LDK 2023

Multimodal Offensive Meme Classification with Natural Language Inference

Shardul Suryawanshi[1] and **Mihael Arcan**[1] and **Suzanne Little**[2] and **Paul Buitelaar**[1]

[1,2] Insight SFI Research Centre for Data Analytics
[1] Data Science Institute, University of Galway, Ireland
[2] Dublin City University, Ireland
{shardul.suryawanshi, mihael.arcan, suzanne.little, paul.buitelaar}@insight-centre.org

Abstract

Multimodal offensive meme classification is a challenging classification task, where a multimodal meme needs to be classified as offensive or not offensive based on the provided text and image. A well-known approach to solving this problem is to fuse the text and image features captured either by the text and image encoder or by a transformer architecture to form a multimodal meme representation. In our work, we argue that the image features captured by the image encoder are unable to capture the abstract representation like language. Hence, we propose to transform the multimodal offensive meme classification task into an unimodal offensive text classification task for which we leverage the Natural Language Inference (NLI) task. Firstly, we carefully generate image captions using an off-the-shelf image captioner and automatically transcribed the meme as if it was explained to a visually impaired individual. Later, these meme transcriptions and labels (image-text-label) have been transformed into NLI format (premise-hypothesis-label). To evaluate our approach, we run benchmark analysis on Memotion, Hateful memes and MultiOFF datasets (in their NLI format) using four baselines finetuned on Emotion Analysis, Sentiment Analysis, Offensive tweet Classification, and NLI task. We achieve state-of-the-art (SOTA) results for the MultiOFF dataset and close to SOTA results for Memotion while achieving competent evaluation scores on the Hateful Memes dataset.

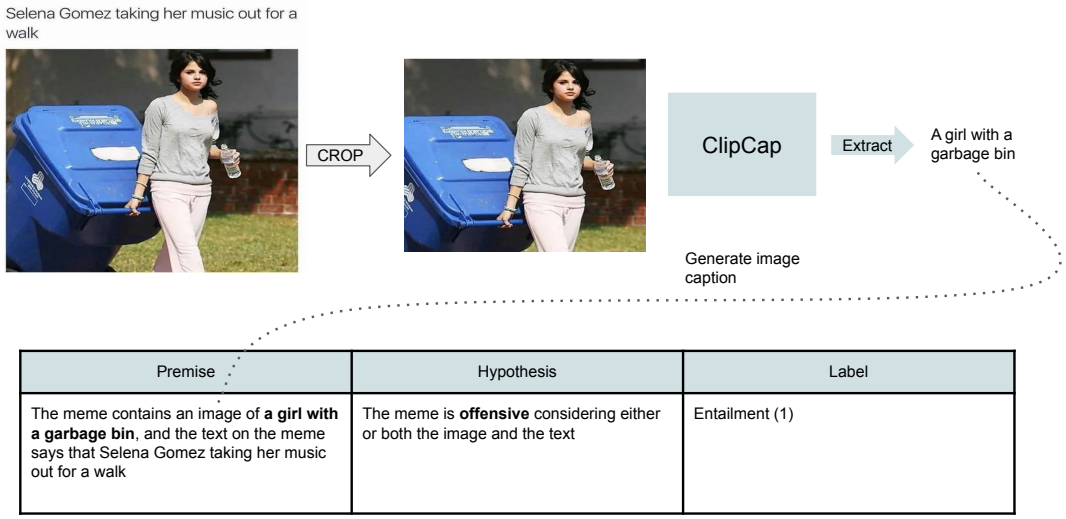
1 Introduction

Memes in the social media context are means of expressing emotions and ideas (Du et al., 2020). They easily propagate across various cultures due to their ability to mutate and spread (Dawkins, 2016). Hence, memes have become an integral part of online communication. But sadly, they have become the means of spreading hatefulness and offensiveness towards an individual or a group based on but not limited to their ethnicity, sexual orientation, and religion (Suryawanshi et al., 2020). A multimodal or Image-with-text (IWT) (Du et al., 2020) offensive meme contains an image embedded with the text with either an image or text or both being offensive. Hence, it is necessary to consider both the image and text modality for the multimodal offensive meme classification.

The multimodal offensive meme classification task (Suryawanshi et al., 2020; Sharma et al., 2020a; Kiela et al., 2020) is a classification task where one needs to classify if the meme is offensive based on the image and text modalities associated with the meme. The nature of the task is multimodal since both the image and text modalities are required for the classification. The research community has been actively organizing shared tasks (Sharma et al., 2020a; Suryawanshi and Chakravarthi, 2021) and competitions (Kiela et al., 2020) to solve this challenging task.

Previous research in this area proposed novel approaches that combined both the image and text modalities using deep learning techniques, most of which leverage VL pre-training, which involves a large corpus of image and text. However, VL pre-trained models are susceptible to domain shifts when finetuned on a small multimodal offensive memes dataset (Singh et al., 2020); Additionally, the quality of global multimodal representations learnt during the VL pre-training might degrade after finetuning on out-of-domain datasets (Singh et al., 2020).

