Towards resource inequities in catching the "dark side" of social media: A hateful meme classification framework for low-resource scenarios

Yuming Li
University of Auckland
yuming.li@auckland.ac.nz

Johnny Chan
University of Auckland
jh.chan@auckland.ac.nz

Gabrielle Peko
University of Auckland
g.peko@auckland.ac.nz

David Sundaram
University of Auckland
d.sundaram@auckland.ac.nz

Abstract

The increasing prevalence of social media platforms has led to the emergence of multimodal information such as memes. Hateful memes poses a risk by perpetuating discrimination, reinforcing stereotypes, and causing online harassment, thereby marginalising certain groups and impeding efforts towards inclusivity and social justice. Detecting hateful memes is crucial for creating a safe and equitable online environment. However, existing research heavily relies on complex and large deep learning models, requiring substantial computational resources for training. This creates a barrier for under-resourced researchers and small companies, limiting their participation in hateful information detection and exacerbating inequalities in the field of artificial intelligence. This paper attempts to tackle the problem by proposing a low-resource-oriented framework of hateful meme classification to address limitations in training data, computing power, and modality integration. Our approach achieves faster performance with reduced computational requirements, while maintaining a 94.7% accuracy comparable to the existing highest-scoring model.

Keywords: Hateful meme classification, knowledge distillation, data augmentation, deep learning, low-resource NLP.

1. Introduction

In light of the exponential growth and pervasive influence of information and communication technologies (ICT), a wide range of people are opting for self-disclosure, communication, and interaction on social media platforms. With the upgrades of these platforms, the information carriers have gradually expanded from single-modal (i.e., text) to multimodal (i.e., text, image, video, audio, etc.). One of the most typical multimodal cases is meme. These units, which carry symbolic meaning representing a particular phenomenon or theme, are often used to replace plain text to express emotions more abundantly. Some of them are purely emotional expositions, while others are insulting. Compared to text-based harmful information, image-based meme is often more impactful and visual (Du et al., 2020). The meme can perpetuate discrimination, reinforce stereotypes, and contribute to online harassment and cyberbullying, which results in the marginalisation of certain groups and undermines efforts for inclusivity and social justice. For instance, a meme that uses derogatory language and imagery to mock the religious practices of a certain faith community can marginalise that group by perpetuating prejudice and fostering a hostile environment for its members. Therefore, detecting hateful meme is an important task in building a bright Internet, and creating a safe and equitable digital space for marginalised communities. It has also attracted widespread attention in recent years (Fharook et al., 2022; Sharma et al., 2022).

On the task of multimodal harmful information classification, existing research predominantly uses deep learning-based models and multimodal approach (Aggarwal et al., 2021). However, with the advent of the era of large models, deep learning models have been increasing in complexity and size over the years, leading to a greater demand for computational resources during training. The size of the model is growing linearly, especially the recent Megatron NLG proposed by Microsoft and Nvidia has exceeded 500 billion of parameters (Smith et al., 2022). To train large-scale models, researchers and organizations often rely on high-performance computing clusters or specialised hardware, such as graphics processing units (GPUs) or tensor processing units (TPUs). The development and optimisation of GPU and TPU architectures involve significant research and development costs due to the complex circuitry, large memory capacity, and power consumption. Even though the emergence of the pre-training model can skip the resource-intensive training from scratch, which saves a certain degree of computing resources than the initial training model, the pre-training model has limitations in domain-specific knowledge. So it still requires certain computational resources, a sufficient amount of high-quality labelled data for further fine-tuning. This eventually leads to the

URI: https://hdl.handle.net/10125/107251
978-0-9981331-7-1

phenomenon of 'deep learning, deep pockets' (Borge, 2022), so that the artificial intelligence (AI) industry will also show a Matthew effect, that is, companies or individuals with more basic resources or willing to invest more costs tend to accumulate and lead to further advantages or success in AI area (Engström & Strimling, 2020). This results in under-resourced researchers or small-scale start-up companies being disadvantaged and marginalised before engaging in harmful information detection tasks, thereby setting a high-cost entry ticket for brainstorming in the field of artificial intelligence.

As individual researchers, beginners, small-scale research institutions or small to medium-sized companies, the inevitable dilemma in the above-mentioned background and trend is the problem of low-resources. These low-resource entities also have a need to identify harmful memes; the most direct reason being to participate in corresponding competition challenges (e.g. Facebook hateful memes challenge) to hone skills or enhance their reputation. In addition, there are some research needs, or reputation needs for individual researchers, small sales platforms and startup brands. This low-resource problems in deep learning arise on account of the high computational demands, limited memory capacity, restricted hardware availability, scalability challenges, and insufficient labelled data. For example, training large models can take weeks or months, and generating predictions in real-time applications can be time-consuming. For individual researcher, using private device to train large models often occurs out of memory (OOM) and out of resource (OOR) error because of insufficient computing power. Another example is the inadequate training data in some niche fields where only a small amount of data can be collected, and some professional fields with high data collection costs and high manual labelling costs.

