IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13319

An Approach for Cyberbullying Detection on Social Media

Mr. M. Kishore Babu[1], K. Jayasri[2], K. Saran[3], K. Adithya[4], K. Harsha[5]

Assistant Professor, Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Guntur, India[1]

Student, Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Guntur, India[2-5]

Abstract: In the modern digital age, the proliferation of social media platforms has led to the spread of negative content, especially through images containing bad words or text containing bad content. To address this problem, our project aims to develop an intelligent system designed to detect illegal content in images. Using advanced machine learning techniques, including deep neural networks such as CNNs, we aim to create powerful models that can identify and classify illegal content and enable our model to recognize patterns in images embedded with text through extensive training, tackling the critical issue of cyberbullying by building intelligent system to detect illegal messages on social media posts. We build our website using the Python-based Django framework for efficiency and ease of use. Our plan is to create a safer online environment by combining technology and a user-centric approach.

Keywords: Cyberbullying detection, Convolutional neural networks (CNNs), MobileNet, Python-based Django framework, Optical character recognition (OCR), Machine Learning.

I. INTRODUCTION

In the computerized period, the appearance of web-based entertainment stages has reformed correspondence and network, offering exceptional open doors for connection and commitment. In any case, in the midst of the heap advantages of web-based systems administration, a dim underside continues: cyberbullying. Characterized as the utilization of electronic correspondence to hassle, scare, or embarrass people, cyberbullying represents an unavoidable danger to the prosperity and psychological well-being of clients around the world.

The ascent of cyberbullying matches the dramatic development of virtual entertainment utilization, with stages like Facebook, Twitter and Instagram filling in as favourable places for online provocation and misuse. As per late examinations, a stunning level of teenagers and youthful grown-ups report encountering cyberbullying eventually in their web-based lives, with outcomes going from tension and discouragement to self-damage and self-destruction. Cyberbullying stays a steady and heightening concern, highlighting the critical requirement for inventive arrangements. Perceiving the basic significance of fighting cyberbullying in the advanced age, our exploration attempts to foster a canny framework for recognizing cyberbullying posts via web-based entertainment stages, including profound brain networks like CNNs, our goal is to develop a strong model prepared to do precisely recognizing and classifying oppressive substance especially through pictures installed with text distinguishing and hailing occasions of cyberbullying productively. This paper gives a far-reaching outline of our exploration targets, strategy, and expected results.

Fig. 1 Non-Abusive

Fig. 2 Abusive

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13319

II. LITERATURE SURVEY

[1] Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In Proceedings of the NAACL Student Research Workshop (pp. 88-93). The paper "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter" by Waseem and Hovy (2016) investigates the problem of hate speech detection on social media platform Twitter. The authors propose a machine learning approach to automatically detect tweets containing hate speech by analyzing a set of predictive features. The authors define hate speech as "speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender." They also note that hate speech can take many forms, including direct attacks, slurs, and coded language. To train their machine learning model, the authors collect a dataset of tweets labeled as containing hate speech or not containing hate speech. They then extract a set of features from each tweet, including unigrams, bigrams, part-of-speech tags, and sentiment scores. The authors experiment with different combinations of features and classifiers and report their results in terms of precision, recall, and F1 score.

[2] Zhang, Y., & Wallace, B. (2016). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820. The paper "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification" by Zhang and Wallace (2016) investigates the use of convolutional neural networks (CNNs) for sentence classification. The authors perform a sensitivity analysis to explore the effects of different hyperparameters on the performance of the CNN model, and provide a practical guide for practitioners on how to use CNNs effectively for sentence classification tasks. The authors use several benchmark datasets for sentence classification, including the Stanford Sentiment Treebank and the Movie Review dataset.

