DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,

SHARDA SCHOOL OF ENGINEERING AND TECHNOLOGY,

SHARDA UNIVERSITY, GREATER NOIDA

Detection of Hate Speech In Multi-Modal Social Posts

A project submitted
in partial fulfillment of the requirements for the degree of
Bachelor of Technology in Computer Science and Engineering

by

Abhishek Goswami (2019645386)

Ayushi Rawat (2019593585)

Shubham Tongaria (2019544008)

Supervised by:

Mr. Sushant Jhingran, Assistant Professor (CSE)

May, 2023

CERTIFICATE

This is to certify that the report entitled **“Detection of Hate Speech in Multi-Modal Social Posts”** submitted by **Mr. Abhishek Goswami (2019645386), Ms Ayushi Rawat (2019593582) and Mr. Shubham Tongaria (2019544008)** to Sharda University, towards the fulfillment of requirements of the degree of Bachelor of Technology is record of bonafide final year Project work carried out by him/her in the Department of Computer Science and Engineering, School of Engineering and Technology, Sharda University. The results/findings contained in this Project have not been submitted in part or full to any other University/Institute for award of any other Degree/Diploma.

Signature of Supervisor

Name: Mr. Sushant Jhingran

Designation: Assistant Professor

Signature of Head of Department

Name: Prof. (Dr) Nitin Rakesh

(Office seal)

Place:

Date:

Signature of External Examiner

Date:

ACKNOWLEDGEMENT

A major project is a golden opportunity for learning and self-development. We consider our self very lucky and honored to have so many wonderful people lead us through in completion of this project.

First and foremost, we would like to thank Dr. Nitin Rakesh, HOD, CSE who gave us an opportunity to undertake this project.

My grateful thanks to Mr. Sushant Jhingran for his guidance in my project work. Mr. Sushant Jhingran, who in spite of being extraordinarily busy with academics, took time out to hear, guide and keep us on the correct path. We do not know where we would have been without his help.

CSE department monitored our progress and arranged all facilities to make life easier. We choose this moment to acknowledge their contribution gratefully.

ABHISHEK GOSWAMI (2019645386)	AYUSHI RAWAT (2019593582)	SHUBHAM TONGARIA (2019544008)

ABSTRACT

Memes are a tool aimed at conveying concepts on social media. Even while most memes are meant to be funny, some can turn into harsh statements when text and images are added. It may be possible to lessen the negative social effects of hateful memes by automatically recognising them. In contrast to old-style multimodal activities, where there is semantic association between the text and the image. The model must understand the information and carry out reasoning across various modalities since the picture and text in memes are out of alignment or irrelevant. These issues' representation, alignment, fusion, co-learning, and translation concerns. 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.

The proposed model aims to analyze social media posts that contain text, images using a combination of natural language processing and computer vision techniques. The dataset used for this project will be multi-modal, meaning it will contain different types of data, such as text, images. The model will employ various deep learning architectures such as Sentiment Analysis, Visual BERT, and transformer-based models to extract features from the data and identify patterns that are indicative of hate speech.

Correspondingly, a bulk of the multi-modal baselines gives hate speech more options. In order to address these problems, 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. In this research, we concentrate on multimodal hostile memes identification and offer a unique technique that combines meme recognition with picture captioning. Also, we carry out in depth evaluation of the above methods stated, supplying compelling motives in want of the methodologies used.

The successful development of the proposed model will have significant implications in the prevention of hate speech on social media platforms. The model will be able to detect hate speech accurately, which will enable social media platforms to take appropriate actions against it. The model can also be used as a tool to raise awareness and educate users about the impact of hate speech on individuals and society as a whole.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER – 1

INTRODUCTION

1. Problem Statement

The problem that the project on hate speech detection in multimodal social media posts seeks to address is the prevalence of hate speech in social media. Hate speech, which refers to any speech or expression that attacks or insults a person or group based on their race, religion, gender, sexual orientation, or other characteristics, is a serious issue that can have harmful consequences for the individuals and communities targeted by such speech.

In recent years, the rise of social media has made it easier for individuals to spread hate speech and target vulnerable groups. 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.

