AI-Powered Cyberbullying Detection And Recommendation System On Social Media

Dasari Sony[1],Dr.V.Kamakshi Prasad[2]
[1]Post Graduate Student, [2]Senoir Professor of CSE & Director,
M.Tech (DS),Department of Computer Science and Engineering,
Jawaharlal Nehru Technological University , Hyderabad , India
[1]Email: dasarisony16@gmail.com
[2]Email: kamakshiprasad@jntuh.ac.in

Abstract- The issue of cyberbullying is rising at the same time as social networking has become ubiquitous, causing social and psychological challenges among people using the Internet. To address these issues, a new adaptive cyberbullying detection and content moderation approach using natural language processing, machine learning, and image-based textual analysis is presented in this study. For the text extraction, the textual features are extracted using TF-IDF, and for the detection of harmful content, the class-balanced Logistic Regression model is used to detect the harmful content in the social media posts. The Optical Character Recognition (OCR) module extends the capabilities of the program to extract and analyze text from memes, screenshots, and images, enabling the analysis of data in this format. This includes a hybrid validation approach to augment the confidence in the results by using the combined validation rules and machine learning to predict validation status. This proposed solution will differ from the two current moderation solutions, which are to just categorize the content and the other is to provide positive suggestions for flagged content. This proposed solution will not be categorical as seen in the existing solutions, but will actively suggest positive alternatives to posts in the flagged area, encouraging positive communication. Experimental results on the Davidson Twitter dataset show that it can achieve an accuracy rate of 94.2%, has low computational requirements, and has a real-time capability. The image-analysis pipeline also achieves good performance in identifying abusive material in visual media. The system affects a few cloud resources, is deployed as a lightweight web app of Flask, and doesn't over-ambitiously require any specific setup of the application (like Amazon AWS Cloud), and fits the needs of any of the common applications and is suitable for community moderation, social media, and/or educational institutions. The results suggest that this strategy involving easily interpretable machine learning techniques and multimodal analysis using OCR works well for realistic 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

Social networking technologies have made tremendous changes in the lives of people by altering how they communicate with each other, share information, and interact in a geographical manner. These platforms are crucial in today's digital age, with billions of users interacting with them every day, from Twitter and Facebook to Instagram and TikTok. These tools help people share information, collaborate, and engage socially, but they are also tools that can be used for the dissemination of inappropriate behaviors. Specifically, cyberbullying is one of these issues, which is the frequent abuse of communication technologies to intimidate, harass, insult, or isolate someone. [1][2].

Cyberbullying has distinct challenges compared to face-to-face bullying, as digital content is linked to other challenges with persistence and pervasiveness. Harmful messages can be sent out in real-time to mass audiences, stay accessible for long periods, and be received by victims anywhere and anytime. In addition, many online forums are anonymous, and the anonymity can foster aggressive behavior that would not take place in face-to-face interactions [3]. Several studies have been conducted that correlate cyberbullying with

negative psychological and social consequences, such as higher levels of stress, anxiety, depression, lower academic achievement, social isolation, and, to date, in extreme cases, self-destructive behaviors.[4].

Manual moderation of the vast amounts of content posted on social media is impossible. This means that automated content moderation systems are crucial for detecting and preventing harmful online activities. In reality, however, the classic keyword filters often fail with disguised language, context variations, abbreviations, and changing expressions of offenses. While significant advancements have been made using machine learning methods, the task of detecting abusive content is still difficult because of the context ambiguity and the complexity of language [5].

Along with text, social media interactions are increasingly using visual media like memes, screenshots, and images with text embedded within the image. These nonverbal forms are often used to send harmful messages, which can be hard to spot with a text-only moderation system. However, by not incorporating OCR, bad content can get into the images [6]. It is therefore important that there is a solution that can deal with both text and image within the same solution in fighting cyberbullying.

To address the above problems, this study introduces an AI system for detecting cyberbullying and moderating content through the application of natural language processing, machine learning, and the image analysis provided by OCR. The proposed system employs TF-IDF feature extraction and class-balanced Logistic Regression (LR) for effective identification of harmful text in a system. EasyOCR has been integrated to enable the extraction and analysis of content from an image; this is useful for multi-modal content moderation. There is also a hybrid decision-making mechanism combining both rule-based validation and machine learning predictions to further improve the accuracy of the predictions. The solution not only helps recognize inappropriate content but also provides constructive guidance on how users can modify their interactions to develop them into more positive and respectful ones.

This work entails several key contributions where we cut down our work on the following points:

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.