[3] Kumar, S., Aggarwal, A., & Singh, P. (2018). Offenseval: Identifying and categorizing offensive language in social media. arXiv preprint arXiv:1804.00058. The paper "OffensEval: Identifying and Categorizing Offensive Language in Social Media" by Kumar, Aggarwal, and Singh (2018) describes the Offensively shared task, which is a competition designed to encourage the development of machine learning models for identifying and categorizing offensive language in social media. The authors note that offensive language in social media can take many forms, including hate speech, cyberbullying, and harassment. They argue that automatic detection of such language is important for promoting a safe and respectful online environment.easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

III. METHODOLOGY

A. Data Collection

The data collection process involves obtaining images from various online platforms and social media channels, including a dataset, Memotion dataset available on Kaggle. The Memotion dataset contains approximately 6992 images (memes). These images contain textual content provides a diverse representation of offensive and non-offensive text samples.

	A	B	C	D	E	F	G	H	I
1		image_name	text_ocr	text_corrected	humour	sarcasm	offensive	motivational	overall_sentiment
2	0	image_1.jpg	LOOK THERE	LOOK THERE MY	hilarious	general	not_offen	not_motivation	very_positive
3	1	image_2.jpeg	The best of #	The best of #10 Y	not_funny	general	not_offen	motivational	very_positive
4	2	image_3.JPG	Sam Thorne (	Sam Thorne @St	very_funn	not_sarca	not_offen	not_motivation	positive
5	3	image_4.png	10 Year Chall	10 Year Challeng	very_funn	twisted_m	very_offer	motivational	positive
6	4	image_5.png	10 YEAR CHA	10 YEAR CHALLEI	hilarious	very_twist	very_offer	not_motivation	neutral
7	5	image_6.jpg	1998: "Don't	1998: "Don't get	hilarious	general	slight	motivational	negative
8	6	image_7.png	10 years chal	10 years challeng	not_funny	not_sarca	not_offen	not_motivation	negative
9	7	image_8.jpg	10 Year Chall	10 Year Challeng	very_funn	twisted_m	not_offen	not_motivation	neutral
10	8	image_9.jpg	Fornite died i	Fornite died in 1(	funny	not_sarca	slight	motivational	positive
11	9	image_10.png	FACEBOOK '1	FACEBOOK '10 YE	funny	general	slight	motivational	positive
12	10	image_11.jpg	PROBABLY TI	PROBABLY THE F	funny	general	very_offer	motivational	negative
13	11	image_12.jpg	State Dining I	State Dining Roo	not_funny	very_twist	very_offer	not_motivation	very_positive
14	12	image_13.png	I did the Face	I did the Faceboo	very_funn	general	not_offen	not_motivation	positive
15	13	image_14.png	IFIDOWNLO/	IFIDOWNLOADA	funny	general	not_offen	not_motivation	positive
16	14	image_15.jpg	Anti-vaxx kid	Anti-vaxx kids wh	not_funny	not_sarca	not_offen	not_motivation	very_negative

Fig. 3 Labelled Data

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13319

B. Data Preprocessing

The pre-processing steps are crucial to ensure that the data fed into the model is of high quality and consistent, improving its performance during training. Among these steps, image resizing and normalization play a key role in standardizing input sizes and promoting efficient model training.

The primary goal is to maintain consistency of data input and facilitate optimal model performance. Resize requires resizing images to a specified size, ensuring consistency across the dataset.