Language is more abstract than image. It condenses information better than the image. For example, when we refer a word “cat", we could imagine cat from cartoons shows such as “Tom and Jerry", “Garfield" to a real world cat. The word in itself condenses all the information. A well documented human knowledge is in text which could be learnt from language models. On the other hand, if we consider the Selena Gomez meme from Figure



Premise	Hypothesis	Label
The meme contains an image of **a girl with a garbage bin**, and the text on the meme says that Selena Gomez taking her music out for a walk	The meme is **offensive** considering either or both the image and the text	Entailment (1)

Figure 1: Step by step method of NLIfication of the multimodal data: First the image is cropped to get rid of the text. Later, the image caption is generated using ClipCap which is incorporated in the meme transcription along with the text associated with the meme. The old label (OFF or NOT) translated into a hypothesis[a]. Lastly, the new label is assigned as 1 (entailment) if the old label was OFF else 0 is assigned.

[a]Hypothesis is identintical for each data sample

1 whereby her music is called out as rubbish, the meme is still offensive even after Selena Gomez is replaced by any other musician. In such case, a VL pre-trained will tune its parameter to accommodate multiple variations. But, by transforming the task to unimodal will reduce such variations. Because the meme captions generated for each such meme would be "a girl/boy/person with a garbage bin". Moreover, research by (Prajwal et al., 2019) shows that memes could be made accessible to a visually impaired individual through meme transcriptions. Hence, we propose to transform the multimodal offensive meme classification task into an unimodal offensive text classification task. Hence, we hypothesize that the transforming images into text could aid in our task.

In our research, we design a systematic framework that utilises NLI task to transform the multimodal task into the unimodal one. Firstly, we transform the data from image-text-label format into premise-hypothesis-label format. Later, the newly transformed data is used for finetuning three RoBERTa models previously finetuned on Emotion Analysis, Sentiment Analysis, Offensive tweet classification and NLI task respectively. We performed three ablations for each model to gain a better understanding of each model's behaviour. Furthermore, we lay out detailed quantitative and qualitative error analysis of the task

2 Related Work

The research community has been actively facilitating supervised datasets (Sharma et al., 2020a; Kiela et al., 2020; Suryawanshi et al., 2020) to contribute towards solving the multimodal offensive meme classification task. However, unlike their text counterparts (Zampieri et al., 2019, 2020; Risch et al., 2021) these supervised multimodal datasets are smaller. In our research, we are utilizing three popular datasets: Memotion, Hateful memes and MultiOFF datasets.

Initially, researchers opted for a sequence to sequence (Seq2Seq) architecture for capturing text and image features with two encoders and later on fusing them to classify if the meme is offensive. But due to the efficiency of transformers over Seq2Seq, and the advent of VL pre-training, the research has been shifted towards transformer-based architectures such as LXMERT (Tan and Bansal, 2019), Visualbert (Li et al., 2019), Vilbert (Lu et al., 2019), UNITER (Chen et al., 2020). A Seq2Seq approach proposed by (Sharma et al., 2020b) fuses image features derived from InceptionNet and text features derived from the GloVe embedding. The feature fusion proposed in their research uses Bi-LSTM initialized with image features as hidden and cell state and calculated attention over the text features. They were able to score first rank with a macro-average F-score of 0.52907 on the Memotion shared task in subtask B: Humour Classification. A winning solution (AUROC: 0.8449, Accu-

racy: 0.7320) for the Hateful memes challenge by (Zhu, 2020) proposes an ensemble model that combines VL-BERT, UNITER-ITM, VILLA-ITMand ERNIE-Vil. Moreover, the authors extracted the entity, gender and race of the individuals from the meme by using face extraction with Mask-RCN. (Zhong et al., 2022) proposed injecting an external knowledge base in the form of entity recognition from the meme text to enhance the semantic representation of the meme. However, they relied on the raw image features captured via VGG. Their approach established new SOTA results (precision: 0.670, recall: 0.671, f-score: 0.671) for the MultiOFF dataset. In summary, all of these top-scoring approaches are multimodal. However, we propose an unimodal approach where we use the image caption of a meme (meme caption) as a text feature as a replacement for the raw image features. We are comparing our results with these current SOTAs and baselines in Section 4.

We take inspiration from (Prajwal et al., 2019), they suggest that memes could be transcribed to the visually impaired individual using carefully generated facial image captions. We argue that one might lose crucial information from the meme by just concentrating on facial image captions. Hence, we crop the meme to get rid of the unnecessary meme text, we consider the cropped meme as whole over just faces while generating meme captions. (Yin et al., 2019) proposes a framework that leverages the NLI task for zero-shot text classification. We closely follow this approach in our work but unlike their research, we use our framework to finetune the text classifier rather than zero-shot classification. Moreover, we just use one hypothesis for each data sample rather than generating true and false hypothesis for each sample.

All the Multimodal SOTA's are complex and computationally heavy due to millions of trainable parameters. Moreover, they ensemble multiple VL models which is less practical since such models are complex to deploy in the real world. Hence to make the solution more simpler and practical, we propose to transform the multimodal offensive meme classification problem into an unimodal offensive text classification problem by leveraging the NLI task.