This paper proceeds as follows: Section II discusses the literature on cyberbullying detection and content moderation. The proposed framework, system analysis, and architecture are explained in sections III and IV. The implementation details and experimental results are included in Section V. Lastly, Sections VI and VII present the results, system limitations, and proposed future improvements to 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] developed the first algorithm to build a collection of tweets with racism and sexism labels, showing that character n-grams are able to detect the patterns of harmful language that extend beyond exact keyword matches, and that consistency of annotation is important for building a dataset.

B. Distinguishing Offensive Language from Hate Speech

Davidson et al. [8] introduced a framework for distinguishing hate speech and offensive language from neutral content from a single classification system applied to 24,783 labeled tweets, revealing that combining both profanity and hate speech results in systematic misclassification. The database they have is the main training benchmark for the current work.

C. Contextual and Conversational Analysis

Van Hee et al. [9] investigated cyberbullying based on fine-grained role distinctions (bully, victim, bystander), thus pointing to the limitations of single message analysis to capture cyberbullying dynamics in conversational threads. Dadvar et al. [10] showed that including user history and demographic context provides further improvements in detection, but places additional privacy constraints.

D. Network and Behavioural Features

User metadata, social network topology, and temporal behavioural patterns were added by Chatzakou et al. [11] when differentiating between aggression, bullying, and spam on Twitter. Yes, it is effective; if services do not have access to the graph data of their platforms, then it is not generalizable.

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 is a practical first step in resource-constrained environments.

F. Deep Learning Approaches

Badjatiya et al. [13] have tested LSTM and GRU for detecting hate speech and found that these models have learned more than a static vocabulary of sentiment words. Neural models, however, are more costly in terms of computation and less interpretable compared to classical linear models, with the latter being more suitable when latency and transparency are important [14].

III. System Analysis

A. Existing System Limitations

Content moderation approaches fall into three broad categories. Rule-based keyword filters are deterministic but not against synonyms, coded language, and phonetic substitutions. Generalisation is learned by supervised ML classifiers, which face distribution shift and class imbalance. Superb performance has been demonstrated by large neural language models, but they require computation costs that are too high to deploy on personal devices in real-time [15]. What is missing is corrective guidance – systems that return only a binary label are not very useful to users who might want to edit undesirable contents. In addition, nearly all text-based systems overlook the act of communicating via images. [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 receives raw CSV data sets, cleans them for Unicode, removes URLs, filters stop words, and generates a consolidated training corpus. The Model Training Layer is using TF-IDF vectorisation with a vocabulary ceiling of 10,000 features and logistic regression with balanced class weights. There are two routes in the Application Layer: one to submit text and one to upload an image. The result is presented on an HTML page using a Jinja template for the classification, confidence score, and recommendation.

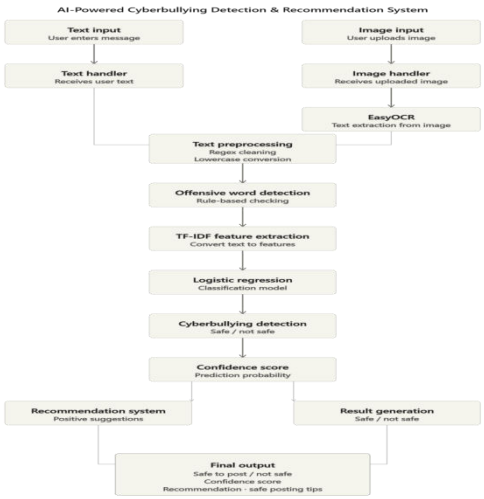

Fig. 1. High-Level System Architecture

B. Data Preparation and Feature Extraction

The Davidson Twitter dataset (24,783 annotated tweets) is lowercased, all URLs and mentions are stripped using regular expressions, non-ASCII is normalized, stopwords are stripped using the NLTK stopwords list, and the labels (hate speech and offensive language) are converted to binary. TF-IDF vectorisation assigns weight to the terms that occur with high frequency in harmful messages but infrequently in the rest of the messages and yields sparse feature matrices (that can be used with linear classifiers). This ceiling is identified as 10,000 in the vocabulary hierarchy, the balance being between representational capacity and inference latency.

C. Classification and Confidence Estimation

Logistic regression was chosen because it is interpretable, efficient to compute, and it demonstrated good empirical performance on TF-IDF features [16]. The class_weight=balanced setting normalizes the loss per class against the class frequency, which balances the contribution of each class, thus reducing the impact of the natural imbalance of the benign social media content classes. The predict_proba method provides a confidence score in [0,1] that is reported along with the classification label to indicate the level of confidence of the model for the end user.

