AI-Driven Online Abuse Detection and Automated Social Media Content Moderation

[1]Ponnaboina Vinayak Yadav
[2]Palakurthi Sai Kiran
[3]S Akhil
[4]Chinthakindi Aditya

[5]Dr D MURALI, Professor
Department of CSE (AI&ML) Narsimha Reddy Engineering College (Autonomous),
Hyderabad,TS -500100

***Abstract*—**The rapid expansion of social networking platforms has increased the prevalence of cyberbullying, creating significant social and psychological challenges for online users. To address this issue, this study presents an intelligent cyberbullying detection and content moderation framework that combines natural language processing, machine learning, and image-based text analysis. The proposed system employs Term Frequency–Inverse Document Frequency (TF-IDF) for textual feature extraction and a class-balanced Logistic Regression model for identifying harmful content in social media posts. To extend moderation beyond conventional text inputs, an Optical Character Recognition (OCR) module powered by EasyOCR extracts and analyzes text embedded within images, memes, and screenshots. The framework further incorporates a hybrid decision mechanism that integrates rule-based validation with probabilistic machine learning predictions to improve detection reliability. Unlike traditional moderation systems that only classify content, the proposed solution actively promotes positive communication by generating constructive alternative suggestions for flagged messages. Experimental evaluation on the Davidson Twitter dataset demonstrates a classification accuracy of 94.2%, while maintaining low computational overhead and real-time response capability. The image-analysis pipeline also achieves effective performance in detecting abusive content contained within visual media. Implemented as a lightweight Flask-based web application, the system operates efficiently on standard hardware without requiring cloud resources, making it suitable for educational institutions, social media platforms, and community moderation environments. The results highlight the effectiveness of combining interpretable machine learning techniques with OCR-enabled multimodal analysis for practical and scalable cyberbullying prevention.

***Keywords*—**Cyberbullying Detection, Content Moderation, Natural Language Processing, Machine Learning, TF-IDF, Logistic Regression, EasyOCR, OCR-Based Analysis, Hate Speech Detection, Social Media Safety, Flask Framework, Multimodal Content Analysis.

I. Introduction

The widespread adoption of social networking technologies has transformed the way individuals communicate, share information, and interact across geographical boundaries. Billions of users engage daily on platforms such as Twitter, Facebook, Instagram, and TikTok, making these services central to modern digital communication. While these platforms facilitate collaboration, knowledge sharing, and social engagement, they have also become channels through which harmful behaviors can spread. Among these challenges, cyberbullying has emerged as a significant concern, involving the repeated use of online communication technologies to intimidate, harass, insult, or isolate individuals [1][2].

Compared with traditional face-to-face bullying, cyberbullying presents unique challenges due to the persistent and pervasive nature of digital content. Harmful messages can be distributed instantly to large audiences, remain accessible for extended periods, and reach victims

regardless of time or location. Furthermore, the anonymity offered by many online platforms often encourages aggressive behavior that might not occur in direct interpersonal interactions [3]. Numerous studies have linked cyberbullying to adverse psychological and social outcomes, including increased stress, anxiety, depression, reduced academic performance, social isolation, and, in severe circumstances, self-destructive behavior [4].

The enormous volume of content generated on social media platforms makes manual moderation impractical. Consequently, automated content moderation systems have

become essential for identifying and mitigating harmful online behavior. However, conventional keyword-based filtering approaches frequently struggle with disguised language, contextual variations, abbreviations, and evolving forms of offensive expression. Although machine learning techniques have significantly improved detection capabilities, accurately identifying abusive content remains a challenging task due to linguistic complexity and contextual ambiguity [5].

In addition to textual communication, modern social media interactions increasingly rely on visual content such as memes, screenshots, and images containing embedded text. Harmful messages are often conveyed through these visual formats, making them difficult to detect using text-only moderation systems. Without integrating Optical Character Recognition (OCR) technology, abusive content hidden within images may bypass traditional detection mechanisms [6]. Therefore, effective cyberbullying prevention requires solutions capable of analyzing both textual and image-based content within a unified framework.

