**Offensive content detection in multimodal memes on social media**

**Aim**

The aim of this project is to develop a multimodal learning model that utilizes text and visual features to accurately detect offensive content in memes from the Multimodal Meme Dataset. By leveraging advanced machine learning and deep learning techniques, the project aims to contribute to the development of efficient methods for identifying and classifying offensive language in multimodal memes, ultimately fostering a safer and more respectful online environment on social media platforms.

**Course-Specific Learning Outcomes**

This project aligns with the following course-specific learning outcomes:

* Apply text and image processing techniques to extract meaningful information from multimodal data.
* Implement machine learning and deep learning models for multimodal data analysis and classification.
* Evaluate and interpret the performance of computational models in offensive content detection.
* Demonstrate proficiency in programming languages and libraries relevant to multimodal data analysis and deep learning.
* Critically assess the effectiveness and limitations of different computational approaches for offensive content detection.

**Project Background**

The emergence of memes as a popular form of communication on the internet has brought about new challenges in detecting offensive content. Memes, which combine textual and visual elements, often employ humor and sarcasm in implicit ways, making it difficult to identify offensive content by considering only the text or image. Consequently, there is a need to combine both modalities to effectively determine whether a meme is offensive or not (Suryawanshi et al., 2020).

The Multimodal Meme Dataset (MultiOFF) has been developed to address this need for offensive content identification in memes. The dataset consists of a diverse collection of multimodal memes, encompassing both textual and visual components. Research has shown that the joint modeling of multimodal features, including text and image, can improve the accuracy of offensive content classification (Hossain et al., 2022). By leveraging the Multimodal Meme Dataset, this project aims to explore the use of advanced machine learning and deep learning techniques to detect offensive content in memes accurately.

Previous studies have focused on the analysis of offensive memes and the development of multimodal models that combine visual and textual features. Weighted ensemble techniques have been employed to assign weights to different models, enhancing the overall performance in offensive content classification (Hossain et al., 2022). Furthermore, lightweight architectures have been proposed to generate object tags and image captions, facilitating the fusion of visual and textual cues for improved detection (Bhat et al., 2023).

The project acknowledges the challenges associated with analyzing memes due to their implicit expression of human emotions and the need for effective multimodal modeling to classify offensive content accurately. By building upon the research findings and methodologies in the field, the project aims to develop a robust multimodal learning model that combines textual and visual features to identify offensive memes effectively.

Through the utilization of the Multimodal Meme Dataset and leveraging advanced machine learning and deep learning techniques, the project seeks to contribute to the development of efficient methods for offensive content detection in memes. The ultimate goal is to enhance content moderation efforts on social media platforms, fostering a safer and more inclusive online environment.

**Objectives**

1. Data Collection and Preprocessing:

1. Collect and curate the Multimodal Meme Dataset, ensuring the availability of labeled samples for offensive and non-offensive memes.
2. Preprocess the dataset by cleaning, normalizing, and transforming the textual and visual components to create a suitable input for the multimodal learning model.

2. Feature Extraction and Representation:

1. Extract relevant features from the textual content of memes, such as word embeddings or sentiment analysis scores.
2. Extract visual features from the image components of memes, such as visual descriptors or deep visual representations.

3. Model Development:

1. Design and implement a multimodal learning model that combines textual and visual features to accurately classify offensive and non-offensive memes.
2. Explore and select appropriate algorithms and architectures for integrating the multimodal features effectively.

4. Training and Optimization:

1. Train the multimodal learning model using the labeled dataset, optimizing the model parameters to achieve high accuracy and robustness in offensive content detection.
2. Employ techniques such as transfer learning or fine-tuning to leverage pre-trained models and enhance the performance of the multimodal learning model.

5. Performance Evaluation:

1. Evaluate the performance of the developed multimodal learning model using appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score.
2. Compare the performance of the model with existing approaches and state-of-the-art methods in offensive content detection in memes.

6. Analysis and Interpretation:

1. Analyze the strengths and limitations of the multimodal learning model in detecting offensive content in memes.
2. Identify potential areas for improvement and propose future directions for enhancing the accuracy and efficiency of offensive content detection in multimodal memes.

