MIMIC: Multimodal Islamophobic Meme Identification and Classification

S M Jishanul Islam[1*] **Sahid Hossain Mustakim**[1*] **Sadia Ahmmed**[1*]
Md. Faiyaz Abdullah Sayeedi[1*] **Swapnil Khandoker**[2*] **Syed Tasdid Azam Dhrubo**[3*]
Nahid Hossain[1]
[1]United International University [2]Johannes Kepler Universität, Linz [3]University of Alberta
{sislam201024,smustakim201274,sahmmed201146,msayeedi212049}@bscse.uiu.ac.bd
k12215556@students.jku.at
syedtasd@ualberta.ca
nahid@cse.uiu.ac.bd

Abstract

Anti-Muslim hate speech has emerged within memes, characterized by context-dependent and rhetorical messages using text and images that seemingly mimic humor but convey Islamophobic sentiments. This work presents a novel dataset and proposes a classifier based on the Vision-and-Language Transformer (ViLT) specifically tailored to identify anti-Muslim hate within memes by integrating both visual and textual representations. Our model leverages joint modal embeddings between meme images and incorporated text to capture nuanced Islamophobic narratives that are unique to meme culture, providing both high detection accuracy and interoperability.

1 Introduction

The widespread use of social media has transformed memes into a popular form of digital communication. While memes are often created for humor, they can serve as powerful vehicles to spread hate speech and reinforcing harmful stereotypes. The field of hate speech on social media platforms has become increasingly sophisticated through the use of memes—multimodal content that combines images and text to spread harmful narratives. While progress has been made in detecting general hate speech (Subramanian et al. [2023]), the specific challenge of identifying and countering anti-Muslim hate memes remains largely unaddressed. Recent advances in multimodal learning have demonstrated promising results in meme classification tasks (Bikram Shah et al. [2024]). However, these developments are hindered by a critical limitation: the absence of datasets focusing on anti-Muslim hate memes. Existing hate speech datasets (Hermida and Santos [2023]) either focus solely on text-based content or address broader categories of direct hate speech, failing to capture the covert form of hate with cultural nuances specific to anti-Muslim prejudice expressed through memes.

To address this gap, we present a novel dataset of anti-Muslim hate memes collected from various online platforms. Our research reveals distinct patterns in how anti-Muslim sentiment is propagated through memes, highlighting the importance of considering both cultural context and multimodal elements in hate speech detection systems. These insights not only advance our understanding of online Islamophobia but also provide practical implications for content moderation strategies. The code and dataset are open-sourced[2].