Therefore, in this paper, we focus on hateful meme classification in limited resource scenarios and propose a low-resource-oriented multimodal information classification framework. The low-resource problem we set out to solve is mainly divided into three aspects. The first aspect is the low-resource of knowledge. Since meme belongs to multimodal data including two modalities of text and image, the processing and analysis of meme requires knowledge in both the field of natural language processing (NLP) and computer vision (CV). As two different domains in deep learning, they belong to different research directions in academia (Goodfellow et al., 2016). Therefore, at the individual level of researchers, it is more difficult to possess and master the domain knowledge of both NLP and CV than to master one of them. At the industry level, it is more difficult to find experts who have both NLP and CV knowledge than to find experts who specialise in one of them. While hiring employees or experts in the two fields at the same time will undoubtedly increase time costs and capital costs. Therefore, in our framework, we convert the image part in the meme into text descriptions based on the image caption technology, thereby transforming the hateful meme recognition task into a pure NLP task, which reduces the resource and knowledge requirements to a certain extent. The second aspect is the low-resource of training data, insufficient training data can lead to poor generalisation and performance of the model. In such cases, the model may struggle to learn complex patterns or exhibit overfitting, which is particularly pronounced in scenarios with limited labelled data or in specialised domains where data collection is challenging. In our proposed framework, we introduce data augmentation techniques, that artificially expands the training dataset by applying various transformations to existing data samples, improving the diversity of the training data and the robustness of the model. In terms of implementation, we adopt a three-level data augmenter to perform data augmentation on limited training data in parallel at the character-level, word-level, and sentence-level. The third aspect is the low-resource of computational power. Training large-scale deep learning models, even fine-tuning on existing pretrained models, can indeed require significant computational resources. These models often demand powerful hardware such as high-end GPUs or even specialised hardware like TPUs to train effectively. Limited resources can result in OOM or OOR errors, slow training processes, or the inability to train models of a desired size. Thus, our proposed model is based on knowledge distillation technology for fine-tuning process. Through this technique, researchers can train smaller models that still capture the knowledge and performance of larger models, leveraging the benefits of pretrained models even with limited computational resources. The objective is to create a more efficient model that maintains or improves performance while reducing computational resources and memory requirements.

The following section delves into the literature on Cyberbullying and hateful meme detection, multi-modal data processing and low-resource approach. Next, in section 3, our proposed hateful meme classification framework for a low-resource scenario is described. Followed by the results and analysis in section 4. Section 5 concludes the paper.

2. Literature review

The Wikimedia foundation conducted a survey on network behaviour and states that 54% of users experience a decline in participation after experiencing malicious online violence (Wulczyn et al., 2017).

Additionally, according to the data report by the Cyberbullying Research Centre marginalised groups (such as transgender, LGBTQ+, and multiracial communities) are more susceptible to the occurrence of harmful information and cyberbullying. These groups are also more vulnerable and experience greater negative effects from such attacks (Li et al., 2022). Therefore, in order to maintain a safe, productive and equitable online environment, research on the identification of harmful information in social media is extremely necessary.

2.1. Hateful meme detection approach

Cyberbullying often involves the use of various forms of digital media, including meme, to target and harm individuals or groups (Pradhan et al., 2020). As meme continues to gain popularity as a means of communication on social platforms, it becomes crucial to develop effective techniques for detecting and mitigating hateful or offensive meme that perpetuates cyberbullying (Maity et al., 2022). In the task of identifying and detecting hateful meme, existing research mostly utilises multimodal approaches (Kiela et al., 2021). Shang et al. (2021) proposed a social media platforms offensive analogy meme detection framework based on multimodal learning. The framework specifically focuses on the implicit connection between visual and textual element in meme. The authors conducted experiments on real-world data they collected, demonstrating significant performance improvements compared to state-of-the-art baselines. Bhowmick et al. (2021) proposed a multimodal deep learning framework for derogatory social media post identification. The framework integrates text analysis, face recognition, and optical character recognition techniques. Through experiments, the research demonstrates the effectiveness of the framework in identifying offensive analogy meme about famous individuals on social media platforms. Additionally, this research expanded meme identification to the multilingual domain, and proved that the expansion of language categories is also an emerging key research direction. Singh et al., (Singh et al., 2022) focuses on the multimodal fusion techniques for hate meme classification. The research based on Facebook hateful meme challenge and achieved a 0.79 score with 63.3% model accuracy by combining image and text using early fusion with Inception v3 and BERT models.

