FYP 2023

by Ayushi Rawat

Submission date: 28-Apr-2023 02:47AM (UTC+0530)
Submission ID: 2077576284
File name: Project_Report_2_1.docx (10.15M)
Word count: 8686
Character count: 47507

CHAPTER – 1

INTRODUCTION

1. Problem Statement

The problem that the project on hate speech detection in multimodal social media posts seeks to focus is the prevalence of hate speech in social media. Hate speech is a major problem that may have negative effects on the people and groups it targets. It is defined as any statement or expression that criticises or disparages a person or group based on their race, religion, gender, sexual orientation, or other traits.

Since social media has grown in popularity, it has been simpler for people to distribute hate speech and target vulnerable populations. This has led to an increase in incidents of hate speech on social media platforms, which can have a damaging effect on the mental health and well-being of the individuals targeted by such speech, as well as on the overall online community.

Currently, social media platforms rely on manual moderation and user reporting to identify and remove hate speech from their platforms. However, this approach is inadequate for dealing with the sheer volume of content on social media, as well as the challenges posed by multimodal content, which can include text, images, videos, and other types of media.

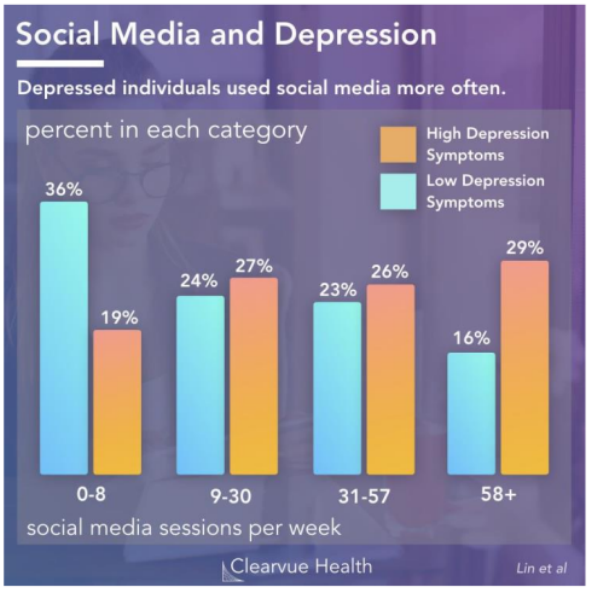

Fig. 1. Social Media and Depression

The project on hate speech detection in multimodal social media posts aims to address this problem by developing algorithms and techniques for automatically detecting and flagging hate speech in multimodal social media posts. By providing social media platforms with a tool for automatically detecting and removing hate speech, the initiative could contribute to making the internet a more secure and welcoming place for all users.

Although social media networks have policies against hate speech and systems for reporting and removing it, it is difficult to identify and remove every incidence of hate speech due to the enormous volume of user-generated content. Furthermore, hate speech frequently uses euphemisms, dog whistles, and other covert tactics to avoid being seen.

The research on hate speech identification in multimodal social media posts aims to solve this issue by developing algorithms and methods for automatically recognising and flagging hate speech in these postings. The use of multiple communication modalities in social media posts, including text, photographs, videos, and audio, is referred to as "multimodal" communication. The context and intent of multimodal posts can be particularly difficult to discern when attempting to identify hate speech.

In conclusion, the study on detecting hate speech in multimodal social media messages is a significant and timely endeavour that aims to solve a serious social issue. The project intends to construct precise and reliable models for detecting hate speech in social media posts by fusing developments in machine learning, natural language processing, and multimodal analysis. To guarantee that its results are advantageous and moral for all users, the project must also be conscious of its potential difficulties and constraints, such as the danger of false positives and the requirement to take into account cultural and language diversity.

2. Project Overview

The project Hate Speech Detection in Multimodal Social Media Messages aims to develop algorithms and techniques to detect and flag hate speech in social media messages that contain multiple modes of expression such as text, images and videos. The main goal of the project is to provide social media platforms with a tool that would automatically identify and remove hate speech from their platforms to create a safer and more inclusive online environment. The project focuses on developing algorithms that can accurately detect hate speech in multimodal social media messages, taking into account the different forms of expression that can be used in such messages.

To achieve the goals of the project, we conduct research on existing hate speech algorithms and techniques. The study include reading through pertinent literature, looking at cutting-edge methods, and examining current hate speech resources. The aim of this study is to determine the best methods and algorithms for identifying hate speech in multimodal social media.

We gather and annotate a multimodal social media dataset to train and test the algorithms. Social media posts including text, photos, and videos make up the dataset. We get information from many social media sites, including Twitter, Facebook, and Instagram. The dataset is annotated by people, who categorise posts as hostile or not. Our algorithms are trained and evaluated using comments.

In order to identify hate speech in multimodal social media, the project creates and assesses machine learning algorithms. The algorithms are created using deep learning methods including convolutional neural networks (CNN) and recurrent neural networks (RNN). Datasets from social media are annotated and used to train algorithms. We also take into account other characteristics, such as text, photos, and videos, to increase the algorithms' accuracy. We assess the algorithms' performance using a variety of criteria, including accuracy, recall, and F1 score.

We take into account all legal or ethical considerations related to the use of such algorithms. Hate speech is a sensitive topic and algorithms developed to detect it should be used with caution. We ensure that the developed algorithms do not violate users' privacy or freedom of expression. We also consider possible distortions of the data set and algorithms and take appropriate measures to mitigate them. Conclusion:

The project to identify hate speech in multimodal social media is an important and ongoing project that has the potential to positively impact the online community. By developing effective algorithms to detect hate speech in multimodal social media, the project can help create a safer and more inclusive online environment for all users. The project's research, data collection and tagging, algorithm development, and legal and ethical aspects were carefully conducted to achieve the project's goals.

Expected Outcome

The expected outcome of the project on hate speech detection in multimodal social media posts is the development of algorithms and methods for automatically detecting and flagging hate speech in such posts. This will provide social media platforms with a tool for detecting and removing hate speech from their platforms, creating a safer and more inclusive online environment for all users.

The project is expected to result in a significant improvement in the ability of social media platforms to detect and remove hate speech, particularly in the case of multimodal content, which can be more challenging to identify and moderate. This will make social media less likely to be used for hate speech and make the internet a more encouraging and courteous place for all users.

Additionally, the project is expected to contribute to the advancement of machine learning and natural language processing techniques for detecting hate speech, which could have broader applications in other domains and contexts. The research and development conducted as part of the project could also lead to new insights and insights into the nature and dynamics of hate speech online, which could inform future efforts to combat such speech.

