Logically at the Constraint 2022: Multimodal role labelling

Ludovic Kun * , Jayesh Bankoti * , and **David Kiskovski**
Logically, London, UK

Abstract

This paper describes our system for the Constraint 2022 challenge at ACL 2022, whose goal is to detect which entities are glorified, vilified or victimised, within a meme . The task should be done considering the perspective of the meme's author. In our work, the challenge is treated as a multi-class classification task. For a given pair of a meme and an entity, we need to classify whether the entity is being referenced as Hero, a Villain, a Victim or Other. Our solution combines (ensembling) different models based on Unimodal (Text only) model and Multimodal model (Text + Images). We conduct several experiments and benchmarks different competitive pre-trained transformers and vision models in this work. Our solution, based on an ensembling method, is ranked first on the leaderboard and obtains a macro F1-score of 0.58 on test set. The code for the experiments and results are available at here .

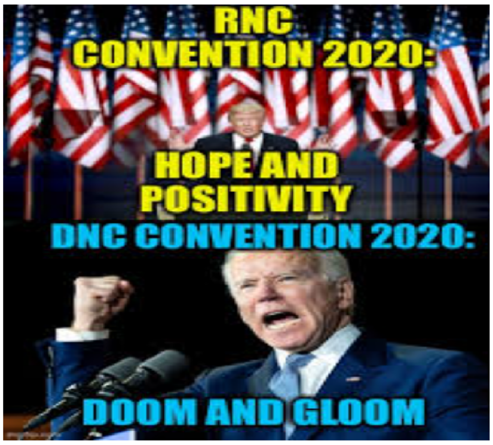
1 Introduction

The rapid rise in the amount of harmful content being spread online is becoming a major societal challenge, with still unknown negative consequences. Large resources have been invested by many actors in the field of social media to shield users from harmful content. It is imperative to understand in a systematic way how information is spread, and be able to scalably monitor existing narratives and flag hateful ones circulating using technology. One way this is done is using entity recognition coupled with entity sentiment (Kiritchenko et al., 2021). The former technique is to support OSINT(open source intelligence) analysts in understanding who or what are the subjects of discussion, and the latter automates the process of analysing if they are coupled with positive or negative feelings, in order to assist with understanding the stance of online users on specific topics. Efforts to tackle this challenge were mainly focused on English-language text-based data formats such as articles (Wankhade et al., 2022). However, the complexity of content being posted online has drastically increased over time, and the challenge of harmful content detection now extends to multimedia, including memes (Alam et al., 2021). The emergence and proliferation of memes on social media have made their analysis a crucial challenge to understand online interactions. A point can also be made about the study of entities sentiment online, as the polarising portrayal of famous (or infamous) personalities or institutions often give rise to inflammatory views and content.

Extracting insights from memes is a novel field and still has a lot of opportunities for growth. The multimodality of text and image adds a layer of complexity which contains more information, but is also harder to extract. Indeed each modality needs to understand their intrinsic properties but also capture cross-modal semantic understanding (Müller-Budack et al., 2021). This paper delves into the field of multimodal semantic role labelling, a new task with particular challenges.

Examples of the multimodal dataset (Sharma et al., 2022) used to tackle this problem and provided as part of the CONSTRAINT competition are presented in Figure 1. The first sample shows a meme image displaying two politicians from opposite parties separated on two sides of the image, with text around them, as well as the associated JSON line input with the extracted text from the image (also known as Optical Character Recognition or OCR), as well as the entities' mentioned labelled roles. In this case, all entities are referenced in the text of the image. In the second sample, however, we notice that not all are mentioned in the text, and visual information is needed to classify all entities.

Depending on the textual information in the image, textual role classification is insufficient as some memes' underlying message requires under-



```
{
‘image’ : memes_1486.png’,
‘OCR’: "AAE RNC\nCONVENTION 2020:\nHOPE AND\nPOSITIVITY\nDNC CONVENTION 2020:\nDOOM
AND GLOOM\n",
‘hero’ : ['Donald Trump'],
‘villain’ : ['Joe Biden'],
‘victim’ : [],
‘other’ : ['Democratic National Convention (DNC)’, 'Republican National Convention
(RNC)']
}
```

Figure 1: CONSTRAINT dataset example

standing of the visual information it contains, especially with the use of humour and sarcasm often associated with the format.