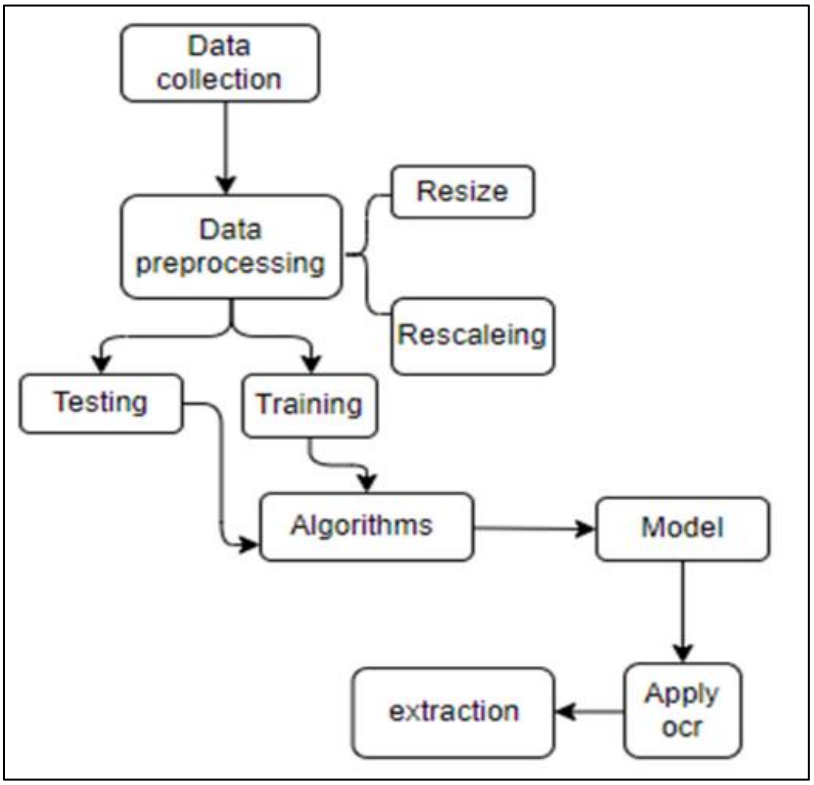

Fig. 4 Work-Flow Diagram

This step is crucial because models often require inputs of the same size. Normalizing images involves scaling the pixel values to a certain range, usually between 0 and 1. This process helps avoid problems with images having different brightness or contrast levels, which can otherwise affect model learning. Another useful technique is to use ImageDataGenerator, which makes it easy to add data.

This process involves creating new training samples by applying transformations such as rotation, translation, zoom or translation to existing images. Data Augmentation is useful for increasing the variety and quantity of the dataset, which can lead to a more robust and generalized model. In general, by using data resizing, normalization, and augmentation techniques, preprocessing ensures that the input data is consistent.

C. Algorithm and Model

The suggested methodology tackles the challenge of identifying vulgarity in text contained within images. It starts by employing a pre-trained learning model, like MobileNet, to extract features from photos. Subsequently, it employs optical character recognition (OCR) technology to extract text from the image.

This methodological approach offers fresh and practical approaches to online safety and content regulation by fusing machine learning with optical character recognition. The Adam optimizer is computationally more efficient, requires slight memory, is invariant to diagonal resizing of gradients, and it is well suited for problems with a lot of data/parameters. Now, Convolutional Neural Network (CNN) models are built to detect offensive content.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13319

