Master Thesis

From Words to Images: Exploring the Crucial Role of Textual Modality in Multimodal Hate Speech Detection

Hisham Alkaed

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Ilia Markov
2^{nd} reader: Antske Fokkens

Submitted: June 30, 2023

Abstract

This research project aims to investigate the role of textual modality in the detection of hate speech in multimodal memes. Previous studies have demonstrated that unimodal BERT performs well in a variety of tasks, including those in multimodal environments. This research seeks to understand the reasons for this high performance through error analysis. Moreover, the study explores the potential of boosting performance by combining BERT with other approaches through late fusion and ensemble strategies. To create models, textual and visual modalities were used as sources of features. These models were then used in late fusion and ensemble techniques. Analysis of errors indicated that BERT is a powerful unimodal Large Language Model that accurately detects hateful content while exhibiting some bias. However, the late fusion method, which leveraged both textual and visual cues by combining BERT with ResNet50, resulted in an increased false negative rate and decreased false positive rate. In contrast, the ensemble approach, which combined multiple models targeting different modalities, outperformed the late fusion method but had a higher false positive rate. The study identified two types of memes that pose challenges for the proposed approaches: those with benign content paired with an offensive image and those containing implicit or deceptive hateful content. Additionally, the dataset used for the Hateful Meme Challenge displayed multiple doubtful labels. The results of this study provide valuable insights into the strengths and limitations of different methods for detecting hate speech in multimodal memes, contributing to the field of multimodal hate speech detection.

Declaration of Authorship

I, Hisham Alkaed, declare that this thesis, titled *From Words to Images: Exploring the Crucial Role of Textual Modality in Multimodal Hate Speech Detection* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Master of Science degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 30/06/2023

Signed: 

Acknowledgments

I like to express my sincere gratitude to my supervisor, Ilia Markov, for his invaluable guidance, support and encouragement throughout this research project. His insightful work and feedback have been instrumental in shaping this study and have greatly contributed to its success.

I extend my heartfelt thanks to Antske Fokkens, for her substantial input, critical analysis and valuable feedback in the final stages of this work. Her constructive feedback has greatly enriched the quality of the error analysis performed in this study.

I genuinely appreciate the authors of the research papers upon which this study was built. Their pioneering work has been a great inspiration and has contributed significantly to the development of this research.

I also like to express my gratitude to all those who participated in the brainstorming sessions and provided essential input to improve the quality of this research.

Finally, I like to thank my family and friends for their continuous support, encouragement and understanding throughout my academic journey.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The pervasive growth of social media platforms and the widespread accessibility of internet communication have revolutionized the way people connect and exchange information. While this digital landscape offers numerous advantages such as increased transparency and information availability, it also harbours significant drawbacks that demand attention. Alongside concerns related to privacy breaches and online scams, there is a pressing issue that has gained prominence in recent years - the proliferation of hate speech.

Hate speech, defined as public discourse that expresses hatred or promotes violence towards individuals or groups based on attributes like race, religion, sex, or sexual orientation, has become a prominent challenge in online environments (MacAvaney et al., 2019). It manifests through offensive, toxic, and abusive language that targets specific individuals or communities. Detecting and addressing hate speech is a complex task, especially considering its implicit nature (ElSherief et al., 2021) and evolving tactics employed by perpetrators to evade detection, such as, intentionally misspelling words to evade detection by AI systems.[1]

In some jurisdictions, hate speech may be constitutionally protected, posing legal limitations on addressing its harmful effects (van Mill, 2021; Stone, 1994; Volokh, 2015). However, victims in other countries may seek redress through civil or criminal law (Mattei and Bussani, 2010). Consequently, the identification and prevention of hate speech necessitate interdisciplinary approaches to tackle this persistent issue effectively.

One mode of communication that necessitates specialized attention is the realm of memes. Memes are multimodal constructs that combine text and images to convey messages and provoke reactions. Identifying and understanding hateful content within memes presents unique challenges compared to text-only or image-only formats. Accurate detection requires a comprehensive analysis of both the image and text components, as well as an awareness of cultural references and contextual knowledge (Zhu, 2020).

The focus of this thesis is the detection of hateful speech in memes, acknowledging their prominent role in online discourse. By addressing this specific phenomenon, this research contributes to the development of effective hate speech detection methods that encompass multimodal content. The research utilizes three main methods: Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), Late Fusion, and Ensemble, to analyze the interplay between text and image modal-

[1] Anti-vaxxers during the COVID-19 pandemic using the term "vachscenes" or "wax seen" instead of "vaccines" (T and Mathew, 2022).

ities, enhance the accuracy of multimodal hate speech detection, and develop robust models for improved performance.

The integration of text and image modalities is crucial for a comprehensive understanding of memes. Relying solely on either text or image analysis can lead to erroneous classifications, resulting in false positives or false negatives (Kiela et al., 2020). To overcome this limitation, this research explores advanced techniques to merge models post-processing. By leveraging the complementary information present in both modalities, the models aim to capture the nuanced cues necessary to accurately identify and classify hateful content in memes.

This research encompasses the application of the BERT model, which has demonstrated exceptional capabilities in natural language processing tasks, to analyze textual components of memes (Devlin et al., 2019). The BERT model, with its pre-training and fine-tuning phases, provides a robust foundation for understanding and classifying textual content in memes. By leveraging the contextual information captured by BERT, the models aim to accurately identify hateful speech within memes. The aim here is to shed light on the good performance of BERT through error analysis.

Furthermore, this research explores the concept of late fusion and ensemble techniques to enhance the overall performance of the unimodal BERT model in the context of hate speech detection in memes. The used late fusion in this thesis combines the outputs of a pair of individual models trained on different modalities, allowing for a more comprehensive analysis and integration of multimodal cues. The ensemble, on the other hand, merges the output of multiple models each targeting a specific aspect of the text and image modalities. By combining the strengths of individual models, this approach aim to compensate for the limitations of the constituent models.

In conclusion, this thesis addresses the challenge of hate speech detection in multimodal memes. By integrating text and image modalities and exploring advanced techniques, the aim is to develop effective models for identifying hateful content and explore the role of textual modality in this context. The significance of this research lies in its potential to contribute to the creation of safer and more inclusive online spaces, promoting respectful and responsible online communication. By answering the research questions, to be introduced later on, the hope is to mitigate the harmful impact of hate speech and foster a more harmonious digital society.

Chapter 2

Related Work

Hate speech has become a prevalent issue in recent years and it has gained increased attention in both academic and public spheres. As evidence of this trend, there has been an exponential increase in the number of paper publications about hate speech since 1992 (Tontodimamma et al., 2021). Moreover, since 2010 especially with the increased usage of artificial intelligence (AI) in all life aspects, hate speech has been a popular research topic and found itself in the second phase of the Price Theory of Productivity on a given subject (Price, 1963). Meaning, this topic has gained more momentum since 2010 and scholars are focusing their attention on exploring ways to identify and detect hate speech online.

Detecting hate speech in text has been the primary focus of research in recent years. Previous studies have analyzed various linguistic attributes, such as word counts, sentence structure, and sentiment, and used traditional machine learning (ML) approaches to identify hate speech (Del Vigna et al., 2017; Gitari et al., 2015; Ali et al., 2022; Chatzakou et al., 2017; Chen et al., 2012; Davidson et al., 2017; Waseem and Hovy, 2016; Nobata et al., 2016). Those studies usually used Term Frequency Inverse Document Frequency (TF-IDF) scores and Bag-of-Words (BoW) vectors as feature representations. Some studies have utilized meta-information from users' profiles and online platforms where hate speech was posted, such as followers, likes and mentions, to enhance the detection accuracy (Papegnies et al., 2017; Singh et al., 2017). These features are then used as input for classifiers such as Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest.

Recent research has achieved near state-of-the-art performances in identifying hate speech in text using deep neural networks. For example, some studies have utilized Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models with word unigram and character bigram embeddings for hate speech detection (Mehdad and Tetreault, 2016; Gambäck and Sikdar, 2017). While Badjatiya et al. (2017) proposed a novel approach that combines LSTM models with Gradient-boosting Decision Trees. And others (Zhang et al., 2018) combined CNN with Gate Recurrent Unit (GRU) to classify hate speech using word embeddings as input. Moreover, new deep learning architectures have proposed combining multi-faceted text representations for the classification of hate speech, thus enhancing the accuracy of detection (Cao et al., 2020).

Though some studies have shown promising results with deep neural networks, it's worth noting that the conventional SVM has demonstrated near state-of-the-art results in recent studies (Markov et al., 2021). Additionally, it's important to distinguish be-

tween neural nets (CNN, RNN, etc.) and transformer models that have emerged as the state-of-the-art in natural language processing. Transformers employ a mechanism called self-attention that allows them to learn multidirectional dependencies between different words in a sentence or sequence of text (Vaswani et al., 2017). Unlike traditional RNNs that process sequences of data sequentially, transformers are able to process the entire input sequence in parallel, which makes them computationally efficient and more effective in capturing long-range dependencies in text. This has made transformers particularly well-suited for tasks such as language translation, text generation, and sentiment analysis.

Apart from textual content, some research has focused on detecting hateful content in images. Studies using CNN and deep learning methods to detect hateful content in images have emerged (Niam et al., 2018; Putra et al., 2022). However, these studies relied on recognizing any hate speech in the image through the existing text. Another study used state-of-the-art models from OpenAI to detect antisemitic/Islamophobic images (González-Pizarro and Zannettou, 2022). Additionally, there are studies that combine computer vision and natural language processing (NLP) to detect hateful content in Twitter messages (Perifanos and Goutsos, 2021). These studies utilize transfer learning and fine-tuning of Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and Residual Neural Networks (ResNet) (Koonce, 2021) to detect hateful content in tweets. They used images of how the tweets would look like in a web browser for the user as input.

Furthermore, a multimodal approach that combines text and image processing is crucial for accurately detecting hateful speech in online environments (Kiela et al., 2020). Much of the research has concentrated on text-based features, mainly due to the absence of annotated datasets in the vision or multimodal domain. Nonetheless, recent attempts have been made to rectify this limitation by introducing multimodal datasets. For example, some studies have tackled this issue through a multimodal dataset of tweets (Perifanos and Goutsos, 2021; Gomez et al., 2020). Additionally, researchers created a dataset of 246K posts from 4chan and /pol/ containing 420 antisemitic/Islamophobic phrases and 21K likely antisemitic/Islamophobic images (automatically annotated) (González-Pizarro and Zannettou, 2022). These resources can help in advancing research on multimodal hate speech detection. For further information about studies in the field of hate speech detection, the reader is advised to look into the survey studies by Fortuna and Nunes (2018); Schmidt and Wiegand (2017); Jahan and Oussalah (2021); Yin and Zubiaga (2021).

In this thesis, the Hateful Meme Challenge (HMC) Dataset (Kiela et al., 2020) is employed to investigate the textual modality in memes. The HMC is a noteworthy task in the domain of hate speech detection as it necessitates the incorporation of multimodal features. The dataset includes "benign confounders," which are alternate images or text that invert the meme's meaning from hateful to non-hateful and vice versa (Kiela et al., 2020). This feature is intended to challenge unimodal solutions and guarantee that any created model is resilient enough to handle complex scenarios, as discussed in depth later in Section 3.1. Interestingly, the WOAH shared task (Mathias et al., 2021) which builds on the HMC and is designed specifically for detecting multimodal hateful memes suggests that text-based models outperform multimodal systems. For example, a submission that relied solely on text achieved significant performance improvements, achieving an AUROC of 0.876 and outperforming the provided VisualBERT baseline

and some multimodal systems. Although these findings are promising, there is some concern that such models may capture unintended biases in the relatively small datasets available for the shared task. In this study, some light is shed on this phenomenon and the reasons behind this unforeseen performance.

While most recent NLP problems are being solved using neural networks and advanced transformers, in some cases, a simpler ML model with the correct choice of features might outperform a neural network. For example, in the paper by Del Vigna et al. (2017), SVM performed better than LSTM in some scenarios. Hence, the current work will also serve as a comparison between a simple ML architecture (SVM) using hand-crafted useful features inspired by previous studies and the more advanced transformer model BERT.

In addition, prior research has emphasized the significant role of textual modality in detecting online hate speech across multiple modalities (Gomez et al., 2020). The significance of combining the modalities, such as text and images, in later stages has been emphasized to enhance the noise correction ability of the textual information, in contrast to the early fusion approach (Fersini et al., 2019). The latter approach combines the modalities' representations in the early stages before training a model and does not allow for such noise correction. Therefore, this study aims to explore the possibility of improving the performance of the unimodal BERT through post-processing fusion with other models that target other modalities. This strategy will supplement BERT with a combination of simpler learners, thereby investigating the model's capability of compensating for the limitations of its constituent models. A comprehensive explanation of adopted strategies is provided in Sections 3.2.4 and 3.2.5.

2.1 Research question

This section outlines the research questions that will be addressed in this thesis. The primary research question concerns the role of textual modality in hateful meme detection. Specifically, the thesis aims to understand why unimodal large language models (LLM) exhibit high performance on multimodal datasets designed for hateful meme detection. In addition, the thesis seeks to investigate whether an ensemble or a late fusion strategy that supplements a unimodal LLM with models targeting other modalities can enhance hateful meme detection accuracy. Addressing these research questions will provide insights into the effectiveness of different modelling approaches and contribute to the advancement of research in multimodal hate speech detection.

Overall, the thesis's primary research question is:

What is the role of textual modality in hateful meme detection?

Which will be answered through answering the following sub-questions:

- *Why do unimodal Large Language Models demonstrate high performance despite the dataset being designed for multimodal approaches?*

  Unimodal LLMs have shown remarkable performance in various natural language processing tasks. However, it is intriguing to understand why these models perform well even on datasets that are explicitly designed for multimodal analysis, such as the HMC dataset. Exploring this question will help uncover the inherent capabilities of unimodal LLMs and shed light on their effectiveness in capturing textual cues in hateful memes. To do this, the state-of-the-art LLM commonly

used for detecting hate speech embedded in the textual modality BERT will be employed, and its performance will be analyzed through an in-depth error analysis.