2.2. Hateful meme detection trend

In terms of general trends, Hermida et al. (2023) reviewed the research on hateful meme detection, emphasising the importance of information size in hateful meme detection, suggesting that using complex models and extracting more data is considered a promising direction to improve the results of this task. Furthermore, in consideration of the aforementioned literature review, it is evident that existing research has placed excessive emphasis on the complexity and accuracy of models, while lacking consideration for computational resource constraints. The existing solutions to address the low-resource challenges in deep learning primarily revolve around knowledge transfer, active learning, data augmentation, and semi-supervised learning methods. These approaches have been applied in various specialised domains such as speech recognition (Yu et al., 2020), sentiment analysis (Kastrati et al., 2021), opinion mining (Al-Sallab et al., 2017), entity resolution (Kasai et al., 2019), etc. However, existing research predominantly focus on single-modal data, and research in the context of multi-modal data applications is relatively scarce. This trend is reflected in the increasing number of multimodal data processing research "surpass state-of-the-art methods" often requiring significant computational power. However, such research tends to originate from well-funded teams or companies, which poses a disadvantage for individual researchers and small-scale institutions.

3. The hateful meme classification framework

In this paper, we propose a hateful meme classification framework based on the work by Zhou et al. (2021) for a low-resource scenario, addressing the three low-resource problems defined in Section 1 (low-resource for knowledge, low-resource for training data, and low-resource for computational power). The framework is shown on Figure 1, which initially splits the input meme data into text and image components. Modality reduction is then performed by converting images into text and integrating the image-to-text output with the original text in the meme, thus addressing the low-resource for knowledge issue. Subsequently, the training data is augmented using character-level, word-level, and sentence-level augmenters in parallel, thus addressing the low-resource for training data issue. The best-performing augmenter is selected through evaluation to enhance the scarce training data, mitigating the low-resource for training data challenge. Finally, in the model training phase, the framework employs a lightweight model based on knowledge distillation techniques, avoiding extensive hardware and funding requirements, thereby resolving the low-resource for computational power issue. Detailed descriptions of each module will be presented in Sections 3.2-3.5.

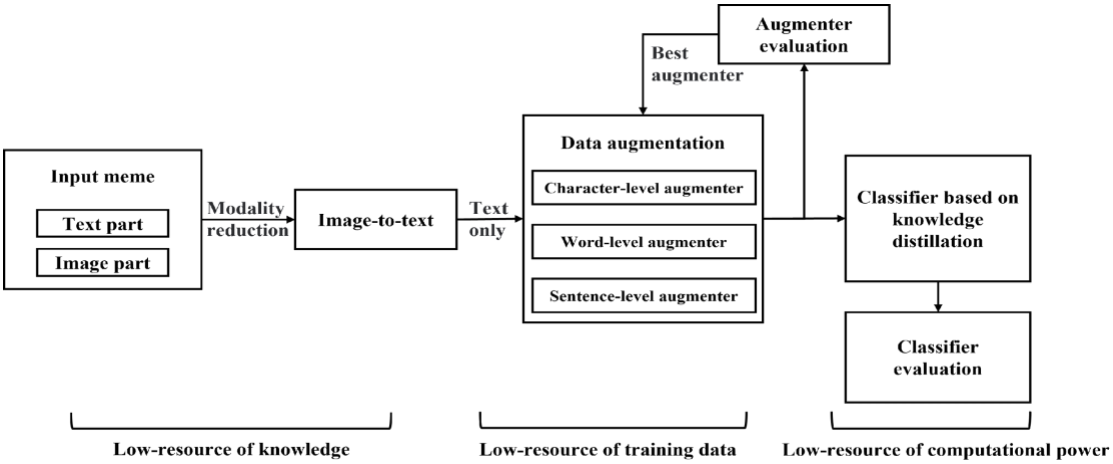

Figure 1. The hateful meme classification framework for low-resource scenario.

3.1. Data description

We use the Facebook hateful meme dataset (Kiela et al., 2020) for experiments, the data sample is shown in Figure 2. The training set and test set are JSON files that contain the corresponding image id, the relative position of the image in the folder, the label (1 is hateful, 0 is non-hateful) and the text part in each meme.

Figure 2. Data sample of Facebook hateful meme dataset.

The original dataset contains 10,000 meme data. Considering the low-resource case simulation, we only randomly select 5% (500 sampled data) as the training set for this experiment.

3.2. Image-to-text

Taking into account the low-resource nature of knowledge, specifically in multimodal data analysis, presents challenges for researchers and industries due to the scarcity of experts in both textual and image domains. Hence, in our proposed framework, we employ the image-to-text technique to convert the image part in meme into textual representations, thereby converting the multimodal task into a single modality task and alleviating the low-resource problem of knowledge. Image-to-text refers to the process of generating a textual sentence that describes the scene depicted in an input image. For example, a randomly selected sample in Facebook hateful meme dataset is shown in Figure 3. The text part of the meme "when is the time to eat i am hungry" is directly extracted text object from the Facebook hateful meme dataset, which includes image id, label and text from the JSON files. The corresponding caption "a pig is eating hay from a trough" will be generated through image-to-text. In this paper, we directly call the application programming interface (API) of HuggingFace (Wolf et al., 2019), with open-source libraries and models that can be readily accessed for free. The accessibility of HuggingFace allows individual researchers and developers with limited resources to leverage cutting-edge NLP models and tools, thus democratising access to cutting-edge NLP technologies, which makes it stand out among large language model APIs. Then we combined the text in the meme with the caption of the image through image-to-text technique, and finally get the combination "In the picture, a pig is eating hay from a trough. And the text says when is the time to eat i am hungry" as the data that is finally passed into the following data augmentation module.