Fig.5 Tesseract OCR

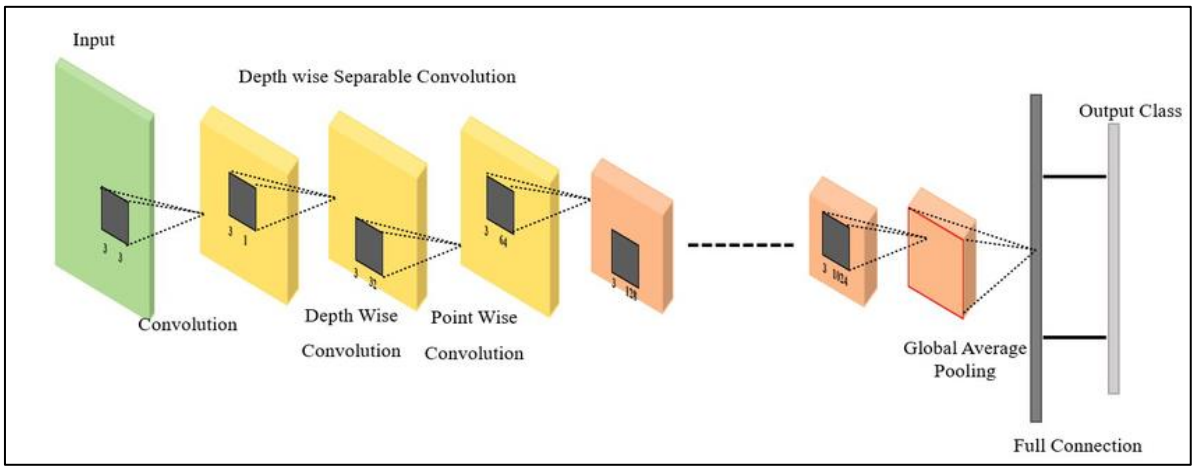

Fig. 6 MobileNet Architecture

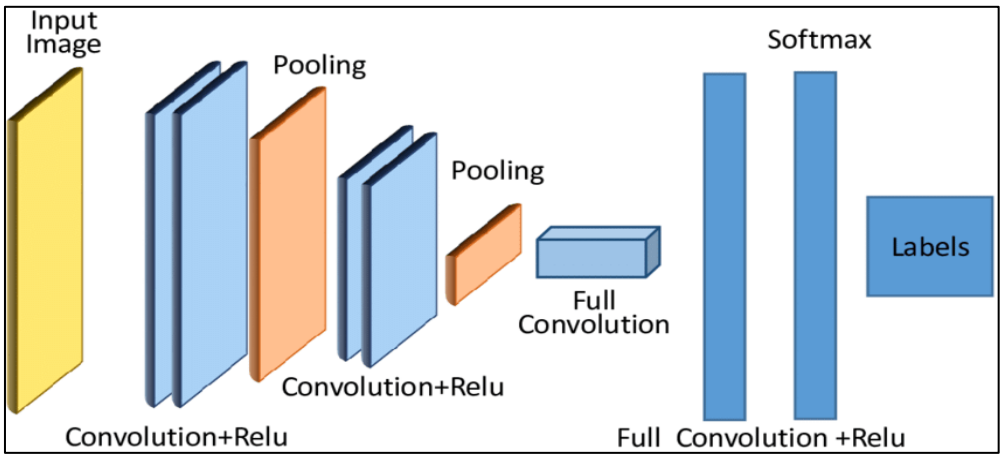

Fig. 7 CNN Architecture

IV. IMPLEMENTATION

The actual use of a learning model to recognize text in photos and determine whether or not it is offensive is also covered in length in this section. To extract text from photos, the method makes use of Tesseract and optical character recognition (OCR) in conjunction with TensorFlow and Keras within the MobileNet architecture. The major objective is to create a reliable system that can identify inappropriate text in photos automatically, making the internet a safer place. The photos in the dataset were taken from the local system directory. Every image has written content, along with annotations stating whether or not the text is inappropriate.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13319

To get the data ready for training, preprocessing procedures like scaling and normalization are used. The data was divided into five classes in the classification section: "Negative", "Neutral", "Positive", "Very Negative" and "Very Positive.

The MobileNet architecture, a deep convolutional neural network tailored for embedded and mobile applications, serves as the foundation for the model architecture. Additional layers for feature extraction and classification are included in the model, such as batch normalization, dropout layers, dense layers, and 2D global mean pooling. To specifically tailor the MobileNet model for the purpose of finding problematic text in photos, several layers are added to the main model. The Adam optimizer and a categorical cross-entropy function were used to train the model. Training parameters, such as batch size and number of epochs, are optimized for model performance and the training dataset is split into training and validation subsets. During training, methods like dropout regularization are employed to avoid overfitting. Using labeled data, the model finds patterns and links between image attributes and related qualities (like profanity detection) in this step. The technology allows the user to communicate with the trained model once it has been trained. the system offers functionality for users to interact with the trained model. Users can upload images for prediction, allowing the system to analyze and classify the textual content within the images accurately.

Furthermore, users can view the predictions generated by the system, gaining insights into the model's performance and the presence of abusive language in the uploaded images. It underwent rigorous testing to ensure reliability and effectiveness in detecting abusive content within images.