- *Is it possible to further improve the results of the unimodal BERT by supplementing it with models targeting other modalities through an ensemble or late fusion strategy?*

  Although unimodal LLMs perform well, there may still be room for improvement. By incorporating models that target other modalities, such as images, it might be possible to enhance the overall performance of the unimodal BERT. The ensemble strategy aims to leverage the strengths of different modalities and compensate for the limitations of individual models. While the late fusion strategy has proved its effectiveness compared to the early fusion strategy in previous works. To achieve this, multiple models, including an SVM targeting a specific aspect of the text modality and a BoW baseline, will be implemented. Those models will be incorporated into the late fusion and ensemble methods. Investigating this will provide insights into the effectiveness of combining multiple models to achieve more accurate and robust hateful meme detection.

Addressing these sub-questions will provide a comprehensive understanding of the role of textual modality in hateful meme detection. The findings will contribute to the advancement of research in hate speech detection and shed light on the strengths and limitations of different modelling approaches.

Chapter 3

Methodology

3.1 Dataset Description

The dataset used in this thesis is known as the Hateful Memes Challenge Dataset [1] (HMCD). The process of constructing and annotating the HMCD is described and explained in detail in Kiela et al. (2020). The HMCD was designed specifically to test models that are of multimodal nature. In other words, Kiela et al. (2020) introduced specific examples that would make unimodal models suffer and end up low on the scoring board. The dataset is currently freely available to be used for research on Kaggle[2]. The following section summarizes the process of constructing and annotating the dataset, along with some information about the resulting splits and the baseline models provided by the authors.

[1] https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/
[2] https://www.kaggle.com/datasets/williamberrios/hateful-memes

3.1.1 Construction

The construction phase was initiated by utilizing a dataset consisting of more than 1 million images. These images were collected through web scraping and manual installation of memes from recognized platforms from within the United States, as a trial to a obtain a representative sample of images from real-world contexts. This dataset was filtered to include only unique memes which resulted in 162k memes. Then the annotators were asked to discard memes with non-English text and violating content. A violating content was defined as a meme containing any of the following:

> "self injury or suicidal content, child exploitation or nudity, calls to violence, adult sexual content or nudity, invitation to acts of terrorism and human trafficking" (Kiela et al., 2020)

Figure 3.1: An example of an image reconstructed using Getty images

Then all memes containing slurs were discarded as well since those are unimodal in nature. The filtering till now resulted in 46k memes. To reduce visual bias and obtain appropriate licences for the underlying content the memes were reconstructed from scratch using a custom-built tool from Getty Images[3] through a partnership. The reconstruction phase included finding an alternative picture to the meme at hand that "preserves the meaning and intent of the original meme" (Kiela et al., 2020). This is illustrated in image 3.1 where both images have the exact same text but different images, however, the images do not change the meaning or the intent of the meme.

3.1.2 Annotation

Related to the annotation process, third-party annotators, trained to recognize hate speech according to the paper's definition, spent an average of 27 minutes per final meme in the dataset (Kiela et al., 2020).

The definition of hate speech used in this thesis is

> "A direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease. We define attack as violent or dehumanizing (comparing people to non-human things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking hate crime is also considered hate speech." (Kiela et al., 2021, 2020)

[3] `https://www.gettyimages.nl/`

Each meme was given 5 scores between 1 and 3 by different annotators. 1 being hateful; 2 not sure; and 3 is not-hateful. The usage of the ranks rather than binary annotation made it easier to discard memes with extreme disagreement. Then out of those ratings/rankings binary labels were obtained by an expert.

3.1.3 Benign Confounders

The final step of creating the dataset was by introducing "benign confounders" which are alternative images or text that flip the label of a meme from hateful to non-hateful. Doing this reduced biases in the data where a specific word such as "black" is associated with hate speech. In addition, it made it more difficult for models of unimodal nature to perform well since depending solely on one aspect of the memes (text or images) would result in poor performance. Also, memes that no benign confounds were found for were considered unimodal hate. The hatefulness in the unimodal hate memes was mostly textual-based.

3.1.4 Splits

After collecting, reconstructing, and annotating the memes and adding the benign confounders, a total of 12540 memes were obtained. 8.5k memes serve as a training set, 1.5K dev and test (seen), and the rest serves as the final test set (unseen). A dev and test set were created from 5% and 10% of the training data (seen dev and test). Those sets are fully balanced and contain five different types of memes: multimodal hate (40%), unimodal hate (10%), benign image and text confounders (20% and 20%) and random non-hateful examples (10%) as can be noted in Table 3.1. While the different types of memes are mentioned in (Kiela et al., 2020), they are not available in the dataset. I.e., the dataset only contains binary labels. The other dev and test data were created to decide the winner of this challenge. Those unseen splits will serve as the final test data in this thesis and will be used for the final evaluation. Information about the size of each split is provided in Table 3.1. Note that the unseen splits are not as balanced as the seen splits. Meaning, accuracy could be used as informative performance metric in the seen splits but not in the unseen ones. A better metric would be to use the F1 score or as advised by the creators of the challenge the Area Under the Receiver Operating Characteristic (AUROC).

	Total	Not-hate	Hate	MM Hate	UM Hate	Img Conf	Txt Conf	Rand Benign
Train	8500	5481	3019	1100	1919	1530	1530	2421
Dev seen	500	253	247	200	47	100	100	53
Dev unseen	540	340	200	200	0	170	170	0
Test seen	1000	510	490	380	110	190	190	130
Test unseen	2000	1250	750	750	0	625	625	0

Table 3.1: Hateful Meme Challenge Dataset Splits

The dataset is divided into five subsets as shown in Table 3.1: Train, Dev seen, Dev unseen, Test seen, and Test unseen. The train contains 8500 memes, out of which 5481 are not hateful and 3019 are hateful. The hateful memes are further divided into multimodal hate (MM Hate) (1100) and unimodal hate (UM Hate) (1919). The table also includes the number of confounding memes (Img Conf and Txt Conf) and random benign memes (Rand Benign). The same information is provided for all splits. It is

worth noting that using the aforementioned unimodal text BERT it is expected to be able to correctly classify all UM hate memes that are text dependent. Nonetheless, it is observable that the unseen splits do not contain any UM Hate memes. Meaning, getting a reasonably good performance in those splits should warrant further investigation.

3.1.5 Baselines

Various baseline models have been provided by the organizer of the HMC (Kiela et al., 2020), which comprise both text and image features. These models entail both unimodal and multimodal systems that rely on different techniques to train and evaluate their accuracy and AUROC performance. Based on the outcomes, the multimodal models surpass all unimodal models. In particular, the most effective models leverage BERT-based architectures to combine both image and text features. These models include ViLBERT (Lu et al., 2019), pre-trained on the Conceptual Captions dataset (Sharma et al., 2018), and VisualBERT (Li et al., 2019), pre-trained on the Common Objects in Context dataset (Lin et al., 2014). Therefore, it is evident that processing both image and text signals is vital to achieve a comprehensive understanding of the content in this dataset.

The ViLBERT model (Lu et al., 2019) is a two-stream multimodal model that extends the BERT architecture by incorporating transformers for both the language and vision domains. These transformers interact through co-attentional transformer layers to capture joint representations of images and text. In contrast, the VisualBERT architecture (Li et al., 2019) is a single-stream model that integrates vision and language inputs and applies self-attention within a transformer layer.

To tackle the challenge of detecting hateful content in multimodal data, researchers have employed two primary methods, namely early and late fusion (Kiela et al., 2021). Both methods require multimodal feature extraction. Early fusion involves extracting independent unimodal features for different modalities and fusing them at the beginning of the classification process. In contrast, late fusion involves creating separate pipelines for image and text and combining the results later in the process. In both methods, the fusion step is critical and can dramatically impact the classifier's performance. In the early fusion, the challenging task is combining the representation of each modality in such a way no or little information gets lost. While the late fusion's performance depends on the manner in which the decisions of the constituent models are merged. Various approaches have been examined in previous research, such as taking the mean probability of each unimodal classifier (Kumar and Nandakumar, 2022), rule-based methods (Chen et al., 2016; Franco et al., 2020; Imran and Raman, 2020; Khaire et al., 2018) or employing another classifier that takes the outputs of the unimodal models as input (Boulahia et al., 2021).

It is hypothesized that late fusion attempts to emulate the human approach in assessing whether the meme is hateful or not (Mercier and Cappe, 2020; Scheliga et al., 2023; Engel et al., 2012). For example, when perceiving a multimedia presentation, if there is a lot of text on the slides, it becomes hard to focus on the auditory information provided by the presenter. This aligns with the fact that in the human body, visual information is processed by the visual system, auditory information is processed by the auditory system, and so on. After initial separate processing, humans integrate information within each modality to form a coherent understanding. For instance, in speech perception, humans combine information from the auditory system, such as speech sounds, with linguistic knowledge stored in the brain to comprehend spoken

words. Therefore this task could be approached by creating different pipelines that each perceives and processes each modality separately, and then combine the knowledge obtained from the pipelines into one. Consequently, the task at hand is considerably complex and interesting, particularly when attempting to detect hateful memes with benign confounders.

The top three submissions that took the first three places on the leaderboard of the challenge are:

- Ron Zhu's solution (Zhu, 2020) won first place, using an ensemble of VL-BERT (Su et al., 2019), UNITER-ITM (Chen et al., 2019), VILLA-ITM (Gan et al., 2020) and ERNIE-Vil (Yu et al., 2020) models. These models were provided with entity, race and gender - that were extracted using transformers- in addition to text and image inputs.

- Niklas Muennighoff's solution (Muennighoff, 2020) won second place, using a uniform framework for vision and language models with specific enhancements such as masked pre-training, visual token type, and Stochastic Weight Averaging (Izmailov et al., 2018). Models used include ERNIE-Vil (Yu et al., 2020), UNITER (Chen et al., 2019), OSCAR (Li et al., 2020) and VisualBERT (Li et al., 2019), ensembled in a loop for the final score.

- Team HateDetectron's solution (Velioglu and Rose, 2020) won third place, employing a simpler approach with a single model - VisualBERT (Li et al., 2019). The model was pre-trained on Conceptual Captions (Sharma et al., 2018) and fine-tuned on an aggregated dataset containing Memotion and Hateful Memes datasets (Sharma et al., 2020). Hyper-parameter tuning resulted in an ensemble of 27 models using the majority voting technique for classification.

In this thesis, the focus is on the role of textual modality in such a multimodal scenario. To do this post-processing fusion methods will be used and the image classifier will be frozen throughout the experiments. More specifically a method similar to the one proposed by (Boulahia et al., 2021) will be incorporated. This is explained in depth along with the experimental setup and the design choices in the next section.

3.2 Experimental Setup

In this section, the experimental setup including design choices, motivation and all details needed to replicate the performed experiment are discussed. First, the performed preprocessing steps are explained. Followed by the engineered features, the established baselines and the subsequent models that were created to include in the ensemble and the late fusion. The code of this thesis is available on GitHub [4].

3.2.1 Preprocessing

In this research, the objective is to distinguish between hateful and non-hateful memes using the HMCD (Hateful Memes Challenge Dataset). As discussed in Section 3.1, the HMCD was created through a rigorous annotation process and comprises high-quality data. The original dataset includes an entry ID (corresponding to the image ID), image

[4] `https://github.com/HishamAlkaed/Master_Thesis`

path, binary label indicating whether the meme is hateful or non-hateful and the meme text. For instance, a sample row of the original dataset is presented below.

> {"id": "46971", "img": "img/46971.png", "label": 1, "text": "bravery at its finest"}

Initially, the dataset was in JSON format, which was converted to CSV format for ease of processing. Subsequently, a variety of basic preprocessing steps were applied, such as tokenization, lemmatization and Universal Part-of-Speech (UPOS) tagging, using the Spacy en_core_web_sm model (spaCy, 2017). Given that the textual data is represented as a single piece of text in each entry, the tokenized and lemmatized sentences are included as additional two columns. Further, the UPOS was represented in the same manner as the other two features, with the UPOS tags separated by a space.

3.2.2 Features Engineering

In addition to utilizing BERT as the primary approach for text classification in memes, this research explores alternative methods to classify the text embedded within memes. To provide a comprehensive comparison and assess the effectiveness of different approaches, an SVM model, as proposed by Markov et al. (2021), was employed. The SVM model was replicated and trained on the dataset, allowing for a comparative evaluation of its performance alongside BERT in different scenarios.

In this context, a novel set of features was developed. Markov et al. (2021) demonstrate the effectiveness of integrating stylometric and emotion-based features in detecting hate speech, achieving near state-of-the-art performance. As a result, three new features are introduced to the dataset, leveraging stylometric and emotion-based characteristics.

Additionally, in an attempt to enhance the performance of the stylometric and emotion-based features, a set of features utilizing high-end transformers is added. These features aim to provide complementary information and improve the classification of hate speech in memes. By incorporating these alternative methods and introducing new features, this research aims to provide a comprehensive analysis and comparison of different approaches for hate speech detection in multimodal memes.

The first feature, denoted as **emotion_association**, leverages the NRC emotion lexicon (Mohammad and Turney, 2013), which is a comprehensive compilation of words and their associated emotions. This feature identifies whether each word in a sentence appears in the NRC emotion lexicon and, if so, adds the associated emotion to this feature. If not, it remains empty.

The second feature, referred to as **count**, is derived from the emotion_association feature and represents the count of emotion words present in the text. This feature has manifested its usefulness in the same research too (Markov et al., 2021).

The third feature, designated as **pos_fw_emo**, represents the text through UPOS tags, function words, and emotion words. This feature utilizes the aforementioned NRC lexicon and a list of function words. Function words are those whose UPOS is one of the following:

```
['ADP', 'AUX', 'CCONJ', 'DET', 'NUM', 'PART', 'PRON', 'SCONJ']
```

pos_fw_emo retains words that have an emotional association and functional roles in the sentence and replaces the rest with their corresponding UPOS. This feature is utilized to construct n-grams (n=1-3). The three previously explained features will be referred to as ***stylometric & emotion-based features***.

Apart from the features inspired by Markov et al. (2021), two additional features were integrated into the dataset. The first feature, named **sentiment scores**, was obtained by utilizing the *sieBERT/sentiment-roBERTa-large-english* model from HuggingFace[5]. The sentiment score is a measure of how positive or negative the text is, with values closer to 1 indicating a more positive sentiment and values closer to 0 indicating a more negative sentiment. The rationale behind using sentiment scores in this hateful meme detection task is to simply capture the sentiment of the text.