D. OCR Integration

Images uploaded through the upload interface are converted to text using EasyOCR, the world's most popular open-source OCR engine that supports more than eighty different languages. If the domain of the classifier is different from the domain in which the extract is used, English-only extraction is called for. EasyOCR utilises a CRAFT-based text detection model to identify the region of text, and then a recognition step that generates character sequences. The extracted string is then fed into the same text classification pipeline, and a common decision architecture for both modalities is kept.

E. Hybrid Detection and Recommendation Engine

The submitted text is evaluated against a carefully created lexicon of high certainty harmful terms prior to invoking the trained model. The matches result in an immediate unsafe classification, which does not involve probabilistic classification. If the explicit term check is passed, the TF-IDF model offers a probabilistic score. The recommendation engine identifies recognized negative linguistic patterns and pairs them with pre-written positive suggestions, alongside specific instructions on what to change in the content.

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

Logistic regression was used for quick experimentation to model training, which took around 0.143 seconds on the reference hardware using the full 24,783 sample corpus. Stratified sampling was used to divide the dataset 80/20, which maintained the same class distribution as the original dataset. The TF-IDF vectoriser was only fitted on the training data, without being fit on the test partition, thus avoiding information leakage from the test statistics into the feature space.

C. Evaluation Metrics and Results

The held-out 20% test partition was used to measure the performance in terms of accuracy, precision, recall, and F1-score. A total of 500 Memotion meme images were manually labeled with binary ground-truth labels for end-to-end evaluation in the OCR pathway. Table III shows the full evaluation results for both 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 accuracy of the text branch is 94.2%, and the F1-score is 0.929, which are comparable with the classical ML performance on the Davidson benchmark [8]. The OCR

pathway has 91.7% accuracy, which is mostly due to the errors in OCR transcription of the typography used in the memes that have low resolution or are highly stylized, not weaknesses in the downstream classifier. Average latency for text prediction is 0.5ms per sample, which is well-suited for real-time use.

D. Recommendation System Evaluation

The relevance and usefulness of generated recommendations were rated by a survey of fifteen computer science students on a five-point Likert scale. Those recommendations that were correctly classified as harmful content were found to be helpful by the users, with a mean helpfulness rating of 4.1 out of 5.0. The proposed system is different from the purely detection-oriented baselines because of this aspect of preventive guidance.

E. Results

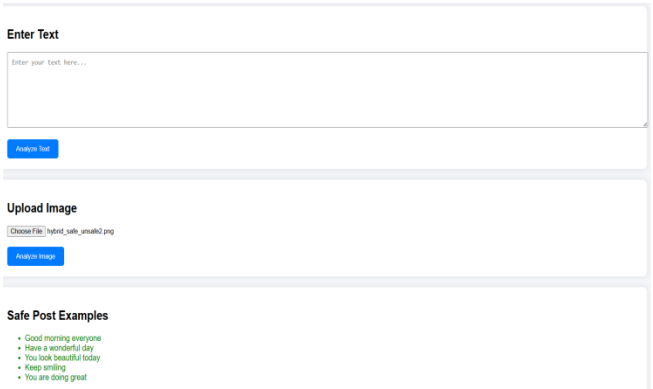

Fig 2: Image Based Input

An image-based input page is provided for the user to upload images with text to be analyzed for content safety. An image can be uploaded via an easy-to-use interface. It supports popular image formats and also optimizes them for OCR. Images uploaded are shown for verification before analysis. This interface offers a simple method to evaluate the content contained in an image.

Fig 3: Image Based Results

The results page for the uploaded image shows the extracted text from the image using the OCR technology. The extracted text is then analysed and categorized as either "Safe To Post" or "Not Safe To Post". A confidence score is given to reflect the confidence level of the model. The system creates suggestions for enhancing the message when the content is unsafe. This will enable the user to grasp the end of the analysis and to decide whether or not to post it.

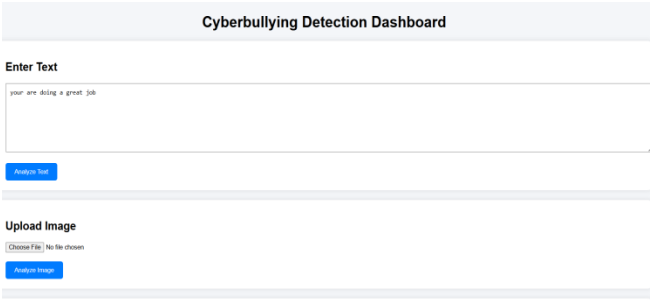

Fig 4: Text Based Input

The page with the text box allows people to type or copy and paste text into the system for analysis. So, there's a text area for content submission. Users can go back and review and edit their text before they start the assessment. The interface will be optimized for rapid and effective evaluation of content. This feature allows text safety checking in real-time.