The work done in this competition aims at finding unique and effective ways of tackling harmful meme classification as seen in the current social media space. An algorithm is designed for the task of role labelling for memes using a twin model (and ensemble) method. This Siamese network is constructed by combining the output of pre-trained State-of-the-Art (SoTA) models for both the visual components in the form of a CNN (Efficientnet-B7 (Tan and Le, 2019)) and for textual components using a transformer (DeBERTa (He et al., 2020)). The feature outputs obtained from both branches are then combined to obtain a final solution. Data analysis and investigation into potential bias in the dataset are also conducted to contextualise the task and present the difficulties of curating accurate multimodal datasets aimed at tackling the task for data in the wild (Gao et al., 2021). In this paper, an overview of past work in the field is presented (section 2), followed by a deep dive into the problem statement as well as the method followed to respond to it (section 3), then data analysis (section 4). Experiments ran are presented in section 5, with results and discussion in section 6, and finally conclusion (section 7).

2 Related Work

There have been some work done with respect to semantic role labelling in text. The idea of ABSA(Aspect Based Sentiment Analysis) works along the same line. Hence, utilisation of DeBERTa has provided the SoTA results (Silva and Marcacini) due to the disentangled attention improving the focus more on the positional embeddings rather than just based on the word embeddings. Hence, improved results were also obtained in various SNLI task for this algorithm(He et al., 2020).They are nowadays very popular in Natural Language Processing (NLP) as they usually get SoTA for a variety of NLP tasks such as classification, sentiment analysis, Named Entity Recognition, Translation, Question Answering, etc.

Classifying memes into relevant classes is a field that has got much more interest over the past few years. The Facebook Hateful meme competition(Kiela et al., 2020) was a very publicised initiative to try and augment the field's capabilities. The task was a binary classification of hateful/not hateful meme based on a dataset curated by META. The winning solutions all comprised of ensembles of multimodal models. The Memotion competitions(Sharma et al., 2020) are another example of work done in the meme space. This time, the classification was based on sentiment (positive, negative, neutral), as well as the strength of the sentiment and the underlying aim of the meme (satirical, humour or harmful). Multimodal models here also obtained the top scores.

Multimodal models have seen a change over the past few years from twin networks like Siamese (Gu et al., 2018) to models pretrained on multiple multimodal tasks such as image captioning and visual question answering using transformers (Devlin et al., 2018). Object detection is used in these models to extract image features thanks to pre-trained two-staged detectors Faster R-CNN model (Ren et al., 2015)), or single-stage detectors (YOLO V3 (Adarsh et al., 2020)). Inspired by BERT (Devlin et al., 2018), models such as Uniter (Chen et al., 2019) and VisualBERT (Li et al., 2019b) use a transformer architecture to jointly encode text and images, while LXMERT (Tan and Bansal, 2019) and ViLBERT (Lu et al., 2019) innovated by splitting their architectures in two, where a different transformer is applied to images and text individually before the features are combined by a third transformer. OSCAR (Object-Semantics Aligned

Pre-training)((Li et al., 2020)) add in the text input the class objects detected from the images by a Faster R-CNN detector called object tags. The use of object tags in images as anchor points, significantly ease the learning of alignments during the pretraining. These models' effectiveness are demonstrated through their SoTA results on different multimodal dataset tasks such as NLVR2. This can be attributed to the models' increased capability to understand cross-modal correlations. However, these models are only as good as the data they've been pretrained on, which will present a challenge for the use case of the competition tackled in this paper. Another point is that the architectures of the textual streams of these models are a few years old (such as BERT) and inferior to the current SoTA (DeBERTa).

3 Methodology

3.1 Problem Statement

The CONSTRAINT competition is a multimodal semantic role labelling multi-class classification problem. The aim is to classify the role of entities present in a meme using the image, its textual information and the entities it contains. The different classes are ("Hero", "Villain", "Victim", "Other"). The label applied for each entity depends on how the entity is presented in the meme:

Hero: The entity is glorified

Villain: the entity is vilified

Victim: the entity is victimised,

Other: none of the above.

3.2 Ensembling :

Our final model is an ensemble of 5 classifiers based on existing pretrained Unimodal (text) and Multimodal (text + images) architectures. (see figure 3) An ensemble combine several models to obtain a better generalised one. It usually gives a boost of performance in exchange for a more time-consuming model compared to more shallow model. Different methods of ensembling exist such as bagging, boosting, stacking, etc. We consider that this strategy will be very helpful to reduce the overfitting given the small number of instances we have, and how imbalanced the dataset is. To combine our models, we average the predictions of our individual models.