3 Data Pre-processing

As shown in the Figure 1, first we crop the image to avoid the text embedded in the meme. The meme is

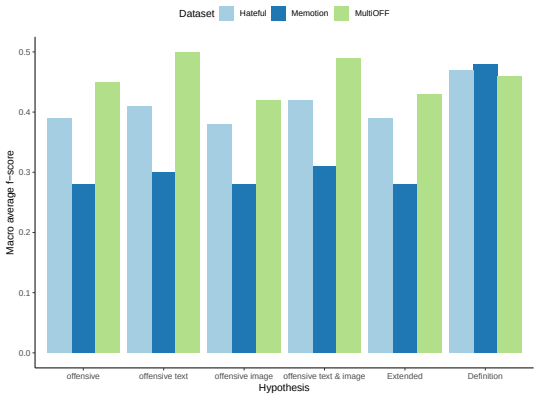


Figure 2: Clustered bar graph of macro f-score of RoBERTa in zero-shot setting for each Memotion, Hateful memes and MultiOFF dataset. Y-axis represent macro f-score while x-axis shows the hypothesis.

cropped on both the length (l) and breadth (b) by ¼ margin based on the manual inspection of random sample drawn from each dataset. Later, we generate the image caption using the ClipCap (Mokady et al., 2021) image captioner. We incorporate the generated captions inside the meme transcription along with the text associated with the meme which is used as a premise. Finally, the offensive label is converted into a natural sentence and a new label i.e. entailment label is assigned to either 1 or 0 if the meme is offensive or not offensive respectively.

3.1 Meme Transcription

CLIP by (Radford et al., 2021) gives competent results for Hateful Memes dataset in the zero-shot setting. Hence, we opted for ClipCap image captioner based on the CLIP image encoder with GPT2 prefix decoder (pre-trained on the MS-COCO dataset (Chen et al., 2015)) to generate meme captions at inference time. We transcribed memes by combining these meme captions with the meme text (the text embedded on the meme provided along with each dataset). The template used to automatically transcribe the meme is "The meme contains an image of `meme caption`, and the text on the meme says that `meme text`". For example, the meme in Figure 1 is transcribed as "The meme contains an image of a girl with a trash can, and the text on the meme says that Selena Gomez is taking her music out for a walk". In this example, the text "a girl with a trash can" is a meme caption, and the text "Selena Gomez is taking her music out for a walk" is a meme text.

3.2 NLI-fication

Figure 1 shows the overview of transforming data from image-text-label to premise-hypothesis-label

Dataset	Train size	Val size	Test size	Epochs	Learning rate	Batch size	Weight decay	Grad acc
Memotion	5,940	660	1,878	10	5.0e-7	8	0.001	4
Hateful	7,650	850	2,000	10	5.0e-5	8	0.001	16
MultiOFF	445	149	149	10	1.0e-5	8	0.001	8

Table 1: On the left side of the vertical line:Data statistics in terms size of training, validation and test for each Memotion, Hateful and MultiOFF dataset. On the right side of the vertical line: Hyper-parameter settings for all the three RoBERTa for Memotion, Hateful memes and MultiOFF dataset in terms of number of epochs, learning rate, batch size, weight decay and gradient accumulation steps .

format. We refer to this procedure as NLI-fication in the rest of the article. The meme transcription acts as a premise in the context of the Natural Language Inference (NLI) task, while the label i.e. “OFF” has been transformed into the hypothesis. We experimented with the hypothesis on three levels: primary, extended, and definition. Primary level does not use the natural sentence to explain the label, rather it just uses the label i.e. “offensive" as a hypothesis. The motivation behind the NLI-fication of the data comes from the fact that the model would get a better understanding of the label once it has been translated into a hypothesis in natural sentence. We tried different versions of the primary hypothesis by adding more words–offensive text, offensive meme, offensive image and text, offensive image or text–to the primary hypothesis. The extended hypothesis is just the natural sentence that describes the label “The meme is offensive considering either or both the image and the text”. At the definition level, we add the definition for the hate or offensive content provided by (Kiela et al., 2020) i.e. “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 dehumanising (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." We report the zero shot results on each dataset in Figure 2, the figure shows the averaged macro f-score of RoBERTa finetuned on NLI data (SNLI and MNLI) across all three (Memotion, Hateful memes, Multioff) datasets. We chose RoBERTa because it achieved SOTA results on NLI task despite being a small compared to bigger language models such as T5, BART, BigBird. The figure shows that the model yielded the highest mean macro f-score at “Definition level". This points out that the hypothesis with more offence related keywords worked the best. Because, the rest hypothesis (other than the definition) just has the word “offensive" as offence-related keyword, while definition has more keywords such as attack, dehumanising, mocking, hate, crime. The fact that “Definition level" works best in the zero shot setting emphasises that the model has a knowledge of the offensive keywords which could be improved upon further finetuning. We maintained identical “Definition level" hypothesis for each data sample across all three datasets. This does not only simplify our approach but also removes manual overhead of hypothesis tuning based on each sample. Hence, making our approach more generalizable to new multimodal offensive datasets.