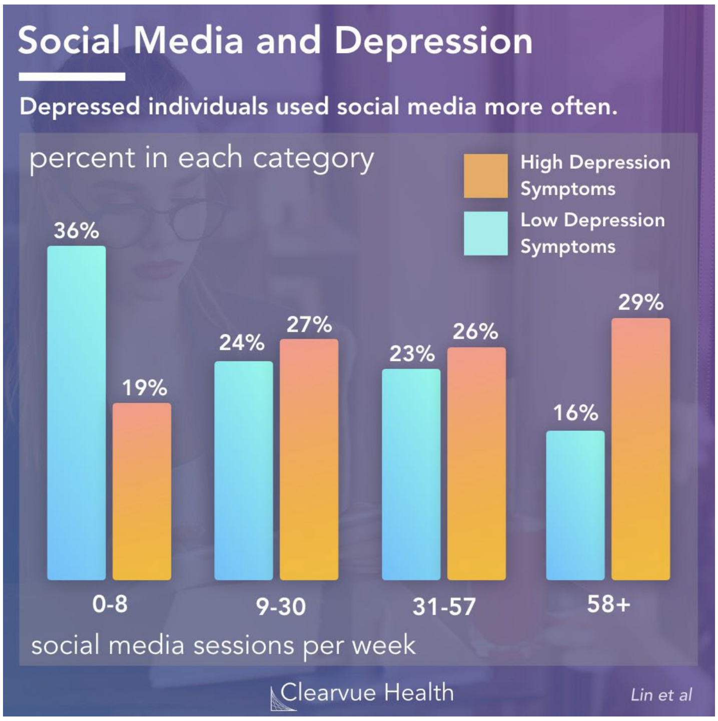


Fig. 1. Social Media and Depression

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 project could help to create a safer and more inclusive online environment 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.

By creating algorithms and approaches for automatically identifying and flagging hate speech in multimodal social media postings, the project on hate speech detection in multimodal social media posts seeks to address this issue. 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 research involves reviewing the relevant literature, researching the latest techniques and analyzing existing hate speech materials. The purpose of this research is to identify the most effective algorithms and techniques for detecting hate speech in multimodal social media.

To train and test the algorithms, we collect and annotate a multimodal social media dataset. The dataset consists of social media that contain text, images and videos. We collect data from various social media platforms such as Twitter, Facebook and Instagram. The dataset is marked by human annotations, which mark posts as hateful or non-hateful. Comments are used to train and evaluate our algorithms.

The project develops and evaluates machine learning algorithms to detect hate speech in multimodal social media. Deep learning techniques such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are used to develop the algorithms. Algorithms are trained on annotated social media datasets. We also consider different features such as text, images and videos to improve the accuracy of the algorithms. We evaluate the performance of the algorithms using various metrics such as precision, 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.

3. Expected Outcome

The expected outcome of the project on hate speech detection in multimodal social media posts is the development of algorithms and techniques 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 help to reduce the prevalence of hate speech on social media and create a more positive and respectful online environment 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.

4. 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.

Python is a popular programming language in machine learning and artificial intelligence research, due to its simplicity and the large number of libraries available for data analysis and machine learning. TensorFlow is a popular deep learning framework for building and training machine learning models, and Keras is a high-level neural network API that runs on top of TensorFlow, making it easier to build and train neural networks.

In addition to these libraries, the project may also require the use of the Natural Language Toolkit (NLTK) library for natural language processing tasks 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. It provides free access to GPUs and TPUs, which can be used to speed up the training of machine learning models. Additionally, Colab comes preinstalled with many of the libraries and tools required for machine learning and data analysis, making it easy to get started with 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

5. 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 detect hate speech in multimodal social media posts, with a low rate 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 a large number of computing resources. 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 detecting hate speech in multimodal memes.

Another example is the WSDM Cup 2021 Task 2: Hate Speech Detection in Multimodal Posts, which is a competition organized by the Association for Computing Machinery's Special Interest Group on Information Retrieval. The competition involves developing algorithms for detecting hate speech in multimodal posts on the Reddit platform, which includes a variety of media types, such as text, images, and videos. The competition has attracted a number of participants and has resulted in the development of a number of algorithms and techniques for detecting hate speech in multimodal social media posts.

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.

For effective photo indexing, searching, and grouping in this study, the following last layer of the pre-trained ResNet neural network on ImageNet is used to create the picture embeddings. 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 Previous research on hate speech has solely used textual 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	Detecting the specific features in hate-speech and if these features provide insights regarding the identification of hate-speech.
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 outperforms state-of-the-art models over multiple V+L tasks by a significant margin.
3	Automatic Hate Speech Detection using Machine Learning: A Comparative Study	2020	SVM and RF algorithms showed better results compared to LR, NB, KNN, DT, AdaBoost, and MLP.
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 multi-modal sentiment analysis and emotion detection in conversation.
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	Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Transformer-based Models	Tokenization, Stemming, Stop-word Removal, Resizing, Normalization	Precision, Recall, F1-Score
Hateful Memes Detection Challenge Baseline Model	Text, Images	Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), 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	Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Bi-directional Encoder Representations from Transformers (BERT)	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	Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BiLSTM), Hierarchical Attention Network	Tokenization, Stemming, Stop-word Removal, Resizing, Normalization	Area Under the Receiver Operating Characteristic Curve (AUC-ROC)
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.