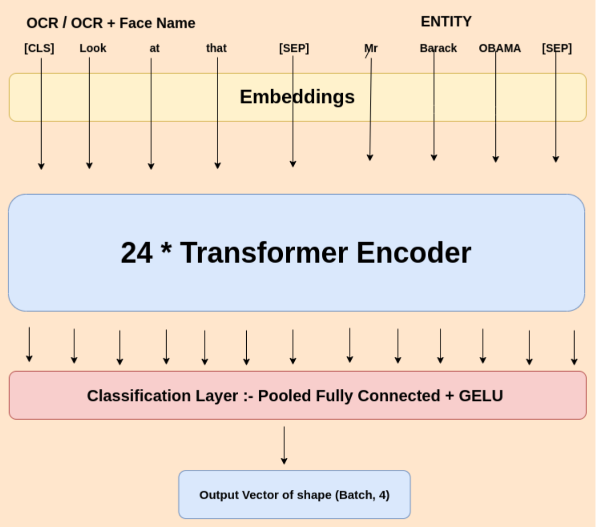


Figure 2: UniModal Model

3.2.1 Unimodal :

We experimented a few unimodal architectures based on transformers (Vaswani et al., 2017) such as DeBERTa and RoBerta (Liu et al., 2019) using only texts (OCR) and entities provided. The idea here was to see how much performance could be obtained just by textual information. These models are based on self-attention layers and an improved version of the BERT method pretrained on millions of sentences (Devlin et al., 2018) for language modelling. We fine-tuned on these models and found DeBERTa to be performing the best among the pretrained BERT models. For the fine-tuning, the last FC layer added over pooler layer of DeBERTa. The last layer was a FC layer of size 4 to provide us with the respective role label. The architecture for this structure is given (see figure 2) .

3.2.2 Multi-Modal :

We also experimented Multi Modal models which include as input data : images and texts (OCR + entity). We tried different approaches:
(1) The "Naive" approach consisted in extracting text features with a strong Language model - DeBERTa - and concatenating it with visual features with Convolutional Neural Network - EfficientNet-B7. We added on top of these concatenated features a Linear Layer to predict the class.
(2) The second approach was based on fine-tuning the whole image-text multimodal model. We experimented with two models: MMBT transformers (Multimodal Bitransformers) (Kiela et al., 2019) and VisualBERT (Li et al., 2019b) which has been pre-trained on classifying multimodal experiments.

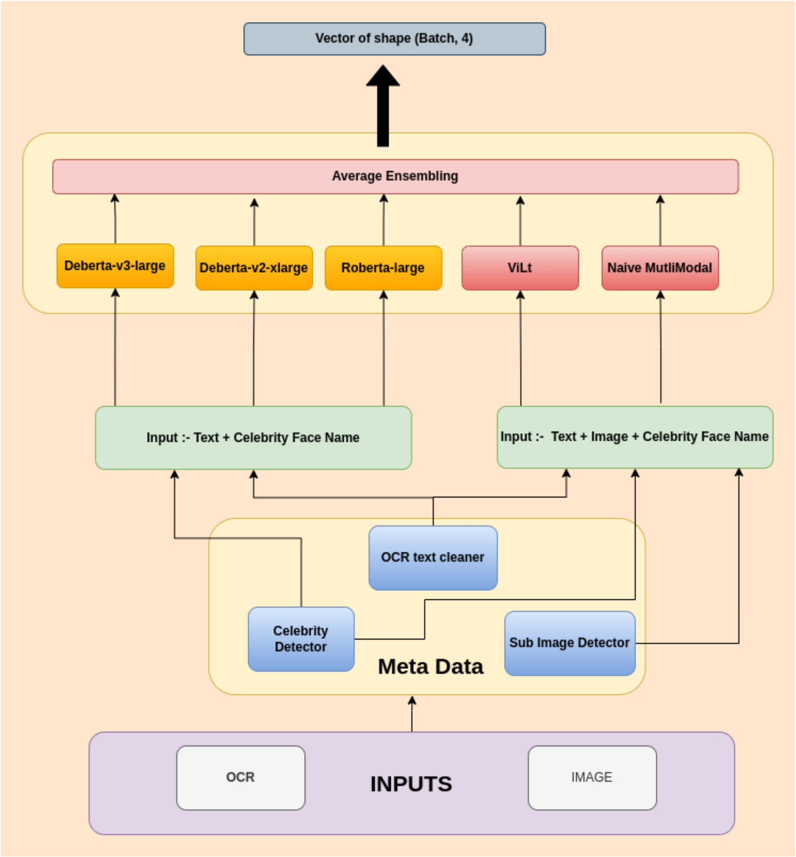


Figure 3: Final model used for the Constraint22 competition.