4 Experimental Settings

4.1 Baselines

The baselines are based on the use of the fine-tuned dataset. Emotion and Sentiment of the text acts as a auxiliary information to offensiveness of the text (Mnassri et al., 2023). Hence, emotions and the sentiment of the text play an important role in identifying the offensiveness of the text. Moreover, the model finetuned on the offensive tweets could prove as a strong baseline due to the inter-training on closely related offensive tweet classification dataset (Choshen et al., 2022). We use RoBERTa fine-tuned on the Emotion, Sentiment and Offensive tweet classification data (Barbieri et al., 2020) as baselines. Specifically, we chose “twitter-roberta-base-sentiment[1]” , “twitter-roberta-base-emotion[2]”, and “roberta-base-offensive[3]”respectively for Emotion, Sentiment, and Offensive RoBERTa baselines. These models are finetuned on the short text i.e. tweets, which is similar to the text captions embedded in the memes. These models are loaded with pop culture knowledge since they are finetuned on 54M tweets before

[1]https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

[2]https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion

[3]https://huggingface.co/cardiffnlp/roberta-base-offensive

finetuning on the Emotion, Sentiment Analysis and Offensive datasets respectively. Hence, we believe that it is easier for these models to adapt to the domain of our task. We deliberately use RoBERTa for each of baseline to maintain the consistency and comparability of the results across the baselines and our approach. Table 1 shows the data statistics and hyperparameters used for each each dataset across all experiments. Moreover, we compare our results with current SOTAs along with mentioned baselines. We already cover their details in Section 2, we collectively call them Multimodal SOTA irrespective of the dataset.

4.2 Significance test

We performed a 5X2 significance test (Dietterich, 1998) for each model pair for the Memotion dataset to show that they are significantly different from each other. The significance test is primarily five-fold cross-validation performed two times. The macro-averaged f-score is recorded for each fold, resulting in 10 macro-averaged f-score which are used later to calculate the p-value and t-statistics for each pair of models: Inference Vs Emotion RoBERTa, Inference Vs Sentiment RoBERTa, Inference Vs Offensive RoBERTa, Emotion Vs Sentiment RoBERTa, Emotion Vs Offensive RoBERTa, and Offensive Vs Sentiment. Here, the null hypothesis is these pairs do not differ significantly from each other.

4.3 Our Approach

Based on the text classification framework proposed by Yin et al. (2019), we finetuned RoBERTa (on binary NLI dataset with 28k samples labelled as “entailment" and “not entailment") which was previously finetuned on the NLI datasets such as Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2018). We chose RoBERTa based on its state-of-the-art performance on the NLI task. Moreover, it was pre-trained not only on masked language modelling (MLM) but also on the sentence prediction objective which we thought would be helpful in our task since we are dealing with two different texts i.e. premise and hypothesis. Since our classification task is binary (labels: OFF or NOT), we decided to finetune RoBERTa on a binary NLI dataset. For this purpose, we sampled 28,000 examples from the combined SNLI and MNLI dataset and converted their labels into *Entailment* (1) or *Not entailment* (0) by encoding the *neutral* and *contradiction* label of the original dataset into Not entailment. Amongst 28,000 examples, 4,000 each were sampled randomly for the validation set and test set. The model was trained for five epochs and the best model with the least validation loss was saved during the training. We used this saved model later on to finetune the multimodal offensive meme datasets.

Models	p-value	t-statistics
Inference Vs Emotion	1.15e-05	17.39
Inference Vs Sentiment	6.00e-06	19.85
Inference Vs Offensive	1.07e-06	28.07
Emotion Vs Sentiment	1.52e-06	26.16
Emotion Vs Offensive	2.70e-09	93.11
Offensive Vs Sentiment	5.47e-08	50.99

Table 2: 5 X 2 significance test results for Inference, Emotion and Offensive RoBERTa (against each other) with respect to Memotion dataset.

4.4 Ablations

We performed three ablations on each of the experiments. The first ablation uses just the meme caption (meme-captions-only), and the second ablation uses just the text from the meme (meme-text-only). In these ablations, we intend to study the impact of each text individually on the performance of the experiments evaluated with precision, recall and f-score. In the third ablation (no-NLI-fication), we removed the NLI-fication of the data from the data pre-processing pipeline and used just the premise as a text by removing the hypothesis altogether. In the last ablation, we intend to study the effect of NLI-fication on each Emotion, Sentiment, Offensive, and Inference RoBERTa model.

5 Quantitative error analysis

We refer to scores reported in (Mokady et al., 2021) for quantitative error analysis of the ClipCap image captioner whereby it is evaluated with 32.15 using Bleau@4, 27.1 using METEOR, 108.35 using CIDEr, and 20.12 using SPICE evaluation scores. These scores are close to that of other image captioning models such as BUTD, VLP, and OSCAR.

Table 2 shows results from 5 X 2 significance test. It could be seen in the table that all the p-values are less than 0.05. Hence, we do not have enough confidence to accept the null hypothesis: all the pairs of the models are not significantly different from each other. Hence, we reject the null hypothesis. Emotion Vs Offensive RoBERTa