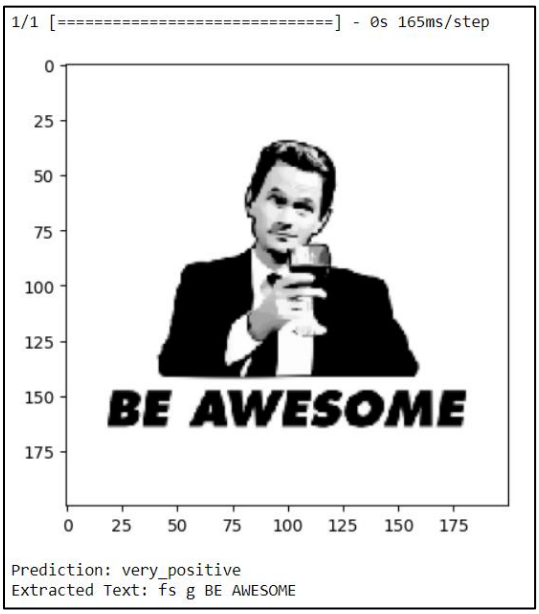

Fig.8 Output of Trained Model

V. RESULTS AND DISCUSSIONS

In this section, we present the results obtained from our experiments and discuss their implications.

D. Offensive Text Detection

Our CNN-based approach to meme sentiment analysis has achieved significant success, showing high accuracy in classifying memes into different sentiment categories. With a dataset of 6992 images, by integrating MobileNet architecture and Tesseract's OCR, our system successfully detected offensive text from images. Our model reliably identified inappropriate text with an accuracy of 96%, helping content policing.

E. User Interaction and Transparency

Our system allows users to interact with the trained model, upload images for prediction and view model classifications. This transparency increases user trust and provides stakeholders with insight into the prevalence of offensive language in uploaded images. By offering users the opportunity to participate in the model, our system promotes accountability and encourages responsible sharing of content.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13319