(i) The MMBT transformer model utilise bert-base-uncased model as text encoder and the CLIP model (Radford et al., 2021) as image encoder. The main idea was to reuse the BERT text model we had fine-tuned for the task and freeze the 12 encoder layers. Further we fine-tuned the MMBT multimodal model by projecting the image embeddings to text token space. (ii) The VisualBERT was pretrained model (Li et al., 2019b) for image-and-language tasks like VQA, VCR, NLVR2, and Flickr30Ks. We used the detectron2 embeddings (Ren et al., 2015) as image encodings with bert-base-uncased as text encoder to finetune the model.
(3) The last architecture used was ViLT (Kim et al., 2021) (Vision and Language Transformers) which is one of the simplest architectures for a vision and language model. ViLT is composed of a transformer module which extracts and processes textual and visual features without using separate embedder as it can be the case for MMBT for instance. That method gave a significant runtime and parameter optimisation. (see figure 5)

3.3 Meta Data extractions :

We attempted to extract meta data information from images in order to improve the insight from those. Indeed, using only the OCR was sometimes insufficient because the entities were not always present in the text. Multiple strategies were investigated for gathering insights from images.

3.3.1 Celebrity Detector :

The first observation made was in the image below (see figure 4) , the MEME is talking about Donald Trump (who is considered as a villain in the author's view). However he is not mentioned explicitly. His face is visible in the MEME though. That is why we decided to use a celebrities face detector which detects if a select famous face is visible in the MEME. The model is composed of two main steps : (i) a face detector based on the popular MTCNN face detector ((Zhang et al., 2016)) (ii) the face recognition part is based on a ResNet Architecture. We consider adding the face in the jsonl provided by the host when the confidence score of the face celebrities was above 0.95. The celebrity detector comes from Giphy's github.

3.3.2 Sub Image Detector

The second observation made was that a MEME can contain multiples "sub images". In fact, as in the figure 4, the MEME contains two images in it. A "sub images" detector was implemented based on YoloV5 (https://github.com/ultralytics/yolov5). We generated an artificial dataset, based on the Hateful MEME competition (Kiela et al., 2020), where we filtered and kept only the MEMEs with one image. Different single images were then combined to create one artificial MEME, with associated bounding boxes of the multiple subimages it contained. For the evaluation, 100 manually labelled images were used. The YOLO checkpoint is shared in our github solution. Our original idea was to extract with our detector each sub images from the MEME and associate each sentence of the OCR to the correct sub image with the name of the famous face if it existed. However, the OCR provided did not contain the coordinate of the sentence. We attempted to make the OCRed text match an open source OCR framework containing word coordinates, which yielded poor results. Therefore, the final multimodal model used the sub image as well as the face name into the text processing. The input of the transformer for text data was then as follows : "[CLS] Sentence OCR [SEP] entity to classify [SEP] face names [SEP]"