The second feature, denoted as **intent detection**, was obtained by utilizing the *mrm8488/t5-base-finetuned-e2m-intent* model from HuggingFace[6]. This feature extracts the intent of the text, such as the sentence *enjoying a day at the beach* having the intent *to relax*. Typically, intent detection is employed in chatbots to identify user intents (Hasani et al., 2022; Fernández-Martínez et al., 2021; Abbet et al., 2018). However, in cases where there is irony or humour in the speech, such a model should also be able to identify it. This feature is utilized later on to construct n-grams. Since the intent detection and sentiment score features have been retrieved using transformers those will be referred to as ***transformer features***.

The enriched dataset comprises all of the aforementioned features, along with the original text, image, and label data. This dataset was utilized to train and evaluate various textual models in our experimental setup. Two rows of the enriched dataset are provided as an example below.

id	25489	53976
img	img/25489.png	img/53976.png
label	1	0
text	brother... a day without a blast is a day wasted	black power comes with a lot of responsibility
tokens	brother ... a day without a blast is a day wasted	black power comes with a lot of responsibility
lemmas	brother ... a day without a blast be a day waste	black power come with a lot of responsibility
upos	NOUN PUNCT DET NOUN ADP DET NOUN AUX DET NOUN VERB	ADJ NOUN VERB ADP DET NOUN ADP NOUN
emotion_associations	positive trust anger fear negative surprise disgust negative	negative sadness
count	3	1
pos_fw_emo	brother PUNCT a NOUN without a blast be a NOUN waste	black NOUN VERB with a NOUN of NOUN
intent	to be lazy	to be a leader
sentiment_score	0.993244	0.992763

Table 3.2: Enriched HMCD

3.2.3 Models

After preprocessing the data and engineering some useful features both from literature and state-of-the-art transformers, it is time to put those features to work and create models that make use of them. Since the focus of this thesis is on the textual modality, but the task requires the use of both visual and textual cues, an image classifier will be used and frozen throughout all experiments.

Image models The deep residual nets (ResNet50) introduced by He et al. (2016) were used in this thesis to retrieve image vectors from the images. This method involves the use of neural networks with a large number of layers to recognize complex patterns in

[5] `https://huggingface.co/sieBERT/sentiment-roBERTa-large-english`
[6] `https://huggingface.co/mrm8488/t5-base-finetuned-e2m-intent`

images. The idea behind deep residual learning is to create a network that can learn from the residual data that is left over after each layer of processing. This residual data is then used to improve the accuracy of the network's predictions.

There are various benefits of using deep ResNet50 (He et al., 2016). For example, it can help to reduce the number of training samples needed to achieve high accuracy levels. It can also improve the ability of neural networks to recognize complex patterns in images, such as those found in medical imaging or satellite imagery. Furthermore, deep residual learning can help to reduce the amount of time and computational resources required to train a neural network. As a result, it has become a popular technique in the field of computer vision and has been successfully applied to a range of image recognition tasks.

The effectiveness of this approach is demonstrated by the authors of He et al. (2016) with empirical evidence on the ImageNet (Russakovsky et al., 2014) and CIFAR-10 (Krizhevsky et al., 2009) datasets. The deep residual nets achieved first place in the ILSVRC & COCO 2015 competitions [7] for tasks such as ImageNet detection, localization, COCO detection, and segmentation.

The model is available for free on Hugging Face under the name *microsoft/resnet-50*[8]. As pointed out in the documentation, each image is processed by the "AutoImageProcessor" from the transformer library and then fed into the model. To extract image tensors, the last hidden state weights from the model's output are used, which resulted in 2048x7x7 image tensors.

These image tensors are then employed in 1) support vector machine (SVM) and 2) convolutional neural network (CNN). In the SVM model, a handful of combinations of parameters such as kernel type and C value were tested. The kernel types utilized were radial basis function (RBF) and linear, while the C values tested were 1 and 10. In the CNN technique, the image tensors were fed into a neural network with convolutional layers to extract relevant features.

The CNN used in this thesis consists of four convolutional layers and two fully connected layers. Each convolutional layer is followed by a Rectified Linear Unit (ReLU) activation layer. The first convolutional layer has 2048 input channels and 1024 output channels, while the second, third, and fourth convolutional layers have 1024, 512, and 256 input channels, respectively, and 512, 256, and 64 output channels, respectively. All convolutional layers have a kernel size of 3x3.

The two fully connected layers have 512 and 2 output features, respectively. Between these layers, there is a dropout layer with a dropout rate of 0.5 to prevent overfitting. Another ReLU activation layer is also applied after the dropout layer. This CNN architecture is designed to utilize relevant features from the image tensors and classify the images into two classes.

Those models have been trained on the training data and tested on the dev and test data (seen). The predicted labels by each one of the models have been saved as separate columns. As discussed earlier the unseen splits will be used as the final evaluation step in this thesis.

Text models During the development stage, various text models were implemented to efficiently identify hate speech. Initially, a Support Vector Machine (SVM) with Bag-of-Words (BoW) approach was utilized to establish a baseline for the text models.

[7] `https://image-net.org/challenges/LSVRC/2015/`
[8] `https://huggingface.co/microsoft/resnet-50`

The BoW model is a representation of text that ignores the order and context of words but focuses on their frequencies. In the BoW approach, each document is represented as a vector where each element corresponds to the count of a particular word in the document. The CountVectorizer class in scikit-learn[9] helps in creating such a representation by converting a collection of text documents into a matrix of token counts. We set the `analyzer='word'` and `ngram_range=(1, 1)`, such that the vectorizer will treat each word as a separate token and considers only individual words (unigrams).

Another basic approach was also implemented, an SVM with character n-grams was developed. The same CountVectorizer was used, but we set the `analyzer='char'` and `ngram_range=(1, 3)`, such that the vectorizer will consider individual characters as separate tokens and it will generate unigrams (single characters), bigrams (sequences of two characters), and trigrams (sequences of three characters) as features.

Additionally, an SVM using the stylometric & emotion-based and transformer features was created, incorporating the BoW as an extra feature. CountVecorizer has been used to vectorize features that are represented as strings. More specifically, word n-grams were created from the pos_fw_emo and intent features where n=1,2,3 and n=1,2,3,4, respectively. All other textual features were vectorized as unigrams. All SVM models were optimized through specific parameter configurations, ensuring convergence. The number of maximum iterations was limited to 1 million to ensure convergence, while the value of C was set to 10 through a grid search. Additionally, a random state of 456 was selected to ensure reproducibility.

In addition to the SVM models, in order to shed light on the unexpectedly high performance of unimodal textual systems for detecting hateful memes in a multimodal scenario (subquestion 1), the state-of-the-art language model, Bidirectional Encoder Representations from Transformers (BERT), was leveraged for the identification of hate speech in text. BERT has demonstrated remarkable results in various NLP tasks, as illustrated in Devlin et al. (2019) where the authors showcase the performance of BERT on 11 NLP tasks.

To incorporate BERT, the Hate-speech-CNERG/dehateBERT-mono-english model, developed by Aluru et al. (2020), described in "Deep Learning Models for Multilingual Hate Speech Detection" and available on Hugging Face[10], was used. This model is a fine-tuned version of the multilingual BERT for detecting hate speech content in the English language. The model achieved a validation score of 0.73 with a learning rate of 2e-5. The training code can be found online as well [11].

Apart from using the pre-trained Hate-speech-CNERG/dehateBERT-mono-english model, word embeddings were retrieved from the same model to train an SVM with a linear kernel. The embeddings were created by taking the weights of the last hidden state of the model which resulted in tensors of size 768. The sentence embeddings were created from the word embeddings by stacking them and padding them to the same size as the longest sentence. The previously explained two BERT models will be referred to as hateBERT for ease of reading.

Lastly, the BERT-base-cased[12] model was fine-tuned specifically for the task of hate speech detection in the HMCD to further explore its effectiveness. The fine-tuning was performed as follows: The input texts are tokenized using the pre-trained tokenizer of

[9] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html`

[10] `https://huggingface.co/Hate-speech-CNERG/dehateBERT-mono-english`

[11] `https://github.com/punyajoy/DE-LIMIT`

[12] `https://huggingface.co/BERT-base-cased`

`BERT-base-cased` and converted into input tensors. The input tensors must contain `input_ids`, `attention_masks` and naturally the `labels`. A TensorDataset and Dataloader were created for training the model. Both of which are from Torch library [13]. Those were utilized to make training in batches easier given the limited resources. A batch size of 32 was used. AdamW optimizer[14] and linear scheduler with a warmup [15] from Transformers were used. The model is then trained for 10 epochs, with the loss being calculated and the model weights being updated after each batch. We kept track of the average training loss per epoch. The fine-tuning code is available in the GitHub repository of this thesis.[16]

Each one of these models has been trained on the training data and tested by predicting the development and test data. The output of each model has then been saved in a separate column. An overview of the all models is provided in 3.2.

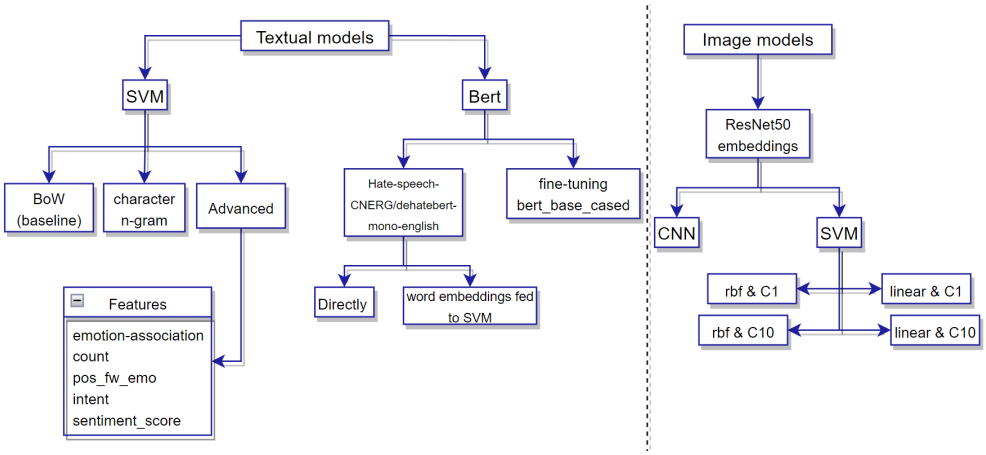


Figure 3.2: Experimental setup overview

3.2.4 Late fusion

In recent years, multiple approaches have been proposed to fuse modalities, including early fusion, intermediate fusion and late fusion. Early fusion involves integrating data from different modalities at a feature level, whereas intermediate fusion combines features from different modalities at an intermediate level. Late fusion, on the other hand, combines the decisions derived from models trained on different modalities.

In a study about multimodal action recognition by Boulahia et al. in 2021, which requires the integration of several image modalities, multiple novel methods for fusing modalities were proposed and compared. They concluded that the intermediate approach performed the best in their settings, followed by late and then early fusion. While their experiment focused on fusing different modalities related to images such as RGB data, skeleton data and depth data, this thesis intends to employ late fusion to combine data related to text and images. The hypothesis is that the late fusion approach used by Boulahia et al. will also be efficient in this setting. The reason for

[13] `https://pytorch.org/docs/stable/data.html`
[14] `https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html`
[15] `https://huggingface.co/docs/transformers/main_classes/optimizer_schedules`
[16] `https://github.com/HishamAlkaed/Master_Thesis`

not choosing the intermediate approach is because that is more applicable to the data used by Boulahia et al..

The proposed approach for late fusion uses neural learning strategies to extract significant features from data without relying on handcrafted rules. This results in better performance compared to existing late fusion solutions. The approach involves stacking the score vectors obtained from models trained on different modalities and feeding them into fully connected layers (MLP) with ReLU and softmax activation functions to obtain the final prediction. To validate its effectiveness, the proposed approach was evaluated on two challenging datasets - NTU RGB-D (Shahroudy et al., 2016) and SBU Interaction (Vicente et al., 2016). These dataset are commonly used to evaluate action recognition systems (SK et al., 2023; Mopidevi et al., 2023; Zhang, 2023; Kong et al., 2023; Gutierrez-Gallego et al., 2023; Reilly et al., 2023). It was found that the proposed approach either performs equivalently or outperforms existing state-of-the-art approaches.

The process of late fusion can be divided into four distinct blocks. The first block is concerned with creating pipelines for both the text and image modalities. This has been achieved in Section 3.2.2. In the second block, a score vector of length N is produced for each modality and for N classes that are considered. This is achieved by using `predict_proba()` instead of `predict()` from sklearn. The third block involves concatenating decision vectors generated by each modality. In the final block, a classification module consisting of four fully connected layers is utilized. The last layer comprises a softmax function used for classification purposes. To avoid a potential bottleneck effect, the size of the first fully connected layer is calculated based on the N number of classes multiplied by K modalities. The second layer's size is four times N, while the third layer includes a dropout mechanism employed for regularization to prevent over-fitting during training and results in two times N neurons. Finally, the fourth and last layer employs a softmax to obtain the predicted class.

This method has been used to merge the frozen ResNet50 model with 4 textual models; BoW, character-n-gram, advanced SVM and fine-tuned BERT (Figure 3.2. Due to the fact that the classifiers are fused in a later stage, the training split cannot be used for training. Hence, the seen and unseen splits serve as training and test sets, respectively, for the MLP in this experiment.

3.2.5 Ensemble

In this thesis, it is explored whether is is possible to improve the performance of a LLM (BERT) through an ensemble strategy that supplements it with models targeting other modalities. To this point, many unimodal models were created. An ensemble is a machine learning approach where multiple models, referred to as "weak learners", are trained to solve the same problem, and then combined to form an optimal predictive model that yields better results. The primary premise is that correctly combining weak models leads to more accurate and/or robust models (Sagi and Rokach, 2018).

A stacked ensemble, also known as "stacking" or "stacked generalization", is a particular type of ensemble machine learning algorithm. It involves combining the predictions of several machine learning models using another machine learning model, called a meta-learner or meta-model, to make the final prediction (Wolpert, 1992). The idea is to leverage the strengths of different base models and combine them optimally to achieve better overall performance. This should also give the meta-model the capability of compensating for the limitations of its constituent learners.