[*]Equal Contribution
[2]Code and Data: https://github.com/faiyazabdullah/MIMIC

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

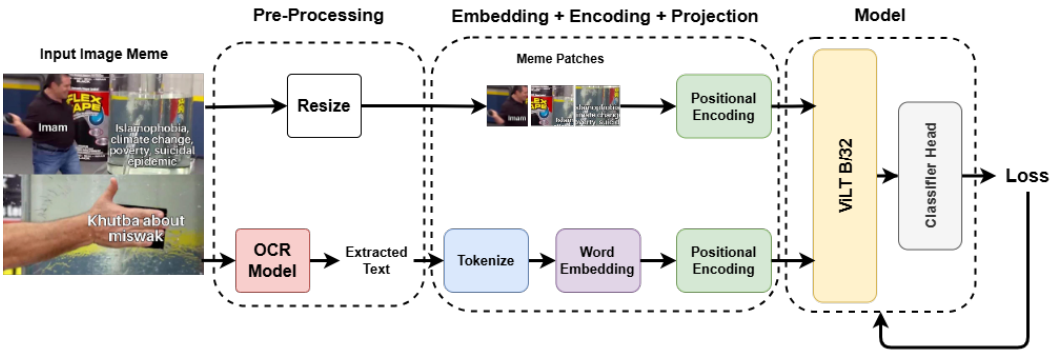


Figure 1: The end-to-end pipeline of our methodology

2 Related Works

Hate speech detection on social media has become a critical area of research, particularly with the rise of multimodal content that combines text and imagery to convey offensive or discriminatory messages (Arya et al. [2024]). While early studies on hate speech detection relied primarily on text-based datasets, advances in deep learning have allowed researchers to expand beyond text, employing multimodal approaches that incorporate visual elements and language models to improve accuracy and contextual understanding (Guo et al. [2023]). In response to the limitations of text-only approaches, recent research has focused on multimodal hate speech detection, particularly in the context of memes (Gandhi et al. [2024]). Memes present a unique challenge, as they often blend image, text, and context-dependent humor to convey subtle or overt hate messages. Visual language models (VLMs) and transformer-based architectures such as VisualBERT (Li et al. [2019]), ViLBERT (Lu et al. [2019]), and CLIP (Radford et al. [2021]) have shown promise in addressing these challenges. MemeCLIP (Shah et al. [2024]) was designed for multimodal hate detection, demonstrating that integrating visual and textual representations improves model performance in identifying hate memes. Although similar models achieve high accuracy in general hate meme classification, they lack specificity for certain types of hate, particularly Islamophobic content.

Recent developments in visual language models (ViLMs) and optical character recognition (OCR) techniques have enhanced multimodal hate speech detection capabilities. (Kim et al. [2021]) introduced ViLT, which is very effective for visual question answering and meme analysis tasks. Fine-grained OCR model by (Pettersson et al. [2024]), has improved text recognition in complex, low-quality images, enabling more accurate text extraction in memes with varying font styles, languages, and image quality.

3 Dataset

Our dataset consists of 953 memes gathered from Reddit, X, 9GAG, and Google Images, capturing diverse examples of potential anti-Muslim content. These memes were carefully curated to represent a range of content with potential anti-Muslim sentiment. We only take the samples which have text incorporated in the images, as we formally define them as "memes". To label the dataset, the annotators comprised of researchers with experience in hate speech detection, and conducted a manual review of each meme to classify it as hateful or non-hateful towards Muslims. Annotation used binary classification (0: non-hateful, 1: hateful), with 545 non-hateful and 408 hateful labels. To reduce bias, we established predefined rubrics based on language, symbols, and context, helping annotators make consistent decisions. Disagreements were addressed through discussions, with a consensus threshold requiring at least 80% annotator agreement to confirm a label. The distribution of labels and statistics are shown in Figure 2, with detailed dataset information presented in Table 5.

4 Methodology

This section presents our methodology for classifying memes into hateful and non-hateful categories. The end-to-end pipeline of our methodology is shown in Figure 1.

4.1 Data Pre-Processing

Initially, we extract the text in the memes using an optical character recognition (OCR) model by (Wei et al. [2024]), which extracts fine-grained OCR from images. To avoid dimension errors, we ensure that the texts are of the same length during batch-wise training by padding the texts to the maximum length of the extracted text from a meme, which is 40. We resize the images to 252×252 to ensure that the pixel values and attention mask generated by the models' preprocessor are consistent for every image data. Additionally, we applied the random rotation data augmentation technique to cover up for the small dataset size. We use this specifically as it does not distort the image or reverse the text in the memes. We record and compare its performances in Table 3.

4.2 Visual Language Model

To learn the representations between the image meme and the OCR-extracted caption, we utilize the Vision-and-Language Transformer (ViLT) base model proposed in (Kim et al. [2021]), primarily used for visual-question answering. We use ViLT because it is a transformer-based architecture (Vaswani [2017]) designed to handle vision-language tasks by simplifying the representation by directly integrating image and text modalities without relying on convolutional neural networks (CNNs) or region-based detectors. It directly projects the raw patches of the meme image and a linear embedding for the OCR-extracted meme text to prepare them for modality interaction. This means there is no preliminary step in our method to extract the image embeddings using a CNN backbone (Huang et al. [2020], He et al. [2016], Xie et al. [2017]). Avoiding this visual backbone reduces the computational overhead of our method. Positional encodings are added to the text and image embeddings. Next, the image and text embeddings are concatenated along the sequence dimension to form a unified input representation for the transformer, where self-attention mechanisms capture relationships within the meme image and the text in it. We record the performance of the ViLT model compared to alternatives in Section 5.