Dataset	Models	Class	Precision	Recall	F-score
Memotion	Multimodal SOTA	macro	-	-	52.90
	Emotion	OFF	63.21	87.87	73.53
		NOT	43.20	15.28	22.57
		macro	**53.20***	**51.57**	**48.05**
	Sentiment	OFF	62.44	84.63	71.86
		NOT	38.14	15.70	22.24
		macro	**50.29**	**50.16**	**47.05**
	Offensive	OFF	62.71	65.50	64.08
		NOT	38.32	35.50	36.86
		macro	**50.52**	**50.50**	**50.47**
	Inference (our approach)	OFF	64.54	56.28	60.13
		NOT	40.26	48.80	44.12
		macro	**52.40**	**52.54**	**52.12**↓
Hateful	Multimodal SOTA	Accuracy/AUC-ROC	73.20	0.8449	
	Emotion	HATE	51.22	39.33	44.49
		NOT	68.05	77.52	72.48
		macro	**59.63**	**58.43**	**58.49**
	Sentiment	HATE	59.00	39.33	47.20
		NOT	69.67	83.60	76.00
		macro	**64.33**	**61.47**	**61.60**
	Offensive	OFF	55.87	52.67	54.22
		NOT	72.54	75.04	73.77
		macro	**64.21**	**63.85**	**64.00**
	Inference (our approach)	HATE	61.93	40.13	48.71
		NOT	70.34	85.20	77.06
		macro	**66.14**	**62.67**	**62.88**
		Accuracy/AUC-ROC	68.23↓	0.3662↓	
MultiOFF	Multimodal SOTA	macro	67.00	67.00	67.00
	Emotion	OFF	54.29	65.52	59.37
		NOT	74.68	64.84	69.41
		macro	**64.48**	**65.18**	**64.39**
	Sentiment	OFF	54.67	70.69	61.65
		NOT	77.03	62.64	69.09
		macro	**65.85**	**66.66**	**65.37**
	Offensive	OFF	45.00	93.10	60.67
		NOT	86.21	27.47	41.67
		macro	**65.60**	**60.29**	**51.17**
	Inference (our approach)	OFF	59.09	67.24	62.90
		NOT	77.11	70.33	73.56
		macro	**68.10**↑	**68.79**↑	**68.23**↑

Table 3: The quantitative results of experiments: The report presents the detailed evaluation results in terms of class-wise and macro-averaged precision, recall and F1 score. Accuracy and AUC-ROC* denotes Accuracy and AUC-ROC scores for Hateful Memes dataset for comparing our approach with SOTA. The ↑ and ↓ in Inference section indicates if our approach surpassed the current SOTA or not.

shows the largest t-statistics which means they are more significantly different than any other pair.

Table 3 shows the detailed classification report –with class-wise and macro averaged precision (p), recall (r) and f-score (f)– of each Emotion, Sentiment, Offensive and Inference RoBERTa on Memotion, Hateful memes and MultiOFF datasets. The highlighted bold cased score shows the macro averaged p, r, f score for each RoBERTa model, and with * denoting the highest macro-averaged score. In this table we are evaluating our approach in two ways. Firstly, we evaluate against the baselines (Emotion, Sentiment, and Inference RoBERTa) whereby we compare macro-averaged p, r, f scores. Hence, these scores highlighted in bold for better readability. Secondly, we evaluate against Multmodal SOTAs whereby we either use ↑ or ↓ to specify if our approach has surpassed the SOTAs or not respectively.

For the first part of the evaluation, it could be seen clearly that the macro averaged evaluation score is increased in the inference RoBERTa over Sentiment and Emotion RoBERTa across all datasets except for the fact that macro averaged precision of Emotion RoBERTa (53.20%) is greater than that of the Inference RoBERTa (52.40%). However, the difference between the macro-average recall of the two models was significant (4.08%). This difference shows that Inference RoBERTa shows more balanced class-wise

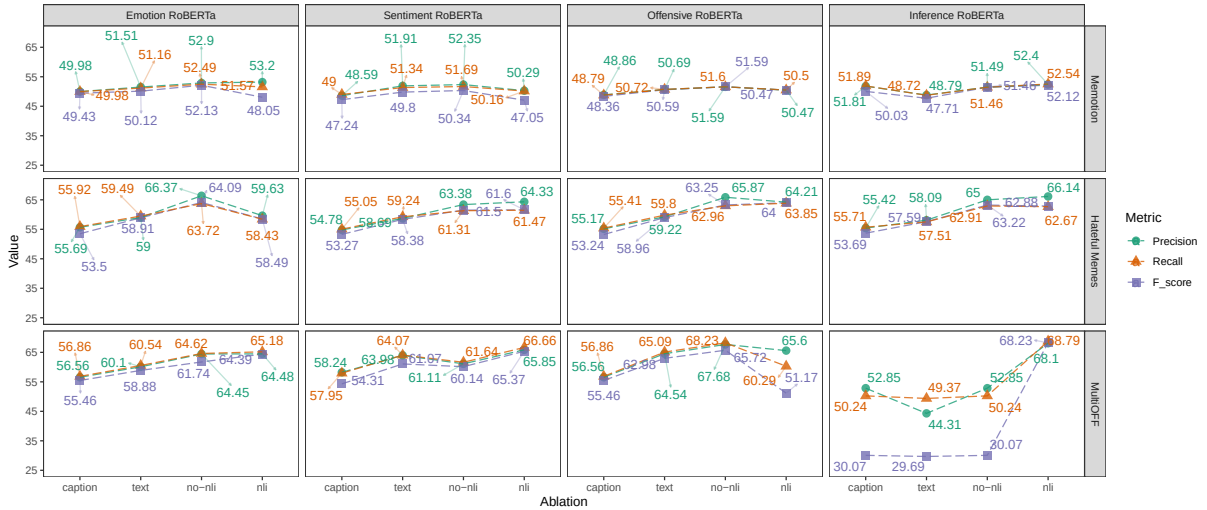


Figure 3: Line plot for the macro-averaged p, f, f-score for each ablation–meme-caption-only (caption), meme-text-only (text), no-NLI-fication (no-nli)– along with original experiment (nli).