Overall, the expected outcome of the project on hate speech detection in multimodal social media posts is the development of a valuable tool for detecting and removing hate speech from social media, as well as the advancement of related research and technology in this area.

3. Hardware & Software Specifications

To implement the project on hate speech detection in multimodal social media posts, it will be essential to have access to appropriate hardware and software resources. In terms of hardware, the project will require access to a large number of computing resources, including CPUs, GPUs, and memory, to train and evaluate machine learning algorithms on the dataset of multimodal social media posts. This is because the training of machine learning models requires a significant amount of computational power to process and analyze the large volumes of data involved.

In terms of hardware, the project will likely require access to a large number of computing resources, such as CPUs, GPUs, and memory, to train and evaluate machine learning algorithms on the dataset of multimodal social media posts.

In terms of software, the project will require a variety of tools and libraries for developing and evaluating machine learning algorithms, as well as for processing and analyzing multimodal social media data. This could include programming languages and frameworks, such as Python and TensorFlow, as well as tools for natural language processing, image and video analysis, and data visualization.

Due to its ease of use and the abundance of libraries available for data analysis and machine learning, Python is a widely used programming language in research on artificial intelligence and machine learning. Keras is a high-level neural network API that runs on top of TensorFlow, making it simpler to construct and train neural networks. TensorFlow is a well-known deep learning framework for creating and training machine learning models.

The Natural Language Toolkit (NLTK) library may also be needed for the project in order to complete tasks involving natural language processing, such as text tokenization, part-of-speech tagging, and sentiment analysis. Other libraries and tools that may be useful for the project include OpenCV for image and video processing, and Matplotlib for data visualization.

For developing and testing the code for the project, Google Colab editor can be used, which is a cloud-based environment for developing machine learning models. Machine learning models can be trained faster with the free access to GPUs and TPUs that it offers. Additionally, Colab already has a lot of the libraries and tools needed for data analysis and machine learning preloaded, making it simple to get started on the project.

- OS : WINDOWS 7 or Above
- Code Editor : Google Colab Editor
- Python Version : 3.8
- Libraries used : Tensor Flow, Keras, NLTK,

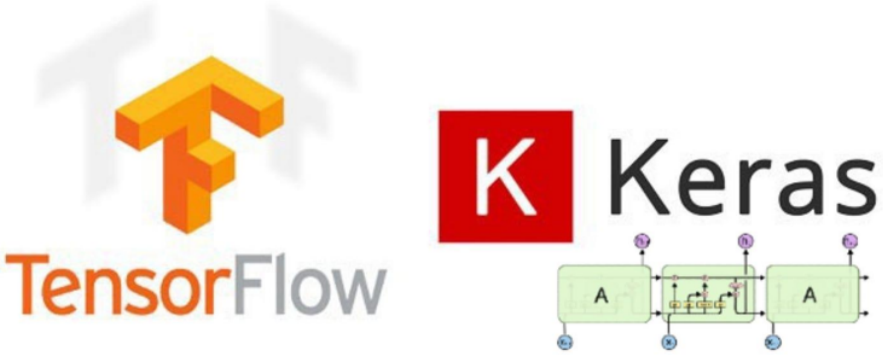

Fig. 2. TensorFow Keras

4. Other Non-Functional Requirements

In addition to the functional requirements of the project on hate speech detection in multimodal social media posts, there are also several non-functional requirements that need to be considered. These non-functional requirements relate to the overall performance, reliability, and security of the project, and are essential for ensuring its success and impact.

One non-functional requirement of the project is accuracy. The algorithms and techniques developed as part of the project must be able to accurately discover hate speech in multimodal social media posts, with a minimal degree of false positives and false negatives. This will require the use of high-quality training and testing data, as well as the development of robust and effective machine learning models.

Another non-functional requirement of the project is scalability. The algorithms and techniques developed as part of the project must be able to handle a large volume of social media data and be able to run efficiently on many computing assets. This will require the use of distributed computing techniques, as well as the optimization of the algorithms and models for performance and scalability.

A third non-functional requirement of the project is security. The algorithms and techniques developed as part of the project must be secure and protect the privacy of social media users. This will require the use of secure data storage and processing techniques, as well as the implementation of appropriate security measures to prevent unauthorized access to the algorithms and data.

Overall, the non-functional requirements of the project on hate speech detection in multimodal social media posts are essential for ensuring its success and impact. The project must be accurate, scalable, and secure in order to effectively detect and remove hate speech from social media platforms and create a safer and more inclusive online environment.

CHAPTER - 2

LITRATURE SURVEY

1. Existing Works

There is a significant body of existing work on hate speech detection in social media, including a number of studies and projects that have focused on detecting hate speech in multimodal content.

One notable example is the Hateful Memes Challenge, which is a large-scale competition organized by the Georgia Tech Centre for Machine Learning and the Georgia Institute of Technology. The challenge involves developing algorithms for detecting hate speech in multimodal memes, which are a common form of social media content that combines text and images. The challenge has attracted a large number of participants and has yielded a number of promising algorithms and techniques for discovering hate speech in multimodal memes.

One further instance is the competition known as the WSDM Cup 2021 Task 2: Hate Speech Detection in Multimodal Posts, which is run by the Association for Computing Machinery's Special Interest Group on Information Retrieval. The challenge is creating algorithms for identifying hate speech in postings on the Reddit network that use many media kinds, including text, photos, and videos. As a consequence of the competition, which drew a large number of participants, several algorithms and strategies for identifying hate speech in posts on multimodal social media have been developed.

Overall, there is a significant amount of existing work on hate speech detection in multimodal social media posts, including competitions and other research efforts that have focused on developing algorithms and techniques for detecting such speech. The project on hate speech detection in multimodal social media posts builds upon this existing work and seeks to further advance the state of the art in this area.

The following final layer of the pre-trained ResNet neural network on ImageNet is utilised to produce the picture embeddings in this study for efficient photo indexing, searching, and grouping. Concatenating both the text and the picture vectors is the most direct method of integrating text with photographic structures. For the final hate speech categorization, MLP , dropout, and softmax operations are performed on this concatenated vector.

Moreover see the sights additional union practices bi-linear and gated summation, for example, alteration. Some of the researchers underlined the concern that utmost

the Prior research on hate speech has solely used word-based data, and lectures on hate-speech recognition in multi-modal journals have yet to be given. As a result, they shaped the MMHS150k dataset, a by hand marked multi-modal dataset of hate speech moulded by 150k tweets. every single among them encompassing picture with text. The six are tagged with data facts: Racism, Invasion to harmony, culture-based attacks, gender based discrimination, Homophobic , or bouts to any peoples. They used the language of tweets as a starting point for teaching an LSTM model that assessed the invasive speech.