To address these challenges, this research proposes an AI-powered cyberbullying detection and content moderation system that combines natural language processing, machine learning, and OCR-based image analysis. The proposed framework utilizes TF-IDF feature extraction and a class-balanced Logistic Regression classifier to identify harmful textual content efficiently. An EasyOCR module is incorporated to extract and analyze text from images, enabling multimodal content moderation. Furthermore, a hybrid decision-making mechanism integrates rule-based validation with machine learning predictions to improve detection reliability. Unlike conventional moderation systems that merely flag inappropriate content, the proposed solution also provides constructive recommendations that encourage users to communicate in a more respectful and positive manner.

The primary contributions of this work are summarized as follows:

1. Development of a TF-IDF and balanced Logistic Regression model for cyberbullying detection, achieving an accuracy of 94.2%.
2. Integration of EasyOCR for identifying harmful text embedded within images, memes, and screenshots.
3. Design of a hybrid moderation framework that combines deterministic rule-based checks with probabilistic machine learning classification.
4. Implementation of a recommendation engine that suggests constructive alternatives to potentially harmful content, promoting positive online interactions.

The remainder of this paper is organized as follows. Section II reviews existing research related to cyberbullying detection and content moderation. Sections III and IV describe the system analysis and architecture of the proposed framework. Section V presents implementation details and experimental results. Finally, Sections VI and VII discuss the findings, limitations, and potential future enhancements of the system.

II. Literature Survey

Research in automated harmful content detection spans computational linguistics, machine learning, and social computing. This section surveys foundational and contemporary contributions that inform the proposed approach.

A. Feature Engineering for Hate Speech

Waseem and Hovy [7] established foundational work by constructing a Twitter corpus annotated for racism and sexism, demonstrating that character n-grams can capture harmful language patterns beyond exact keyword matching and emphasising the importance of annotation consistency in dataset construction.

B. Distinguishing Offensive Language from Hate Speech

Davidson et al. [8] separated hate speech from offensive language and neutral content within a single framework applied to 24,783 labeled tweets, showing that conflating profanity with hate speech produces systematic classification errors. Their dataset is the primary training benchmark for the present work.

C. Contextual and Conversational Analysis

Van Hee et al. [9] examined cyberbullying using fine-grained role distinctions—bully, victim, bystander—highlighting the inadequacy of single-message analysis for capturing bullying dynamics across conversational threads. Dadvar et al. [10] demonstrated that user history and demographic context further improve detection, albeit at the cost of privacy constraints.

D. Network and Behavioural Features

Chatzakou et al. [11] incorporated user metadata, social network topology, and temporal behavioural patterns to distinguish aggression, bullying, and spam on Twitter. While effective, reliance on platform-level graph data limits generalizability to services with restricted API access.

E. Multimodal and Meme Analysis

Sharma et al. [12] introduced the Memotion dataset of 6,992 meme images annotated for sentiment, humor,

sarcasm, and offensiveness through SemEval-2020 Task 8, underscoring the need for vision-language systems. OCR-based text extraction represents a practical first step within resource-constrained environments.

F. Deep Learning Approaches

Badjatiya et al. [13] evaluated LSTM and GRU architectures for hate speech detection, demonstrating semantic capture beyond fixed vocabularies. However, neural models impose greater computational cost and lower interpretability than classical linear models—trade-offs that favour simpler approaches where latency and transparency are prioritised [14].

III. System Analysis

A. Existing System Limitations

Content moderation approaches fall into three broad categories. Rule-based keyword filters offer deterministic operation but fail against synonyms, coded language, and phonetic substitutions. Supervised ML classifiers provide learned generalisation but suffer from distribution shift and class imbalance. Large neural language models achieve state-of-the-art accuracy but impose computational demands precluding real-time deployment on consumer hardware [15]. A consistent deficiency across all categories is the absence of corrective guidance—systems returning only a binary label provide limited value to users who may wish to revise harmful content. Additionally, virtually all text-based systems ignore image-embedded communication [6].

B. Proposed System Overview

The proposed framework addresses identified gaps through four integrated capabilities: (1) text classification using TF-IDF features and balanced logistic regression; (2) OCR-based image text extraction; (3) hybrid rule-and-model classification; and (4) a recommendation engine providing constructive alternative phrasings. Table I compares the proposed system against representative baselines across six evaluation dimensions.