precision and recall with less false positive and false negative count than that of Emotion RoBERTa. The increment in the evaluation scores of Inference RoBERTa compared to the other two could be credited to the NLI finetuning performed on binary NLI dataset (28k samples labelled as "entailment" and "not entailment"). Offensive RoBERTa shows better macro averaged recall and f-score compared to that of Inference RoBERTa for Hateful Memes dataset. This shows that Offensive RoBERTa while retains more offensive memes compared to the Inference RoBERTa at the expense of lesser macro averaged precision. It is interesting to observe that Offensive RoBERTa performs better than Emotion and Sentiment RoBERTa in the case of Memotion and Hateful memes dataset, but it fails to beat the two in case of MultiOFF dataset. This shows that even after offensive tweet classification being closely related to our downstream task, less training data leads to poor generalisation.

For the second part of the evaluation, in the case of the Memotion dataset, we achieve close to Multimodal SOTA performance (our macro-averaged f-score: 52.12%, SOTA: 52.90%). Since the official evaluation metric is accuracy and AUC-ROC for the Hateful memes challenge, we calculated both the metric for the best performing RoBERTa. Inference RoBERTa with the highest macro-averaged p, r, f score showed an accuracy of 68.30% and AUC-ROC of 0.3662 which is less than that of the Multimodal SOTA with an accuracy of 73.20% and AUC-ROC of 0.8449. The difference in the performance could be attributed to the reduced complexity of our approach since the winning solution used a complex ensemble technique that leveraged complex pre-trained models such as VL-BERT, UNITER-ITM and VILLA-ITM. If this complexity is taken into account then the difference in the accuracy (4.9%) is not more. However, the less AUC-ROC of Inference RoBERTa shows that the model is more susceptible to threshold change when compared with the SOTA. In case of the MultiOFF dataset, we beat the Multimodal SOTA (Zhong et al., 2022) (p: 67.10%, r: 67.00%, f: 67.10%). Since the MultiOFF dataset consists of only 743 examples (# train: 445, # validation: 149, # test: 149), our approach shows robust performance in terms of new SOTA results even with fewer training samples.

Figure 3 shows the detailed evaluation report – across all of the three datasets and three models– in terms of macro-averaged p, r, f score on the three ablations (meme-caption-only, meme-text-only, no-NLI-fication). One common trend amongst graphs of Emotion, Sentiment and Offensive RoBERTa showed the least macro-averaged p, r, f score in meme-caption-only ablation compared to the rest. Moreover, it could be seen that the p, r, f scores for meme-text-only ablations are better than that of the meme-caption-only ablations for the three models. This shows that the meme text plays a more vital role than the meme caption at identifying offensive memes in the case Emotion, Sentiment and Offensive RoBERTa across all three datasets. However, the Inference RoBERTa trained on the Memotion and MultiOFF dataset show contradictory trend where meme-caption-only ablation shows better evaluation score than that of meme-text-only. But the same model in meme-text-only ablation shows improvement in evaluation scores over the meme-caption-only ablation in case of Hateful Memes dataset. This indicated that meme captions are more reliable features for Inference

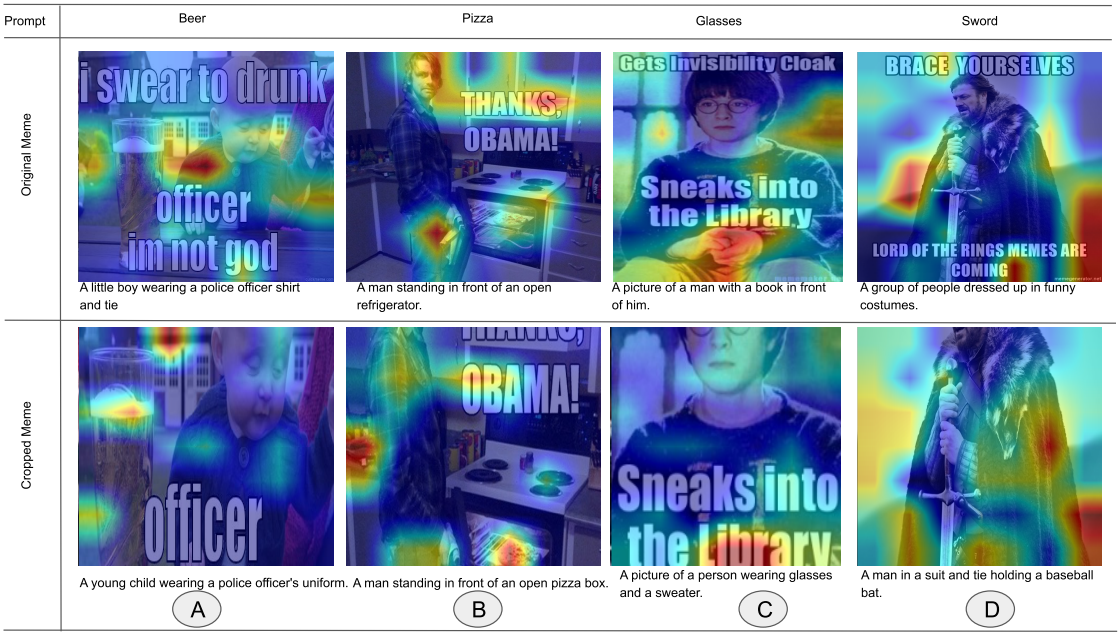


Figure 4: The first row represents original memes, and the second row represents cropped memes. The captions of the meme are mentioned below each image .