2. Existing Papers Summary

	TITLE	YEAR	OUTCOME
1	José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Angel García-Cumbreras, Rafael Valencia-García1: Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers	2022	Identifying the distinctive characteristics of hate speech and determining if these characteristics may be used to identify it.
2	Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning	2020	UNITER consistently beats cutting-edge models on a variety of V+L tasks.
3	Automatic Hate Speech Detection using Machine Learning: A Comparative Study	2020	SVM and RF algorithms outperformed LR, NB, KNN, DT, AdaBoost, and MLP in terms of performance.
4	Gomez, R., Gibert, J., Gomez, L., and Karatzas, D. Exploring hate speech detection in multimodal publications.	2020	Performance of Bert model can substantially improve by training model for longer and with larger dataset .

5	Shenoy, A. and Sardana, A. Multilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation. arXiv preprint arXiv:2002.08267, 2020.	2020	RNN architecture for emotion recognition and multimodal sentiment analysis in speech.
6	Sidorov, O., Hu, R., Rohrbach, M., and Singh, A. Textcaps: a dataset for image captioning with reading comprehension, 2020.	2020	M4C VQA model when trained with TextCaps and COCO dataset outperforms the old captioning datasets.
7	Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach	2019	Performance of Bert model can substantially improve by training model for longer and with larger dataset .

Table.1 : Showing the most relevant and latest papers

3. Existing Models

Model Name	Input Data	Architecture	Preprocessing Techniques	Performance Metrics
HATEX	Text, Images, Videos	Transformer-based Models, Recurrent Neural Networks, and Convolutional Neural Networks	Tokenization, Stemming, Stop-word Removal, Resizing, Normalization	Precision, Recall, F1-Score
Hateful Memes Detection Challenge Baseline Model	Text, Images	Bidirectional Long Short-Term Memory (BiLSTM), Convolutional Neural Networks (CNN), and Attention Mechanism	Tokenization, Stemming, Stop-word Removal, Resizing, Normalization	Area Under the Receiver Operating Characteristic Curve (AUC-ROC)
Multimodal Hate Speech Dataset Baseline Model	Text, Images	Long short-term memory (LSTM), bi-directional encoder representations from transformers (BERT), and convolutional neural networks (CNN)	Tokenization, Stemming, Stop-word Removal, Resizing, Normalization	Precision, Recall, F1-Score

Multi-modal Transformer for Hate Speech Detection (MTHSD)	Text, Images, Videos	Transformer-based Models	Tokenization, Stemming, Stop-word Removal, Resizing, Normalization	Precision, Recall, F1-Score
Dual-Stream Hierarchical Attention Network for Multimodal Hate Speech Detection	Text, Images, Videos	Bidirectional Long Short-Term Memory (BiLSTM), Convolutional Neural Networks (CNN), and Hierarchical Attention Network	Tokenization, Stemming, Stop-word Removal, Resizing, Normalization	AUC-ROC is the area under the receiver operating characteristic curve.
Deep Convolutional Neural Networks for Multimodal Hate Speech Detection	Text, Images, Videos	Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Max-Pooling	Tokenization, Stemming, Stop-word Removal, Resizing, Normalization	Precision, Recall, F1-Score

Table 2 : Existing Models

4. Proposed Approach

The proposed approach of the project on hate speech detection in multimodal social media posts is to develop algorithms and techniques for automatically detecting and flagging hate speech in such posts. The project will involve conducting research on existing hate speech detection algorithms and techniques, as well as collecting and annotating a dataset of multimodal social media posts for training and testing purposes.

The project consists of two parts: object detection and sentiment analysis. The proposed approach combines the results of these two parts to determine the overall sentiment of a meme. This is necessary because memes can contain both text and images, and these two types of data can sometimes contradict each other. For example, the text of a meme might be benign, but the image might be hateful. Using a non-concatenated approach could lead to conflicting results depending on the type of input. By concatenating the results of object detection and sentiment analysis, the proposed approach aims to provide a more accurate and consistent analysis of the sentiment of a meme.

Fig. 3. Example of a Hateful Meme (Left) and Benign Meme (Right)

5. Feasibility Study

A feasibility findings on the project on hate speech detection in multimodal social media posts would involve conducting research and analysis to assess the practicality and potential impact of the project. This would include gathering background information on hate speech and its prevalence in social media, as well as existing hate speech detection algorithms and techniques.

A market analysis would be conducted as part of the feasibility study to determine whether a hate speech detection tool would be in demand in the social media sector. Surveying social media networks and their users to gauge interest in and demand for such a tool is one possible method of doing this.

- Technical feasibility: The technical feasibility of the project on hate speech detection in multimodal social media posts would involve assessing the availability of appropriate data and resources for developing a hate speech detection algorithm, as well as the potential challenges and limitations of such an algorithm.

  In terms of challenges and limitations, the project would need to consider the potential difficulties of developing a hate speech detection algorithm for multimodal social media posts. This could include challenges related to the complexity and diversity of the media types in such posts, as well as the potential for hate speech to be expressed in subtle or implicit ways. The project would also need to take into account any moral or legal difficulties pertaining to the use of hate speech detection algorithms, such as concerns about free speech and privacy.

  Overall, the technical feasibility of the project on hate speech detection in multimodal social media posts would depend on the availability of appropriate data and resources, as well as the potential challenges and limitations of developing a hate speech detection algorithm for such posts.

- Legal feasibility: This aspect evaluates the legal implications of developing and implementing the project. The project would need to comply with data privacy laws and ensure that the data collected and analyzed do not violate any user rights. Additionally, the project would need to comply with the terms and conditions of the social media platforms from which the data is being collected.

- Data privacy: The project must abide with data privacy rules, such as the Indian Penal Act (IPC) in the India or the General Data Protection Regulation (GDPR). The project would need to ensure that the data collected and analyzed do not violate any user rights.

- User consent: The project would need to obtain user consent to collect and investigate their data. The consent should be informed, and users should have the decision to opt-out of data gathering and analysis.

- Liability: The project would need to ensure that it does not create any legal liability for itself or the social media platforms from which the data is being collected. The project should also ensure that the developed models do not create any legal liability for the users or the social media platforms.