4.3 Classifier Head

The features learned from ViLT are pooled and passed to a classifier head with sequential multi-layer perceptions to refine and map representations to the output space. It starts with layer normalization, dropout (0.3), and a fully connected layer projecting to 768 dimensions, matching the ViLT output. Another normalization layer, followed by ReLU activation and dropout, introduces non-linearity and regularization. Finally, a fully connected layer outputs predictions, with a sigmoid activation bounding the output between 0 and 1 for hateful/non-hateful classification.

5 Experiments

This section presents the experiments conducted to test the performance of our model in hateful Islamic meme classification. We describe the experimental setup and record the median performance of our model with two dataset split settings.

5.1 Experimental Setup

The experimental setup outlines the hyperparameter and device configurations, and the evaluation metrics used to validate the effectiveness of our proposed approach. The experiments are carried out in a Kaggle environment with an NVIDIA P100 GPU with 16 GB memory. The model is trained and evaluated with two different independent techniques: splitting the dataset into train:validation:test set, and conducting a k-fold cross-validation. The training was done for 10 epochs using the Adam (Kingma [2014]) optimizer with a learning rate of $1e^{-4}$with a batch size of 16 for training samples, and 2 for the validation and test samples. The loss function used is the binary cross-entropy loss. A batch size of 16 was selected to balance computational efficiency with model performance during training. The issue of overfitting was mitigated by implementing early stopping and regularization techniques during training. The execution time averaged 3 hours, underscoring the computational demands of multimodal analysis.

5.2 Evaluation Metrics

The model's performance is evaluated using the F_1 score. Moreover, the macro and micro average scores are also recorded. We select the F_1 score due to the class imbalance present in the dataset.

Table 1: Model analysis recorded from 3 visual-language models as the base model

Model	Loss	Precision	Recall	F1-micro	F1-macro	F1-weighted
VisualBert [10]	0.681	0.845	0.585	0.585	0.482	0.482
CLIP ViT B/32 [13]	0.682	0.882	0.574	0.574	0.496	0.496
ViLT B/32	**0.621**	**0.872**	**0.617**	**0.617**	**0.511**	**0.581**

The model's performance is further assessed using $Precision$, which measures the proportion of correctly identified positive instances among all instances predicted as positive. $Recall$ indicates the proportion of correctly identified positive instances out of all actual positive instances in the dataset.

5.3 Results

Table 1 shows that the ViLT performs better than alternative visual-language models such as CLIP (Radford et al. [2021]) and VisualBERT (Li et al. [2019]). The results shown in Table 2 summarize the model's performance on the test set, following a train-validation-test dataset split. The model achieves a median loss of 0.621. The $precision$ is relatively high, with a median score of 0.872, indicating that 87.2% of the positive predictions made by the model are correct. However, the $recall$ is recorded to have a median value of 0.617, exhibiting room for improvement. This suggests that the model correctly identifies 61.7% of the actual positive instances. The median weighted F_1-weighted score is 0.581, reflecting a moderate overall performance. While this score indicates reasonable performance, it also underscores the model's challenges in achieving perfect generalization. The training and validation curves in Figures 3 and 4 further illustrate signs of overfitting, as the model exhibits significantly better performance on the training data compared to the test set. This can be due to several issues, such as insufficient data and noise within the dataset.

To address potential overfitting and assess the model's generalization ability, we evaluate the model using the K-fold cross-validation technique. We begin by splitting the dataset into a 90:10 ratio, where 90% is divided into K-folds for training and evaluation, while the remaining 10% serves as a holdout set to test the model on unseen data. The results are shown in Table 4. Overall, K-fold cross-validation outperforms the traditional train-validation-test split approach. Specifically, the model achieves an F_1-weighted score of 0.716 for K=5 and 0.738 for K=10, both surpassing the highest performance recorded in the standard split. These results indicate that the model demonstrates an above-average generalization ability, effectively distinguishing between hateful and non-hateful memes. The loss curves for this K-Fold evaluation method are shown in Figures 5 and 6.