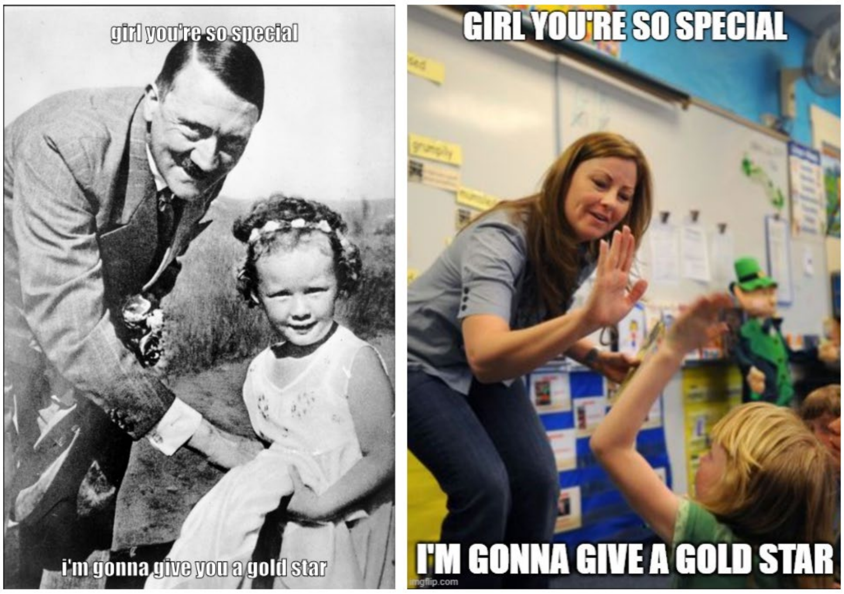


Fig. 3. Example of a Hateful Meme (Left) and Benign Meme (Right)

5. Feasibility Study

A feasibility study 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.

The feasibility study would also involve conducting a market analysis to assess the potential demand for a hate speech detection tool in the social media industry. This could include surveying social media platforms and users to determine their interest in and need for such a tool.

- 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 consider any legal or ethical considerations related to the use of hate speech detection algorithms, such as issues of privacy and free speech.

  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 needs to comply with data privacy laws such as the General Data Protection Regulation (GDPR) in the European Union or the California Consumer Privacy Act (CCPA) in the United States. 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 analyze their data. The consent should be informed, and users should have the option to opt-out of data collection 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 considerations that need to be taken into account:
  a. Availability of data: The project's success depends on the availability and quality 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 ensure 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 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 use of social media platforms like Facebook has grown 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.

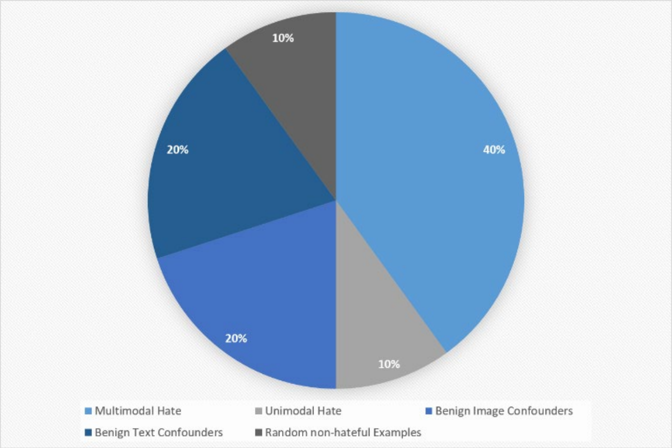


Fig. 4: Types of Social Media Posts in the Dataset

To address these challenges, machine learning models can be trained on the dataset using multimodal approaches that combine both visual and textual information. The dataset can be split into training, validation, and testing sets, with the training set used to train the model, the validation set used to tune model hyperparameters, and the testing set used to evaluate the model's performance.

NLP techniques can also be used to analyze the textual content in the images, which can further improve the model's performance in identifying and classifying disrespectful content. For example, techniques like sentiment analysis and topic modeling can be used to extract meaningful information from the text.

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 create a safer and more respectful online environment 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: The algorithms and techniques developed as part of the project must be able to accurately detect hate speech in multimodal social media posts, with a low rate of false positives and false negatives. This would require the development of models with high precision, recall, AUCROC, and F1 score, using standard evaluation metrics for hate speech detection. The models must also be capable of detecting various forms of hate speech and address the issue of false positives and false negatives.
- Scalability: The algorithms and techniques 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 data protection and privacy regulations, such as the GDPR and the CCPA.
- 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 create a safer and more respectful online environment for all 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: YOLO (You Only Look Once) is a popular object detection algorithm that is widely used in computer vision applications. It is a deep learning-based approach that is able to quickly and accurately identify objects in images and videos. YOLO works by dividing an input image into a grid of cells, and using a convolutional neural network to predict the presence and location of objects within each cell. The algorithm is able to process an entire image in a single forward pass through the network, making it fast and efficient. YOLO has several advantages over other object detection algorithms, including its speed and accuracy. It is also able to handle a wide range of object sizes and shapes, and is able to detect multiple objects in an image simultaneously.