- Intellectual property: The project needs to comply with intellectual property laws such as copyright and trademark laws. The project would need to ensure that it does not violate any intellectual property rights while collecting and analyzing data from social media platforms.
    a) Terms and conditions: The project would need to comply with the terms and conditions of the social media platforms from which the data is being collected. The terms and conditions may impose restrictions on data collection, analysis, and usage, which the project would need to comply with.

- Financial feasibility: A financial feasibility study of the project on hate speech detection in multimodal social media posts would involve assessing the potential costs and revenue streams associated with developing and selling a hate speech detection tool. This would involve several key steps, including the following:
    a. Gather information on the costs associated with developing and testing a hate speech detection algorithm, including the costs of data collection and annotation, hardware and software, and personnel.
    b. Estimate the potential revenue streams associated with selling a hate speech detection tool to social media platforms, including the potential price of the tool and the number of potential customers.
    c. Conduct a cost-benefit analysis to assess the financial feasibility of the project, including the potential return on investment and the payback period.
    d. Consider potential risks and uncertainties associated with the financial feasibility of the project, such as changes in the market for hate speech detection tools or shifts in the social media industry.
    e. Draw conclusions and make recommendations based on the research and analysis conducted, including suggestions for

improving the financial feasibility of the project and addressing any potential risks or challenges.

f. Overall, a financial feasibility study of the project on hate speech detection in multimodal social media posts would involve conducting research and analysis to assess the potential costs and revenue streams associated with developing and selling a hate speech detection tool, and provide recommendations for moving forward with the project.

- Operational feasibility: Operational feasibility is a crucial aspect of the project, including the Detection of Hate Speech in Multi-Modal Social Posts. It assesses the practicality of implementing the project in a real-world setting. Below are some operational factors that need to be taken into account:
    a. Availability of data: The project's success depends on the accessibility and superiority of data. The project would need to ensure that the data collected from social media platforms are representative and diverse enough to train the models effectively.
    b. Scalability: The project would need to ensure that the developed models are scalable enough to handle the vast amount of data generated by social media platforms daily. The models should also be adaptable to new data sources and languages.
    c. Accuracy: The project's success depends on the accuracy of the developed models in detecting hate speech in multi-modal social posts. The project would need to guarantee that the models are accurate enough to minimize false positives and false negatives.
    d. User-friendliness: The project would need to ensure that the system's user interface is user-friendly and easy to use. The project should also ensure that the system's output is interpretable and actionable, allowing users to take appropriate actions.
    e. Integration: The project would need to ensure that the developed models can be easily integrated with the existing social media platforms. The project should also ensure that the integration does not impact the performance or usability of the social media platforms.

CHAPTER - 3

SYSTEM DESIGN & ANALYSIS

1. Project Perspective

One approach to this problem is to use a combination of sentiment analysis and Visual BERT. Sentiment analysis can be used to determine the emotional tone of the text and identify whether it contains hate speech or not. Visual BERT, on the other hand, can be used to extract visual features from images and videos and integrate them with the text analysis to improve the accuracy of hate speech detection.

From a project perspective, the first step would be to collect and annotate a large dataset of multimodal social posts that contain hate speech. This dataset would be used to train and validate the machine learning model. The dataset should be diverse and representative of different cultures and languages to ensure that the model can generalize well to new data.

Finally, the model should be deployed to a production environment where it can be used to automatically detect hate speech in multimodal social posts. The system should also be monitored and updated regularly to ensure that it remains accurate and up-to-date with the latest trends and patterns of hate speech.

2. Dataset Analysis

The utilization of social media like Facebook has intensified rapidly in recent years. While social media platforms offer a convenient means of communication and information sharing, they are also prone to the misuse of language, including the use of disrespectful language.

To address this issue, researchers and data scientists have developed machine learning models that can automatically identify and flag images and text with disrespectful content. These models require large and diverse datasets to train on, and the Facebook dataset described in the scenario could be a valuable resource for this purpose.

The dataset contains various types of images, including images with text overlays, which can be challenging to analyze using traditional computer vision techniques. The presence of non-disrespectful confounding factors in the images further complicates the task of identifying and classifying disrespectful content.

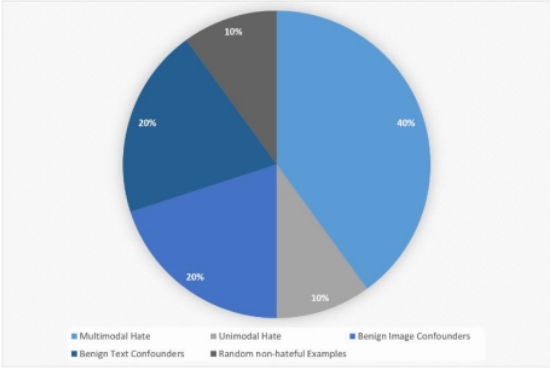

Fig. 4: Types of Social Media Posts in the Dataset

Machine learning models may be trained on the dataset utilising multimodal techniques that use both visual and textual input to address these issues. The dataset may be divided into three sets: a training set, a validation set, and a testing set. The training set is used to train the model, the validation set is used to fine-tune its hyperparameters, and the testing set is used to assess the model's effectiveness.

The linguistic information contained in the photos may also be examined using NLP approaches, which will help the model perform better at spotting and categorising offensive content. To extract important information from the text, for instance, topic modelling and sentiment analysis might be employed.

In conclusion, the Facebook dataset described in the scenario is a valuable resource for developing machine learning models that can identify and classify disrespectful content on social media platforms. The dataset's size, diversity, and

multimodal nature make it a challenging yet rewarding task for data scientists and researchers working on this problem. By developing more accurate and efficient models, we can make the internet a more secure and civil place for all users.

Overall, the Facebook dataset described in the scenario can be a valuable resource for researchers and data scientists working on developing machine learning models for identifying and classifying disrespectful content on social media platforms.

3. Performance Requirements

Detecting and addressing hate speech in social media is a critical challenge for ensuring a safe and respectful online environment. Hate speech is often disguised or concealed within multimodal social media posts, making it challenging to detect using traditional approaches. To address this challenge, machine learning models can be developed to automatically identify and flag hate speech in social media posts, including multimodal posts with both visual and textual information.

