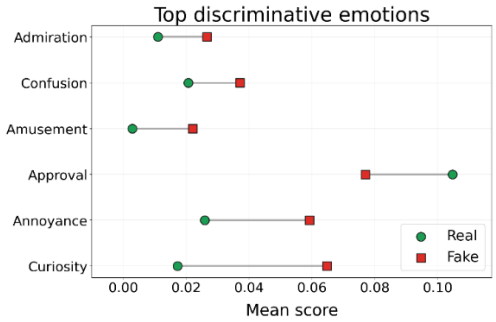

Figure 2. Six emotion categories with largest average softmax score differences between real and fake articles.

Curiosity showed the largest absolute difference: fake scored approximately 0.063 compared to 0.020 for real ($\Delta \approx +0.043$). These were followed by annoyance and amusement, where fake expression ratings exceeded real ones by approximately 0.036 and 0.026, respectively. Admiration and confusion showed moderate differences ($\Delta \approx 0.018$ and 0.016). In contrast, approval was the only category in which real articles yielded significantly higher ratings than fake ones (real ≈ 0.104 vs. fake ≈ 0.079, $\Delta \approx -0.025$). This suggests that real articles of approval convey a distinct signal that fake articles fail to replicate.

Fig. 3 presents the mean differences in softmax scores between fake and real articles for six Ekman-style basic emotions: surprise, anger, disgust, fear, joy, and sadness.

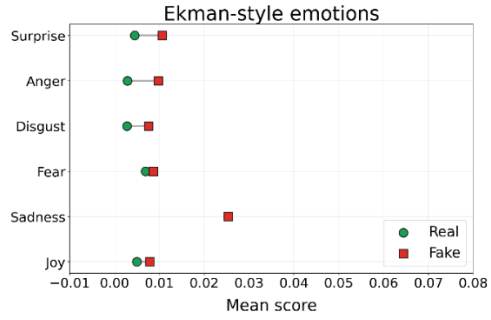

Figure 3. Mean differences in softmax scores for six Ekman-style basic emotions.

For most of Ekman's emotions, the differences between fake and real were relatively small, generally limited to the range of 0.002–0.005, with fake articles slightly outperforming real for emotions such as anger, disgust, surprise, fear, and joy. This suggests that Ekman's category classifiers are less sensitive. Sadness showed no meaningful difference between conditions, with fake and real receiving virtually identical mean softmax scores (≈ 0.026).

As shown in Fig. 4, real articles received significantly higher neutral ratings ($M \approx 0.72$) than fake ($M \approx 0.60$), yielding a difference of Δ (Fake − Real) = −0.122.

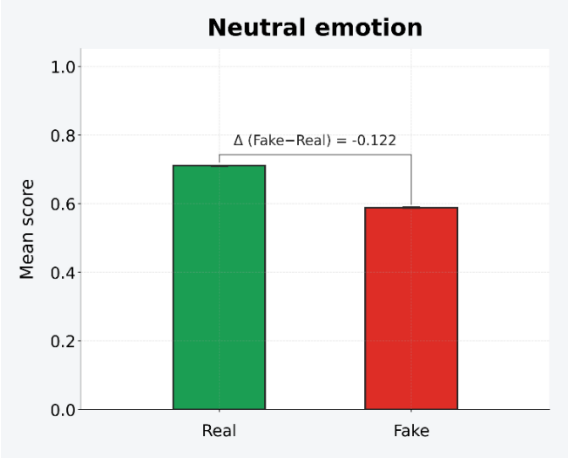

Figure 4. Neutral score comparison between real and fake articles

This result has important theoretical implications.

Classification performance

Table 7 reports the accuracy, macro-averaged precision, recall, and F1-score for each classifier.

TABLE VII. CLASSİFİCATİON PERFORMANCE OF THE THREE SUPERVİSED MODELS

Model	Accuracy	Macro Precision	Macro Recall	Macro F1
LR	0.7040	0.7154	0.7068	0.7017
RF	**0.8241**	**0.8251**	**0.823349**	**0.823675**
SVM	0.6997	0.7135	0.7028	0.6967

The Random Forest classifier performed best across all metrics, achieving an accuracy of 82.4% and a macro-F1 score of 0.824, indicating a strong and balanced ability to distinguish real from fake. Logistic regression demonstrated competitive results as a linear algorithm, with a macro-F1 score of 0.702. A significant portion of the discriminative signal is linearly separable in the feature space. Linear SVM, despite its theoretical suitability for high-dimensional input data, performed slightly worse than LR (macro-F1 = 0.697). The consistent gap between RF and the two linear models indicates the presence of nonlinear interactions between emotion features.