RoBERTa than meme texts when it comes to offensive dataset such as Memotion and MultiOFF. Moreover, the contradictory trend shows the difference in the working of Inference RoBERTa against the other two corroborated by the significance test results mentioned in the Table 2. The inference RoBERTa shows the peak performance in NLI settings than any other ablation. This illustrates that Inference RoBERTa is getting a bump in the performance due to the finetuning of the binary NLI data. Similarly, Emotion RoBERTa –in the case of MultiOFF dataset– Sentiment RoBERTa–in the case of the Hateful memes dataset and MultiOFF dataset– showed improvement in the performance in NLI settings compared to no-NLI-fication ablation. However, Emotion RoBERTa showed a decline in the performance in NLI settings in the case of the Memotion and Hateful Memes dataset while Sentiment RoBERTa showed similar results in the case of the Memotion dataset. Irrespective of these differences in the performance of each model in Hateful memes and Memotion datasets, we could see that all three models show improvement in macro-averaged p, r, f score in NLI settings for MultiOFF dataset. On the one hand, Emotion RoBERTa in no-NLI-fication ablation beats the Inference RoBERTa in the NLI setting in the case of the Memotion and Hateful Memes dataset. On the other hand, Offensive RoBERTa showed highest evaluation scores across all the three datasets in no-NLI-fication settings. This highlights the competency of the RoBERTa finetuned on the emotion and offensive tweet classification data. This shows that the emotion and offensive tweet classification could also be leveraged for offensive meme classification. It could be seen that the performance of the inference RoBERTa in NLI settings shows the highest evaluation scores which beat the current SOTA (p:67%, r:67%, f:67%). Moreover, our Offensive RoBERTa in no-NLI-fication settings comes close to SOTA (p:67.08%, r:68.23%, f:65.72%). This shows the robustness of our approach in low sample settings. The significant marginal difference between no-NLI-fication ablation and the original NLI shows that the Inference RoBERTa can perform better while leveraging the NLI knowledge gained after finetuning on the NLI dataset specially for small dataset.

6 Qualitative error analysis

In this section, we analyse the inference for the examples from the test samples of each Memotion, Hateful Memes and MultiOFF dataset. Firstly, we would like to highlight some of the examples from the ClipCap prefix image captioner in Figure 4. The first row in the Figure shows the prompt used to generate the heatmap. In the context of computer vision, heatmaps are used to identify the regions in

Image				
Meme captions	A man in a suit and tie standing next to another man in a suit and tie	A close up of a person wearing a suit and tie	a close up of a small animal near a field of grass	An old picture of a woman holding an umbrella
Dataset	Memotion	Memotion	MultiOFF	Hateful Memes
Inference RoBERTa	OFF	NOT	NOT	NOT
Emotion RoBERTa	OFF	OFF	NOT	HATE
Sentiment RoBERTa	OFF	OFF	NOT	NOT
Gold Label	**OFF**	**OFF**	**NOT**	**HATE**

Figure 5: Inference result on examples from Memotion, MultiOFF and Hateful memes for each Inference, Emotion and Sentiment RoBERTa. Please note that the meme captions mentioned here are generated after cropping the meme.

an image that are likely to contain objects of interest. Higher intensity or warmer colors (e.g., red or yellow) in the heatmap indicate higher confidence or probability of the object being present at those locations, while lower intensity or cooler colors (e.g., blue or green) indicate lower confidence. We used gScoreCAM (Chen et al., 2022) to generate heatmap from CLIP. These maps are generated top 1000K channels out of total 3072K channels. The centre of the map is dark red and turns to the lighter shade outwards. This indicates the confidence of the heatmap which is higher at the centre and lowers in the outward direction. The second row in the figure represents original memes along with their caption stated below the meme. Similarly, the last row represents cropped memes with their captions stated below the meme. We chose prompt based on wrong noun predicted in either original or cropped meme caption.

The first example circled (A) shows that the image captioner correctly identified the young child in the meme in both the original and cropped meme. However, it falsely identified the word `officer` from the meme text being printed on the child's top. Furthermore, the important object here to be detected was `beer`. We prompted CLIP with the word `beer` on both original and cropped meme. It can be seen that no heatmap is present near the original, but the cropped meme shows two such heatmaps on the object beer. This shows that the CLIP is more confident at selecting the required object in the cropped version. The third example circled (C) shows that the image captioner correctly identified meme captions after cropping unlike the meme caption for the original meme which emphasized the word `library`. Moreover, heatmaps generated for the prompt `Glasses` is present on the glasses on the cropped meme, but nowhere seen near the glasses in original meme. The example circled (B) shows improvement in the quality of the meme caption after cropping the meme as it could be seen that the object falsely recognized as `refrigerator` has been replaced by the correct one i.e. `pizza box`. Furthermore, if we prompt both memes the word `pizza`, the original meme shows bigger heatmap concentrated around meme text `THANKS`, and a small less confident heatmap on the pizza. However, heatmap on the cropped meme is concentrated on the object `pizza` as well as pizza-like object on the left side of the meme. All the examples (A), (B) and (C) shows that the cropped meme not only generated better captions but also helps CLIP to capture useful image feature. In this case, the meme text is is acting like an adversary while capturing useful image feature. However, in example circled (D), the image cap-