  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: Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. We try to implement a Twitter sentiment analysis model that helps to overcome the challenges of identifying 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 contextual information of the text in relation to the image it appears in.

Finally, we concatenated the results from both models and used them to classify the text as either hateful or non-hateful. This step involved training a classifier on a labeled dataset of hateful and non-hateful text samples, using the combined outputs from the BERT and VisualBERT models as input features. By combining the results from both models, we were able to leverage the strengths of each to improve the overall accuracy of our classification.

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.

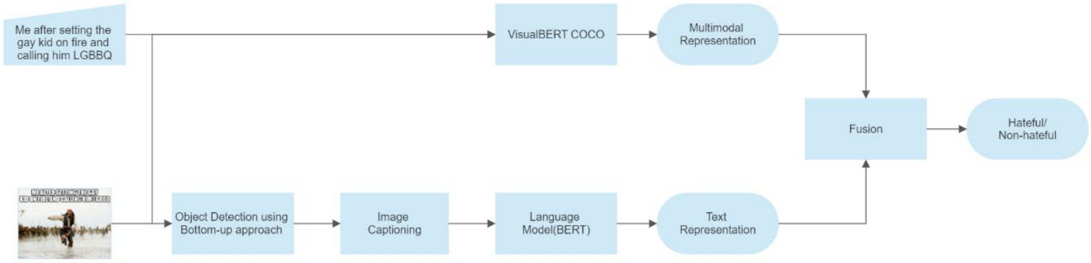

Me after setting the gay kid on fire and calling him LGBBQ
VisualBERT COCO
Multimodal Representation
Fusion
Hateful/ Non-hateful
ME AFTER SETTING THE GAY KID ON FIRE AND CALLING HIM LGBBQ
Object Detection using Bottom-up approach
Image Captioning
Language Model(BERT)
Text Representation

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.

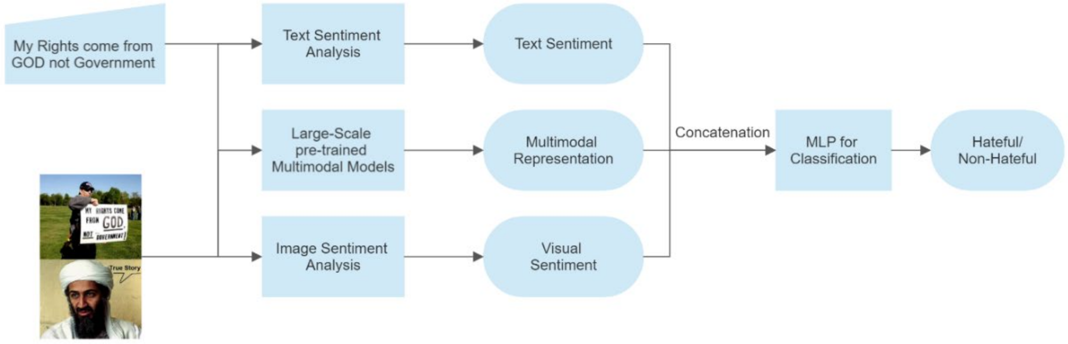


Fig. 6: Model Architecture 2

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 can be especially problematic in the context of hate speech, as it can allow individuals to disguise their true intentions and spread hateful messages under the guise of humor 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 field 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: Train a multi-modal classification model (e.g., a neural network) on the extracted features to classify the posts as either hate speech or non-hate speech.

Evaluation: Evaluate the performance of the model using various metrics (e.g., precision, recall, F1-score) and fine-tune the model as necessary.

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

The Recipient Working Attributes bend is a diagram that compares the True Positive Rate (TPR) to the False Positive Rate (FPR). It evaluates how effectively the parallel classifier recognizes classes as the choice edge is changed. Bradley (1997).

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.