V. DİSCUSSİON

Positioning Against Existing Reviews

Despite the broad scope of sentiment analysis research, this paper identifies a significant and important gap. No existing review comprehensively addresses how sentiment and emotion analysis can be applied specifically to identify cybersecurity

threats in text. The most comprehensive review in this area, Rodriguez-Ibáñez et al. (2023), covers over 2,300 articles across six fields. However, cybersecurity is completely absent from its taxonomy [10]. Similarly, Nandwani and Verma (2021) acknowledge several limitations, including sarcasm and informal language, but do not consider hostile texts such as phishing or disinformation [7]. This paper attempts to partially fill this gap by explicitly examining sentiment and emotion analysis through the lens of cyberthreats, providing empirical evidence that emotional features carry a discriminatory signal in this context.

Main Gaps Identified in the Literature

Several recurring limitations were identified in the reviewed literature. First, most datasets are derived from general social media content and are not specifically designed for cyberthreat scenarios. This discrepancy is particularly problematic for phishing and social engineering, where the emotional impact is deliberate and subtle. Second, most studies conflate sentiment polarity with emotional content, treating them as interchangeable proxies. As discussed in Section 1, this distinction carries meaningful implications for cybersecurity applications. Third, comparison of results across studies remains unreliable due to inconsistencies in metrics, dataset partitioning, and annotation schemes. Finally, transformer-based models are increasingly dominating performance benchmarks. However, their interpretability in the context of security and forensic analysis remains unexplored.

Empirical Findings and the Role of Emotion Features

The empirical analysis presented in Section 4 is conducted using only emotional features, 28-dimensional probability vectors derived from the RoBERTa model. Lexical, syntactic, or structural features are not included. This is a deliberate choice, intended to isolate the discriminatory contribution of purely emotional content.

The results show that these features are sufficient to achieve a classification accuracy of up to 82.4% and a macro-F1 value of 0.824, which is significantly above chance. A particularly important finding is that fake news demonstrates systematically lower neutral ratings ($\Delta = -0.122$). This suggests that fabricated content is structurally more emotionally charged than real content, consistent with the theoretical concept of disinformation as a form of emotional manipulation. Furthermore, the key discriminatory categories are curiosity, annoyance, amusement, and approval. However, these are not canonical emotions in the Ekman style. This suggests that fine-grained emotion models like GoEmotions may be more informative for deception detection than classic six-emotion models. Therefore, they capture more subtle emotional characteristics and are more suitable for identifying manipulative text.

Limitations

Several limitations of this study should be noted. First, the analysis is limited to English-language content. The WELFake dataset excluded articles in other languages, and the GoEmotions model was trained primarily on English-language Reddit comments. Second, the empirical analysis is conducted on a dataset of fake news, not phishing emails or social engineering messages. Although fake news and phishing share common mechanisms of emotional manipulation, their structure and intent differ significantly. Third, the study uses the first 512 tokens of each article due to the length limitation of RoBERTa's input data, which may result in the rejection of emotionally salient content present in longer texts.

CONCLUSİON

This article presents a structured review of sentiment and emotion analysis methods and their limitations, as well as an empirical study of the usefulness of features for detecting deceptive text. The review confirms that, although high technical results have been achieved in this field across various domains, the explicit application of emotion analysis to cybersecurity threat detection remains understudied. Empirical results indicate that emotion features themselves carry a significant discriminatory signal. Fake news exhibits systematically different emotional profiles from real news, with higher scores in most categories and lower neutrality. These patterns support the hypothesis that emotional manipulation is a structural property of deceptive content.

The obtained results suggest several directions for further research. The most important priority is to integrate emotional characteristics with additional types of features, including syntactic and metadata cues, to improve performance. A second direction concerns extending the empirical analysis to other threat categories, including phishing emails and social engineering campaigns. Third, systematic evaluation in different languages is necessary before making statements about the applicability of emotion-based detection. Finally, greater attention should be paid to the interpretability of emotion-based models. Understanding which specific emotional dimensions most reliably distinguish real content from manipulative content could facilitate automated detection systems.