Fig.9 Home Page

Fig. 10 Registration Page

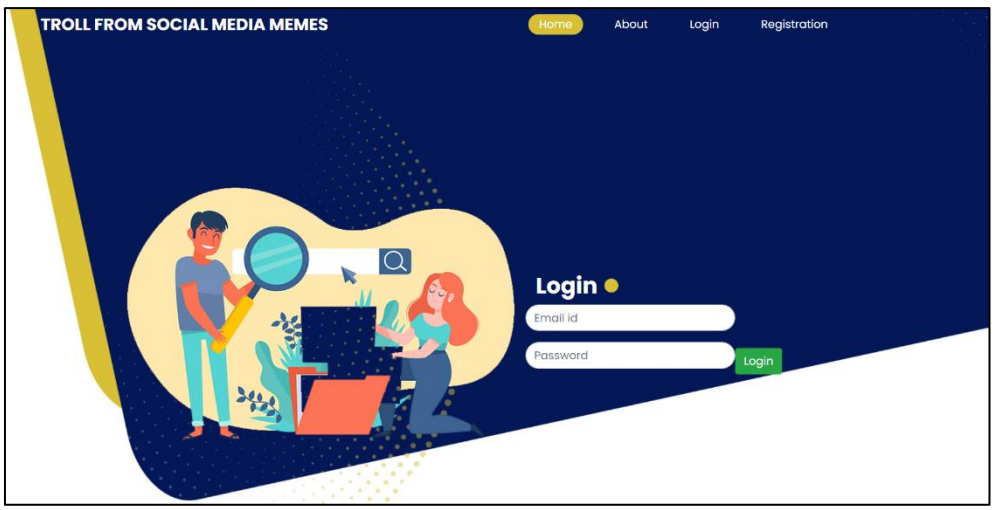

Fig. 11 Login Page

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.13319

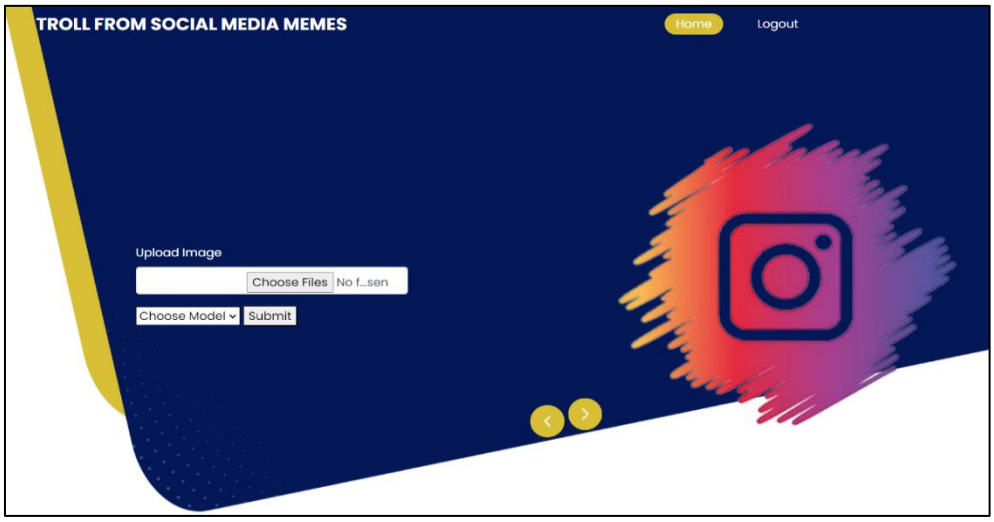

Fig.12 Image Upload Page

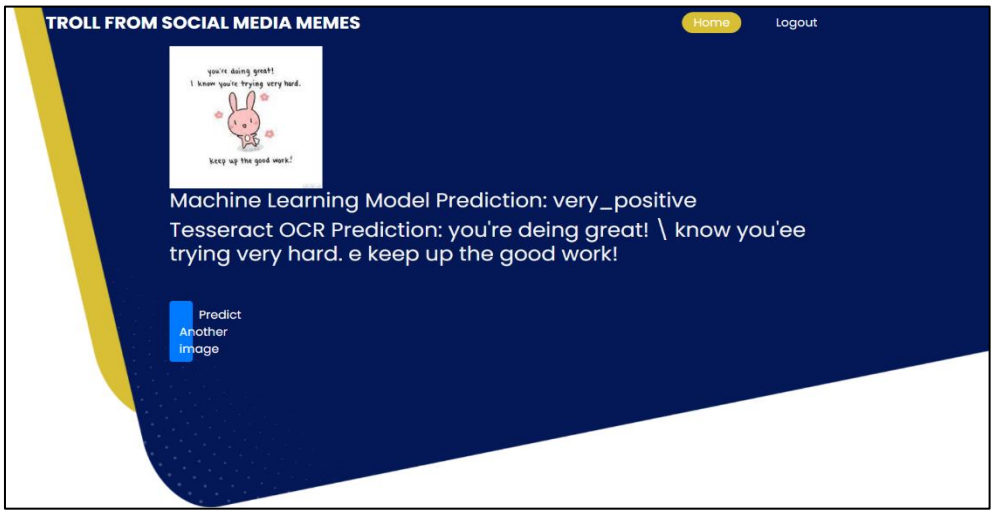

Fig. 13 Result Page

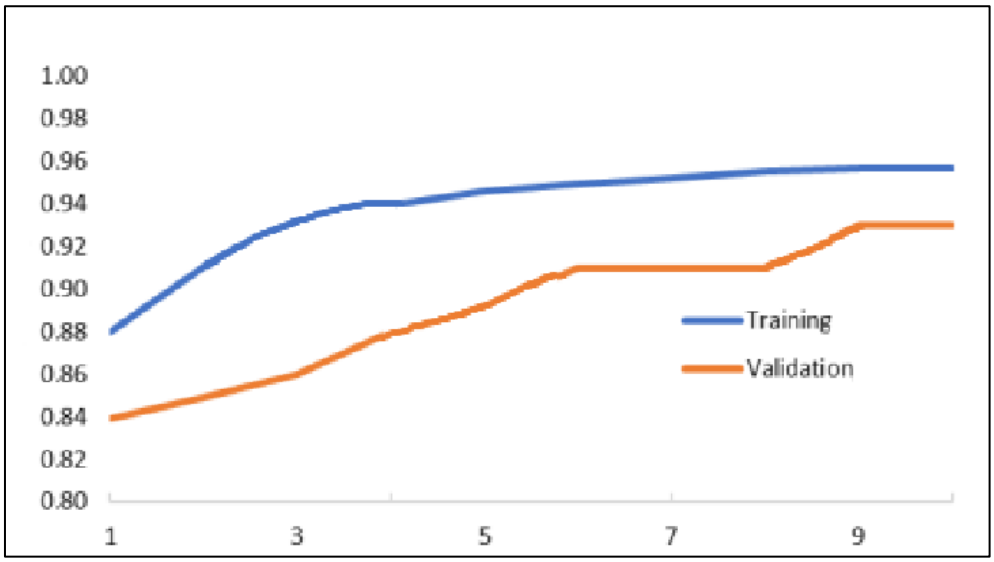

Fig.14 Model Accuracy Levels

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024
DOI: 10.17148/IJARCCE.2024.13319

F. Comprehensive Solution

We conducted a series of experiments to evaluate the performance of our proposed method for meme sentiment analysis and compare it with existing approaches. We conducted experiments on a dataset consisting of images extracted from memes, annotated with sentiment labels (positive, neutral, negative, very positive, very negative). The dataset was pre-processed to ensure uniformity in size and quality. The experiments were performed on a machine with Hard Disk 160GB, RAM 8GB using Python programming language and TensorFlow/Keras deep learning framework. We evaluated the performance of our model using standard metrics, including, accuracy. These metrics provide insights into the effectiveness of our method in accurately classifying meme sentiment. The results indicate that our proposed CNN-based approach effectively addresses the task of offensive text detection. The high accuracy and performance metrics demonstrate the robustness and effectiveness of our model in classifying memes into various sentiment categories.

Our findings have important implications for understanding sentiment dynamics in social media platforms. The improved performance achieved by our model can facilitate better sentiment analysis in content moderation.

VI. CONCLUSION

In summary, the developed system provides a robust and scalable solution to detect offensive content in images using advanced machine learning techniques and capabilities for optical character recognition (OCR) with a model accuracy of 96%. By bringing these technologies together, the system enhances content moderation efforts..