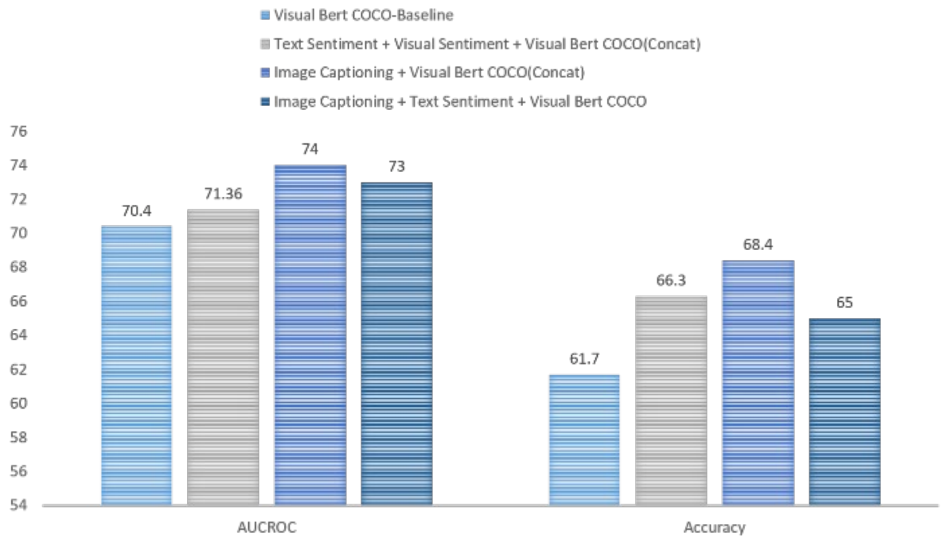


Fig. 7: AUCROC

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

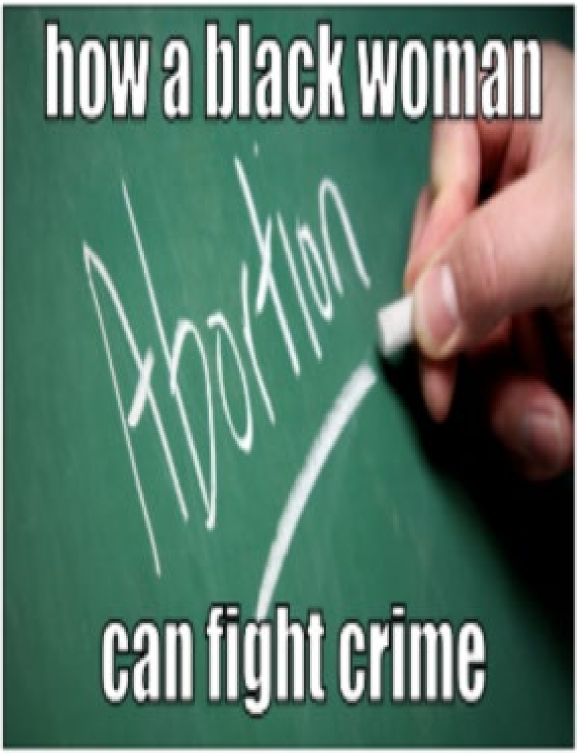


Fig. 8: Test Case 1

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

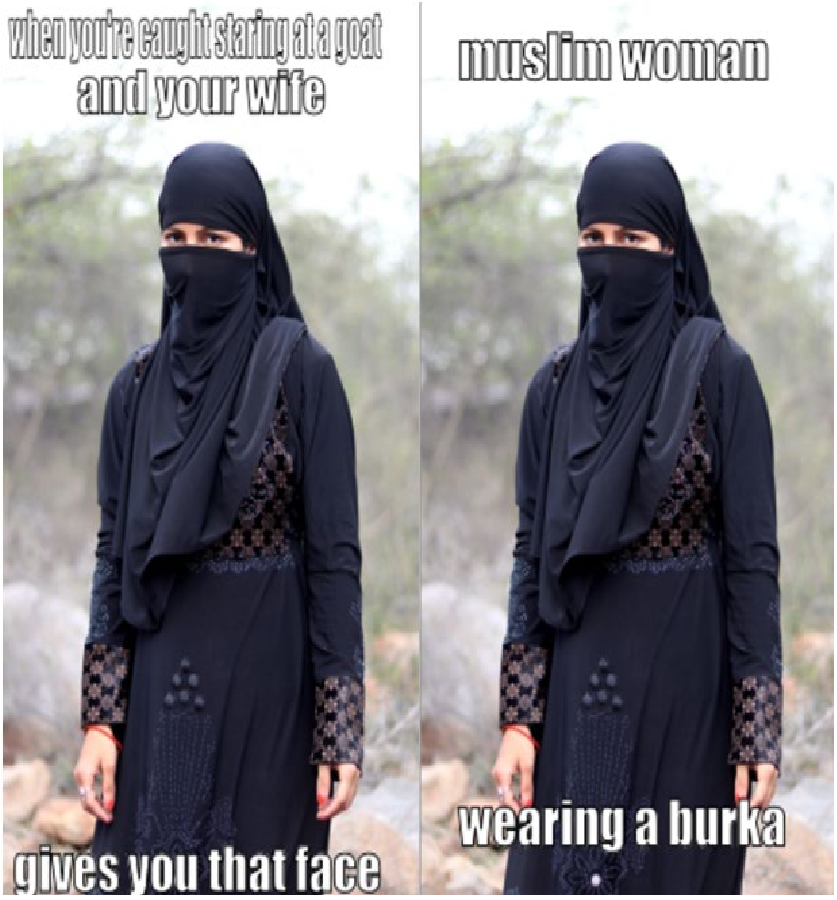


Fig. 9: Test Case 2

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

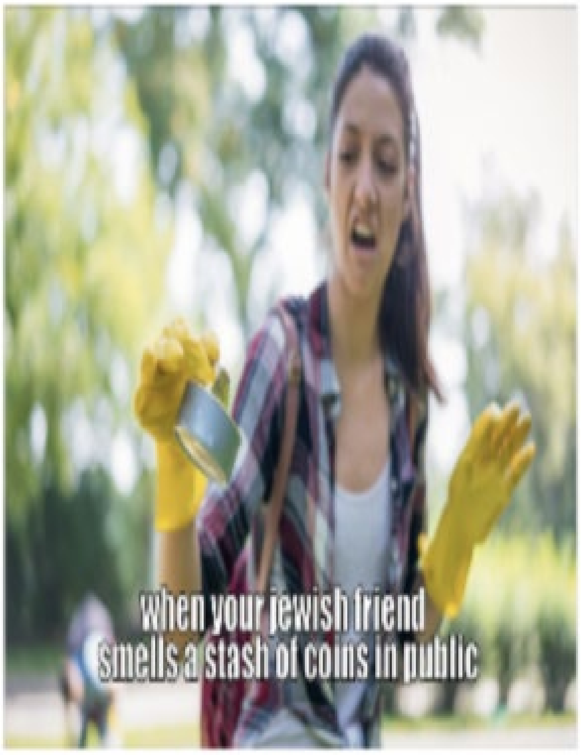


Fig. 10: Test Case 3

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

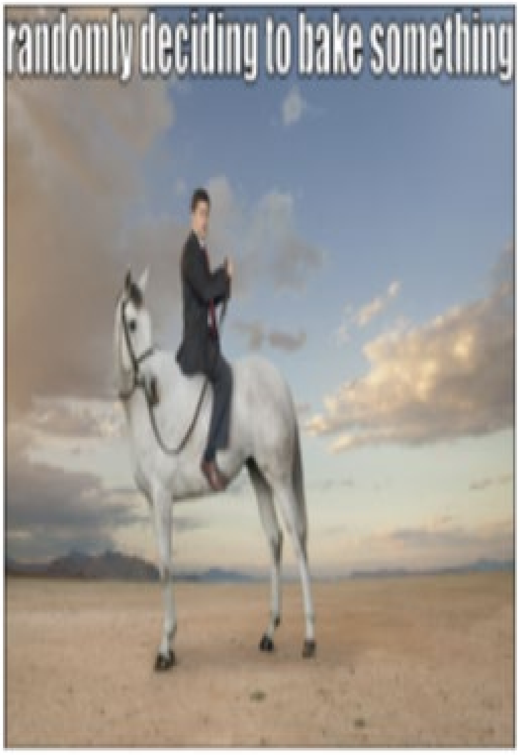


Fig. 11: Test Case 4

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