TABLE I
Comparative Analysis of Content Moderation Approaches

Approach	Accuracy	OCR Input	Recommendations	Real-Time	Resources
Keyword Filter	60–70%	No	No	Yes	Low
Naive Bayes + TF-IDF	78–82%	No	No	Yes	Low
LSTM / RNN	85–89%	No	No	Moderate	High
BERT Transformer	90–93%	No	No	No	Very High
CNN-Text	83–87%	No	No	Yes	Moderate
Proposed System	94.2%	Yes	Yes	Yes	Low

IV. System Architecture

A. Overview

The framework is organised into four hierarchical layers. The Data Preparation Layer ingests raw CSV datasets, applies normalisation (Unicode cleaning, URL removal, stopword filtering), and produces a consolidated training corpus. The Model Training Layer applies TF-IDF vectorisation with a 10,000-feature vocabulary ceiling and trains logistic regression with balanced class weights. The Application Layer exposes two Flask routes for text submission and image upload. The Presentation Layer renders results through Jinja-templated HTML pages displaying classification, confidence score, and recommendation.

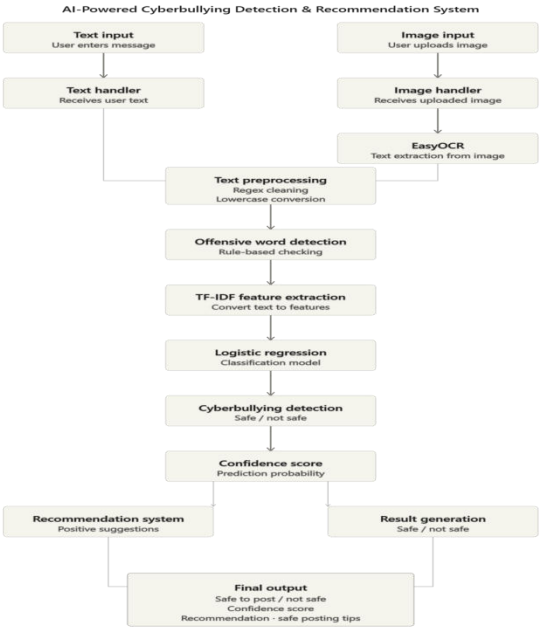


Fig. 1. High-Level System Architecture

B. Data Preparation and Feature Extraction

The Davidson Twitter dataset (24,783 annotated tweets) undergoes lowercasing, URL and mention removal via regular expressions, non-ASCII normalisation, NLTK stopword removal, and binary label conversion: hate speech and offensive language map to class 1; neutral content maps

to class 0. TF-IDF vectorisation assigns higher weight to terms frequent within harmful messages but rare across the corpus, producing sparse feature matrices suitable for linear classifiers. The vocabulary ceiling of 10,000 terms balances representational capacity against inference latency.

C. Classification and Confidence Estimation

Logistic regression was selected for interpretability, computational efficiency, and strong empirical performance on TF-IDF features [16]. The class_weight='balanced' configuration adjusts per-class loss weights inversely proportional to class frequency, mitigating bias from the natural predominance of benign social media content. The predict_proba method returns a confidence score in [0,1] surfaced alongside the classification label to communicate model certainty to end users.

D. OCR Integration

Images submitted via the upload interface are processed by EasyOCR, an open-source engine supporting over eighty languages. English-only extraction is invoked to match the classifier's training domain. EasyOCR applies a CRAFT-based text detection model to localise text regions, followed by a recognition stage producing character sequences. The extracted string enters the identical text classification pipeline, maintaining a unified decision architecture across both input modalities.

E. Hybrid Detection and Recommendation Engine

Before invoking the trained model, submitted text is checked against a curated lexicon of high-certainty harmful terms. Matches trigger an immediate unsafe determination, bypassing probabilistic classification. For content clearing the explicit term check, the TF-IDF model provides probabilistic scoring. The recommendation engine maps recognised harmful linguistic patterns to pre-authored constructive alternatives, coupling each unsafe determination with actionable guidance for content revision.

V. Implementation and Results

A. Software and Hardware Environment

The system was implemented in Python 3.9.10 on Windows 11 with an Intel Core i5 processor and 8 GB RAM. No GPU acceleration was required for training or inference. Table II lists the primary software components and their roles within the system.