The reason we chose text as the final unified modality because text processing generally requires fewer computational resources compared to image processing. In most cases, text data have lower-dimensional input data, simpler architectures, and lighter feature representations. Since text data is represented as sequences, while images have high-dimensional pixel data; and text models use word

embeddings or contextual embeddings, which requires fewer iterations and shorter training times than image features. In addition, utilising the API for tasks that do not excessively pursue precision in the initial stage such as image captioning, is faster than training a model from scratch. APIs offer efficient and scalable inference capabilities, leveraging optimised implementations and high-performance infrastructure.

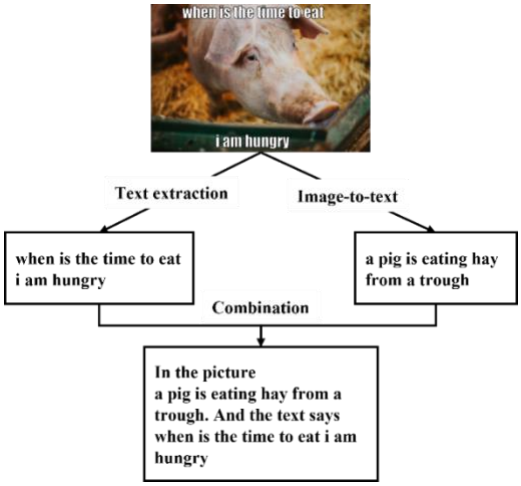

Figure 3. A random sample (04569.png) of Facebook hateful meme dataset.

3.3. Data augmentation

Data augmentation is a technique used in machine learning / deep learning to increase the diversity of training data by applying transformations such as rotations, translations, and distortions. It helps improve model generalisation, reduce overfitting, and enhance performance by artificially expanding the dataset without collecting new labelled examples. In the data augmentation module, we use character-level, word-level and sentence-level data augmenters to process data in parallel.

3.3.1. Character-level augmenter. Character-level NLP data augmentation focuses on manipulating the original text at the character level, such as replacing, deleting, inserting, or swapping characters, to generate new text. In this paper, we use random swap (RS) method as our character-level data enhancement approach. RS refers to a text data augmentation technique used to enhance the diversity of the training data by swapping 2 characters within a word randomly. It helps in creating new sentence variations while preserving the overall structure and meaning of the original sentence. For example, for the original sentence "The cat is sitting on the mat.", randomly exchange characters "a" and "t", and finally get Augmented Sentence: "The tac is sitting on the mat." while maintaining the original sentence's structure. This augmentation technique introduces small perturbations that can help improve the generalisation of natural language processing models. The advantage of character-level augmentation is its simplicity and its ability to enhance the robustness of models to noisy text, such as text generated by social media users. It can handle common noise sources like typos or leetspeak to some extent. However, a drawback is that the generated data may have lower quality due to introduction of syntactic or semantic errors during augmentation. Therefore, additional restrictions and evaluations are necessary to control the degree of augmentation and ensure data quality.

3.3.2. Word-level augmenter. Word-level data augmentation refers to the technique that operate at the individual word level to expand textual data, like word replacement, insertion, deletion, swap or perturbation. Word-level data augmenter is usually relied on WordNet (Miller, 1995) or the paraphrase database (PPDB) (Ganitkevitch et al., 2013) as semantic support. In this paper, we replace words with synonyms based on WordNet, a lexical database that provides semantic relationships between words and organises English words into a series of synsets. For example, for the original sentence "The dog chased the ball.", word "chased" is replaced with its synonym "pursued." in WordNet, so as to get augmented sentence "The dog pursued the ball." Word-level augmenter addresses the semantic errors associated with character-level data augmenter, as the generated words are correct and existing. However, since WordNet does not capture contextual or domain-specific meanings of words, when a word has multiple meanings, such as "apple" referring to both a fruit and a company, WordNet based synonym replacement may introduce ambiguity.

3.3.3. Sentence-level augmenter. Sentence-level augmentation refers to the technique that utilises pre-trained contextual word embeddings models to generate augmented sentences. The pretrained language models can capture the contextual information and meaning of words in a sentence. For example, for the original sentence “I love to go hiking in the mountains.”, sentence-level data augmenter modifies the sentence context and uses the same pre-trained model to decode and transform it into augmented sentence: "I enjoy hiking in the beautiful mountains." Unlike character-level and word-level augmenters that operate on a per-character or per-word basis, sentence-level augmenters often retain the original language context but generate new sentences with different structures. This increases the flexibility and diversity of training data, contributing

to improved model generalisation. Additionally, given that the pre-trained language model can capture rich linguistic knowledge and context, using sentence-level data augmenter can avoid the ambiguity dilemma in word-level data augmenter. However, compared with character-level and word-level data augmenter, the computational complexity and time consumption of sentence-level augmenter is slightly larger.