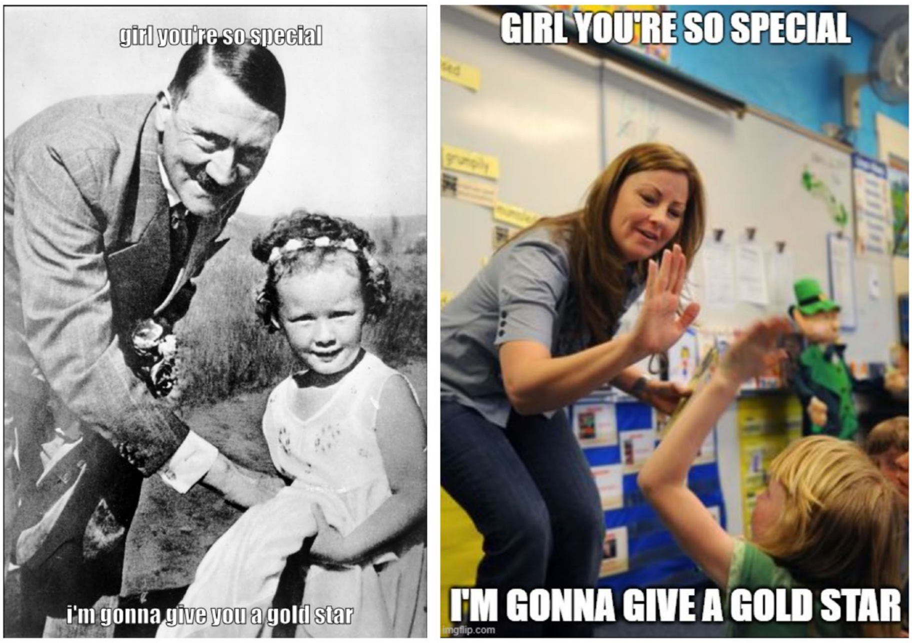


Fig. 12: Test Case 5

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

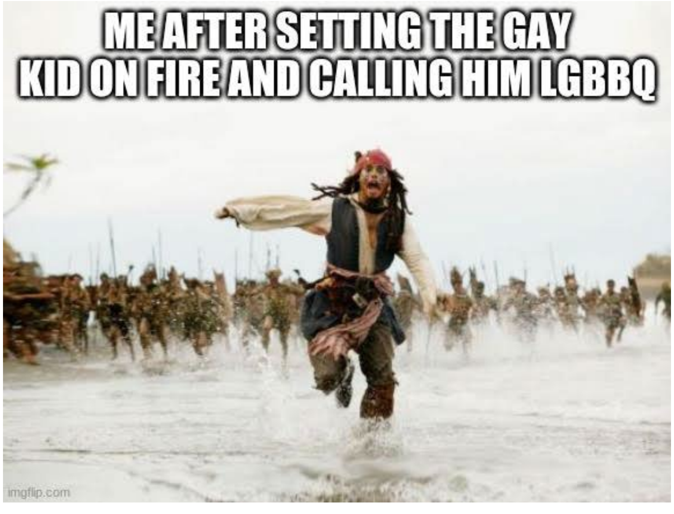


Fig. 13: Test Case 6

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

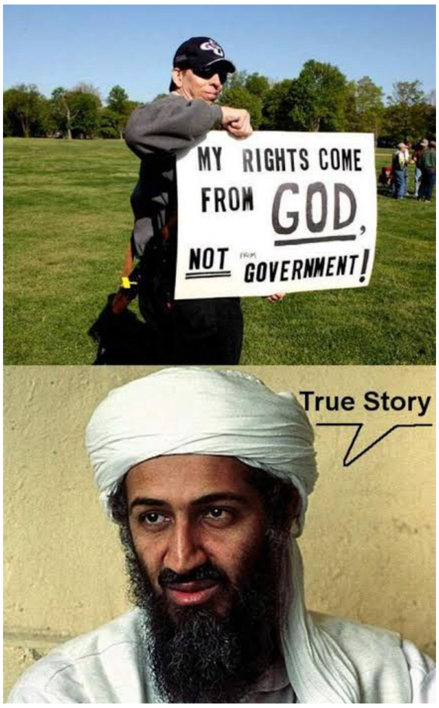


Fig. 14: Test Case 7

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

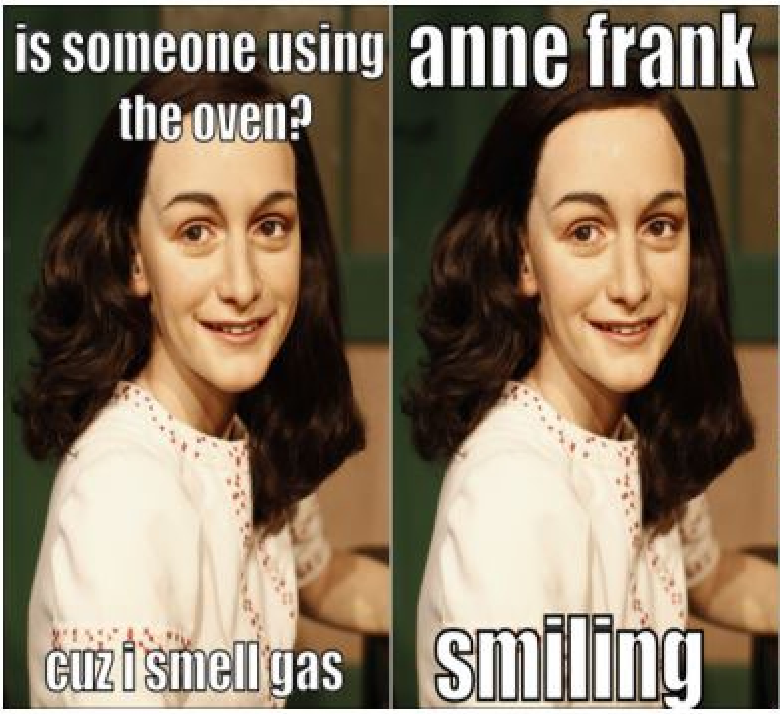


Fig 15 : Test Case 8

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

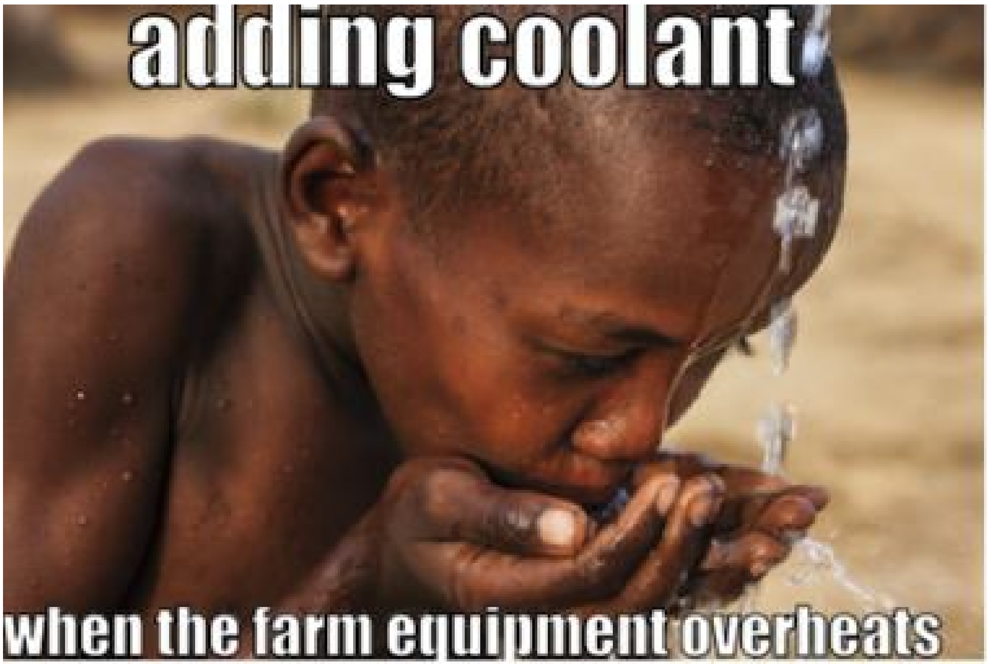


Fig 16 : Test Case 9

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