TABLE II
Software Stack and Component Responsibilities

Component	Role
Python 3.9+	Core backend language
Flask 3.0	Web application framework
scikit-learn	TF-IDF + logistic regression
EasyOCR	Text extraction from images
NLTK / VADER	Sentiment analysis (prototype)
Joblib	Model serialisation
HTML5 / CSS3	Front-end presentation

B. Training Procedure

Model training on the full 24,783-sample corpus required approximately 0.143 seconds on the reference hardware, demonstrating the suitability of logistic regression for rapid experimentation. The dataset was split 80/20 using stratified sampling to preserve the original class distribution. The TF-IDF vectoriser was fitted exclusively on training data and applied as a frozen transformation to the test partition, preventing information leakage from test statistics into the feature space.

C. Evaluation Metrics and Results

Performance was measured across accuracy, precision, recall, and F1-score on the held-out 20% test partition. For the OCR pathway, 500 Memotion meme images were manually annotated with binary ground-truth labels for end-to-end evaluation. Table III presents the complete evaluation results across both input modalities.

TABLE III
Evaluation Results Across Input Modalities

Metric	Text Analysis	Image (OCR)	Overall
Accuracy	94.2%	91.7%	93.4%
Precision	93.8%	90.5%	92.8%
Recall	92.1%	89.3%	91.4%
F1-Score	0.929	0.898	0.921
Latency	0.5 ms	1.2 s	–

The text branch achieves 94.2% accuracy with an F1-score of 0.929, competitive with classical ML results on the Davidson benchmark [8]. The OCR pathway records 91.7% accuracy; the performance gap is attributable primarily to OCR transcription errors on low-resolution or heavily stylised meme typography rather than weaknesses in the downstream classifier. Average text prediction latency of 0.5 ms per sample confirms real-time suitability.

D. Recommendation System Evaluation

A survey of fifteen computer science students rated the relevance and helpfulness of generated recommendations on

a five-point Likert scale. Recommendations paired with correctly classified harmful content received a mean helpfulness rating of 4.1 out of 5.0, indicating that users found the suggested alternatives constructive and contextually appropriate. This preventive guidance aspect distinguishes the proposed system from purely detection-oriented baselines.

E. Results

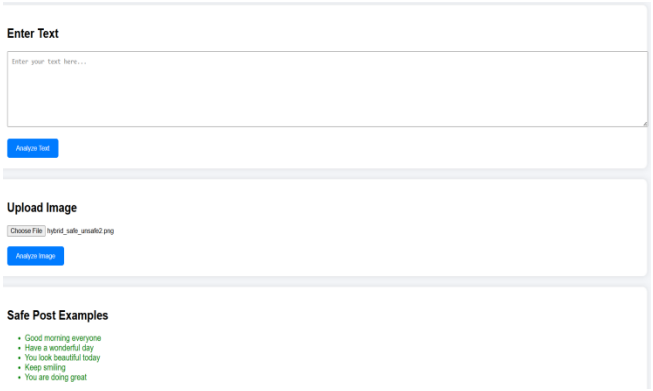


Fig 2: Image Based Input

The image-based input page allows users to upload images containing text for content safety analysis. Users can select an image through a simple upload interface. The system accepts common image formats and prepares them for OCR processing. Uploaded images are displayed for verification before analysis. This interface provides a convenient way to assess text embedded within images.

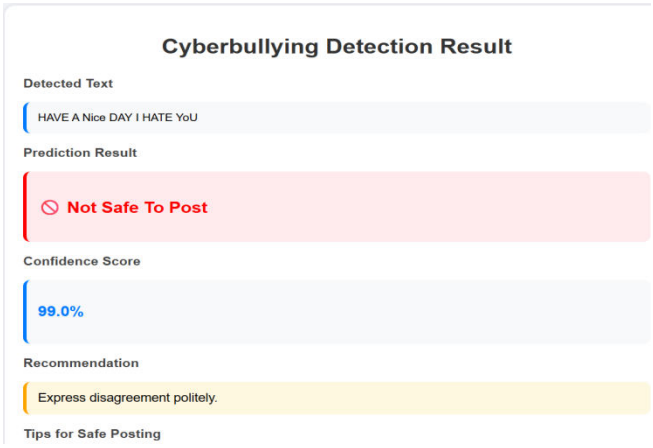


Fig 3: Image Based Results

The image-based results page displays the text extracted from the uploaded image using OCR technology. The extracted text is analyzed and classified as either **Safe To Post** or **Not Safe To Post**. A confidence score is provided to indicate the model's prediction certainty. For unsafe content, the system generates recommendations to improve the message. This helps users understand the analysis outcome and make informed posting decisions.