However, the success of such a project would depend on the establishment of specific goals and metrics for the performance of the algorithms and techniques developed as part of the project. The performance requirements of the project would be critical in ensuring the effectiveness and impact of the developed models. The following are some of the key performance requirements that should be considered:

- Accuracy: With a low percentage of false positives and false negatives, the algorithms and methods created as part of the project must be able to properly identify hate speech in multimodal social media posts. This would necessitate the creation of models employing common assessment measures for hate speech detection that have good accuracy, recall, AUCROC, and F1 scores. Additionally, the models must be able to identify various types of hate speech and deal with false positives and false negatives.
- Scalability: The algorithms and methodologies developed as part of the project must be able to handle a large volume of social media data and run efficiently on a large number of computing resources. This would require the use of scalable and distributed computing techniques to handle large volumes of data. The models must be able to process data in real-time or near-real-time, and must be able to handle a growing volume of data over time.
- Security: The algorithms and techniques developed as part of the project must be secure and protect the privacy of social media users. This would require the implementation of various security measures to prevent unauthorized access to the algorithms and data, as well as the ability to handle sensitive data in a secure and responsible manner. The models must comply with information safety and privacy regulations, such as the GDPR and the IPC.
- Explainability: The algorithms and techniques developed as part of the project must be interpretable and explainable. This would require the use of transparent machine learning models that can provide explanations for their predictions. This would help build trust in the models and improve their adoption.

- Multilingual support: The algorithms and techniques developed as part of the project must be capable of handling multiple languages. This would require the use of NLP techniques that can handle different languages and dialects, and the ability to scale across multiple languages.

Overall, the performance requirements of the project on hate speech detection in multimodal social media posts would be critical in ensuring the success of the project. The establishment of specific goals and metrics for the performance of the algorithms and techniques developed as part of the project would help ensure their effectiveness and impact. By developing models with high accuracy, scalability, security, explainability, and multilingual support, we can design a protected and more respectful online environment for the users.

4. Methodology

As discussed in the proposed approach the model consists of two 2 parts: Object detection and Sentiment analysis. In order to achieve highest order of accuracy the model adopts several state-of-the-art techniques. For our object detection YOLO v7 is implemented and for the text sentiment analysis SVM model is trained to achieve high accuracy.

Both the models are then concatenated with some adjustments to obtain an overall sentiment score for the said meme. Concatenating the results gives us the more insights in the message the meme is trying to convey, thus significantly decreasing possibility of miss-classification.

- Object Detection: Popular object identification technique YOLO (You Only Look Once) is frequently applied in computer vision applications. It uses deep learning to detect things in pictures and videos with high speed and accuracy. A convolutional neural network is used by YOLO to divide an input image into a grid of cells and predict the presence and placement of items within each cell. The method is quick and effective since it can process an entire image in a single forward run across the network. In comparison to other object identification algorithms, YOLO provides a number of benefits, including speed and accuracy. Additionally, it can recognise numerous things in an image at once and can handle a broad variety of item sizes and forms.

  Overall, YOLO is a powerful and widely-used object detection algorithm that has proven effective in a variety of computer vision applications, including hate speech detection in multimodal social media posts.

- Sentiment Analysis: The method of 'computationally' assessing whether a piece of text is good, negative, or neutral is known as sentiment analysis. We attempt to put into practise a Twitter sentiment analysis model that aids in overcoming the difficulties associated with determining the sentiments of the tweets. The dataset provided is the Sentiment Dataset from twitter which consists of 12,800 tweets. The model classifies the tweets in three classes : Negative ,Neutral ,Positive.

5. APPROACH 1

We began by using a bottom-up approach to extract text from the image. This involved detecting individual words or characters in an image and using OCR technology to convert them into machine-readable text. By leveraging this approach, we were able to capture text information that may not have been available through traditional OCR methods, which is particularly useful when dealing with complex images that contain text in a variety of orientations and sizes.

Once we had extracted the text from the image, we passed the resulting image captions to a BERT model. BERT is a pre-trained language model that uses a transformer-based neural network architecture to encode natural language text. By using a BERT model, we were able to represent the extracted text as a high-dimensional vector that captures its semantic meaning.

Additionally, we fed the dataset to the VisualBERT model for further analysis and processing. VisualBERT is a neural network architecture that combines image and text processing in a single model. It uses a transformer-based architecture similar to BERT, but is designed to encode both image features and textual information in a joint representation. By using VisualBERT, we were able to capture the appropriate knowledge of the text in relation to the image it appears in.

Finally, we combined the outcomes from the two models and classified the text as either being hateful or not as a consequence. This phase included utilising the combined outputs from the BERT and VisualBERT models as input features to train a classifier on a labelled dataset of hateful and non-hateful text samples. We were able to take use of each model's advantages by merging the data from the two, which increased the classification's overall accuracy.

Overall, our approach is an example of how computer vision and natural language processing can be combined to tackle complex tasks such as text classification. The increase in accuracy that we achieved by combining these techniques is a testament to the power of these tools and their potential to drive real-world applications in areas such as hate speech detection and content moderation.

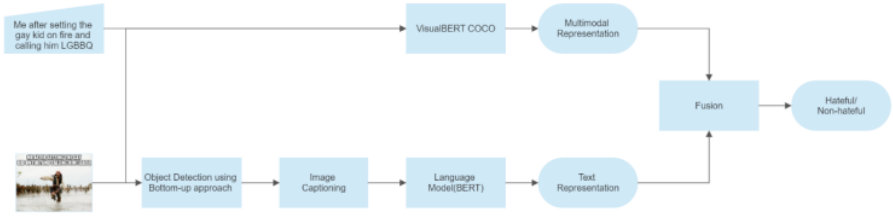

Fig. 5 Model Architecture 1

6. APPROACH 2

VisualBERT is a cutting-edge model that excels in understanding multimodal data, particularly visual content and language. However, it may not always be the most suitable option for every task and dataset, as it has some limitations. For example, VisualBERT does not have the ability to identify the sentiment or emotion behind the text or object.

To overcome this limitation, we have added two sentiment analysis models - one for text and one for images - to improve the overall accuracy of the model. The addition of sentiment analysis has resulted in an increase in accuracy by more than 4%. The sentiment analysis models are designed to detect the emotional tone of the text and images, whether it is positive, negative or neutral.

The sentiment analysis approach is particularly useful in detecting irony in multimodal data, such as in the case of memes where the captions can often contradict the sentiment conveyed by the image. By analyzing both the text and the image, the model is able to gain a more nuanced understanding of the sentiment being expressed and provide more accurate predictions. This makes our approach particularly effective in applications such as social media monitoring, where understanding the sentiment behind user-generated content is critical for businesses and organizations.

Overall, the addition of sentiment analysis to the VisualBERT model has significantly enhanced its accuracy and made it more robust for a wider range of tasks and datasets.