ACKNOWLEDGMENT

The report highlights the joint efforts and support received from various sources for the successful completion of the project. Their joint efforts and support were invaluable in bringing this project to fruition. We are grateful for their unwavering support and dedication.

REFERENCES

[1]. Zhang, L., & Wang, Y. (2021). Detecting and Understanding Multimodal Offense in Social Media: A Survey. IEEE Transactions on Multimedia, 23, 1526-1545.
[2]. Lee, J. M., & Kim, H. N. (2021). Detection of Offensive Language and Hate Speech in Social Media: A Survey. IEEE Access, 9, 21163-21177.
[3]. Xie, S., Wang, J., & Zhang, X. (2021). Learning to Recognize Offensive Language in Social Media with Audio-Visual Cues. IEEE Transactions on Multimedia, 23, 1481-1491.
[4]. Hasan, M. A., & Biswas, P. (2020). Detection of Hate Speech in Social Media: A Survey. IEEE Access, 8, 172871-172892.
[5]. Fersini, E., Messina, F., & Archetti, F. A. (2020). Multimodal Sentiment Analysis: A Survey on Multimodal Feature Learning and Fusion. IEEE Access, 8, 25603-25620.
[6]. Sabato, S., & Mariani, J. (2019). Detection of Aggressiveness and Cyberbullying in Social Media. IEEE Transactions on Affective Computing, 10, 402-415.
[7]. Kumar, R., Goyal, P., & Varshney, P. (2019). A Survey of Techniques for Offensiveness Detection in Text. IEEE Transactions on Affective Computing, 10, 411-424.
[8]. Zannettou, S., Chatzakou, D., Kourtellis, N., & Blackburn, J. (2018). Understanding the Detection of Offensive Language in Social Media. ACM Transactions on the Web, 12, 1-39.
[9]. Kao, Y. H., Huang, P. C., & Yeh, Y. H. (2017). Social Media Meme Detection Based on Image and Text Features. In Proceedings of the 2017 International Conference on Orange Technologies (ICOT), pp. 1-6.