tioner falsely recognized `sword` as a `baseball bat` instead after cropping the meme. If word `sword` is prompted to CLIP, the original as well as cropped meme fails to capture the useful image features as shown in the heatmap. This could be attributed to the fact that word `baseball bat` has been observed more in MS-COCO dataset, hence the captioner is biased towards such words. All examples show the sensitivity of the image captioner towards the meme text. Hence, although our approach works optimally with the current state of the image captioner, meme caption quality could be improved by completely removing the meme text from the meme.

Figure 5 shows the detailed report on the inference results for each Inference, Emotion and Sentiment RoBERTa on examples from the test set of Memotion, MultiOFF and Hateful memes datasets. The first column in the table illustrates an example from the Memotion dataset. This example is offensive as it intends to demean Obama [4]. All the models were able to correctly identify the given example as OFF. On the other hand, the example from the second column which is labelled as OFF has been incorrectly classified as NOT by Inference RoBERTa. But if we take a look at the meme from the example, it does not mean to harm or attack anyone. Hence, it could be labelled as NOT. This shows the noisiness of the dataset which could be attributed to the annotation process. The example from the third column belongs to the MultiOFF dataset. Here, all the models were able to correctly classify the given meme into the NOT category. The example in the fourth column showed interesting results since the meme has been incorrectly captioned which led to the failure of Inference and Sentiment RoBERTa while Emotion RoBERTa succeeded. This difference in the performance of the models shows a difference in their pattern recognition ability which has already been proven by the 5X2 significance test shown in Table 2.

7 Conclusion

All the experiments and their ablation suggest that the transforming multimodal offensive meme classification into unimodal offensive text classification problem not only simplifies the approach but also achieves SOTA results. It could also be seen that the NLI-fication of the multimodal data could improve the evaluation metric, especially in the case of a smaller dataset (MultiOFF). Emotion RoBERTa outperformed its counterpart after removing the NLI-fication of the data while Sentiment RoBERTa fell short by a minute margin. This shows that the models finetuned on the Emotion and Sentiment Analysis task could prove useful in the offensive meme classification task. Overall, it is a viable option to translate the multimodal offensive meme classification into a unimodal (text) classification problem to get competent evaluation scores. Moreover, NLI-fication is not only simple but also effective at training on smaller out of domain dataset.

Limitations

In the qualitative error analysis, we observed the sensitivity of the CLIP prefix image captioner towards the meme text. This approach may generate an out of context meme caption which later could harm the performance of the model. Moreover, Figure 4 (D) shows inferior image captions upon cropping. Hence, fixed cropping ¼ margin along length and breadth could lead to information loss which results in incorrect captions. To tackle this issue in future, we plan to use an in-house image captioner model which will ignore the noise generated from the meme text without cropping it. To better understand the image captioning errors, we plan to train our model on a small subset of manually human-generated image caption.

Ethics Statement

The definition of "offensive" content is highly subjective and can vary across different cultures and communities. Hence, the same content that is deemed for certain group or community might not be offensive to others. Therefore, marginalised groups may be disproportionately affected by the model's decisions.

Acknowledgements

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 for the Insight SFI Research Centre for Data Analytics, co-funded by the European Regional Development Fund.

[4] 44th President of the United States

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Peijie Chen, Qi Li, Saad Biaz, Trung Bui, and Anh Nguyen. 2022. gscorecam: What objects is clip looking at? In *Proceedings of the Asian Conference on Computer Vision*, pages 1959–1975.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Leshem Choshen, Elad Venezian, Shachar Don-Yehia, Noam Slonim, and Yoav Katz. 2022. Where to start? analyzing the potential value of intermediate models. *arXiv preprint arXiv:2211.00107*.

Richard Dawkins. 2016. *The selfish gene*. Oxford university press.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. 2020. Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 153–164.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2023. Hate speech and offensive language detection using an emotion-aware shared encoder. *arXiv preprint arXiv:2302.08777*.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: Clip prefix for image captioning.

K R Prajwal, C V Jawahar, and Ponnurangam Kumaraguru. 2019. Towards increased accessibility of meme images with the help of rich face emotion captions. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 202–210, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020a. SemEval-2020 task 8: Memotion analysis- the visuolingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Mayukh Sharma, Ilanthenral Kandasamy, and W.b. Vasantha. 2020b. Memebusters at SemEval-2020 task 8: Feature fusion model for sentiment analysis on memes using transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1163–1171, Barcelona (online). International Committee for Computational Linguistics.

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are we pretraining it right? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744*.

Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the shared task on troll meme classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 126–132, Kyiv. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the*

Second Workshop on Trolling, Aggression and Cyberbullying, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447.

Qi Zhong, Qian Wang, and Ji Liu. 2022. Combining knowledge and multi-modal fusion for meme classification. In *MultiMedia Modeling*, pages 599–611, Cham. Springer International Publishing.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *CoRR*, abs/2012.08290.