A stacked ensemble is preferred because it has the potential to enhance the accuracy and robustness of predictions by utilizing the strengths of multiple base models. By using a meta-model to combine predictions, the stacked ensemble can learn which model's predictions are more reliable or accurate in specific situations, leading to better overall predictions. Furthermore, stacking can help mitigate overfitting, as the final prediction is based on the output of multiple models rather than relying on a single one (Jangam and Annavarapu, 2021).

To implement a stacked ensemble, all previously mentioned models in Section 3.2.3 were incorporated. Namely, the BoW and character n-gram baselines, the advanced SVM, the fine-tuned BERT, the hateBERT (both directly and word-embeddings fed to SVM) and the different configurations of the ResNet50. Additionally, a grid search was performed to identify the optimal combination of models that would yield the highest performance, as measured by accuracy. Multiple meta-models were tried within the grid search, including Logistic Regression[17], Random Forest[18] and Gradient Boosting[19], all of which are available through scikit-learn. This approach should ensure the optimal use of each model's strengths and improve the overall performance of the predictive model. Due to the fact that the models are ensembled in a later stage, the training split cannot be used for training. Hence, the seen and unseen splits serve as training and test sets for the meta-models in this experiment. However in the grid search, to avoid overfitting the unseen splits are not used, rather, the seen test and the seen dev splits are used as training and test sets for the meta-models.

Hardware specification This experiment has been run locally on an Intel Core i7-11700F CPU, NVIDIA GeForce RTX 3060 (12GB) GPU, and 16GB of RAM memory.

[17] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

[18] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`

[19] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html`

Chapter 4

Results

After conducting exploratory analysis, testing various architectures using state-of-the-art pre-trained models and self-engineered features, developing a late fusion pipeline and performing a grid search to identify the ideal combination of learners for the stacked ensemble approach, the obtained results will be presented in this chapter. The results are divided into three sections, each related to one of the main methods introduced in 2.1. The performance of the unimodal BERT systems is introduced in Section 4.1 to answer subquestion 1, which is to investigate the reasons for the high performance of the unimodal LLM BERT in a multimodal setting. Then, the results of the late fusion and ensemble experiments are provided in Section 4.2 and 4.3, respectively, to assist in finding out what is the impact on the performance of BERT when combined with other approaches using late fusion and ensemble strategies (subquestions 2). An error analysis of the results is performed in Chapter 5.

To supplement this study and provide a basis for the ensemble and late-fusion models, the performance of the unimodal classifiers on the dev and test splits (seen and unseen) is provided in Appendix B. Five ResNet configurations, two SVMs baselines and one Advanced model (utilizing the stylometric & emotion-based and transformer features) were used in this evaluation. Though those results are not relevant to the research questions, they were used in the ensemble and late-fusion models.

4.1 BERT - Unimodal

Table 4.1 shows the performance of the individual BERT models on both seen and unseen splits for various evaluation metrics. The fine-tuned BERT model outperformed the other two hateBERT models in all metrics and splits, with an f1-score of 0.704 and 0.635 for the unseen dev and test splits, respectively. The model achieved an AUROC of 0.704 in the unseen dev set, the highest of all BERT models. This finding indicates that the performance of the unimodal BERT can be improved with fine-tuning, which enables the model to learn task-specific representations and achieves better performance on the task at hand. It also indicates that the fine-tuned BERT-base-cased model can be used effectively in the multimodal setup. The gap in performance between the fine-tuned bert-base-cased and the other two hateBERTs could be attributed to the fact that the hateBERT was fine-tuned on a different dataset than the one used in this research, which might have had a different style of language than the dataset at hand. In Section 5, error analysis will be performed on the fine-tuned BERT to understand its performance better. From now on, the fine-tuned BERT-base-cased model will be

model	split	f1_score	precision	recall	accuracy	AUROC
hBert_direct	dev_seen	0.454	0.590	0.535	0.540	
hBert_direct	test_seen	0.426	0.549	0.517	0.525	
hBert_direct	dev_unseen	0.486	0.624	0.536	0.644	0.536
hBert_direct	test_unseen	0.453	0.555	0.515	0.622	0.515
hBert_we	dev_seen	0.506	0.604	0.558	0.562	
hBert_we	test_seen	0.505	0.587	0.552	0.558	
hBert_we	dev_unseen	0.528	0.572	0.544	0.626	0.544
hBert_we	test_unseen	0.530	0.580	0.548	0.628	0.548
fine-tuned BERT	dev_seen	**0.513**	0.579	0.553	0.556	
fine-tuned BERT	test_seen	**0.547**	0.600	0.574	0.579	
fine-tuned BERT	dev_unseen	**0.704**	**0.704**	**0.704**	**0.724**	**0.704**
fine-tuned BERT	test_unseen	**0.635**	**0.634**	**0.641**	**0.648**	**0.641**

Table 4.1: Individual performance of unimodal BERTs

referred to as BERT.

Despite the promising performance, the gap between the AUROC on seen and unseen splits is substantial, with the biggest difference of 0.148 between the seen and unseen dev data. Therefore, the performance of the fine-tuned BERT model on the unseen set might deteriorate, indicating that there could be a noteworthy impact of the dataset shift between the seen and unseen splits. By dataset shift, it is meant the differences in the statistical properties of the data between the seen and unseen splits. This was illustrated in Table 3.1. However, since the interest, in this experiment, is to explore the reason behind the high performance of BERT, the best-performing model and the best split will be used for error analysis later in Chapter 5.

4.2 Late Fusion

Split (unseen)	**model**	**f1_score**	**precision**	**recall**	**accuracy**	**AUROC**
Development	BoW + ResNet50	0.516	0.541	**0.540**	0.541	**0.540**
Test	BoW + ResNet50	0.542	0.564	**0.559**	0.571	**0.559**
Development	Char + ResNet50	0.488	0.496	0.516	0.550	0.516
Test	Char + ResNet50	0.515	0.528	0.540	0.576	0.540
Development	advanced + ResNet50	0.466	0.526	0.519	0.605	0.519
Test	advanced + ResNet50	0.478	0.553	0.531	0.613	0.531
Development	BERT + ResNet50	**0.522**	**0.554**	0.536	**0.608**	0.536
Test	BERT + ResNet50	**0.546**	**0.583**	**0.559**	**0.626**	**0.559**
Development	fine-tuned BERT	0.704	0.704	0.704	0.724	0.704
Test	fine-tuned BERT	0.635	0.634	0.641	0.648	0.641

Table 4.2: Average performance of Late Fusion: training on dev+test (seen) and evaluation on dev+test (unseen)

In this section, the performance of the late fusion models is presented to address subquestion 2 of the research question. Table 4.2 shows the average performance of the late fusion models on the unseen development and test sets. The first column indicates which split was used for evaluation (either development or test). The second column

indicates the fused models. The remaining columns indicate the performance metric values (f1_score, precision, recall, accuracy, and AUROC) for the corresponding split and models.

It is observable that the best-performing model in terms of the f1-score metric is *BERT + ResNet50*. It is also noticeable that the performance of *Advanced + ResNet50* and *Character-n-gram + ResNet50* models were consistently lower than the rest for both development and test splits. Additionally, the *Character-n-gram + ResNet50* model had the lowest f1-score among all the models for both development and test splits.

In terms of accuracy, *BERT + ResNet50* had the highest accuracy of 0.608 and 0.626, respectively, for development and test splits. All the other models had accuracy values below 0.61 for both splits.

In terms of AUROC, all the models had values above 0.5, indicating that they performed better than random guessing. *BERT + ResNet50* and *BoW + ResNet50* had the highest AUROC values for both development and test splits, followed by *Character-n-gram + ResNet50* which had an AUROC of 0.516 and 0.540 for development and test splits, respectively.

All of the values presented in this table are averaged by running the MLP 10 times and saving the output at each time. This has been done to tackle the extremely random nature of neural networks and provide trustworthy results.

On the one hand, our results indicate that the *BERT + ResNet50* model performed the best among all late fusion models for both development and test splits. This provides evidence that BERT performs better than other text classification approaches explored thus far since the image model has been frozen throughout this experiment. In addition, it also suggests that using the stylometric & emotion-based and transformer features is not as useful in this task as its rival BERT. On the other hand, there is a drop of 0.168 and 0.082 in the AUROC for the dev and test splits, respectively, between using BERT on its own (Table B.3) and fusing it with ResNet50 (Table 4.2). This implies that the inclusion of the image classifier by late fusing the probability scores through an MLP architecture was harmful to the performance. In Section 5, more insights into the performance of Bert + ResNet50 are provided.

4.3 Stacked Ensemble

The table below presents the results of the applied grid search to discover the best combination of learners for the stacked ensemble. The performance was evaluated on the dev_seen split after training on the test_seen split. The learners used here are BoW and char-n-grams baselines, the advanced SVM, the fine-tuned BERT, the hateBERTs (both directly and word-embeddings fed to SVM), and the different configurations of the ResNet50. For readability, the ResNet50 models were abbreviated as "rn", the radial basis function kernel as "rbf", the linear kernel as "lin", and the neural network as "nn". The number at the end of the ResNet50 models refers to the value of C. The meta-models used are Random Forest Classifier, Gradient Boosting Classifier and Logistic Regression.

The results of the grid search show that the Gradient Boosting meta-model with all three BERTs, both baselines, advanced SVM, rn_rbf1, and rn_lin10 learners resulted in the best performance. This combination achieved an f1-score of 0.590, precision of 0.617, recall of 0.602, accuracy of 0.604, and AUROC of 0.6018. Followed by the

meta-model	learners	F1-score	Precision	Recall	Accuracy	AUROC
Random Forest	BERTs, SVMs, rn_rbf10	0.577	0.608	0.592	0.594	0.594
GradientBoosting	BERTs, SVMs, rn_rbf1, rn_lin10	**0.590**	**0.617**	**0.602**	**0.604**	**0.6018**
Logistic Regression	hBERTs, advanced, char-n-gram rn_rbf10, rn_rbf1, rn_nn	0.579	0.594	0.586	0.588	0.586

Table 4.3: Performance of Stacked Ensemble grid search with Different Meta-Models and learners: training on test (seen) and evaluation on dev (seen)

Random Forest meta-model with all three BERTs, both baselines, advanced SVM, and rn_rbf10 learners. And finally the Logistic Regression meta-model with the hateBERTs (both directly and word-embeddings fed to SVM), advanced, char-n-gram baseline, rn_rbf10, rn_rbf1, and rn_nn learners. It is worth mentioning that the difference is marginal and might change in different runs due to the random nature of the process and the models.

Split (unseen)	meta-model	f1_score	precision	recall	accuracy	AUROC
Development	GradientBoosting	0.677	0.676	0.686	0.689	0.686
Test	GradientBoosting	0.623	0.624	0.631	0.634	0.631
Development	fine-tuned BERT	0.704	0.704	0.704	0.724	0.704
Test	fine-tuned BERT	0.635	0.634	0.641	0.648	0.641

Table 4.4: Performance of Stacked Ensemble with the optimal learners: training on dev+test (seen) and evaluation on dev+test (unseen)

Table 4.4 presents the performance of the stacked ensemble model with optimal learners on the unseen development and test sets. The results show that the stacked ensemble model achieved an f1_score of 0.677 on the development split and 0.623 on the test split. We observe that the precision and recall values for each one of the development and test splits are similar, indicating that the model's classification performance is balanced between the two classes. The accuracy of the model on the development split is 0.689, outperforming the accuracy of the model on the test split at 0.634.

Furthermore, the stacked ensemble model achieved an AUROC of 0.686 and 0.631 for the development and test splits, respectively. These results indicate that the model performed better than random guessing in classifying sentiment. Moreover, the performance of our model on both development and test splits demonstrates the generalizability of the proposed approach to unseen dataset splits.

Overall, the results demonstrate that the stacked ensemble model with optimal learners was effective in combining the predictions of various machine learning models into one final prediction, leading to improved performance on the task of multimodal hate speech detection compared to its constituent learners. The successful performance of the stacked ensemble model can be attributed to the complementary strengths of the underlying learning models, which are learned from the representations of both textual and visual features. The proposed stacked ensemble model's performance suggests its potential in detecting hateful speech in an online environment since it reaches an accuracy of 0.689 for development.

In conclusion, this chapter presented the results of the experiment conducted to answer the research questions. The unimodal BERT model outperformed the other two BERT models, indicating that fine-tuning BERT can improve its performance on the task of hate speech detection in a multimodal setting. The late fusion models showed that the BERT + ResNet50 model performed the best among all models for both development and test splits. However, the inclusion of the ResNet50 model appeared to be harmful compared to the performance of BERT on its own. The grid search showed that the stacked ensemble with the optimal combination of learners was effective in combining textual and visual features to classify sentiment, achieving an AUROC of 0.686 on the unseen development data. Yet, unimodal BERT with its 0.704 AUROC score on the unseen development data warrants further investigation. The next section will perform an error analysis of the best-performing models to obtain a better understanding of their strengths and weaknesses.

Chapter 5

Error Analysis

This chapter presents a qualitative analysis to understand the errors committed by the best-performing model and explore the potential causes behind them. Additionally, a contrastive error analysis is provided for the ensemble and late fusion methods and the disagreements with BERT are focused on. The goal is to identify the strengths and weaknesses of each model, understand the types of errors they make, and gain insights into their performance. The unseen dev set is used in this chapter to perform the error analysis. First, the performance of each of BERT, Late Fusion and Ensemble is evaluated in Sections 5.1, 5.2 and 5.3, respectively. Second, collective error analysis of all models is provided in Section 5.4.

5.1 Fine-tuned BERT

In this section, the performance of a fine-tuned BERT model in classifying hateful memes is analyzed. The focus is on cases where the model correctly identified hateful memes as well as misclassifications to gain a thorough understanding of the model's performance. The BERT-base-cased model used was unimodal pre-trained and underwent unimodal fine-tuning using only the textual feature, resulting in high performance, exceeding the unimodal baseline provided by previous work (Kiela et al., 2020) and all other models in this thesis. This makes it interesting to uncover the inherent capabilities of unimodal LLM BERT and shed light on its effectiveness in capturing textual cues in hateful memes. Figure 5.1 provides the confusion matrix of the BERT model.