REFERENCES

[1] S. A. Murtaza, A. Alasadi, and E. Molnár, "Scroll, Pause, and Repeat: How Social Media Fuels Procrastination and Disrupts Work-Life Balance," International Journal of Organizational Leadership , vol. 14, pp. 9–28, Feb. 2025, doi: 10.33844/ijol.2025.60456.

[2] M. Masson, "Cadre générique pour le traitement et l'analyse multidimensionnelle des réseaux sociaux : une approche proxémique," Ludwig-Maximilians-Universität München, 2024. Accessed: Oct. 2025. [Online]. Available: http://www.theses.fr/2024PAUU3080/document

[3] M. Tomsů, V. Rosíková, and M. Madlenak, "Social Engineering as a Problem in Small Organizations," in Annals of DAAAM for ... & proceedings of the ... International DAAAM Symposium , DAAAM International Vienna, 2024, pp. 171–176. doi: 10.2507/35th.daaam.proceedings.022.

[4] K. Nagata, "Establishing Information Security Policy as an Organizational Risk Management," in IntechOpen eBooks , IntechOpen, 2024. doi: 10.5772/intechopen.1004563.

[5] Y. Weng and J. Wu, "Fortifying the Global Data Fortress: A Multidimensional Examination of Cyber Security Indexes and Data Protection Measures across 193 Nations," International Journal of Frontiers in Engineering Technology , vol. 6, no. 2, Jan. 2024, doi: 10.25236/ijfet.2024.060203.

[6] Z. Liu, T. Zhang, K. Yang, P. Thompson, Z. Yu, and S. Ananiadou, "Emotion Detection for Misinformation: A Review." Jan. 01, 2023. doi: 10.2139/ssrn.4625569.

[7] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," Social Network Analysis and Mining , vol. 11, no. 1, pp. 81–81, Aug. 2021, doi: 10.1007/s13278-021-00776-6.

[8] M. Tahir et al. , "Non-Acted Text and Keystrokes Database and Learning Methods to Recognize Emotions," ACM Transactions on Multimedia Computing Communications and Applications , vol. 18, no. 2, pp. 1–24, Feb. 2022, doi: 10.1145/3480968.

[9] J. Duan, Y. Tian, and Y. Liao, "The Emotional Lens of Cybersecurity: A Study on Sentiment Analysis Enhanced Phishing Detection via Large Language Models," Lecture notes in electrical engineering , pp. 217–225, Jan. 2025, doi: 10.1007/978-981-96-4245-8_18.

[10] M. Rodríguez-Ibáñez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," Expert Systems with Applications , vol. 223, pp. 119862–119862, Mar. 2023, doi: 10.1016/j.eswa.2023.119862.

[11] A. Chrysanthou, Y. Pantis, and C. Patsakis, "The anatomy of deception: Measuring technical and human factors of a large-scale phishing campaign," Computers & Security, vol. 140, p. 103780, 2024, doi: 10.1016/j.cose.2024.103780.

[12] J. R. Saura, D. Palacios‐Marqués, and D. R. Soriano, "Privacy concerns in social media UGC communities: Understanding user behavior sentiments in complex networks," Information Systems and e-Business Management , vol. 23, no. 1, pp. 125‐145, Mar. 2023, doi: 10.1007/s10257-023-00631-5.

[13] K. Machová, M. Mach, and M. Vasilko, "Comparison of Machine Learning and Sentiment Analysis in Detection of Suspicious Online Reviewers on Different Type of Data," Sensors , vol. 22, no. 1, pp. 155–155, Dec. 2021, doi: 10.3390/s22010155.

[14] M. F. Almufareh, N. Z. Jhanjhi, M. Humayun, G. N. Alwakid, D. Javed, and S. N. Almuayqil, "Integrating Sentiment Analysis With Machine Learning for Cyberbullying Detection on Social Media," IEEE Access , vol. 13, pp. 78348–78359, Jan. 2025, doi: 10.1109/access.2025.3558843.