3.4. Knowledge distillation

With the development of the NLP pre-training model in recent years, the size and parameters has become larger, and has resulted in implementation difficulties due the limitations of computing power. Especially for individual researchers or small companies, the deployment of large models requires high equipment resources and slow inference speed, which often creates budgetary and computing power bottlenecks. In order to alleviate this situation, the concept of knowledge distillation was proposed by Hinton and Salakhutdinov (2006). As a method of model compression, knowledge distillation refers to using a small model (student model) to learn the output of a large or ensemble model (teacher model), and finally it is the flexible and lightweight student small model that is actually deployed online for prediction tasks. Its purpose is to transfer the knowledge learned from a large model or multiple model ensembles to another lightweight model. The structure of knowledge distillation is shown in Figure 4.

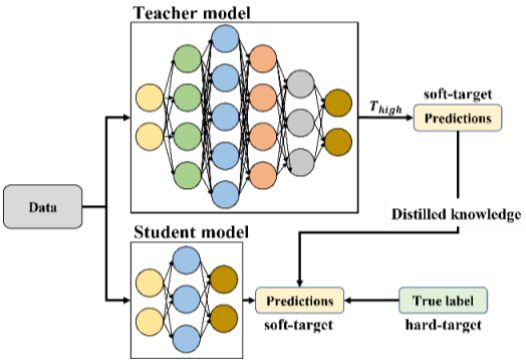

Figure 4. Structure of knowledge distillation.

In the context of general supervised learning for text classification, the deep neural network undergoes nonlinear transformations, resulting in the logit z_i representing the unnormalised output of the input text belonging to class i. Subsequently, the $Softmax$ function is applied to normalise the logits, converting them into probabilities that distinguish and represent the differences between different classes. Finally, the probability p_i of the input text belonging to a certain category i is obtained through:

$$p_i = \frac{exp\ (z_i)}{\sum_j exp\ (z_j)}$$

In knowledge distillation, a new hyperparameter T is added to the calculation of $Softmax$, thus the calculation of class probabilities is improved as:

$$p_i^D = \frac{exp\ (z_i/T)}{\sum_j exp\ (z_j/T)}$$

When $T = 1$, it represents the original $Softmax$. In traditional neural network training, the original dataset for text binary classification tasks often uses one-shot labels, where positive labels are represented as 1 and negative labels as 0, known as hard-target labels. However, apart from positive examples, negative labels also carry valuable information about the inductive reasoning of the model, such as certain negative labels having much higher probabilities than others. Hence, traditional hard-target training process, which treats all negative labels uniformly, can result in information loss.

Knowledge distillation approach uses the $Softmax$ distribution generated by the Teacher model at a high temperature T_{high} as soft-target, where each class is assigned a probability and the positive label has the highest probability. As temperature T increases, the output probability distribution of $Softmax$ becomes smoother, with a higher entropy, amplifying the information carried by negative labels and making the model focus more on them. When training the Teacher model using soft-target, the Student model can quickly learn the reasoning process of Teacher model. Compared to traditional hard-target training, soft-target provides more information to the Student model, and when the soft-target distribution has higher entropy, it contains richer knowledge. Moreover, when training with soft-target, the gradient variance is smaller, allowing for larger learning rates, requiring fewer samples, and enhancing generalisation capabilities.

In this paper we choose DistilBERT (Sanh et al., 2019) as the classifier after data augmentation. DistilBERT is a compact version of BERT (Devlin et al., 2018) using knowledge distillation techniques. It incorporates the concept of soft-targets from the Teacher model as part of the total loss function to guide the training of the Student model. This approach enables knowledge transfer from the BERT model, resulting in a reduction in model size by 40% while retaining 97% of its performance. Compared to the traditional BERT model, DistilBERT reduces the number of network layers to construct the Student model, eliminates the token type embedding and pooler, and leverages the soft-targets generated by the Teacher model as well as the hidden layer parameters of Teacher model to train the Student model. DistilBERT enables compressing the

model size and improving inference speed without significant loss in accuracy, making it more suitable for low-resource cases.