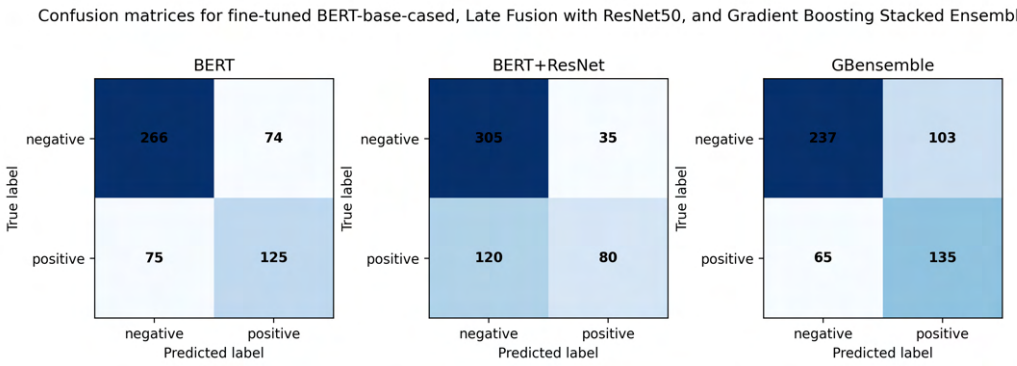


Figure 5.1: BERT, Late Fusion and Stacked Ensemble confusion matices

5.1.1 True Positives

In the context of binary classification, True Positives (TPs) occur when a model correctly predicts a positive outcome for a sample. In this environment, a TP would happen if the model predicts that a meme is hateful and its gold label is hateful. There are, in total, 125 TPs that were examined manually, with memes containing hateful textual cues being discarded as the model is expected to classify those accurately. The interesting examples found are categorized into three main categories: memes related to racism (mostly against black people), memes related to sexism (mostly against women), and memes related to discrimination based on ethnicity.

Figure 5.2: BERT True Positives. Category: Racism

Figure 5.2 provides some examples of the first category, racism. The interesting phenomenon that deserves some light to be shed on is that non of the memes has hateful text that makes it obvious that it is hateful. If those texts were provided to humans they will most probably classify them as non-hateful. Yet, the model was able to correctly classify those memes. More examples from this category are provided in appendix A. This could be attributed to the overfitting that happened during the fine-tuning phase.

Figure 5.3: BERT True Positives. Category: Discrimination based on sex (sexism)

Similarly, in Figure 5.3 three examples of memes where the text did not contain any cues about the hatefulness of the meme are shown. In this case, the words *sandwich maker*, *dishwasher*, and *cooker* appear in the memes. Although there is no direct connection between those words and hate speech (language-wise), those words have been used in this dataset to refer to women. It is possible that the model was able to pick up on that knowledge during the fine-tuning phase, or the model had already been faced with those words in such contexts during the pre-training phase and it was able

Figure 5.4: BERT True Positives. Category: Discrimination based on ethnicity

to efficiently transfer that knowledge to this task. On the one hand, this could be seen as over-fitting since the model is learning properties that are specific to the dataset and this model would not correctly classify the same memes if the images were swapped with an actual dishwasher and sandwich maker. On the other hand, this could also be justified that the model was able to reach the fine line where things get blurry. In other words, the model was able to find the words that could potentially be used for hate speech. Figure 5.4 also provides some examples of non-hateful words used in a hateful context; hence, the meme was (correctly) classified as hateful. The most interesting example in Figure 5.4 is the first one to the left. The text is *"Jamal is practicing for class sport"*. The only word that the model could have seen associated with hate speech is the Arabic name "Jamal". This again reflects the bias the model has gained from either pre-training or fine-tuning.

5.1.2 Misclassifications

In order to gain a complete understanding of the capabilities of the model it is of utter importance to also examine cases where the model was not able to correctly classify a meme. It is worth mentioning, that the unseen development split used for the analysis here does not contain any unimodal hate speech as can be seen from Table 3.1. With this in mind, False Positives (FPs) and False Negatives (FNs) were manually examined to reveal the interesting cases that deserve attention.

Figure 5.5: BERT False Negatives with hateful textual cues

False Negatives occur when the model predicts that a meme is non-hateful while its actual gold label is hateful. FNs are extremely dangerous as they can cause hateful content to go undetected. Most FNs included multimodal hate, which required analysis of both the image and text modalities to be accurately classified. However, there were instances where the BERT model did not perform as expected. The model failed to identify hateful textual cues in some instances, despite being present in the text, leading to misclassification.

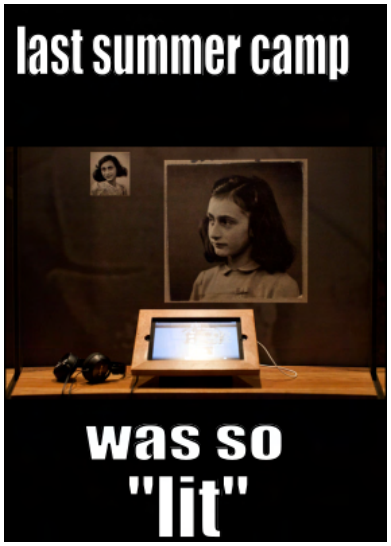
Hateful Meme (Correctly classified)	Benign Confounder (non-Hateful) missclassified
	
	
	
	

Table 5.1: Examples of memes with benign confounders, BERT.

For instance, in Figure 5.5, the texts provided clear hateful cues targeting the Muslim religion, but BERT classified them as non-hateful. Moreover, the previously noted association of words like "dishwasher" and "sandwich maker" with women did not contribute much to the model's performance in this category. As a result, multiple memes containing hate speech were classified as non-hateful. Additionally, multiple memes had ambiguous/**doubtful labels** that led to misclassifications. Those cases are discussed in Section 5.4.

False Positives occur when a model predicts a positive outcome for a sample that genuinely belongs to the negative class. In this environment, an FP would happen if the model predicts that a meme is hateful while its actual gold label is non-hateful. Identifying the reasons behind an FP can disclose the features that the models are excessively sensitive to while performing hateful meme detection. FPs were typically the result of benign confounders in the dataset. In particular, many instances arose where BERT unexpectedly correctly classified a meme as hateful (TP), but only because of the benign confounder. In other words, BERT gave both the meme and its benign confounder the same label. These examples are provided in Table 5.1 (previous page) and in Appendix A and highlight BERT's biases during the fine-tuning phase of the model.

In this section, the performance of a fine-tuned BERT model in classifying hateful memes was analyzed. The BERT model exceeded the unimodal baseline provided by previous work and all other models in the thesis. The analysis of the TPs showed interesting examples of racism, sexism, and discrimination based on ethnicity, where the model correctly identified hateful memes where the text did not explicitly contain any hateful cues. However, in the FNs, the model failed to identify some memes with hateful textual cues. Additionally, the model showed bias in associating words such as *dishwasher* and *sandwich maker* with women and Arabic names with hate speech. The analysis of the benign confounders showed that the model (naturally) cannot handle benign confounders well as it gives both the meme and its respective benign confounders the same label. In the following section, the performance and potential biases of a model that combines BERT with ResNet through late fusion will be analyzed.

5.2 Late Fusion

In this section, the efficacy of late fusion in enhancing the performance of the best-performing system thus far, fine-tuned BERT, is investigated. Specifically, the second research subquestion concerns the possibility of improving the performance of BERT by leveraging models targeting the other modality through late fusion. To achieve this, a robust image classifier, ResNet50, is integrated with BERT, and the results are analyzed.

The confusion matrices of BERT and BERT+ResNet are presented in Figure 5.1. Upon perusal, it is observed that compared to BERT alone, BERT+ResNet generates more negative predictions, thereby improving BERT's accuracy in correctly identifying non-hateful memes. However, it has a negative impact on the overall performance.

Manual inspection of the errors made by the Late Fusion model revealed that most **FPs** resulted from an over-reliance on one modality where either the text or image contained hateful cues while the other modality was benign. This behaviour implies

that the integration of the two modalities through late fusion was suboptimal. On the other hand, the **FNs** indicated that the model needs to improve its ability to recognize and integrate both modalities in detecting hate speech. Multiple examples, where hateful cues existed in both the textual and image modalities but were not captured by the model, suggest an urgent need for improvement in this regard. Such examples are provided in Figure 5.6.

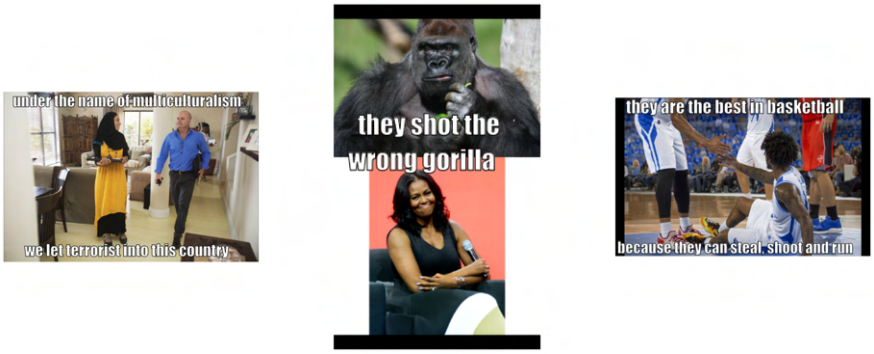


Figure 5.6: Late Fusion: False Negatives with hateful cues

Moreover, disagreements with BERT are examined thoroughly in order to uncover the capabilities and limitations of the late fusion method. There were 105 instances where the fusion resulted in faulty prediction compared to 44 cases where the fusion was useful. This illustrates the drawbacks of fusing BERT with ResNet50 using the aforementioned MLP architecture. Further examination of the 44 memes where the late fusion approach proved superior to BERT revealed a common theme. As mentioned previously, BERT struggled with memes that have benign confounders (Table 5.1). Namely, it gave both memes the same label resulting in one correct and one faulty prediction. The late fusion flipped both labels to fix the prediction of one meme but incorrectly labelled the other.

On the other hand, the manual inspection of the errors made by the Late Fusion model emphasizes a particular interest in instances where BERT's predictions were initially accurate but became erroneous following late fusion. The 105 instances, were categorized into two groups: FPs and FNs. FNs, wherein BERT initially correctly predicted a hateful meme but the merged model classified it as non-hateful, and FPs, where BERT had initially correctly identified a non-hateful meme, but the BERT+ResNet model misclassified it to be hateful. The investigation reveals that of the total instances, there were 88 occurrences of such FNs and 17 cases of FPs.

5.2.1 False Positives - Disagreement with BERT

FPs were mostly attributed to the doubtful annotation or the presence of tricky memes in the dataset. Doubtful annotation refers to memes that are annotated as negative but contain implicit hateful content. On the other hand, tricky memes are those that are noisy and not easily interpretable, thereby posing a challenge to accurate classification.

Tricky memes In Figure 5.7 some examples of such memes are provided. The first meme appears potentially hateful if used in the wrong context, but on its own, it does not directly attack any protected group. The second meme connects the Islamic

Figure 5.7: BERT: non-hateful, BERT+SVM: hateful, gold: non-hateful. Tricky Memes

name Mohammad to a goat, but it does not have explicitly hateful content. The third meme contains the Nazi sign and makes the joke that you are trying to fight racism by prompting the Nazi sign (by burning it in public).

Figure 5.8: BERT: non-hateful, BERT+SVM: hateful, gold: non-hateful. Doubtful annotations

Doubtful annotations In Figure 5.8, three memes with doubtful annotations are presented. The first meme implies that the wife is a one-year-old child; ("paedophile"... big word for one year old). The second meme compares a young black boy to apes, which is an explicit form of hate speech. The third meme could be construed as mocking Mexicans and Hispanics due to its reference to playing Uno and collecting green cards [1].

Overall, FPs can occur due to doubtful annotation or the presence of tricky memes in the dataset. These memes can be challenging to classify accurately, even for experienced annotators. While FPs can be concerning, FNs, where the model initially classifies a hateful meme as non-hateful and it is hateful, can be even more damaging. The next section focuses on FNs in the BERT+ResNet model and their implications for hate speech detection.

5.2.2 False Negatives - Disagreement with BERT

There were in total 88 cases where BERT unimodal correctly classified a hateful meme but after fusion with ResNet50, it got them wrong. Those FNs were manually explored. Most of the memes contain straightforward hate speech that was not detected correctly

[1] Green cards, or lawful permanent resident status, are available to individuals who fulfil certain eligibility criteria, regardless of their nationality or ethnicity. Unfortunately, lately, it has been commonly used to refer to immigrants from Mexican and Hispanic backgrounds.

after the fusion. Figures A.3 in appendix A show some examples of those memes categorized into three groups, Islamophobic memes, racist memes and sexist memes, respectively.

In this section, we examined the effectiveness of late fusion in enhancing the performance of BERT by integrating a robust image classifier, ResNet50. Our analysis revealed that integrating the two modalities improved BERT's accuracy in detecting non-hateful memes. However, the overall performance of the model was adversely affected. We also observed that the main issues with this strategy were tricky memes, doubtful annotations, and the inability of the model to identify straightforward hate speech. These limitations underline the importance of exploring other ensemble strategies, which are addressed in the following section.

5.3 Stacked Ensemble

From the confusion matrix in Figure 5.1 it is observable that the stacked ensemble has a higher count of FPs than FNs. Meaning that the model is more likely to incorrectly identify an instance as positive even though it is actually negative (Type 1 error). This can lead to unwanted real-life scenarios such as flagging some users and/or specific messages as hateful while it is not, negatively affecting the experience of users.

Overall Performance The Gradient Boosting stacked ensemble model had lower performance than BERT, yet it was able to correctly predict more TPs on the cost of more FPs. Similarly, it had fewer FNs than BERT on the cost of fewer True Negatives. The ensemble method showed improvement compared to the Bow + ResNet50 baseline presented in Table 4.2. This indicates that the model is reasonably proficient in correctly categorizing memes in the dataset.

The second research subquestion examined the possibility of improving the performance of BERT by leveraging models targeting the other modality through an ensemble. The ensemble was beneficial in 28 cases and made the predictions of BERT faulty in 47 cases. The errors committed by the ensemble model are analyzed and classified into two types: FPs and FNs.