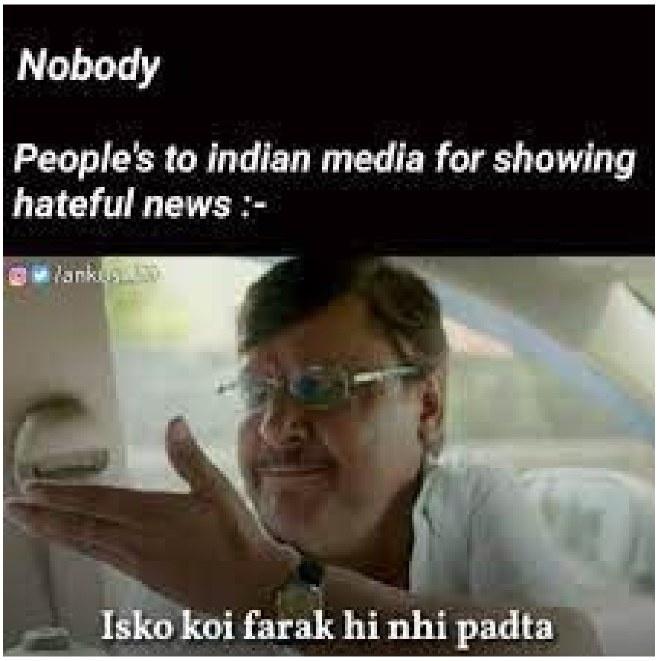

Nobody
People's to indian media for showing hateful news :-
/ankus_
Isko koi farak hi nhi padta

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 conclusion of the project would be that the algorithms and techniques developed as part of the project are able to accurately and efficiently detect hate speech in multimodal social media posts, with a low rate of false positives and false negatives. These algorithms and techniques could be integrated into social media platforms to improve the online environment and promote inclusivity and diversity.