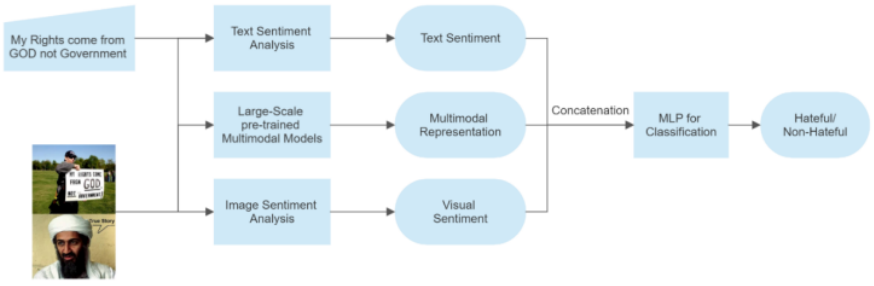

Fig. 6: Model Architecture 2

7. BENEFITS OF APPROACH 2

In particular, the use of irony can make it difficult to accurately identify hateful content. Irony involves saying something that is opposite to what is meant, often with the intent to mock or ridicule. This may be particularly troublesome when it comes to hate speech since it gives people the opportunity to dissimulate their genuine motives and promote divisive ideas under the guise of humour or satire.

To address this challenge, our approach incorporates sentiment analysis into the VisualBERT model, which is a state-of-the-art language and vision model. By analyzing the sentiment of both the textual and visual components of a post, our model is able to identify instances of irony, even when the individual elements of a post appear benign. This can help to flag posts that may contain hidden hate speech, leading to more effective detection and prevention of harmful content.

In addition to detecting irony, our approach also has the ability to distinguish between non-hateful memes and actual instances of hate speech. This is important because not all multimodal social media posts are intended to be harmful or offensive. For example, memes can be used to convey humor or cultural references, without any malicious intent. By analyzing the sentiment of both the textual and visual components of a post, our model can accurately identify non-hateful content, leading to a more nuanced and accurate analysis of multimodal data.

Overall, the addition of sentiment analysis to the VisualBERT model represents a significant advancement in the sphere of hate speech detection for multimodal social media posts. By detecting irony and identifying non-hateful content, our approach can help to create a more inclusive and respectful online community.

8. Algorithm

Here is a high-level algorithm for detecting hate speech in multi-modal social posts using sentiment analysis and VisualBERT:

Data collection: Collect multi-modal social media posts (e.g., text, images, videos) from various social media platforms (e.g., Twitter, Facebook, Instagram) and annotate them with relevant labels (e.g., hate speech, offensive, non-offensive).

Pre-processing: Pre-process the data by cleaning and formatting the text, images, and videos. This can include steps such as removing stop words, stemming or lemmatizing words, and resizing images and videos.

Sentiment analysis: Perform sentiment analysis on the text portion of the posts using a pre-trained sentiment analysis model. This will classify the text as positive, negative, or neutral.

VisualBERT encoding: Encode the visual content (e.g., images, videos) of the posts using VisualBERT, a pre-trained visual language representation model that can encode visual inputs into a textual representation. This will convert the visual content into a language representation that can be analyzed alongside the text.

Multi-modal feature extraction: Combine the textual and visual representations of each post to extract relevant features that can help identify hate speech. This can include features such as the sentiment of the text, the presence of certain words or phrases, and the content and context of the visual components.

Classification: Use the collected characteristics to train a multi-modal classification model (such a neural network) that will categorise the postings as hate speech or non-hate speech.

Evaluation: Examine the model's performance using a variety of measures (such as accuracy, recall, and F1-score) and make any required adjustments.

Deployment: Deploy the trained model to automatically detect and flag hate speech in real-time social media posts.

Note that this algorithm is a high-level overview and the exact implementation may vary depending on the specific use case and available resources. Additionally, detecting hate speech is a complex and nuanced task, and any automated system should be used in conjunction with human moderators to ensure accuracy and fairness.

CHAPTER - 4

RESULTS AND OUTPUTS

1. AUCROC

A diagram that contrasts the True Positive Rate (TPR) and False Positive Rate (FPR) is known as the Recipient Working Attributes bend. It assesses how well the parallel classifier distinguishes between classes while the decision edge changes. (1997, Bradley).

An ideal classifier will have a region underneath the bend of one, with the ideal point in the upper left corner of the plot having a TPR of one and a FPR of nothing. To improve TPR and decline FPR, each classifier ought to have a more noteworthy region under the curve.

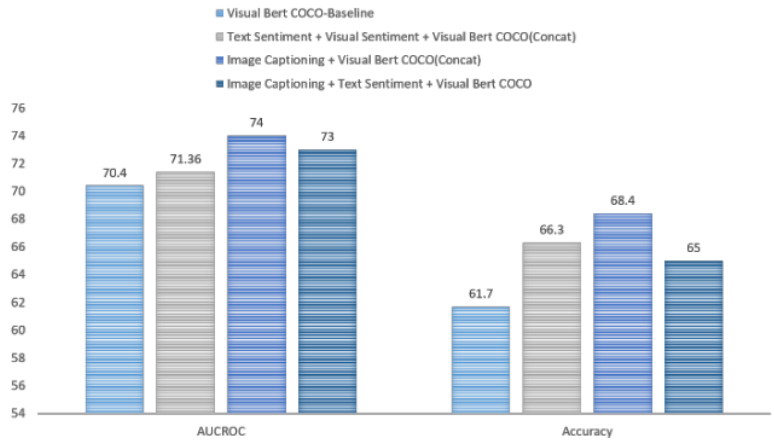

Fig. 7: AUCROC

2. Classification Accuracy

Since it is more straightforward to understand, we decide the precision of the gauges as the proportion of right expectations to the all out number of forecasts delivered. Subsequently, for each test, we yield the names 0 and 1, as well as the likelihood with which the classifier predicts that the example is loathed. The AUCROC bend is plotted utilizing this likelihood.