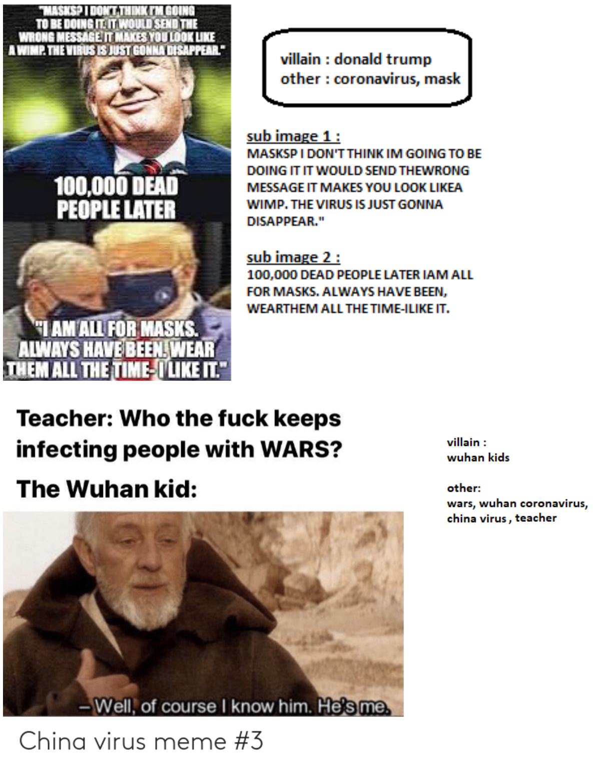

"MASKS? I DON'T THINK I'M GOING TO BE DOING IT. IT WOULD SEND THE WRONG MESSAGE. IT MAKES YOU LOOK LIKE A WIMP. THE VIRUS IS JUST GONNA DISAPPEAR."
100,000 DEAD PEOPLE LATER
"I AM ALL FOR MASKS. ALWAYS HAVE BEEN. WEAR THEM ALL THE TIME- I LIKE IT."
villain : donald trump
other : coronavirus, mask
sub image 1 :
MASKSP I DON'T THINK IM GOING TO BE DOING IT IT WOULD SEND THEWRONG MESSAGE IT MAKES YOU LOOK LIKEA WIMP. THE VIRUS IS JUST GONNA DISAPPEAR."
sub image 2 :
100,000 DEAD PEOPLE LATER IAM ALL FOR MASKS. ALWAYS HAVE BEEN, WEARTHEM ALL THE TIME-ILIKE IT.
Teacher: Who the fuck keeps infecting people with WARS?
The Wuhan kid:
– Well, of course I know him. He's me.
China virus meme #3
villain :
wuhan kids
other:
wars, wuhan coronavirus, china virus , teacher

Figure 4: Constraint dataset example : The first MEME contains two sub images whereas the second MEME don't have the entity we are looking for.

split	other	villain	hero	victim
train	13702	2427	475	910
train (ratio)	0.782	0.139	0.027	0.052
val	1589	305	54	121
val (ratio)	0.768	0.147	0.026	0.058

Table 1: distribution class of Constraint22 dataset

4 Dataset

The competition dataset consists of 2 memes subsets, one about US politics, and the other about Covid-19, totalling 5552 images with associated OCR and entity annotation in the training set, and 650 in the validation set. This size is very small to expect to build any robust SoTA vision or multimodal capabilities, training from scratch.

The distributions of the 4 labels are heavily imbalanced (see table 1). Over three quarters of the entities belong to the "other" class, and of the remaining classes, "villain" appears around twice as much as both the "hero" and "victim" class combined. An analysis of the entities in the dataset was undertaken and they were observed to be well balanced amongst the 4 classes. Indeed, as can be expected of using data from the political domain over the past few years, examples of common mentions were of "Donald Trump", "Barrack Obama", "The Republicans", "The Democrats". The fact that they were all amongst the most cited entities in each label indicates the sources used to curate the dataset was unbiased politically. Table 2 shows the top 5 most common entity per class.

The OCRed text was obtained by running the Google OCR API on the images, which in some examples leads to imperfect text detection or extraction. These two issues materialise in the form of either poorly clustered text paragraphs into the appropriate text boxes, meaning sentences from two separate paragraphs would be concatenated together midway through, but also through more basic spelling mistakes.

Another point relevant to meme analysis is the presence of sub images inside each image. An image might itself contain two separate images which tell a different story, often contrasting between sentiments of entities in each, such as in figure 4.

A big challenge with this task of entity classification is detecting where the entity is mentioned whether in the OCR or in the image. Table 3 shows

top-n entity	other	villain	hero	victim
1	donald trump	donald trump	donald trump	donald trump
2	coronavirus	joe biden	barack obama	america
3	joe biden	democratic party	green party	people
4	barack obama	republican party	joe biden	barack obama
5	mask	barack obama	libertarian party	democratic party

Table 2: Top 5 most common entities per class in training dataset

split	ratio matching
multimodal heighttrain	0.572
val	0.602

Table 3: Ratio of entities which are present in OCR provided

the percentage of entities present in the OCR of the image in the dataset. Some examples, such as in figure 4, have one of the entities to classify not present in neither the OCR nor the image, and must be classified from understanding of context, which makes the task more difficult.

5 Experiments

5.1 Experimental Setting :

To train and evaluate our different models, we used the Google Cloud Service with VM using the V100 GPU (16GB) and A100(40GB). We use the famous Pytorch framework with the Huggingface library in python. All our training used mixed precision and gradient accumulation in order to speed up some training time and allow larger model training.