6 Discussion

The primary limitation of this study is the dataset size, which may limit the model's scope of learning. Expanding the dataset would enhance the model's ability to generalize across diverse contexts. Additionally, our study uses binary classification for labeling; however, adding categories such as misinformation, covert hate, and overt hate could improve the analysis's depth and accuracy. Incorporating additional modalities, such as text, audio, or video features from meme-based content, along with a combined analysis of captions and content, could provide a more comprehensive understanding of Islamophobic content.

7 Conclusion

In conclusion, this study presents a targeted approach for detecting anti-Muslim hate speech in memes using our own custom dataset and the vision-language transformer (ViLT) model. The model achieved a strong 0.738 F_1-weighted score via 10-fold cross-validation, demonstrating effective generalization, though a standard split yielded a moderate 0.581 F_1-weighted score due to overfitting. This highlights the challenges posed by subtle and complex visual hate content. Future work should expand the dataset and explore additional modalities to enhance capabilities, advancing detection strategies for more robust content moderation in digital platforms.

References

[1] Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M Ghazal. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEE Access*, 2024.

[2] Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. Memeclip: Leveraging clip representations for multimodal meme classification. *arXiv e-prints*, pages arXiv–2409, 2024.

[3] Ankita Gandhi, Param Ahir, Kinjal Adhvaryu, Pooja Shah, Ritika Lohiya, Erik Cambria, Soujanya Poria, and Amir Hussain. Hate speech detection: A comprehensive review of recent works. *Expert Systems*, page e13562, 2024.

[4] Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573. IEEE, 2023.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Paulo Cezar de Q Hermida and Eulanda M dos Santos. Detecting hate speech in memes: a review. *Artificial Intelligence Review*, 56(11):12833–12851, 2023.

[7] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[8] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

[9] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[11] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[12] Tobias Pettersson, Maria Riveiro, and Tuwe Löfström. Multimodal fine-grained grocery product recognition using image and ocr text. *Machine Vision and Applications*, 35(4):79, 2024.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[14] Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. Memeclip: Leveraging clip representations for multimodal meme classification. *arXiv preprint arXiv:2409.14703*, 2024.

[15] Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G Deepalakshmi, Jaehyuk Cho, and G Manikandan. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, 80:110–121, 2023.

[16] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[17] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024.

[18] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

A Appendix / supplemental material

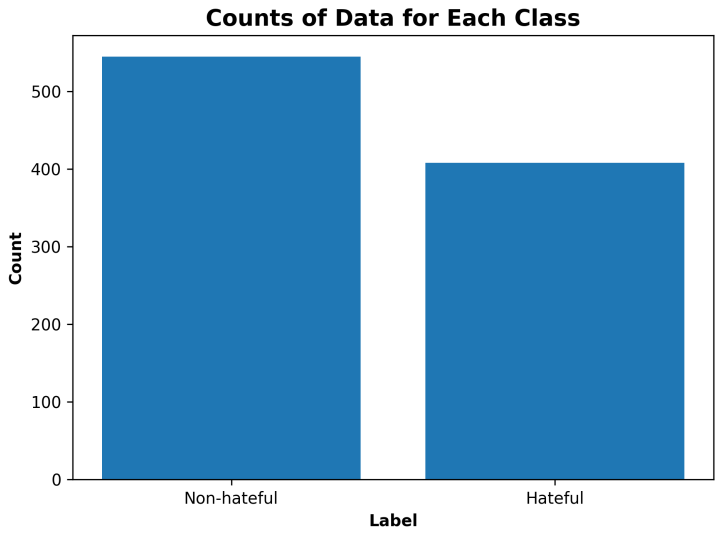


Figure 2: The distribution analysis of the classes (hateful: 1 and non-hateful: 0) in the dataset