3. Test Cases

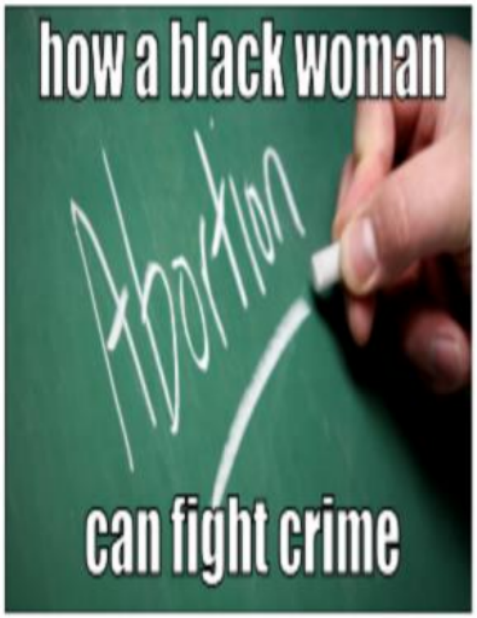

Fig. 8: Test Case 1

Test Case ID	1.
Test Case Name	Abortion
Test Case Description	Abortion is written on a blackboard with a caption
Analysis	Abortion on Blackboard shows it is written for education purposes, but the caption attached with the image impacts the dark colored people which can impact a lot on their mental condition, in a way it is racist to some set of people signifying geographical domain
Expected Result	Extreme Hateful

Actual Result	Hateful

Fig. 9: Test Case 2

Test Case ID	2.
Test Case Name	Woman wearing a burkha
Test Case Description	A sensitive caption is written on an image of a woman who is wearing a burkha due to religious practices
Analysis	The image is normal but the caption attached to the pic made it sensitive to certain religious people as it is part of religious attire
Expected Result	Hateful
Actual Result	Hateful

Fig. 10: Test Case 3

Test Case ID	3.
Test Case Name	Woman
Test Case Description	Woman wearing gloves like in a picnic kind of thing
Analysis	Normal image with woman holding some utensil like structure but the caption attached targetted a certain demographic location people and the caption isn't related to the picture
Expected Result	Benign
Actual Result	Benign

Fig. 11: Test Case 4

Test Case ID	4.
Test Case Name	Man and horse
Test Case Description	Man on a horse in a desert
Analysis	Caption attached is somehow depicting to make a food out of thing but the picture shows man is going to harm the animal for its sake of benefit
Expected Result	Benign
Actual Result	Hateful

Fig. 12: Test Case 5

Test Case ID	5.
Test Case Name	Hitler and a small girl
Test Case Description	Hitler is holding a girl and is having viscious smile on his face
Analysis	Caption here used is oxymoron situation in which anything given by Hitler is consider as punishment to the citizen
Expected Result	Left one is Hateful and Right one is Benign
Actual Result	Left Hateful, Right Benign

Fig. 13: Test Case 6

Test Case ID	6.
Test Case Name	Jack sparrow
Test Case Description	People running after Jack Sparrow in a motive to hurt him over his actions
Analysis	Caption attached to the pic is very sensitive to the certain people in the society and diminishes the morals of the people.
Expected Result	Very Hateful
Actual Result	Extreme Hateful

Fig. 14: Test Case 7

Test Case ID	7.
Test Case Name	Terrorist
Test Case Description	Banner to spread message with a photo of terrorist
Analysis	Caption to spread a wrong message to the society about any action done by terrorist in the name of God and inviting members to the tasks

Expected Result	Extreme Hateful
Actual Result	Hateful

Fig 15 : Test Case 8

Test Case ID	8.
Test Case Name	Anne Frank
Test Case Description	Anne Frank Smiling
Analysis	It could be interpreted as a potentially dangerous situation as the smell of gas could indicate a gas leak, which can be harmful or even deadly.
Expected Result	Benign
Actual Result	Benign

Fig 16 : Test Case 9

Test Case ID	9.
Test Case Name	African Boy
Test Case Description	Boy drinking water from the tap
Analysis	Dark coloured boy is compared with the farm in the caption which is considered as hot due to extreme weather condition in the African region
Expected Result	Hateful
Actual Result	Hateful

Fig 17: Test Case 10

Test Case ID	10.
Test Case Name	Actor
Test Case Description	Actor in a car to show disgrace to someones' comment
Analysis	Caption signifies that large number of audience doesn't pay any heed to the current happening the world.
Expected Result	Hateful
Actual Result	Benign

CHAPTER - 5

CONCLUSION

The project on hate speech detection in multimodal social media posts aims to develop algorithms and techniques for automatically detecting and flagging hate speech in such posts. The project involves conducting research on existing hate speech detection algorithms and techniques, as well as collecting and annotating a dataset of multimodal social media posts for training and testing purposes.

The project's finding would be that, with a low rate of false positives and false negatives, the algorithms and approaches created as part of the research are capable of reliably and effectively detecting hate speech in multimodal social media posts. To enhance the online environment and encourage tolerance and diversity, these algorithms and strategies might be included into social media platforms.

In terms of future scope, there are several potential directions for the project to take in the future. For example, the project could be extended to include additional media types, such as audio or video, and to handle a wider range of languages and cultural contexts. The project could also be extended to include additional types of hate speech, such as cyberbullying or harassment, and to address other challenges and biases in the data.

Overall, the project on hate speech detection in multimodal social media posts has the potential to make a significant impact on the online environment and to promote a more positive and respectful online community

CHAPTER - 6

FUTURE SCOPE

There is a lot of room for further study and development with the suggested concept. The performance of the model can be improved in the future by using more sophisticated deep learning architectures and algorithms. To more correctly identify hate speech, the model may be expanded to include other contextual data, such as user profiles and network information.

Another future scope is to expand the dataset to include more diverse and representative data from different languages, cultures, and regions. This can enable the model to generalize better and be more effective in detecting hate speech across various social media platforms.

Additionally, the proposed model can be used to develop an automated system for detecting and removing hate speech from social media platforms. This system can be integrated with existing moderation systems, enabling platforms to respond more efficiently to hate speech and create a safer online environment for users.

Another potential future scope of this project is to extend the model to detect other types of harmful content, such as cyberbullying, harassment, and misinformation. This can provide a comprehensive solution for detecting and mitigating various forms of harmful content on social media platforms.

In conclusion, the proposed project has significant potential for future research and development, and can be extended to enhance its performance, expand its scope, and provide a comprehensive solution for detecting and mitigating various forms of harmful content on social media platforms.

.