5.2 Data Analysis :

Data Analysis was performed in order to understand the underlying problem better and find potential imbalances that could be leveraged for higher performances. The distribution of the number of entities per class, as well as each individual entity for each class was computed. Based on an a given entity, the aim was to try and predict which class it would most likely belong. An issue we came across was that some entities were mentioned in different ways: "americans" vs "american people". A rule-based approach was incorporated in an attempt to group these similar terms together.

Analysis was running on the OCR as well as the output of the celebrity detection model to determine if the entity was mentioned inside the text, in the image, both or neither. References to single entities in the textual format would vary, one example being for the entity "Donald Trump", which would be referenced as "Trump", "donald", "Donald Trump" to name a few. A rule based classifier was implemented to group these terms together for the entities that showed up most frequently.

A prediction was made based on the heuristics of the imbalances found to establish a baseline model, by classifying all the entities as "other", which is the class which contains over 75% of entities. Learning models would have to beat the accuracy of this rule based baseline to add value.

5.3 Augmentations :

Only one augmentation was used during the training. The augmentation was applied to the entity which needed to be classified. In fact, the entities provided were all without any punctuation and in lowercase format. We created a simple script which found the entity in the original text. The original text could contain punctuation and/or uppercase letter. We used this augmentation for the training, not the inference of the test set.

5.4 Unimodal NLP :

We trained a few competitive transformer architectures on text-only data, DeBERTa-v3 and RoBERTa.

5.4.1 DeBERTa

Two experiements were conducted for DeBERTa
(1) The first was a direct approach where we found the role for the entity based on the OCR extracted by the google model. The input of the transformer was as follows : "[CLS] Sentence OCR [SEP] entity to classify [SEP]"
(2) The second approach consisted of incorporating image signals in the unimodal training. We ran the celebrity face detection algorithm and further added these faces names text with the extracted OCR. The input of the transformer was as follow : "[CLS] Sentence OCR "\n" face name [SEP] entity to classify [SEP]"

We utilized both DeBERTa-small and DeBERTa-large for these experiments. During the training, a batch size of 16 was used, with a sequence length of 128 and a linear scheduler where the learning rate was reduced linearly during the training. The initial learning rate was $1e-5$, gradient accumulation is set at 3 epochs, and the optimizer used was AdamW. We trained these models for 6-7 epochs.

5.4.2 RoBERTa large

A batch size of 8 was used, with a sequence length of 275 and a linear scheduler where the learning rate was reduced linearly during the training. The initial learning rate was $5e-6$, and the optimizer used was AdamW. We trained these models for 6-7 epochs.

5.5 MultiModal

5.5.1 Naive Merging:

We used a batch size of 4 (A100 GPU), with a sequence length of 275. As a unimodal model, we use the face name in the text input processing. We use 4 sub images when they exist and the MEME image. We use an attention system inspired by the Word Attention in (Li et al., 2019a) , before concatenating the image features with the text features. We use a linear scheduler where the learning rate is reduced linearly during the training. The initial learning rate is $5e-6$, gradient accumulation is set at 3 epochs, and the optimizer used is AdamW. We trained these models for 7-8 epochs with early stopping of 2 epoch.

5.5.2 ViLT:

We use a batch size of 4, with a sequence length of 275. As unimodal model, we use the face name in the text input processing. We don't use here a linear scheduler, but ReduceLROnPlateau where the learning rate is reduced by a factor of 0.5 when there is no improvement during 5 epochs. The initial learning rate is $2e-5$, and the optimizer used is Adam. We trained these models for 7-8 epochs with early stopping of 2 epoch.

5.5.3 MultiModal : MMBT and VisualBERT

We use a batch size of 16, with a sequence length of 128. As for multimodal model, we use the image embeddings obtained from CLIP(Radford et al., 2021) and detectron2 (Ren et al., 2015) model individually for MMBT and VisualBERT. The text model used in both the architecture is bert. We use a linear scheduler where the learning rate is reduced linearly during the training. The initial learning rate is $1e-5$, gradient accumulation is set at 3 epochs, and the optimizer used is AdamW. We trained these models for 7-8 epochs with early stopping of 2 epoch.

5.6 Ensembling :

To improve the robustness of our solution we decide to combine 5 of our models (table 4). We chose the models to combine based on the results of the validation score and also the diversity they could bring. For instance, we did not select DeBERTa-v3-small because it is just a smaller version of DeBERTa-v3-large. We select only two multimodal models, as most of them perform quite badly compared to the unimodal. Otherwise they would just harm the ensemble.