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

The proposed project has significant potential for future research and development. One future scope of this project is to enhance the model's performance by incorporating more complex deep learning architectures and algorithms. The model can be extended to incorporate more contextual information to detect hate speech more accurately, 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)

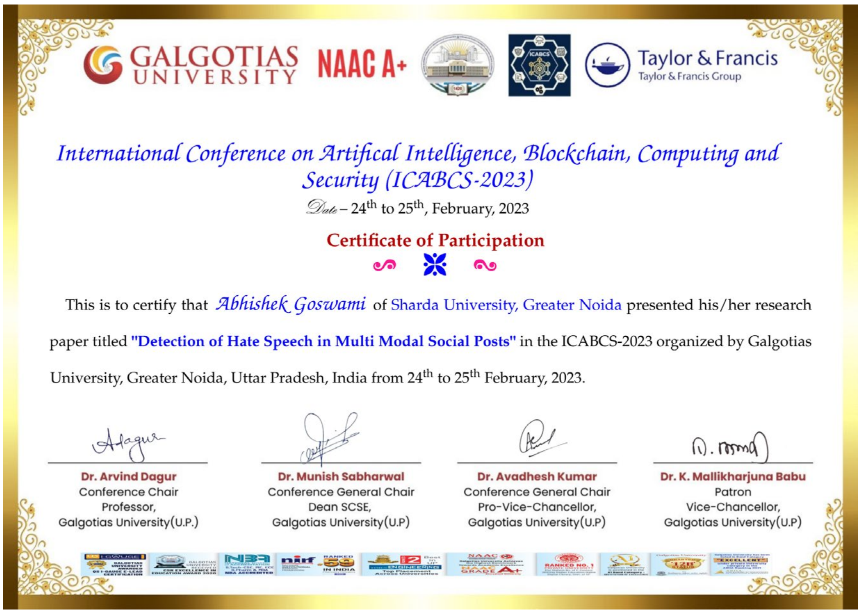
Paper Title: Detection of Hate Speech in Multimodal Social Posts

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)