REFERENCES

1. Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, & Jingjing Liu. Uniter: Universal image-text representation learning. In ECCV, 2020.
2. Shenoy, A. & Sardana, A. Multilogue-net: A context aware rnn for multi-modal emotion detection & sentiment analysis in conversation. arXiv preprint arXiv:2002.08267, 2020
3. Raul Gomez, Jaume Gibert, Lluis Gomez, & Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In WACV, 2020.
4. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. Bottom-up & top-down attention for image captioning & visual question answering. In Proceedings of the IEEE conference on computer vision & pattern recognition, pp. 6077–6086, 201
5. A manpreet Singh, Vedanuj Goswami, & Devi Parikh. Are we pretraining it right? digging deeper into visio-linguistic pretraining, 2020. arXiv:2004.08744
6. Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, & Jianfeng Gao. Unified vision-language pre-training for image captioning & vqa. In AAAI, 2019.
7. Hao Tan & Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In EMNLP, 2019.
8. Yang, F., Peng, X., Ghosh, G., Shilon, R., Ma, H., Moore, E., & Predovic, G. Exploring deep multimodal fusion of text & photo for hate speech classification. In Proceedings of the Third Workshop on Abusive Language Online, pp. 11–18, Florence, Italy, August 2019.
9. Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord, “Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection,” in AAAI, 2019.
10. L iu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
11. Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang, “Unpaired image captioning via scene graph alignments,” in ICCV, 2019.
12. Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., & Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-based systems, 161:124–133, 2018.
13. Singh, A., Goswami, V., Natarajan, V., Jiang, Y., Chen, X., Shah, M., Rohrbach, M., Batra, D., & Parikh, D. Mmf: A multimodal framework for vision & language research
14. Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., & Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-based systems, 161:124–133, 2018.
15. Balaji Lakshminarayanan, Alexander Pritzel, & Charles Blundell. Simple & scalable predictive uncertainty estimation using deep ensembles. In NIPS, 2017
16. Shervin Malmasi and Marcos Zampieri, “Detecting hate speech in social media,” CoRR, vol. abs/1712.06427, 2017.
17. Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao, “Multi-modal factorized bilinear pooling with coattention learning for visual question answering,” in ICCV, 2017.

18. Oriol Vinyals, Alexander Toshev, Samy Bengio, & Dumitru Erhan. Show & tell: Lessons learned from the 2015 mscoco image captioning challenge. In PAMI, 2016.
19. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. Show, attend & tell: Neural image caption generation with visual attention. In International conference on machine learning, pp. 2048–2057, 2015.

ANNEXURE I

Review Paper for the said project has been accepted in International Conference on Artificial Intelligence, Blockchain, Computing and Security (ICABCS-2023)

Paper Title: Detection of Hate Speech in Multimodal Social Posts

Abstract:

It has been observed in the past few years, multi-modal problems have been capable of attaining the interest of a large number of people. The core challenges faced in such problems are its representation, alignment, fusion, co-learning, and translation. The focus of this paper is on the analysis of multimodal memes for hate speech. On the evaluation of the dataset, we found out that the common statistics factors which were hateful initially became benign simply by unfolding the picture of the meme. Correspondingly, a bulk of the multi-modal baselines gives hate speech more options. In order to deal with such issues, we discover the visible modality through the use of item detection and image captioning fashions to realize the "real caption" after which we integrate it with multi-modal illustration to carry out binary classification. The method challenges the benign textual content co-founders present in the dataset to enhance the enactment. The second method that we use to test is to enhance the prediction with sentiment evaluation. It includes a unimodal sentiment to complement the features. Also we carry out in depth evaluation of the above methods stated, supplying compelling motives in want of the methodologies used.

Authors:

Abhishek Goswami, Ayushi Rawat, Shubham Tongaria, Sushant Jhingran


GALGOTIAS UNIVERSITY NAAC A+ Taylor & Francis Taylor & Francis Group

International Conference on Artifical Intelligence, Blockchain, Computing and Security (ICABCS-2023)

Date – 24th to 25th, February, 2023

Certificate of Participation

This is to certify that *Abhishek Goswami* of Sharda University, Greater Noida presented his/her research paper titled **"Detection of Hate Speech in Multi Modal Social Posts"** in the ICABCS-2023 organized by Galgotias University, Greater Noida, Uttar Pradesh, India from 24th to 25th February, 2023.

Dr. Arvind Dagur Conference Chair Professor, Galgotias University(U.P.)

Dr. Munish Sabharwal Conference General Chair Dean SCSE, Galgotias University(U.P.)

Dr. Avadhesh Kumar Conference General Chair Pro-Vice-Chancellor, Galgotias University(U.P.)

Dr. K. Mallikharjuna Babu Patron Vice-Chancellor, Galgotias University(U.P.)

FYP 2023

ORIGINALITY REPORT

10%	8%	7%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	1%
2	Submitted to University of Teesside Student Paper	1%
3	Ahshanul Haque, Md Naseef-Ur-Rahman Chowdhury. "Hate Speech Detection in Social Media Using the Ensemble Learning Technique", Institute of Electrical and Electronics Engineers (IEEE), 2023 Publication	1%
4	www.researchgate.net Internet Source	1%
5	export.arxiv.org Internet Source	<1%
6	"ECAI 2020", IOS Press, 2020 Publication	<1%
7	link.springer.com Internet Source	<1%
8	Submitted to Coventry University Student Paper	

		<1%
9	www.mdpi.com Internet Source	<1%
10	github-wiki-see.page Internet Source	<1%
11	aclanthology.org Internet Source	<1%
12	Submitted to Liverpool John Moores University Student Paper	<1%
13	repository.kaust.edu.sa Internet Source	<1%
14	www.slideshare.net Internet Source	<1%
15	www.sweetstudy.com Internet Source	<1%
16	Submitted to University of Wales Institute, Cardiff Student Paper	<1%
17	"Advanced Information Networking and Applications", Springer Science and Business Media LLC, 2020 Publication	<1%

18	Submitted to University of Technology, Sydney Student Paper	<1%
19	www.science.gov Internet Source	<1%
20	publichealth.jmir.org Internet Source	<1%
21	"Computational Intelligence and Data Analytics", Springer Science and Business Media LLC, 2023 Publication	<1%
22	ijmrap.com Internet Source	<1%
23	Submitted to University of Southampton Student Paper	<1%
24	docplayer.net Internet Source	<1%
25	www.mbie.govt.nz Internet Source	<1%
26	www.multi-mania.be Internet Source	<1%
27	"Advances in Soft Computing", Springer Science and Business Media LLC, 2021 Publication	<1%
	Submitted to Covenant University	

28	Student Paper	<1%
29	rua.ua.es Internet Source	<1%
30	Submitted to The Robert Gordon University Student Paper	<1%
31	internationalhatestudies.com Internet Source	<1%
32	www.dummies.com Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

FYP 2023

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

PAGE 35

PAGE 36

PAGE 37

PAGE 38

PAGE 39

PAGE 40

PAGE 41

PAGE 42

PAGE 43

PAGE 44