**Evaluation Plan**

The project will be evaluated based on the following criteria:

* Successful implementation of a multimodal learning model for offensive content detection in memes.
* Accuracy and performance of the model in correctly classifying offensive and non-offensive memes.
* Comparison of the developed model with existing approaches, highlighting improvements or novel contributions.
* Analysis of the strengths, limitations, and potential future directions of the multimodal learning approach.

**Anticipated Problems**

1. Limited Availability of Labeled Data: The availability of a comprehensive and well-labeled dataset for offensive and non-offensive memes may be limited. Obtaining a diverse and sufficiently large dataset that covers various types of offensive content could be a challenge. This scarcity of labeled data may affect the training and evaluation of the multimodal learning model.
2. Complexity of Offensive Language: Offensive language in memes can be nuanced and context dependent. Capturing the intricacies and subtleties of offensive content within both textual and visual components can be challenging. Variations in language use, cultural references, and meme-specific humor may lead to difficulties in accurately identifying and classifying offensive content.
3. Computational Resource Constraints: Handling large-scale multimodal datasets and training complex multimodal learning models can be computationally demanding. Limited computing resources, including processing power and memory, may pose challenges in efficiently training and optimizing the multimodal learning model, leading to potential delays or resource limitations.
4. Interpretation of Visual Elements: Extracting meaningful visual features from meme images can be challenging due to variations in image quality, diverse visual content, and the interpretation of visual cues. Ensuring that the model can effectively capture and utilize relevant visual information to improve offensive content detection is a potential hurdle.
5. Ethical and Legal Considerations: Offensive content detection involves navigating sensitive topics and potentially infringing on users' privacy and freedom of expression. Ensuring that the project adheres to ethical guidelines, respects user privacy, and complies with legal regulations is of utmost importance. Striking a balance between effective content regulation and maintaining user rights can be a complex and critical challenge.
6. Generalization and Adaptation: The developed multimodal learning model may face difficulties in generalizing to new and unseen types of offensive content or adapting to evolving trends and language usage. Continual monitoring and updating of the model may be necessary to maintain its effectiveness in detecting offensive content over time.

**Required Resources**

To successfully carry out this project, the following resources are needed:

1. Multimodal Meme Dataset: A publicly available dataset consisting of labeled samples of offensive and non-offensive memes.
2. Machine Learning and Deep Learning Frameworks: Python libraries such as TensorFlow, PyTorch, Keras for model development and training.
3. Computing Resources: Sufficient computing resources with suitable processing power and memory to handle the training and evaluation of the models.
4. Research Materials: Relevant research papers, articles, and publications on multimodal data analysis, offensive content detection, and related topics.

**Project Timeline**

|  |  |  |
| --- | --- | --- |
| **Activity** | **Start Date** | **Estimated Duration** |
| **Project Planning and Dataset Collection** | 10th July 2023 | 1 week |
| **Data Preprocessing and Feature Extraction** | 18th July 2023 | 1-2 weeks |
| **Model Development and Training** | 30th July 2023 | ~ 3 weeks |
| **Model Optimization and Evaluation** | 15th Aug 2023 | 1-2 weeks |

**Ethics Number:** 57900

**References:**

Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020). Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (pp. 32-41). Marseille, France: European Language Resources Association (ELRA).

Hossain, E., Sharif, O., Hoque, M. M., Dewan, M. A. A., Siddique, N., & Hossain, M. A. (2022). Identification of Multilingual Offense and Troll from Social Media Memes Using Weighted Ensemble of Multimodal Features. Journal of King Saud University - Computer and Information Sciences, 34(9), 6605-6623. https://doi.org/10.1016/j.jksuci.2022.06.010

Bhat, A., Vashisht, V., Sahni, V. R., & Meena, S. (2023). Hate Speech Detection using Multimodal Meme Analysis. In 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 1137-1142). Salem, India: IEEE. https://doi.org/10.1109/ICAAIC56838.2023.10140393