6 Results and discussion

Just the simple experiment classifying all entities as "other" yielded 0.21 f1 score. We experimented with various models starting with just the text-based model, further adding image signals to using the image embeddings and finally a fully image-and-language based multimodal model to evaluate the model architecture efficiency in predicting a low resource multimodal problem. Here are some observations :-
(1) Unimodal - We can see the difference in results moving from "DeBERTa-v3-small" to "DeBERTa-v3-large" in Table 4. We can also see 2% improvement in the model when we tried to add image signal naively by adding the celebrity face name in text.
(2) Multi-Modal - We can see that multimodal model under performed a lot as seen in Table 4. We tried to fine-tune the Visual-BERT model and the mmbt model i.e. pre-trained vision-and-language model but they seem to under perform due to the lack of pre-training data. As they had been pre-trained on much less data and very different problem like VQA , it failed to capture the model understanding required for the transfer learning. So as to solve this issue we went ahead and utilised trained "DeBERTa-v3-large" model final output layer embeddings and concatenated them with pooled sub-image embedding with EfficientNetB7. Thus we utilised the transfer learning from both the models to give us the optimum results.
(3) Ensemble - The ensemble approach was our final approach where we combined all the different

Model	F1-score val (macro)	F1-score test (macro)
(a) DeBERTa-v2-xlarge w/o face's name	0.54	0.53
(b) DeBERTa-v3-small w/o face's name	0.46	0.46
(c) DeBERTa-v3-small w face's name	0.48	0.47
(e) DeBERTa-v3-large w/o face's name	0.55	0.55
(f) DeBERTa-v3-large w/ face's name	0.56	0.57
(g) RoBERTa-large w/ face's name	0.53	0.51
(h) ViLT w face's name	0.42	0.42
(i) Naive Multi Modal (DeBERTa-v3-large + EfficientNetB7) w/ face's name	0.525	0.55
(j) MMBT (BERT + CLIP) w/ face's name	0.48	0.46
(k) VisualBERT w/ face's name	0.43	0.44
Ensembling Mean(a, f, g, h, i)	**0.578**	**0.583**

Table 4: Experiments Results

Rank	Team	Final accuracy
1	Logically	**58.671%**
2	c1pher	55.240%
3	zhouziming	54.707%
4	smontariol	48.483%
5	zjl123001	46.177%
6	amanpriyanshu	31.943%
7	fharookshaik	23.855%
8	rabindra.nath	23.717%

Table 5: Constraint22 Leaderboard

model outputs . We tried various ensembles and blending techniques but we got the best LB score with averaging of ViLT, RoBERTa large, DeBERTa large, naive multimodal and DeBERTa-xlarge models. Final test set results and competition leaderboard are presented in Table 5. Our best model ("Ensemble") outperforms all competition systems and best baseline models. Test result of $Ensemble$ model achieved 0.58 avg. F1.

7 Conclusion

We described our participation in the CONSTRAINT 2022 Shared Task on "Detecting the Hero, the Villain, and the Victim in Memes" with the implementation of various models. Ensemble model based system outperforms all the models on val set and test set. A challenge in this task is the low resource of data available for training models. Hence, transfer learning provides the best results.
The best performing model in this competition combines the simple averaging of ViLT, RoBERTa large, DeBERTa large, naive multimodal and DeBERTa xlarge models. The ensemble seems to perform the best as the data size is small and we use a large model to allow for better transfer learning, This ultimately leads to some overfit of models but applying the averaging improves the results, like the boosted trees systems.

We found that there were two major challenges with the problem :- (i) The entities were sometimes not present in the image or the text. (ii) The size of data required to learn this implicit learning was not sufficient. This ultimately undermines the performance of our deep learning architecture.
Creating a dataset for real-word multimodal problems, particularly the natural language inference problem of role labelling is challenging (Le Bras et al., 2020). We appreciate the work by the CONSTRAINT 2022 organizers, yet, a more elaborate and extensive data would make this dataset more suitable for benchmarking. As an emergent research field, we hope our extensive model analysis and proposed solutions can act as baseline and inspire further work.

References

Pranav Adarsh, Pratibha Rathi, and Manoj Kumar. 2020. Yolo v3-tiny: Object detection and recognition using one stage improved model. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 687–694. IEEE.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021. A survey on multimodal disinformation detection. *arXiv preprint arXiv:2103.12541*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jie Gao, Hella-Franziska Hoffmann, Stylianos Oikonomou, David Kiskovski, and Anil Bandhakavi. 2021. Logically at the factify 2022: Multimodal fact verification. *arXiv preprint arXiv:2112.09253*.

Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. 2018. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7181–7189.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*.

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. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088. PMLR.

Jianping Li, Yimou Xu, and Huaye Shi. 2019a. Bidirectional lstm with hierarchical attention for text classification. In *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 1, pages 456–459. IEEE.

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

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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*.

Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, Sherzod Hakimov, and Ralph Ewerth. 2021. Multimodal news analytics using measures of cross-modal entity and context consistency. *International Journal of Multimedia Information Retrieval*, 10(2):111–125.

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

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020. Semeval-2020 task 8: Memotion analysis–the visuolingual metaphor! *arXiv preprint arXiv:2008.03781*.

Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Findings of the constraint 2022 shared task on detecting the hero, the villain, and the victim in memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations - CONSTRAINT 2022, Collocated with ACL 2022*.

Emanuel H Silva and Ricardo M Marcacini. Aspect-based sentiment analysis using bert with disentangled attention.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

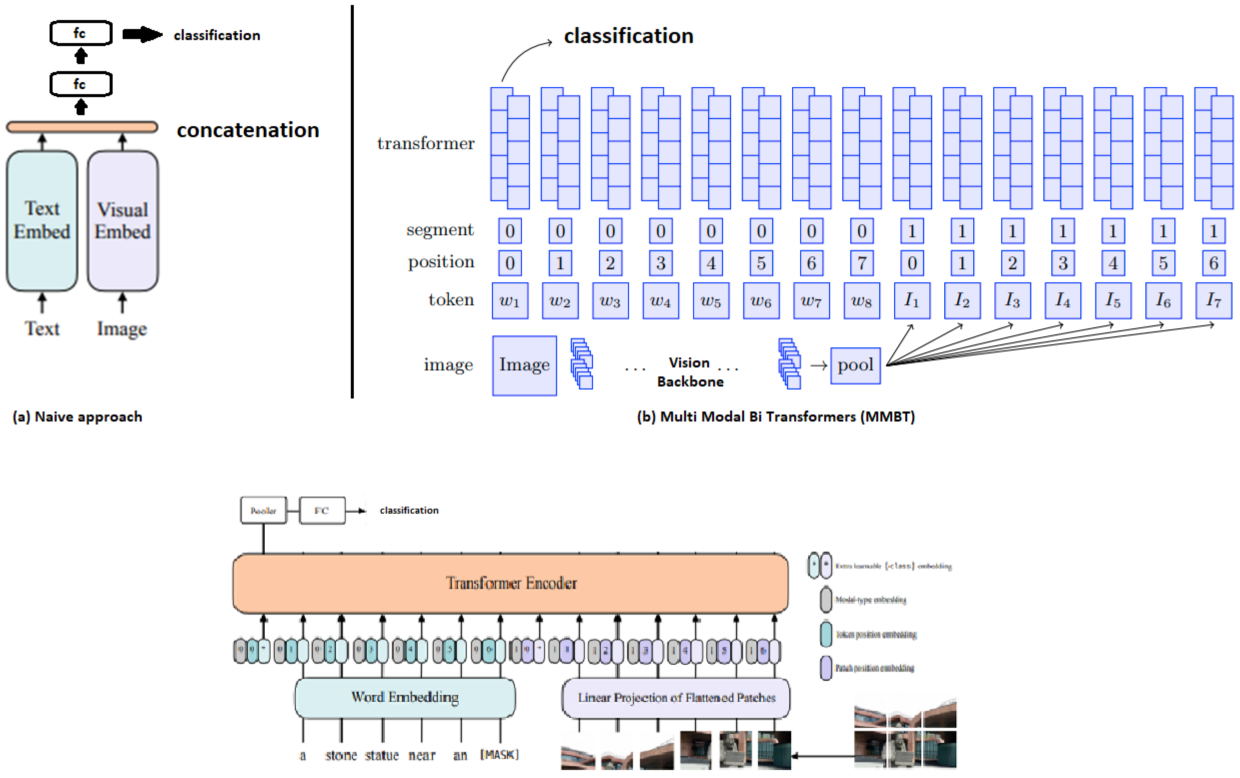
Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, pages 1–50.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.

A Appendix I



(a) Naive approach

(b) Multi Modal Bi Transformers (MMBT)

(c) ViLT

Figure 5: Example of Multimodal Architecture used