5.3.1 False Positives - Type I errors

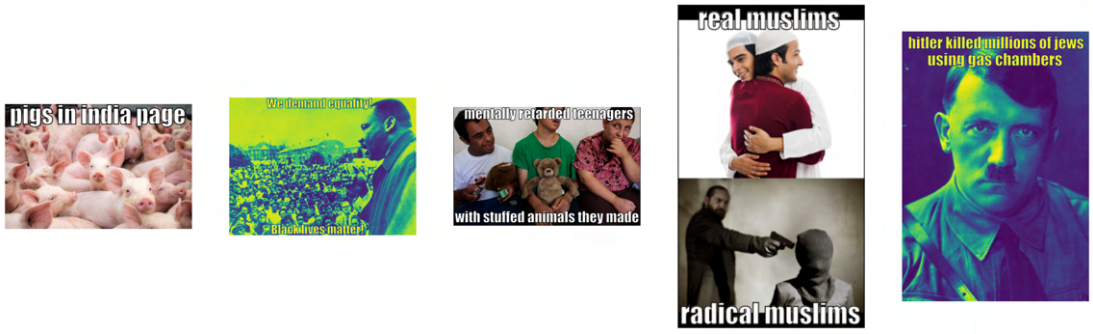


Figure 5.9: Ensemble False Positives where the text is hateful and the image is not

After a manual inspection of all the FPs that occurred in this setting, two main reasons for having an FP were identified (Figure 5.9 and Figure 5.10). Figure 5.9 shows some FPs due to the nature of the text. In other words, the ensemble was likely unable to classify those accurately because the text is hateful, or contains hateful words/slurs, while the images are not (over-reliance on one modality). For example, we note that words like pigs, black, retarded, radical, killed, Jews, and gas chambers contribute to this issue. Although passing such sentences through a text-based hatefulness classifier would result in true predictions, the combination of these sentences with the images makes them benign.

Figure 5.10: Ensemble False Positives where the image is hateful and the text is not

On the other hand, Figure 5.10 shows FPs due to the nature of the image. In other words, the ensemble was unable to classify those memes correctly because the image is hateful, but the text is benign. For instance, we can see that the first meme on the left in Figure 5.10 has a picture of a dead woman with blood, while the second one is a picture of the World Trade Center, which is quickly associated with the 911 event, and the third one is a picture of a woman who has a condition called vitiligo. Notice that none of the memes contains hateful text. This indicates that the ensemble in this case heavily relied on the decisions made by the image learners to produce an output.

5.3.2 False Negatives - Type II errors

FNs would occur if the ensemble predicts that a meme is non-hateful when in reality, the meme contains hateful content. This can be problematic in real-life scenarios where hateful content is not detected online and can reach more people. This could both perpetuate hate and cause targeted individuals to feel less safe online, which might lead to many problems, starting with depression and ending with suicide. Hence, reducing FNs is crucial to improving the accuracy and effectiveness of the ensemble if it were to be implemented online.

Figure 5.11: Ensemble False Negatives where the text and the image are benign but the combination is hateful

Again, FNs have been categorized based on the reason why they were misclassified. First, we have Figure 5.11, which introduces memes where both the text and the images are benign, but the combination of both is hateful. For instance, the first image to the

right depicts two animals having intercourse, where one of them is a goat. There is a reference to a Muslim party in the text making the combination of the two modalities hateful. The phrases "goat-humper" and "goat-f*cker" are commonly used to target the Muslim religion (Hee et al., 2022).

Figure 5.12: Ensemble False Negatives where there is an implicit hateful content

Similarly, Figure 5.12 provides multiple memes where the combination of modalities introduces the hateful content. However, those memes also contain implicit hateful content. The first meme compares white people to crackers that get soggy when wet (racism). In the second meme, the woman holding the couch is indirectly compared to a dishwasher (sexism). The third meme implicitly suggests that a dark-coloured man picking up his friends in a car is a getaway driver in a robbery (racism). In the fourth meme, a racist joke is made, implying that you would run over a dark-coloured person when you are driving. Those memes need a thorough understanding of both the image and the text and the implications of combining the two.

Doubtful labels In some cases, memes can be challenging to classify, either because the content is unclear or because it could be interpreted in more than one way.

Figure 5.13: Ensemble False Positives where the gold label is doubtful

We start by introducing Figure 5.13. In some cases, the gold labels are doubtful, as the memes have been labelled as not hateful, but they are very likely to be hateful. In the first meme to the left, the scenario describes one undermining women's gender equality rights and addressing her as a "washing machine." The second meme suggests that the car is better because it is white (racism), and the third meme contains the word "rape" and suggests that we could use it to help future generations have sex. Rape should never be thought of as a tool to help anything because it is a crime.

Similarly, Figure 5.14 shows examples of doubtful gold labels. None of the memes provided contains hateful content. The first meme uses the body and facial expressions of Hitler to show the emotions in the text. The second meme depicts a white man helping young dark-coloured children. And the last meme serves as a statement.

Figure 5.14: Ensemble False Negatives where the gold label is doubtful

In this section, the errors committed by the Gradient Boosting ensemble model in the context of hateful meme classification were analyzed and classified into FPs, FNs, and doubtful labels. Several reasons behind these errors were identified. FPs were attributed to the combined contribution of text and images, while FNs often resulted from the memes' implicit hateful content. Understanding these errors and their potential causes can improve the accuracy and effectiveness of the ensemble model for future use, such as the inclusion of recent world events in the model.

5.4 Collective evaluation

The error analysis of hate speech detection models for multimodal memes requires a thorough examination of cases where all models fail to correctly classify a meme. These instances not only reveal the limitations and challenges faced by the models but also reflect the inherent difficulty of accurately detecting hate speech in multimodal content. This section presents a collective evaluation of memes that were misclassified by all models, including both FPs and FNs. Through manual inspection, we aim to identify patterns and reasons behind these misclassifications, providing insights into the challenging cases present in the dataset.

Additionally, this section includes a subsection that explores the performance of different models in scenarios involving benign confounders. Analyzing examples of benign confounders alongside the corresponding models' predicted labels sheds light on the performance disparities among models in various contexts. These examples contribute to our understanding of the nuanced challenges faced by multimodal hate speech detection models when confronted with unimodal benign content that may influence the classification outcome.

Figure 5.15: False Positives (all models): over-reliance

False Positives - Over-reliance Figure 5.15 presents examples of FPs where all the models incorrectly classified a meme as hateful. These misclassifications often stemmed from the over-reliance on specific visual or textual cues present in the memes. The models may have excessively weighted these cues while overlooking other contextual information, resulting in erroneous classifications. For instance, the memes here include the word "down" which could be used to refer to individuals with Down syndrome. However, the image modality here nullifies the hatefulness embedded in the text modality. By examining these instances, we gain insights into the need for models to consider a broader range of multimodal cues and strike a better balance between different modalities.

Figure 5.16: False Positives (all models): overfitting

False Positives - Overfitting The examples depicted in Figure 5.16 highlight FPs where biases played a considerable role in the misclassifications. These memes were wrongly labelled as hateful due to inherent biases present in the training data or the models themselves. The models here exhibited signs of over-fitting on the dataset as they incorrectly used textual cues such as *sandwich maker* and *slow cooker* to identify hate speech. The misclassifications underscore the challenges of developing unbiased hate speech detection models and emphasize the importance of addressing and mitigating biases during the training process.

Figure 5.17: False Negatives (all models): doubtful-labels

False Negatives - Doubtful labels Figure 5.17 presents examples of FNs where all models failed to identify hateful content in the memes. These instances fall into the category of doubtful labels that was particularly challenging to classify accurately. The content in these memes may have been ambiguous, containing elements that could be interpreted as hateful but also have potential alternative interpretations. Through the analysis of these cases, we gain insights into the complexities and nuances of multimodal hate speech detection.

5.4.1 Benign Confounders

The examination of benign confounders and the behaviour of different models in various scenarios adds another dimension to the error analysis. Benign confounders, in this case, are alternative images that flip the label of the meme. By analyzing these instances alongside the predictions of different models, we can gain insights into how models respond to benign content and how their performance varies in different scenarios.

Table 5.2: Examples of memes with benign confounders, Performance of All Models

Table 5.2 presents examples where all models yielded the same prediction for both the meme and its corresponding benign confounder, resulting in one correct prediction and one incorrect prediction. These instances demonstrate situations where the models struggle to differentiate between the meme and its benign confounder, leading to

misclassification. While this was clear in unimodal BERT since it only relied on one modality, Late Fusion and Ensemble exhibited the same behaviour here.

Hateful Meme & Benign Confounder (non-Hateful)

Table 5.3: Examples of memes with benign confounders, Performance of All Models

Table 5.3, on the other hand, showcases examples where the ensemble model differs in its prediction for the meme and its benign confounders, while BERT and Late Fusion models yield the same prediction for both. Specifically, out of the six instances, the ensemble model provides four correct predictions and two incorrect predictions. Opposed to three correct predictions and three incorrect predictions for BERT, and one correct prediction and five incorrect predictions for Late Fusion. The stacked ensemble method shows promising behaviour in compensating for the limitation of its constituent learners. These discrepancies in predictions highlight the variations in performance among the models and their sensitivity to different aspects of the meme and its context.

By examining the examples of benign confounders and comparing the predictions of different models, we can gain insights into their behaviour and performance in the presence of non-hateful elements that may influence the classification decision. This

analysis provides valuable information for understanding the strengths and weaknesses of each model and their response to challenging scenarios.

In this chapter, we have examined the errors made by BERT, Late Fusion (BERT + ResNet) and the Gradient Boosting Stacked Ensemble model in classifying hateful memes. The analysis revealed that the fine-tuned BERT model demonstrated superior performance compared to other models and previous baselines. It exhibited the ability to correctly identify hateful memes even when explicit hateful cues were absent in the text. However, the model had limitations in identifying some memes with hateful textual cues and exhibited biases in associating certain words with specific groups.

Late fusion of BERT with ResNet50 improved the accuracy of non-hateful meme classification at the cost of worsening the classification of hateful memes. FPs occurred when one modality contained hateful cues while the other was benign, highlighting suboptimal integration. FNs indicated the need for better recognition and integration of both modalities for detecting hate speech.

The stacked ensemble model using gradient boosting showed improvement compared to the BoW+ResNet baseline but had trade-offs in terms of FPs and FNs. FPs were attributed to the nature of the text or the image, leading to over-reliance on a single modality. FNs resulted from the combination of modalities introducing hateful content, as well as the presence of implicit hate speech.

One key addition to the error analysis chapter was the collective evaluation section, where we examined the cases where none of the models was able to correctly classify the memes. Through manual inspection, 31 FPs and 55 FNs were identified. The FPs were primarily attributed to over-reliance on certain cues and the presence of bias in the models' associations. On the other hand, the FNs were often challenging due to multimodal hate or vague content, and some instances raised doubts about the accuracy of the annotations.

Furthermore, exploring discrepancies between models' predictions when it comes to benign confounders yielded insights into the models' responses in different scenarios. The tables showcasing examples of benign confounders and the predictions of different models highlighted the need for further improvement in accurately distinguishing between hateful content and non-hateful elements.

In the upcoming chapter, we will highlight the results and constraints of our study, along with their theoretical and practical implications for further research and applications. Additionally, we will explore ethical issues associated with hate speech identification and potential biases that automated models may exhibit.

Chapter 6

Discussion

The discussion chapter aims to analyze and interpret the findings of the study, addressing the research questions and sub-questions related to the role of textual modality in hateful meme detection and the effectiveness of late fusion and ensemble methods. In this discussion, we delve into the implications of our findings, discuss the challenges faced in multimodal hate speech detection, and explore potential avenues for improvement in future research.

6.1 Fine-tuned BERT-base-cased

The analysis of the performance of BERT - a unimodal Large Language Model - in detecting hateful memes sheds light on the ability of such models to comprehend the textual modality. The study reveals that BERT exhibited high accuracy in identifying hate speech cues in memes when compared to other models. It's possible that the model learned specific patterns from the dataset, such as the association of certain words with hate speech, or was able to correctly transfer the knowledge gained in the pre-training phase to the task at hand. However, while these results are promising, it is crucial to consider the limitations associated with using specific patterns learned solely from a particular dataset. The presence of certain word associations with hate speech implies potential biases against specific groups, as found in the literature (Fersini et al., 2022), which could impact the generalization of the model to unseen data.

In the context of multimodal hate speech detection, the model's proficiency in recognizing specific marginalized communities, ethnicities, and genders is crucial. However, there are limitations to the model's performance related to overfitting and dataset-specific biases. An example was found in the error analysis of this thesis where the unimodal BERT classified a meme as hateful solely due to the existence of an Arabic name in it (5.4). In addition, the application of this model in this dataset showed the challenging aspects of benign confounders. Consequently, the proficiency of the model in detecting hate speech cues may depend on the representation of those cues in the training data. These findings emphasize the potential dangers of relying too heavily on dataset-specific cues and highlight the need for developing models that generalize to a wide range of scenarios and data.

6.2 MLP Late Fusion and Stacked Ensemble

The second sub-question of this study explored the possibility of enhancing the performance of unimodal BERT by incorporating models targeting other modalities through an ensemble or late fusion strategies. Specifically, the analysis focused on the late fusion and the stacked ensemble methods that combined BERT with ResNet50 and supplemented the classifiers with several "learners" via Gradient Boosting, respectively. The findings indicate that both strategies resulted in decreased performance compared to unimodal BERT in isolation. The error analysis conducted on the late fusion and stacked ensemble methods revealed insights into why they performed worse than unimodal BERT. These findings shed light on the specific challenges and limitations associated with these approaches.

One of the reasons for the decreased performance of the late fusion method and ensemble model is the suboptimal integration of different modalities. In cases where one modality contained hateful cues while the other was benign, the late fusion method often struggled to effectively combine the information from both modalities. This resulted in an increased number of false positives, where the model incorrectly classified instances as hateful when only one modality contained offensive content. The stacked ensemble model also exhibited a higher false positive rate, indicating a tendency to misclassify negative instances as positive. These findings highlight the difficulties in achieving optimal integration and fusion of multimodal features, which can impact the accuracy of hate speech classification.

Additionally, the error analysis identified instances where the combination of modalities introduced deceptive or implicit hateful content. Detecting nuanced and implicit hate speech indicators in multimodal memes is a complex task, and both the late fusion and ensemble methods struggled with accurately classifying these instances. The inherent challenges in recognizing and understanding implicit hate speech contributed to a higher false negative rate, where the models failed to identify instances of hateful content. This indicates the need for further advancements in capturing and integrating multimodal cues to effectively detect and classify implicit hate speech in memes.