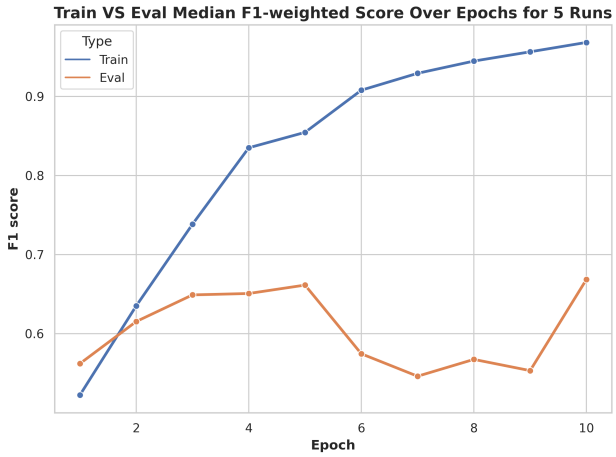


Figure 3: Median F1-weighted score curve

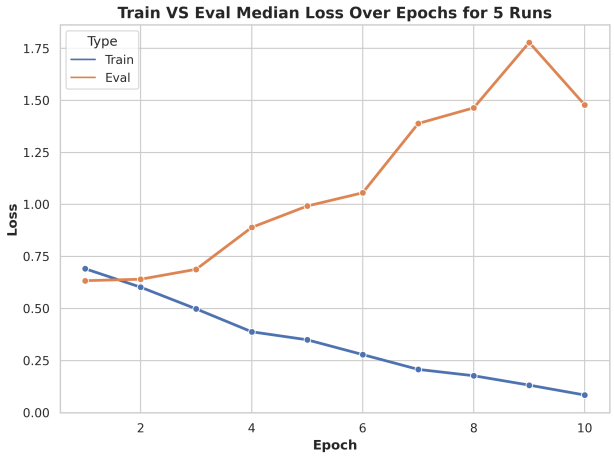


Figure 4: Median loss curve

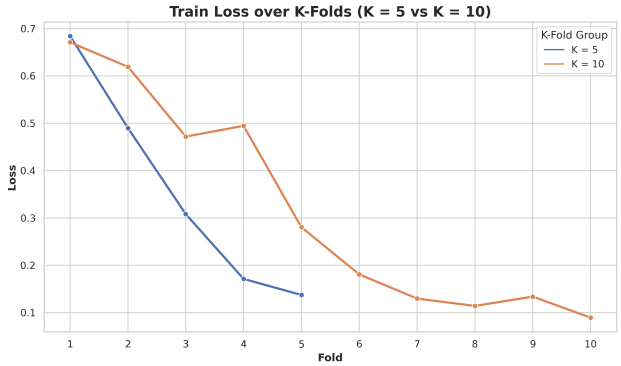


Figure 5: K-Fold train loss curves

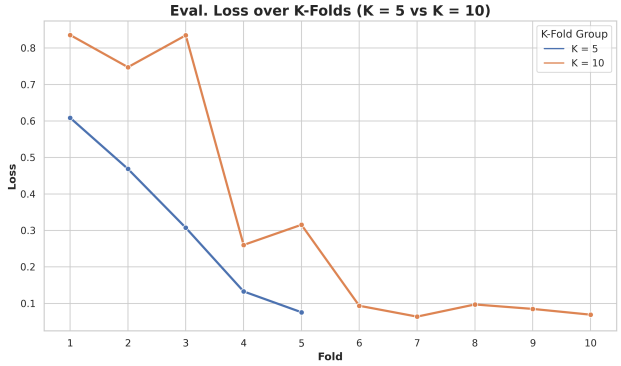


Figure 6: K-Fold validation loss curves

Table 5: Overview of Selected Dataset Samples

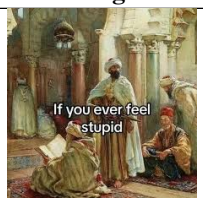
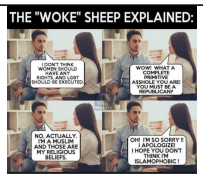
Image	Text	Label
	If you ever feel stupid	1
	THE "WOKE" SHEEP EXPLAINED: I DON'T THINK WOMEN SHOULD HAVE ANY RIGHTS, AND LGBTQ SHOULD BE EXECUTED. WOW! WHAT A COMPLETE PRIMITIVE A$$HOLE YOU ARE! YOU MUST BE A REPUBLICAN? NO, ACTUALLY I'M A MUSLIM AND THOSE ARE RELIGIOUS BELIEFS. OH! I'M SO SORRY!! I APOLOGIZE! I HOPE YOU DON'T THINK I'M ISLAMOPHOBIC!	1
	HEY ISLAMOPHOBES ISLAM IS THE RELIGION OF PEACE	1
	Me before Going through hardtimes Me after "Indeed hardships come with ease", Surah Ad-Duha. And "Indeed Allah is with the patient	0
	Sahabah ask Muhammad what they will get in heaven. Muhammad:	1