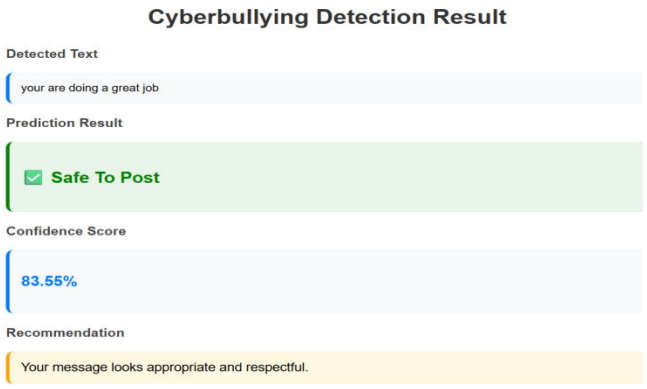

Fig 5: Text Based Results

The results page is text-based, with the text you had entered and the security rating shown. The centre is obviously clearly marked for posting or not. A prediction reliability bar is displayed that represents a prediction confidence percentage. Unsafe messages are provided with safe and constructive advice and tips for communication. It can be used by users to help improve their content and foster positive online interactions.

VI. Discussion

Through the experimental results, it can be seen that the TF-IDF extracted by the balanced logistic regression method performs as well as more complex neural architectures in reality when detecting cyberbullying. These are some of the key advantages of being a completely cloud-agnostic key moderation platform with sub-millisecond inference time, CPU-only, complete transparency of features and decision boundaries, and no dependence on the large language model APIs [17] - these are attractions for educational institutions and not-for-profit moderation platforms with limited budgets for computation and data privacy concerns that make relying on the large language model APIs impractical.

Regarding OCR, no literature showed how to integrate prior to our work. Previous studies on the moderation of social media have primarily been conducted in relation to text-based communication channels and failed to take into consideration image-embedded channels. Experimental results indicate that it is of significant accuracy for text embedded in images, equivalent to the accuracy of the OCR on artistic or heavily filtered typography. Additional studies might include the exploitation of visual context features together with extracted text in order to progress towards a more multimodal classification.

One thing that distinguishes the work from existing works is the recommendation module, which presents every unsafe determination paired with a constructive method of the same, also reflecting the philosophy of prevention rather than punishment, which focuses on digital citizenship studies [18]. Limitations can be: it only works with Twitter-specific data, so it might not be as easily expanded to other social media platforms; flattening of labels has lost a small level of granularity; the static list of recommendations must be continually updated; balanced weighting might yield a higher number of false positives for ambiguous cases.

VII. Conclusion

This study proposed an intelligent cyberbullying detection and content moderation framework that brings together machine learning, natural language processing, and image-based text analysis in a single platform. The proposed system comprises a class-balanced logistic regression classifier, a lightweight Flask-based web app, a text extraction mechanism using the EasyOCR library, and a recommendation-based moderation mechanism, which is integrated with the TF-IDF feature extraction technique. The framework allows for analysis of text and text-in-image formats, and can provide a holistic framework for detecting harmful online communication in a range of content types.

Experimental analysis showed that the proposed model proved to be 94.2% accurate on the Davidson Twitter dataset and to have an extremely low inference latency on standard consumer-grade hardware. The results showed that with proper optimization, classical machine learning methods can be just as effective and efficient as deep neural networks without the need for complex computations. Adding OCR functionality adds another level of effectiveness when it comes to catching abusive content in memes, screenshots, and other visual media. The recommendation engine also enhances the capabilities of conventional platforms by suggesting constructive communication, which promotes positive digital interactions in addition to moderation.

To sum up, the proposed framework is feasible and scalable, and could be used within educational institutions, community forums, social networking sites, content moderation scenarios, etc., and is low-cost. It operates on a self-contained infrastructure and is an ideal fit for private deployment in resource-poor environments.

Other languages and OCR support, lightweight transformer architectures to better understand the semantic content, long-term user studies to evaluate the effectiveness of the

recommended interventions in the development of healthier online communication, and detection of repeated or relationship-oriented bullying situations could be implemented in the future. These enhancements can significantly better equip the system to be flexible, precise, and effective in the real world of fighting cyberbullying in various digital environments.

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.