3.5. Experiment setting

In the experiment, we employed the Facebook hateful meme dataset and conducted a binary classification task to determine the harmfulness of meme, following the guidelines of the Facebook Hateful meme classification challenge. In order to simulate a low-resource environment, we randomly sampled only 5% of the training set as the training data for this paper. Subsequently, we applied data augmentation techniques to the sampled small dataset and evaluated the quality of the augmented data. We then compared the performance of widely used models in the field of natural language processing, using the area under the receiver operating characteristic (AUROC) as the evaluation metric. AUROC quantifies the classifier's ability to distinguish between different classes by measuring the area under the ROC curve, providing a comprehensive evaluation criterion. We also compared the average time in a training epoch, system RAM, and GPU RAM of the models to assess their resource consumption in terms of time and computational power. In the following section we review the results of the experiment using the proposed hateful meme classification framework shown in Figure 1.

4. Result and discussion

In our evaluation, we compared the accuracy of the three levels of augmenter as the evaluation criteria for the quality of the generated augmented data. In this paper we randomly sample a small amount (20%) of augmented data to train a simple logistic regression model and obtain the accuracy from the test on the original data. From Table 1, it can be concluded that the quality of sentence-level augmenter is better than that of character-level augmenter and word-level augmenter. Because the sentence-level augmenter preserves the overall structure and meaning of the original sentence while generating new variations. it also provides greater flexibility and variability than character-level and word-level augmenter since it leverages the rich language knowledge and contextual understanding as well as avoids the ambiguity issues that may arise from word-level augmenter.

Table 1. The evaluation of data augmentation.

Augmenter	Accuracy
Character-level augmenter	0.6792
Word-level augmenter	0.7075
Sentence-level augmenter	0.7642

Subsequently, we performed classification (hateful or not hateful) on the augmented texts. In selecting the classifier, we referred to the latest survey on natural language processing (P. Liu et al., 2023) and compared the most popular text classification models based on BERT (Devlin et al., 2018), GPT (Radford et al., 2019), and XLNet (Yang et al., 2019). Considering that BERT has the fewest parameters among these models (around 110 million parameters for BERT-based, around 117 million parameters for GPT-2 Small, and around 116 million parameters for XLNet Base), we focused on BERT-based variants in our model comparison. Among them, we selected ALBERT (Lan et al., 2019) and DistillBERT (Sanh et al., 2019) as lightweight optimisation variants of BERT, and RoBERTa (Liu et al., 2019) for robustness optimisation. ALBERT improves upon BERT by reducing parameter size while maintaining or surpassing its performance on various NLP tasks, and DistillBERT is a smaller and faster version of BERT that achieves similar performance with fewer parameters and computational resources.

In Table 2, we observe a significant improvement in the effectiveness of data augmentation, with almost a doubling of accuracy in the comparative experiments. For instance, in the case of insufficient training with the original dataset, the final accuracy achieved by fine-tuning the BERT classifier is only 0.4643. However, with Sentence-level data augmentation, the accuracy can reach as high as 0.8856, which is the best result in the comparative experiments. However, from a computational resource perspective, DistilBERT exhibits significantly lower time consumption. The average time per epoch for DistilBERT is only 1 second, which is one-fourth of BERT and one-ninth of RoBERTa's. Additionally, we recorded the usage of system RAM and GPU RAM. System RAM is the main memory utilised by the CPU to store and access data during model training. It stores the model architecture, input data, and gradients computed during backpropagation, which are used to update the model's parameters during optimisation. GPU RAM, on the other hand, is the dedicated memory on the GPU that stores intermediate computations during deep learning model training. It is used to store the model's parameters, activations, and gradients during forward and backward passes. We also observe that DistilBERT requires the least amount of system RAM and GPU RAM during training. DistilBERT utilises 74% of the system RAM compared to BERT and 52% of the GPU RAM compared to BERT while maintaining a 97% accuracy after sentence-level data augmentation. Hence, we can conclude that DistilBERT exhibits excellent cost-effectiveness in low-resource environments.

After selecting DistillBERT as the classifier based on the result in Table 2, we compared the top 5 results from the 2020 Facebook hateful meme classification challenge on DataDriven, as well as new approaches after 2020. Since these top 5 methods are all based on multimodal approach, and our proposed method only based on single modality (text), this fundamental difference in task definition prevents us from directly comparing them with our model in terms of average epoch time and RAM as shown in Table 2. Therefore, in Table 3, we chose to focus on the ultimate goal of the challenge (AUROC score on the test set) and compared the final scores of our low-resource harmful meme detection model with the top 5 models from the challenge to demonstrate the feasibility of our model in practice. From Table 3, we can find that the initial score is very low, only 0.4430, because insufficient training data hampers pattern learning, leading to suboptimal generalisation and increasing risk of overfitting. However, after data augmentation, the final score of the model increased to 0.8616, maintaining 94.7% of the performance of the existing highest scoring model, but using less training data, computing resources and training time.

Table 2. The comparison results of different language model with augmenter on hateful text classification.