6.3 Implications and Challenges

The analysis of errors and limitations in hate speech detection models for multimodal memes highlights several important implications and challenges in this research domain. Firstly, the nuanced nature of multimodal hate speech and the interplay between different modalities (text and image) present significant challenges in accurately identifying and classifying hateful content. Multimodal hate speech often manifests through complex and context-dependent cues, making it difficult to define clear-cut boundaries for classification. Additionally, multimodal memes combine textual and visual elements, necessitating the effective integration of both modalities for reliable detection.

Furthermore, the presence of benign confounders in the particular dataset used and implicit hate speech poses challenges for multimodal hate speech detection models. Benign confounders refer to instances where the presence of hateful cues in one modality (text or image) is offset by the benign nature of the other modality. This highlights the importance of developing models that can effectively capture and integrate information from multiple modalities to accurately classify hateful content. Implicit hate speech, which relies on subtle associations or stereotypes, further complicates the detection

process, as it requires models to capture and understand the underlying meaning and connotations in memes.

The presence of doubtful labels in the dataset adds another layer of complexity to hate speech detection. Memes with unclear or ambiguous content challenge both human annotators and machine learning models, as they require subjective judgments and may lead to inconsistencies in labelling. Addressing the issue of doubtful labels requires establishing clearer guidelines and criteria for annotation, as well as developing robust models that can handle ambiguity and make more nuanced classifications.

The challenges in multimodal hate speech detection extend beyond technical aspects and also encompass ethical considerations. The automated classification of hate speech in multimodal memes requires careful attention to ensure fairness, transparency, and accountability. Bias in the training data or the models themselves can perpetuate or amplify existing societal biases, leading to discriminatory outcomes. It is crucial to address and mitigate biases throughout the development process, from data collection and annotation to model training and evaluation.

6.4 Ethical Concerns

The development and deployment of hate speech detection models for multimodal memes raise several important ethical concerns that need to be carefully considered. While these models have the potential to mitigate the spread of hateful content online and promote a safer digital environment, they also pose risks and challenges that must be addressed to ensure responsible and fair use. The following ethical concerns deserve particular attention:

6.4.1 Bias and Discrimination

Hate speech detection models can inherit and amplify biases present in the training data, as well as introduce new biases during the learning process (Sap et al., 2019). If the training data contains systemic biases, such as racial or gender biases, the model may learn to discriminate against certain groups or reinforce existing stereotypes. This can lead to unfair or discriminatory outcomes, where certain individuals or communities are disproportionately targeted or marginalized. It is crucial to address bias at all stages of model development, including data collection, annotation, and algorithm design, to minimize the risk of discriminatory practices.

Furthermore, biases can also arise due to the lack of diversity in the training data. If the dataset primarily consists of content from specific demographics or cultural contexts, the model may struggle to accurately detect hate speech in underrepresented communities or languages. It is important to ensure the inclusion of diverse voices and perspectives in the training data to mitigate bias and improve the model's generalizability.

6.4.2 Freedom of Speech and Expression

Balancing the detection of hate speech with the principles of freedom of speech and expression is a complex ethical challenge. The definition and interpretation of hate speech vary across jurisdictions, cultures, and social contexts. The automated classification of memes as hateful or non-hateful may inadvertently suppress legitimate speech, satire, or forms of expression that are protected by the principles of freedom of speech. It is

crucial to strike a balance between preventing the harmful effects of hate speech and preserving the right to express diverse opinions and engage in meaningful discourse.

Transparent and inclusive processes for defining hate speech guidelines and developing annotation protocols are essential to mitigate the risk of stifling free expression. Involving diverse stakeholders, including experts in law, ethics, sociology, and human rights, can help establish guidelines that reflect a broad range of perspectives and promote fairness in the detection process.

6.4.3 Unintended Consequences

The deployment of hate speech detection models for multimodal memes can have unintended consequences that need to be anticipated and addressed. For instance, malicious actors may attempt to exploit the model's vulnerabilities by crafting memes that bypass the detection system, leading to a cat-and-mouse game between model developers and those seeking to circumvent detection. Additionally, there is a risk of over-reliance on automated systems, where human judgment and critical thinking are sidelined. Relying solely on machine learning algorithms without human oversight and intervention may result in erroneous classifications, false positives, or false negatives, leading to unjust consequences for individuals.

To mitigate these unintended consequences, a holistic approach is necessary. Continuous monitoring and evaluation of the model's performance, along with human-in-the-loop processes, can help identify and address emerging challenges. Regular updates and improvements to the model's architecture, training data, and algorithmic techniques should be undertaken to adapt to evolving strategies employed by those seeking to propagate hate speech.

Collaboration between researchers, policymakers, industry stakeholders, and civil society organizations is essential to foster discussions around the ethical implications of hate speech detection models. Multidisciplinary teams should work together to develop robust guidelines, policies, and frameworks that prioritize fairness, accountability, and the protection of individuals' rights.

Furthermore, ongoing public education and awareness campaigns can empower users to better understand the limitations and potential biases of hate speech detection models. Promoting media literacy and digital citizenship can help individuals critically evaluate online content and engage in constructive dialogue while being aware of the risks associated with hate speech.

In conclusion, the development and deployment of hate speech detection models for multimodal memes necessitate a comprehensive consideration of ethical concerns. Addressing biases, preserving freedom of speech, protecting privacy, ensuring transparency, and anticipating unintended consequences are critical to developing responsible and fair models that effectively combat hate speech while upholding fundamental ethical principles. By engaging in an ongoing dialogue and collaboration, stakeholders can collectively work towards harnessing the potential of these models while minimizing their potential negative impact on society.

6.5 Future Directions

This study provides insights into areas for future research and development of hate speech detection models for multimodal memes. To improve the performance and

robustness of these models, several directions can be pursued.

Improved integration of text and image modalities: Enhancing the synergy between text and image processing in multimodal models can help capture and leverage the complementary information present in memes. Exploring advanced techniques such as joint representation learning (Jia et al., 2021), attention mechanisms (Fang et al., 2023), and cross-modal fusion (Wang et al., 2022) can enhance the model's ability to effectively combine and interpret multimodal cues.

Optimal Use of Data: Currently, the models are trained on a subset of the data to predict a portion of the data, and these predicted labels are then used for late fusion or ensemble creation. However, this approach may result in the potential misuse of some data since the training split in this thesis is not used to train the late fusion and ensemble systems. Rather, the dev and test seen splits are. Further investigation is needed to explore alternative strategies for utilizing the available data more effectively to improve the overall performance of the models.

Data Augmentation and Expansion: The performance of the models improved with the availability of more data. Therefore, future efforts could focus on collecting and curating a larger and more diverse dataset of hateful memes. Additionally, data augmentation techniques, such as image transformations (Jain et al., 2021) and text paraphrasing (Turkerud and Mengshoel, 2021), can be employed to further enhance the dataset's richness and variability.

Probabilistic Ensemble: Instead of making binary decisions in the ensemble, incorporating probabilities can provide more nuanced and informative results. Future work can focus on developing ensemble techniques that consider the confidence or probability estimates from individual models, allowing for more refined decision-making and potentially improving the overall performance of the ensemble.

Contextual understanding of multimodal hate speech: Hate speech is highly contextual, and its interpretation often relies on understanding social and cultural nuances (Pavlopoulos et al., 2020). Developing models that can capture and analyze the broader context surrounding memes, including cultural references, historical events, and social dynamics, can improve the accuracy of multimodal hate speech detection. (Zhu, 2020)

Handling ambiguous and doubtful labels: Addressing the challenge of doubtful labels requires the development of models that can handle ambiguity and make more nuanced classifications. Techniques such as active learning (Settles, 2009), human-in-the-loop approaches (Wu et al., 2022), and consensus-based annotation (Vedantam et al., 2015) can be explored to improve the quality and consistency of labelling in datasets.

Generalizability and bias mitigation: Future research should focus on developing models that are more robust and generalizable to diverse datasets and real-world scenarios. This includes addressing dataset-specific biases, mitigating biases in model predictions, and ensuring fairness, transparency, and accountability in hate speech detection systems.

Comparison of more LLMs: In this thesis, a detailed error analysis of the fine-tuned BERT-base-cased was provided to answer the main research questions. It is also possible to utilize other LLMs to investigate whether other language models exhibit the same behaviour.

Ethical considerations: It is crucial to consider the ethical implications of hate speech detection models and ensure their responsible deployment. This includes ongoing evaluation and monitoring of model performance, regular audits for bias and

fairness, and active engagement with stakeholders to understand the impact of these systems.

Multilingual and cross-cultural perspectives: Hate speech detection models should be extended to handle multilingual and cross-cultural contexts, as hate speech manifests differently across languages and cultures. Adapting and training models on diverse datasets can improve their effectiveness in detecting hate speech in a global context.

In conclusion, the error analysis shed light on the strengths and limitations of hate speech detection models for multimodal memes. The challenges identified in accurately classifying hateful content, handling doubtful labels, and addressing biases emphasize the need for continued research and development in this field. By addressing these challenges and considering ethical implications, we can advance the state-of-the-art in multimodal hate speech detection and contribute to creating safer and more inclusive online spaces.

Chapter 7

Conclusion

In this study, we delved into the role of text classification approaches in multimodal hate speech detection. This research aimed to uncover the capabilities and limitations of unimodal text-based models, explore the potential for improvement through ensemble methods, and investigate the impact of late fusion techniques. Through rigorous analysis and evaluation, valuable insights have been gained that contribute to our understanding of multimodal hate speech detection and its real-life applications.

First, the performance of the fine-tuned unimodal BERT-base-cased revealed its remarkable performance in identifying hateful memes. Surprisingly, these models exhibited high accuracy even on datasets explicitly designed for multimodal analysis. Although it has been found that unimodal LLMs are effective in capturing textual cues in hateful memes, the model developed bias against some groups. This finding has significant implications for real-life scenarios where textual content plays a crucial role in multimodal hate speech identification, such as online platforms and social media monitoring systems.

Additionally, the impact of late fusion techniques has been examined by supplementing BERT with ResNet50, an image classifier. The BERT+ResNet50 model exhibited improved accuracy in identifying non-hateful memes but suffered from a decline in recognizing hateful memes. False positives were primarily attributed to tricky memes and questionable annotations, while false negatives were prevalent in cases involving explicit hate speech that the fusion process did not accurately detect. These findings emphasize the challenges and complexities associated with hate speech detection, where nuanced interpretation and contextual understanding are crucial.

Furthermore, we explored the potential for enhancing the performance of unimodal BERT through ensemble methods. Our results indicated that the ensemble approach, which combines multiple models targeting different modalities, did not outperform BERT in identifying hateful memes. While the ensemble strategy aimed to leverage the strengths of various modalities and compensate for individual model limitations, it did not yield significant improvements. This finding suggests that unimodal text-based models like BERT already capture crucial textual cues effectively, and the inclusion of additional modalities may not always lead to enhanced performance in multimodal hate speech detection, yet it could be used to reduce bias.

In conclusion, our research highlights the effectiveness of unimodal text-based models like BERT in identifying hateful content within memes, even when explicit hate cues are absent in the text. The study sheds light on the importance of textual cues and the inherent capabilities of unimodal LLMs in multimodal hate speech detection. While

the ensemble and late fusion approaches did not yield significant performance improvements, they provide valuable insights into the challenges and limitations of multimodal hate speech detection.

Moving forward, it is essential to continue exploring and refining text classification approaches, leveraging their strengths in multimodal hate speech detection while addressing biases and limitations. Further research could focus on enhancing the generalizability of models, improving the detection of subtle and context-dependent hate speech, and addressing the challenges posed by multimodal datasets. By combining the power of text classification approaches with advancements in multimodal analysis, we can foster more accurate and robust multimodal hate speech detection systems, contributing to a safer and more inclusive online environment.

Appendix A

Appendix A

Figure A.1: Fine-tuned BERT: True positives: category sexism

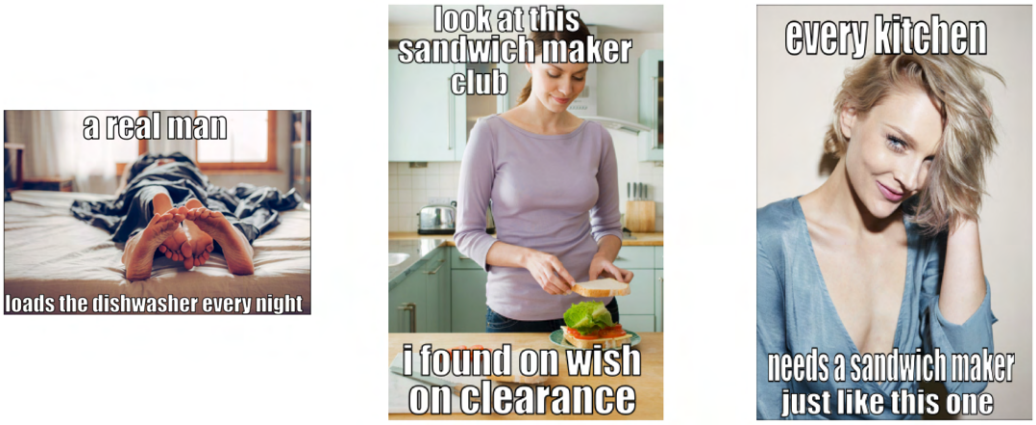
Figure A.2: Fine-Tuned BERT: True positives: category Discrimination based on ethnicity

Hateful Meme (Correctly classified)	Benign Confounder (non-Hateful) missclassified
	
	
	
	

Table A.1: Examples of harmful and benign memes with correct and incorrect predictions by BERT.

(a) Islamophobic memes

(b) Racist memes



(c) Sexist memes

Figure A.3: BERT prediction: hateful, BERT+SVM: non-hateful, gold: hateful. Examples of misclassified posts for (a) Islamophobic memes, (b) Racist memes, and (c) Sexist memes.