Table 2: Scores from the test set recorded from 5 runs using the train:validation:test split

Run No.	Loss	Precision	Recall	F1-micro	F1-macro	F1-weighted
1	0.625	0.878	0.691	0.691	0.603	0.645
2	0.623	0.872	0.574	0.574	0.489	0.489
3	0.598	0.846	0.617	0.617	0.511	0.574
4	0.621	0.910	0.606	0.606	0.511	0.581
5	0.619	0.867	0.649	0.649	0.574	0.631
Median	0.621	0.872	0.617	0.617	0.511	0.581

Table 3: Median Scores from the test set recorded for augmentation vs non-augmentation for the baseline ViLT model via train:validation:test split

Technique	Loss	Precision	Recall	F1-micro	F1-macro	F1-weighted
Non-augmentation	0.621	0.872	0.617	0.617	0.511	0.581
Augmentation	0.543	0.941	0.702	0.702	0.645	0.709

Table 4: Scores from the holdout test set recorded from k-fold cross validation

K	Loss	Precision	Recall	F1-micro	F1-macro	F1-weighted
5	0.698	0.909	0.723	0.723	0.666	0.716
10	0.691	0.899	0.755	0.755	0.695	0.738

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction accurately reflect the paper's contributions by detailing the development of an anti-Muslim hate meme dataset and a ViLT model for multimodal hate detection, with claims that align with the results and scope discussed throughout the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discussed issues such as the small dataset size, which may impact generalizability, and the use of a single vision-language model, suggesting that incorporating additional models could enhance analysis.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not present any theoretical results, it focuses on empirical findings from a dataset and model implementation without involving theoretical assumptions or formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper includes detailed descriptions of the experimental setup, dataset, model architecture, and evaluation metrics, enabling reproducibility to support its main claims.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper will provide open access to both the dataset and code via a GitHub repository, allowing others to replicate the experimental results with adequate instructions. The links will be added upon acceptance to adhere to the anonymity guidelines of NeurIPS, as the commits have been made using the authors' verified accounts.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, the paper provides the training and test details. The setup includes information on Kaggle environments, data splits, training epochs, optimizer(Adam Optimizer with 1e-4 learning rate), and the binary cross entropy binary loss.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars or statistical significance tests were not reported, as the study primarily focused on model performance metrics without an in-depth statistical analysis of experimental variability.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, the paper includes information on the compute resources required for reproducibility. It includes the use of a Kaggle environment with an NVIDIA P100 GPU(16Gb Memory) which is consistent across experiments. It took an average of 3 hours to train the model on this GPU with our experimental setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows the NeurIPS Code of Ethics, addressing ethical concerns in data use and model application, with attention to responsible hate speech detection and mitigation.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Yes, the paper discusses potential positive societal impacts. It emphasizes the positive impact of enhancing the detection of anti-Muslim hate speech within memes, which could foster safer and more inclusive online spaces.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The paper outlines safeguards for the dataset release, including filtering for harmful or explicit content and implementing guidelines for responsible use. Additionally, access to the data is restricted to prevent misuse, with clear terms on ethical usage.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All external assets used, including code, datasets, and models, are properly credited with citations to the original sources. Each asset's license type (e.g., MIT, CC-BY

4.0) and terms of use were respected and explicitly mentioned in the paper. For scraped data, we adhered to each website's terms of service, ensuring compliance with copyright and usage guidelines.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: New assets, including the dataset of anti-Muslim hate memes, are documented with details on data collection, annotation, and intended use.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human subjects; the dataset was collected from existing online sources.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Since no new human subjects were involved in the dataset collection, Institutional Review Board (IRB) approval was not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.