[15] A. A. Maruf, Z. M. Ziyad, M. M. Haque, and F. Khanam, "Emotion Detection from Text and Sentiment Analysis of Ukraine Russia War using Machine Learning Technique," International Journal of Advanced Computer Science and Applications , vol. 13, no. 12, Jan. 2022, doi: 10.14569/ijacsa.2022.01312101.

[16] C. Iwendi, S. Mohan, S. Khan, E. Ibeke, A. Ahmadian, and T. Ciano, "Covid-19 fake news sentiment analysis," Computers & Electrical Engineering, vol. 101, pp. 107967–107967, Apr. 2022, doi: 10.1016/j.compeleceng.2022.107967.

[17] S. Kh. Hamed, M. J. A. Aziz, and M. R. Yaakub, "Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments," Sensors , vol. 23, no. 4, pp. 1748–1748, Feb. 2023, doi: 10.3390/s23041748.

[18] A. Bouchekif, P. Joshi, L. Bouchekif, and H. Afli, "EPITA-ADAPT at SemEval-2019 Task 3: Detecting emotions in textual conversations using deep learning models combination," pp. 215–219, Jan. 2019, doi: 10.18653/v1/s19-2035.

[19] D. P. Dash, M. Kolekar, C. Chakraborty, and M. R. Khosravi, "Review of Machine and Deep Learning Techniques in Epileptic Seizure Detection using Physiological Signals and Sentiment Analysis," ACM Transactions on Asian and Low-Resource Language Information Processing , vol. 23, no. 1, pp. 1–29, Aug. 2022, doi: 10.1145/3552512.

[20] A. D. Yacoub, A. E. Aboutabl, and S. O. Slim, "Multilingual Sarcasm Detection for Enhancing Sentiment Analysis using Deep Learning Algorithms," Journal of Communications Software and Systems , vol. 20, no. 4, pp. 278–289, Jan. 2024, doi: 10.24138/jcomss-2024-0071.

[21] W. S. Ismail, "Emotion Detection in Text: Advances in Sentiment Analysis Using Deep Learning," Journal of Wireless Mobile Networks Ubiquitous Computing and Dependable Applications , vol. 15, no. 1, pp. 17–26, Mar. 2024, doi: 10.58346/jowua.2024.i1.002.

[22] L. S. Zaabar et al. , "Sentiment and Emotion Analysis with Large Language Models for Political Security Prediction Framework," International Journal of Advanced Computer Science and Applications , vol. 16, no. 1, Jan. 2025, doi: 10.14569/ijacsa.2025.0160192.

[23] D. Rozado, R. Hughes, and J. Halberstadt, "Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models," PLoS ONE , vol. 17, no. 10, Oct. 2022, doi: 10.1371/journal.pone.0276367.

[24] R. Chandra, B. Zhu, Q. Fang, and E. Shinjikashvili, "Large language models for newspaper sentiment analysis during COVID-19: The Guardian," Applied Soft Computing , vol. 171, pp. 112743–112743, Jan. 2025, doi: 10.1016/j.asoc.2025.112743.

[25] A. Assiri, A. Gumaei, F. Mehmood, T. Abbas, and S. Ullah, "DeBERTa-GRU: Sentiment Analysis for Large Language Model," Computers, materials & continua/Computers, materials & continua (Print) , vol. 79, no. 3, pp. 4219–4236, Jan. 2024, doi: 10.32604/cmc.2024.050781.

[26] O. Haddad and M. N. Omri, "An Innovative Method for Improving Sentiment Analysis in Large Language Models," Procedia Computer Science , vol. 270, pp. 4605–4614, Jan. 2025, doi: 10.1016/j.procs.2025.09.586.

[27] S. M. Mathews, "Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review" in *Proc. Intell. Comput. (CompCom)*, Springer, Cham, 2019, pp. 1269–1292, doi: 10.1007/978-3-030-22868-2_90.

[28] Z. Zhang, A. Y. Al Hammadi, S. Yoon, E. Damiani, and C. Y. Yeun, "Explainable Data Poison Attacks on Human Emotion Evaluation Systems Based on EEG Signals," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3245813.

[29] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection" IEEE Transactions on Computational Social Systems , vol. 8, no. 4, pp. 881–893, Apr. 2021, doi: 10.1109/tcss.2021.3068519.

[30] D. Demszky, D. Movshovitz‐Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," pp. 4040‐4054, Jan. 2020.