Appendix B

Appendix B

model	f1_score	precision	recall	accuracy
BoW	0.521	0.541	0.536	0.538
advanced	0.508	0.543	0.533	0.536
char-n-gram	0.499	0.521	0.518	0.520

Table B.1: Individual performance of SVMs on Development Set (seen)

model	f1_score	precision	recall	accuracy
advanced	0.513	0.551	0.539	0.544
char-n-gram	0.505	0.533	0.527	0.531
BoW	0.503	0.535	0.527	0.532

Table B.2: Individual performance of SVMs on Test Set (seen)

model	f1_score	precision	recall	accuracy
tuned_Bert	0.513	0.579	0.553	0.556
hBert_we	0.506	0.604	0.558	0.562
hBert_direct	0.454	0.590	0.535	0.540

Table B.3: Individual performance of Berts on Development Set (seen)

model	f1_score	precision	recall	accuracy
tuned_Bert	0.547	0.600	0.574	0.579
hBert_we	0.505	0.587	0.552	0.558
hBert_direct	0.426	0.549	0.517	0.525

Table B.4: Individual performance of Berts on Test Set (seen)

Model	**F1-Score**	**Precision**	**Recall**	**Accuracy**
rn_rbf1	0.367	0.518	0.502	0.508
rn_rbf10	0.437	0.504	0.502	0.506
rn_lin1	0.485	0.507	0.506	0.508
rn_lin10	0.492	0.505	0.504	0.506
rn_nn	0.475	0.476	0.476	0.476

Table B.5: Individual performance of ResNet on Development Set (seen)

Model	**F1-Score**	**Precision**	**Recall**	**Accuracy**
rn_rbf1	0.389	0.554	0.510	0.519
rn_rbf10	0.458	0.543	0.522	0.529
rn_lin1	0.460	0.473	0.477	0.481
rn_lin10	0.481	0.493	0.494	0.497
rn_nn	0.502	0.502	0.502	0.502

Table B.6: Individual performance of ResNet on Test Set (seen)

model	**f1_score**	**precision**	**recall**	**accuracy**	**AUROC**
tuned_Bert	0.704	0.704	0.704	0.724	0.704
advanced	0.669	0.676	0.665	0.700	0.665
char-n-gram	0.650	0.652	0.648	0.678	0.648
BoW	0.650	0.654	0.648	0.680	0.648
rn_lin10	0.605	0.610	0.617	0.613	0.617
rn_rbf1	0.541	0.563	0.548	0.615	0.548
hBert_we	0.528	0.572	0.544	0.626	0.544
hBert_direct	0.486	0.624	0.536	0.644	0.536

Table B.7: Individual performance of the optimal learners (for ensemble) on the Development data (unseen)

model	**f1_score**	**precision**	**recall**	**accuracy**	**AUROC**
tuned_Bert	0.635	0.634	0.641	0.648	0.641
advanced	0.604	0.607	0.603	0.636	0.603
BoW	0.590	0.596	0.589	0.629	0.589
char-n-gram	0.583	0.582	0.583	0.606	0.583
hBert_we	0.530	0.580	0.548	0.628	0.548
rn_lin10	0.521	0.521	0.521	0.544	0.521
hBert_direct	0.453	0.555	0.515	0.622	0.515
rn_rbf1	0.476	0.529	0.514	0.606	0.514

Table B.8: Individual performance of the optimal learners (for ensemble) on the Test data (unseen)

Bibliography

C. Abbet, M. M'hamdi, A. Giannakopoulos, R. West, A. Hossmann, M. Baeriswyl, and C. Musat. Churn intent detection in multilingual chatbot conversations and social media, 2018.

R. A. R. Ali, U. Farooq, U. Arshad, W. Shahzad, and M. F. Beg. Hate speech detection on twitter using transfer learning. *Computer Speech & Language*, 2022. doi: 10.1016/j.csl.2022.101365.

S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee. Deep learning models for multilingual hate speech detection. *CoRR*, abs/2004.06465, 2020. URL `https://arxiv.org/abs/2004.06465`.

P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, 2017.

S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6):121, 2021.

R. Cao, R. K.-W. Lee, and T.-A. Hoang. Deephate: Hate speech detection via multi-faceted text representations. In *12th ACM Conference on Web Science (WebSci '20)*, pages 11–20, 2020. doi: 10.1145/3394231.3397890.

D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 13–22, 2017.

C. Chen, R. Jafari, and N. Kehtarnavaz. Fusion of depth, skeleton, and inertial data for human action recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2712–2716. IEEE, 2016.

Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 71–80, 2012.

Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95, 2017.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. D. Choudhury, and D. Yang. Latent hatred: A benchmark for understanding implicit hate speech, 2021.

A. K. Engel, D. Senkowski, and T. R. Schneider. Multisensory integration through neural coherence, 2012. URL `http://europepmc.org/books/NBK92855`.

M. Fang, S. Peng, Y. Liang, C.-C. Hung, and S. Liu. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82:104561, 2023. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2022.104561. URL `https://www.sciencedirect.com/science/article/pii/S1746809422010151`.

F. Fernández-Martínez, D. Griol, Z. Callejas, and C. Luna-Jiménez. An approach to intent detection and classification based on attentive recurrent neural networks. *Proceedings of the IberSPEECH*, pages 46–50, 2021.

E. Fersini, F. Gasparini, and S. Corchs. Detecting sexist meme on the web: a study on textual and visual cues. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 226–231. IEEE, 2019.

E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, and J. Sorensen. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, 2022.

P. Fortuna and S. Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

A. Franco, A. Magnani, and D. Maio. A multimodal approach for human activity recognition based on skeleton and rgb data. *Pattern Recognition Letters*, 131:293–299, 2020.

B. Gambäck and U. K. Sikdar. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, 2017.

Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020.

N. D. Gitari, Z. Zuping, H. Damien, and J. Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

R. Gomez, J. Gibert, L. Gomez, and D. Karatzas. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478, 2020.

F. González-Pizarro and S. Zannettou. Understanding and detecting hateful content using contrastive learning. *arXiv preprint arXiv:2201.08387*, 2022.

J. Gutierrez-Gallego, S. Martín, and V. Rodriguez. Human stability assessment and fall detection based on dynamic descriptors. *IET Image Processing*, pages n/a–n/a, 06 2023. doi: 10.1049/ipr2.12847.

M. F. Hasani, F. L. Gaol, B. Soewito, and H. L. H. S. Warnars. Deep learning and threshold probability for out of scope intent detection in task oriented chatbot. In *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, pages 322–327, 2022. doi: 10.1109/AiDAS56890.2022.9918764.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

M. S. Hee, R. K.-W. Lee, and W.-H. Chong. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*, pages 3651–3655, 2022.

J. Imran and B. Raman. Evaluating fusion of rgb-d and inertial sensors for multimodal human action recognition. *Journal of Ambient Intelligence and Humanized Computing*, 11:189–208, 2020.

P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

M. S. Jahan and M. Oussalah. A systematic review of hate speech automatic detection using natural language processing, 2021.

A. Jain, P. R. Samala, P. Jyothi, D. Mittal, and M. K. Singh. Perturb, predict & paraphrase: Semi-supervised learning using noisy student for image captioning. In *IJCAI*, pages 758–764, 2021.

E. Jangam and C. S. R. Annavarapu. A stacked ensemble for the detection of covid-19 with high recall and accuracy. *Computers in Biology and Medicine*, 135:104608, 2021.

X. Jia, K. Han, Y. Zhu, and B. Green. Joint representation learning and novel category discovery on single- and multi-modal data, 2021.

P. Khaire, P. Kumar, and J. Imran. Combining cnn streams of rgb-d and skeletal data for human activity recognition. *Pattern Recognition Letters*, 115:107–116, 2018.

D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *ArXiv (Cornell University)*, 2020.

D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, C. Fitzpatrick, P. Bull, G. Lipstein, T. Nelli, R. X. Zhu, N. Muennighoff, R. Velioglu, J. Rose, P. Lippe, N. Holla, S. Chandra, S. Rajamanickam, G. Antoniou, E. Shutova, and D. Parikh. The hateful memes challenge: Competition report. In *NeurIPS 2020 Competition and Demonstration Track*, pages 344–360, 2021.

J. Kong, S. Wang, M. Jiang, and T. Liu. Multi-stream ternary enhanced graph convolutional network for skeleton-based action recognition. *Neural Computing and Applications*, pages 1–18, 06 2023. doi: 10.1007/s00521-023-08671-1.

B. Koonce. *ResNet 50*, pages 63–72. Apress, Berkeley, CA, 2021. ISBN 978-1-4842-6168-2. doi: 10.1007/978-1-4842-6168-2_6. URL `https://doi.org/10.1007/978-1-4842-6168-2_6`.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

G. K. Kumar and K. Nandakumar. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features, 2022.

L. H. Li, M. Yatskar, D. Yin, C. J. Hsieh, and K. W. Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 13, pages 740–755. Springer International Publishing, 2014.

J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 2019.

S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152, 2019.

I. Markov, N. Ljubešić, D. Fišer, and W. Daelemans. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, 2021.

L. Mathias, S. Nie, A. Mostafazadeh Davani, D. Kiela, V. Prabhakaran, B. Vidgen, and Z. Waseem. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 201–206, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.woah-1.21. URL `https://aclanthology.org/2021.woah-1.21`.

U. Mattei and M. Bussani. The project - delivered at the first general meeting on july 6, 1995 - the trento common core project. In *The Common Core of European Private Law*, Turin, Italy, 2010. Common Core Organizing Secretariat, The International University College of Turin. URL `http://www.coe.int/t/dghl/standardsetting/procivilsoc/Common%20Core%20Project%20-%20Presentation%20by%20Ugo%20Mattei%20and%20Mauro%20Bussani.pdf`.

Y. Mehdad and J. Tetreault. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, 2016.

M. R. Mercier and C. Cappe. The interplay between multisensory integration and perceptual decision making. *NeuroImage*, 222:116970, 2020.

S. M. Mohammad and P. D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

S. Mopidevi, P. Kishore, M. Prasad, and D. Anil Kumar. Meta triplet learning for multiview sign language recognition. *International Journal of Intelligent Engineering and Systems*, 16:2023, 06 2023. doi: 10.22266/ijies2023.0430.30.

N. Muennighoff. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*, 2020.

I. M. A. Niam, B. Irawan, C. Setianingsih, and B. P. Putra. Hate speech detection using latent semantic analysis (lsa) method based on image. In *2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC)*, pages 166–171, 2018. doi: 10.1109/ICCEREC.2018.8712111.

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

E. Papegnies, V. Labatut, R. Dufour, and G. Linares. Graph-based features for automatic online abuse detection. In *International Conference on Statistical Language and Speech Processing*, pages 70–81. Springer, 2017.

J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, and I. Androutsopoulos. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*, 2020.

K. Perifanos and D. Goutsos. Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*, 5(7):34, 2021. doi: 10.3390/mti5070034.

D. Price. *Little Science, Big Science*, volume 149. Columbia University Press, New York, 1963.

B. P. Putra, B. Irawan, C. Setianingsih, A. Rahmadani, F. Imanda, and I. Z. Fawwas. Hate speech detection using convolutional neural network algorithm based on image. In *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, pages 207–212, 2022. doi: 10.1109/ISMODE53584.2022.9742810.

D. Reilly, A. Chadha, and S. Das. Seeing the pose in the pixels: Learning pose-aware representations in vision transformers. 06 2023.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. URL `http://arxiv.org/abs/1409.0575`.

O. Sagi and L. Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL `https://aclanthology.org/P19-1163`.

S. Scheliga, T. Kellermann, A. Lampert, R. Rolke, M. Spehr, and U. Habel. Neural correlates of multisensory integration in the human brain: an ale meta-analysis. *Reviews in the Neurosciences*, 34(2):223–245, 2023. doi: doi:10.1515/revneuro-2022-0065. URL `https://doi.org/10.1515/revneuro-2022-0065`.

A. Schmidt and M. Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, 2017.

B. Settles. Active learning literature survey. 2009.

A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis, 2016.

C. Sharma, D. Bhageria, W. Scott, S. PYKL, A. Das, T. Chakraborty, V. Pulabaigari, and B. Gamback. Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor! *arXiv preprint arXiv:2008.03781*, 2020.

P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, July 2018.

V. K. Singh, S. Ghosh, and C. Jose. Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2090–2099, 2017.

A. A. SK, P. MVD, K. PVV, et al. Pose based multi view sign language recognition through deep feature embedding. *International Journal of Intelligent Engineering and Systems*, 16:2023, 06 2023. doi: 10.22266/ijies2023.0630.02.

spaCy. English multi-task cnn trained on ontonotes, 2017. URL `https://spacy.io/models/en`. Computer software.

G. R. Stone. Hate speech and the u.s. constitution. *East European Constitutional Review*, 3:78–82, 1994. URL `https://web.archive.org/web/20180427213042/http://www.law.nyu.edu/sites/default/files/upload_documents/HateSpeech_001.pdf`.

W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

S. T and S. Mathew. The disaster of misinformation: a review of research in social media. *International Journal of Data Science and Analytics*, 13:1–15, 05 2022. doi: 10.1007/s41060-022-00311-6.

A. Tontodimamma, E. Nissi, A. Sarra, and L. Fontanella. Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics*, 126:157–179, 2021.

I. R. Turkerud and O. J. Mengshoel. Image captioning using deep learning: Text augmentation by paraphrasing via backtranslation. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 01–10, 2021. doi: 10.1109/SSCI50451.2021.9659834.

D. van Mill. Freedom of Speech. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition, 2021.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

R. Velioglu and J. Rose. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*, 2020.

T. F. Y. Vicente, L. Hou, C.-P. Yu, M. Hoai, and D. Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 816–832. Springer, 2016.

E. Volokh. No, there's no "hate speech" exception to the first amendment. *The Washington Post*, May 2015. URL `https://www.washingtonpost.com/news/volokh-conspiracy/wp/2015/05/05/no-theres-no-hate-speech-exception-to-the-first-amendment/`.

Y. Wang, Y. Xie, J. Zeng, H. Wang, L. Fan, and Y. Song. Cross-modal fusion for multi-label image classification with attention mechanism. *Computers and Electrical Engineering*, 101:108002, 2022. ISSN 0045-7906. doi: https://doi.

org/10.1016/j.compeleceng.2022.108002. URL `https://www.sciencedirect.com/science/article/pii/S0045790622002701`.

Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, 2016.

D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 2022.

W. Yin and A. Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions, 2021.

F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020.

W. Zhang. Scene context-aware graph convolutional network for skeleton-based action recognition. 06 2023. doi: 10.21203/rs.3.rs-2978684/v1.

Z. Zhang, D. Robinson, and J. Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference*, pages 745–760, 2018.

R. Zhu. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint*, 2020.