Model & criteria	**Data augmentation level**			
	Original	**Character-level**	**Word-level**	**Sentence-level**
BERT (Devlin et al., 2018)				
ACC	0.4643	0.7633	0.8757	0.8856
Ave time/epoch	4s	8s	8s	8s
System RAM	5.0 / 12.7 GB	*GPU RAM*	4.6 / 15.0 GB	
RoBERTa (Y. Liu et al., 2019)				
ACC	0.4219	0.8696	0.8759	0.8580
Ave time/epoch	9s	18s	18s	18s
System RAM	4.9 / 12.7 GB	*GPU RAM*	5.8 / 15.0 GB	
ALBERT (Lan et al., 2019)				
ACC	0.4844	0.8723	0.8089	0.8777
Ave time/epoch	9s	18s	18s	20s
System RAM	4.4 / 12.7 GB	*GPU RAM*	5.7 / 15.0 GB	
DistilBERT (Sanh et al., 2019)				
ACC	0.4430	0.8239	0.8553	0.8616
Ave time/epoch	1s	4s	4s	4s
System RAM	3.7 / 12.7 GB	*GPU RAM*	2.4 / 15.0 GB	
GPT-2 (Radford et al., 2019)				
ACC	0.5113	0.7132	0.5660	0.5585
Ave time/epoch	9s	13s	12s	12s
System RAM	7.3 / 12.7 GB	*GPU RAM*	9.0 / 15.0 GB	
XLNet (Yang et al., 2019)				
ACC	0.5119	0.9172	0.9053	0.8876
Ave time/epoch	5s	10s	10s	10s
System RAM	4.8 / 12.7 GB	*GPU RAM*	5.2 / 15.0 GB	

Tables 1 and 2 highlight our solutions for low-resource challenges in hateful meme detection. Table 1 shows improved accuracy through data augmentation. Table 2 reveals DistilBERT reduces computational needs. While solutions for 'low training data' and 'low computational power' are validated. An intriguing phenomenon is observed in Table 3. In the proposed framework in this paper, we use image-to-text technology to transform multimodal meme into a single-modality, namely textual representation. It stands to reason that the information loss caused by the dimensionality reduction of image information should reduce the final score. Nevertheless, the experimental results show that this simplified low-resource approach paradoxically gets a higher final score than most multimodal methods. This is attributed to the fact that textual data often offers more detailed and explicit information compared to image data, and the focus of text is easier to capture than images. Additionally, traditional multimodal classification often involves

complex fusion methods to integrate information from diverse modalities, which introduces challenges in feature representation and model training. In contrast, converting the image into text homogenises the input data, thus avoiding additional losses caused by modality fusion.

Table 3. Comparison of final scores on the overall task of hateful meme classification.

Source	Approach	Score
Baseline	DistillBERT	0.4430
Top 5 result of Facebook Hateful meme classification challenge on DataDriven	VL-BERT based (Zhu, 2020)	0.8450
	Vilio based (Muennighoff, 2020)	0.8310
	VisualBERT based (Velioglu & Rose, 2020)	0.8108
	UNITER-based (Lippe et al., 2020)	0.8053
	Visual-linguistic Transformer based (Sandulescu, 2020)	0.7943
Recent new approach	PromptHate based (Cao et al., 2023)	0.9096
	Hate-CLIPper based (Kumar & Nanadakumar, 2022)	0.8580
	ERNIE-VisualBERT based (Leyva Massagué, 2022)	0.8424
Our approach	***DistillBERT + augmenter***	***0.8616***

5. Conclusion

Online platforms and multimodal information, like memes, have proliferated, making hateful information detection crucial. While memes add entertainment value to online interactions, some propagate discrimination, bias, and aggression, underscoring the need for a safe, inclusive online environment. Current research often employs deep learning techniques with large neural networks, necessitating significant computational, data, and knowledge resources. Such complexity is problematic for individuals and entities with limited resources. This paper introduces a low-resource framework for hateful meme detection, addressing challenges of limited training data, computing power, and knowledge gaps between modalities. Using image-to-text conversion, data augmentation, and knowledge distillation, our model efficiently achieves a 94.7% accuracy of top-performing models. Through experiments, we validate the data augmentation module's effectiveness in enhancing model performance under data scarcity. Moreover, our knowledge distillation-based model optimises training time and system resource usage without compromising accuracy. By offering this framework, we hope to enable under-resourced entities to engage actively in hateful meme detection, fostering a safer, equitable online space. This adaptable framework allows simple device utilisation for faster fine-tuning on limited datasets, ensuring effective classification of hateful memes or even other multimodal information types. Our results underline the benefits of data augmentation and knowledge distillation in delivering superior performance while navigating resource constraints.

6. References

Agrawal, C., Singh, D., Mishra, V., & Gritli, H. (2021). Two-way feature extraction using sequential and multimodal approach for hateful meme classification. Complexity, 2021, 1–7.

Al-Sallab, A., Baly, R., Hajj, H., Shaban, K. B., El-Hajj, W., & Badaro, G. (2017). Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *16*(4), 1–20.

Bhowmick, R. S., Ganguli, I., Paul, J., & Sil, J. (2021). A multimodal deep framework for derogatory social media post identification of a recognized person. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(1), 1–19.