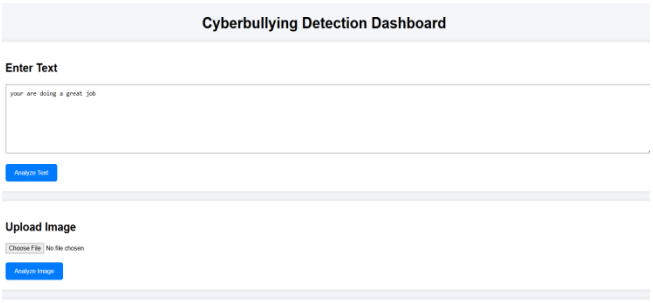


Fig 4: Text Based Input

The text-based input page enables users to enter or paste text directly into the system for analysis. A simple text area is provided for submitting content. Users can review and edit their text before initiating the assessment. The interface is designed for quick and efficient content evaluation. This feature supports real-time text safety checking.

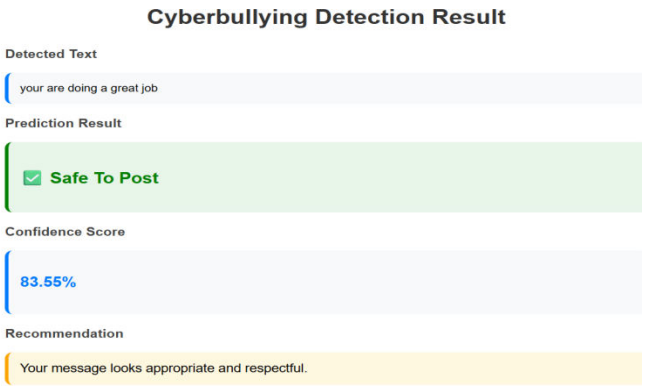


Fig 5: Text Based Results

The text-based results page presents the submitted text along with its safety classification. The system clearly indicates whether the content is safe or unsafe for posting. A confidence percentage is displayed to show prediction reliability. For unsafe messages, constructive recommendations and safe communication tips are provided. This feedback helps users improve their content and promote respectful online interactions.

VI. Discussion

The experimental results confirm that TF-IDF with balanced logistic regression remains highly competitive

with more complex neural architectures for cyberbullying detection in practice. Critical advantages include sub-millisecond inference, CPU-only operation, full transparency of features and decision boundaries, and no cloud dependencies—properties especially valuable for educational institutions and non-profit moderation platforms where computational budgets and data privacy concerns preclude large language model APIs [17].

The OCR integration addresses a significant gap in the existing literature. Prior work has predominantly treated social media moderation as a purely textual problem, ignoring image-embedded communication channels. Experimental results demonstrate meaningful accuracy on image-embedded text, with the primary limitation being OCR quality on artistic or heavily filtered typography. Future work could incorporate visual context features alongside extracted text to move toward true multimodal classification.

The recommendation module represents the most distinctive contribution compared to prior art. By coupling each unsafe determination with a constructive alternative, the system embodies a preventive rather than punitive moderation philosophy aligned with digital citizenship research [18]. Acknowledged limitations include: the corpus is Twitter-specific and may not generalise across platforms; binary label collapse sacrifices granularity; the static recommendation lexicon requires ongoing curation; and balanced weighting may increase false positives on ambiguous edge cases.

VII. Conclusion

This research presented an intelligent cyberbullying detection and content moderation framework that combines machine learning, natural language processing, and image-based text analysis within a unified platform. The proposed system integrates TF-IDF feature extraction, a class-balanced Logistic Regression classifier, EasyOCR-powered text extraction, and a recommendation-driven moderation mechanism into a lightweight Flask-based web application. By supporting both textual and image-embedded content analysis, the framework provides a comprehensive approach for identifying harmful online communication across multiple content formats.