Borge, N. J. (2022). *Deep pockets: The economics of deep learning and the emergence of new AI platforms*. Massachusetts Institute of Technology.

Cao, R., Lee, R. K.-W., Chong, W.-H., & Jiang, J. (2023). Prompting for Multimodal Hateful Meme Classification. *ArXiv Preprint ArXiv:2302.04156*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Du, Y., Masood, M. A., & Joseph, K. (2020). Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 153–164.

Engström, E., & Strimling, P. (2020). Deep learning diffusion by infusion into preexisting technologies–Implications for users and society at large. Technology in Society, 63, 101396.

Fharook, S., Ahmed, S. S., Rithika, G., Budde, S. S., Saumya, S., & Biradar, S. (2022, May). Are you a hero or a villain? A semantic role labelling approach for detecting harmful memes. In Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations (pp. 19-23).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2013). PPDB: The paraphrase database. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 758–764.

Hermida, P. C. de Q., & Santos, E. M. dos. (2023). Detecting hate speech in memes: A review. *Artificial Intelligence Review*, 1–19.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504–507.

Kasai, J., Qian, K., Gurajada, S., Li, Y., & Popa, L. (2019). Low-resource deep entity resolution with transfer and active learning. *ArXiv Preprint ArXiv:1906.08042*.

Kastrati, Z., Ahmedi, L., Kurti, A., Kadriu, F., Murtezaj, D., & Gashi, F. (2021). A deep learning sentiment analyser for social media comments in low-resource languages. *Electronics*, *10*(10), 1133.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., & Testuggine, D. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in Neural Information Processing Systems, 33, 2611–2624.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Fitzpatrick, C. A., Bull, P., Lipstein, G., Nelli, T., & Zhu, R. (2021). The hateful memes challenge: Competition report. *NeurIPS 2020 Competition and Demonstration Track*, 344–360.

Kumar, G. K., & Nanadakumar, K. (2022). Hate-CLIPper: Multimodal Hateful Meme Classification based on Cross-modal Interaction of CLIP Features. *ArXiv Preprint ArXiv:2210.05916*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *ArXiv Preprint ArXiv:1909.11942*.

Leyva Massagué, J. (2022). *Hybrid models for hateful memes classification*. Universitat Politècnica de Catalunya.

Li, Y., Chan, J., Peko, G., & Sundaram, D. (2022). *A Personalized Harmful Information Detection System Based on User Portraits*.

Lippe, P., Holla, N., Chandra, S., Rajamanickam, S., Antoniou, G., Shutova, E., & Yannakoudakis, H. (2020). A multimodal framework for the detection of hateful memes. *ArXiv Preprint ArXiv:2012.12871*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *ArXiv Preprint ArXiv:1907.11692*.

Maity, K., Jha, P., Saha, S., & Bhattacharyya, P. (2022). A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1739–1749.

Miller, G. A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38(11), 39–41.

Muennighoff, N. (2020). Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *ArXiv Preprint ArXiv:2012.07788*.

Pradhan, A., Yatam, V. M., & Bera, P. (2020). Self-attention for cyberbullying detection. *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, 1–6.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Sandulescu, V. (2020). Detecting hateful memes using a multimodal deep ensemble. *ArXiv Preprint ArXiv:2012.13235*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv Preprint ArXiv:1910.01108*.

Shang, L., Zhang, Y., Zha, Y., Chen, Y., Youn, C., & Wang, D. (2021). Aomd: An analogy-aware approach to offensive meme detection on social media. *Information Processing & Management*, *58*(5), 102664.

Sharma, S., Alam, F., Akhtar, M., Dimitrov, D., Martino, G. D. S., Firooz, H., Halevy, A., Silvestri, F., Nakov, P., & Chakraborty, T. (2022). Detecting and understanding harmful memes: A survey. ArXiv Preprint ArXiv:2205.04274.

Singh, B., Upadhyay, N., Verma, S., & Bhandari, S. (2022). Classification of hateful memes using multimodal models. In *Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2021* (pp. 181–192). Springer.

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., & Korthikanti, V. (2022). Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. ArXiv Preprint ArXiv:2201.11990.

Velioglu, R., & Rose, J. (2020). Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *ArXiv Preprint ArXiv:2012.12975*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., & Funtowicz, M. (2019). Huggingface's transformers: State-of-the-art natural language processing. ArXiv Preprint ArXiv:1910.03771.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*.

Yu, C., Kang, M., Chen, Y., Wu, J., & Zhao, X. (2020). Acoustic modeling based on deep learning for low-resource speech recognition: An overview. *IEEE Access*, *8*, 163829–163843.

Zhou, Y., Chen, Z., & Yang, H. (2021, July). Multimodal learning for hateful memes detection. In 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 1-6). IEEE.

Zhu, R. (2020). Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *ArXiv Preprint ArXiv:2012.08290*.