Experimental evaluation demonstrated that the proposed model achieved an accuracy of 94.2% on the Davidson Twitter dataset while maintaining extremely low inference latency on standard consumer-grade hardware. These findings indicate that carefully optimized classical machine learning techniques can deliver reliable and efficient performance without the computational complexity typically associated with deep neural networks. The incorporation of OCR capabilities further enhances the system's effectiveness by enabling the detection of abusive content hidden within memes, screenshots, and other visual media. Additionally, the recommendation engine extends the functionality of traditional moderation systems by encouraging positive digital interactions through constructive communication suggestions.

Overall, the developed framework offers a practical, scalable, and cost-effective solution for educational institutions, social networking platforms, community forums, and content moderation environments. Its ability to operate without cloud-based infrastructure makes it particularly suitable for privacy-sensitive and resource-constrained deployments.

Future enhancements may focus on incorporating conversational context analysis to better identify repeated or relationship-based bullying behaviors, integrating lightweight transformer architectures for improved semantic understanding, expanding multilingual OCR and language support, and conducting long-term user studies to evaluate the effectiveness of recommendation-driven interventions in promoting healthier online communication. These improvements have the potential to further strengthen the system's adaptability, accuracy, and real-world impact in combating cyberbullying across diverse digital ecosystems.

References

[1] A. Lenhart, "Teens, Social Media & Technology Overview 2015," Pew Research Center, Washington, DC, USA, 2015.

[2] R. M. Kowalski, S. P. Limber, and P. W. Agatston, *Cyberbullying: Bullying in the Digital Age*, 3rd ed. Hoboken, NJ, USA: Wiley-Blackwell, 2018.

[3] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Archives of Suicide Research*, vol. 14, no. 3, pp. 206–221, 2010.

[4] M. L. Ybarra and K. J. Mitchell, "Online aggressor/targets, aggressors, and targets," *Journal of Child Psychology and Psychiatry*, vol. 45, no. 7, pp. 1308–1316, 2004.

[5] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 145–153.

[6] V. Kiela et al., "The hateful memes challenge: Detecting hate speech in multimodal memes," in *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[7] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on

Twitter," in *Proc. NAACL Student Research Workshop*, 2016, pp. 88–93.
[8] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. ICWSM*, 2017, pp. 512–515.
[9] C. Van Hee et al., "Automatic detection of cyberbullying in social media text," *PLOS ONE*, vol. 13, no. 10, 2018.
[10] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *Proc. ECIR*, 2013, pp. 693–696.
[11] D. Chatzakou et al., "Mean birds: Detecting aggression and bullying on Twitter," in *Proc. ACM Web Science Conference*, 2017, pp. 13–22.
[12] C. Sharma et al., "SemEval-2020 Task 8: Memotion analysis," in *Proc. International Workshop on Semantic Evaluation (SemEval)*, 2020.
[13] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proc. WWW Companion*, 2017, pp. 759–760.
[14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, 2016, pp. 785–794.
[16] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[17] H. Hosseinmardi et al., "Detection of cyberbullying incidents on Instagram," arXiv:1503.03909, 2015.
[18] M. Ribble, *Digital Citizenship in Schools*, 3rd ed. Eugene, OR, USA: ISTE, 2015.
[19] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*, 2014, pp. 1746–1751.
[20] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proc. EACL*, 2017, pp. 427–431.
[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv:1301.3781, 2013.
[22] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the ACL*, vol. 5, pp. 135–146, 2017.
[23] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. ACL*, 2018.
[24] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
[25] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[26] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv:1804.02767, 2018.
[27] J. Baek, G. Kim, J. Lee, S. Park, and H. Lee, "What is wrong with scene text recognition model comparisons? Dataset and model analysis," in *Proc. ICCV*, 2019.
[28] M. Liao et al., "Real-time scene text detection with differentiable binarization," in *Proc. AAAI*, 2020.
[29] A. Rosebrock, *Practical Python and OpenCV*. PyImageSearch, 2019.
[30] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Pearson, 2023.
[31] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Sebastopol, CA, USA: O'Reilly Media, 2009.
[32] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based NLP," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.
[33] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, *A Practical Guide to Sentiment Analysis*. Springer, 2017.
[34] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Univ. Waikato, New Zealand, 1999.
[35] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.