Hero, Villain and Victim: Dissecting harmful memes for Semantic role labelling of entities

Hanzalah Firdausi (MT21027)
Harsh Vardhan Bhadauriya (MT21122)
Shreyansh Jain (MT21089)

Abstract

In the present political and social scenario, memes play a pivotal role in swaying public opinion. It has almost become a necessity to analyze how memes impact us deeply. The latest work in this field is to figure out who is being glorified or victimized in a meme. We have proposed our multimodal architecture that helps to tackle this problem. We're hoping to shed some light on this aspect of memes with our initiative. We hope that by alerting readers when an entity is being glorified or victimised, they will be able to make unconscious bias explicit, allowing readers to apply their social media literacy abilities better.

1 Introduction

Today's social media is a juggernaut, and the information it contains comes in various forms or modalities. Some of the modalities are images, text, and audio. The information present in today's social media is very complex. It comes in the form of fused modalities like fused images and text, which we call memes.

In the past, Computer Vision and NLP communities have only considered a single form of modality to produce results and generate inferences.

However, the growing popularity of mixed modalities such as memes has forced us to see the information in our hands as both images and text rather than only text or image because mixed modalities such as memes are often misleading if we leverage the information present in them in the form of image or text separately.

1.1 Motivation

In the past few years, many works have been done in mixed modalities such as sentiment analysis of memes, emotion analysis of memes and categorising memes into harmful, humorous, or neutral classes.

However, none of the works has tapped into the side of the memes, which uses the content, context, and narrative structures to present people, organisations, entities in the memes. These contents, narrative structures, context allow readers to associate elements or entities present in the memes with familiar stereotypes, well-known characters, and recognisable outcomes. Meme creators can use this technique to cast real people or organisations as heroes, villains, or victims which was proposed by Gomez *et al.* [3].

We're hoping to shed some light on this aspect of memes with this initiative. We hope that by alerting readers when an entity is cast in one of these roles, they will be able to make unconscious bias explicit, allowing readers to apply their social media literacy abilities better.

1.2 Problem Statement

The task emphasizes detecting which entities are glorified, vilified, or victimized within a meme. The objective is to classify a given combination of a meme and an entity into Hero, Villain, Victim, or others.

1.3 Project Pipeline Summary

We started out with a baseline model which was a combination of LSTMs. After that we tried to implement a co-attention model which fuses all three context, entity and visual feature at the same time. In this model, we tried with different word embedding such as Glove and DistilBert. Also, we tried the above architecture with transformer. Later, we also tried an architecture in which context and entity were aligned first and then this context-entity alignment was fused with visual features extracted from pre-trained model such as ResNet-152. We have mentioned our architecture in detail in subsequent headings and deep dived into the results achieved and why they are such.

	Loss	Accuracy	Macro F1
Train Set	1.11	72.38	72.23
Validation Set	3.97	47.42	28.73
Test Set	3.37	56.03	30.95

Table 1: Table for Co-attention (Distilbert) Macro F1 Score

	Loss	Accuracy	Macro F1
Train Set	1.87	72.31	71.9
Validation Set	4.31	51.86	33.23
Test Set	3.79	54.96	32.27

Table 2: Table for Co-attention (Transformer) Macro F1 Score

2 Dataset

We are given a set of memes divided into training and validation sets for training the model and keeping in check if the model is learning well. Each meme will have an image associated with it and a JSON description. JSON description will have OCR text, a list of entities and their role (hero, villain, victim and others).
We also have an unseen test set which has an image and JSON description. Here, JSON description consists of OCR text and list of entities but does not have any role or tags associated with them. The task is to predict the role/tag associated with all the entities specified in each test example.

The original dataset is merged and each entity present in a datapoint is made into a new data point on the basis of the role it has. After doing this step we have approx 17k data points but the data is hugely imbalanced with the role of "other" comprising more than 70 percent of data points. The images are made text free using Keras functions. The combined dataset is used for training, validation and testing.

3 Related Work

We explored different fusion techniques to combine text and image data in early, late, and hybrid fusion. In Early-fusion (Poria *et al.*) [5], the text and image data are integrated into a single feature vector. Still, early fusion techniques cannot extract useful information from different modalities and often generates a large vector with redundant data.

Late fusion techniques (Erik Cambria *et al.*) [1] treat both image and text modality separately, which are processed independently, and later the feature vectors are combined. These methodologies treat both modalities as independent of one another and lead to poor performance when different modalities are inter-dependent and connected, which may not be accurate in practice.

Hybrid fusion techniques (Cao *et al.*) [2] integrate the different modalities using heuristics and create a shared representation and are better than the early and late fusion techniques in practice.

Zhang *et al.* [9] proposed the methodology to extract the image feature vector by capturing the spatial features of different regions in an image using the deep learning model. Character-level embeddings are used to capture useful morphological information. Both the modalities are concatenated together to represent a multimodal representation of the data.

Pramanick *et al.* [6] proposed the Multihop attention model (MHA) to understand the correlation between the background image and the textual segment in the present in different spatial locations. A correspondence is created between the background image and the foreground text to create a feature vector with different attention model mechanisms.

After figuring out the multimodal representation of the meme, our next task is to create a vector representation for each entity-meme pair. This task can be referred to as entity disambiguation or entity linking.

Nie *et al.* [4] have proposed a Context-Entity Co-Attention model that works on textual context provided. Our aim in this project would be to extend the model and use it when the context is multimodal. This model can be further modified to account for the type of the entity and add greater details to the vector generated. General architecture would be to have vector representation of context and entity (using bi-LSTM) act as input. We would generate an attention probability of each vector in context and take the weighted sum of all contextual vectors.

Zhang *et al.* [9] also proposed a co-attention model for entity representation in multimodal context. Adaptive Co-Attention Network is able to filter out the noise and pinpoint the regions which are highly related to the current entity. Word context generated from text and image context generated from meme are fed into a single layer neural network and then a softmax layer is applied to generate attention distribution. First we generate a

new image vector using weighted sum of attention probabilities with the original image vector. Then this new image vector is used to generate a new text vector using the method described previously. Both the visual and textual attention are fed to a Gated Multimodal fusion model to get a combined representation.

The combined representation received from above co-attention models can be fed to a deep architecture along with entity labels to train. This would give us a model capable of assigning hero, villain or victim tags to entities.

Wang *et al.* [7] proposed a multi-perspective approach to match all the time steps of a sentence to all the time steps of another sentence. They have proposed the use of a Bi-LSTM encoder, four different matching techniques, and finally an LSTM to aggregate the result of all matching techniques. We have used this approach to find attention between context text and entity.

As we can see in Figure 1, a single time step of one sentence is matched with all the time steps of sentence two to bring all different kinds of alignments. The complete proposed architecture can be seen in Figure 2. We try to employ the matching/alignment technique mentioned in the paper and use their aggregation layer technique to find the share representation of both context and entity.

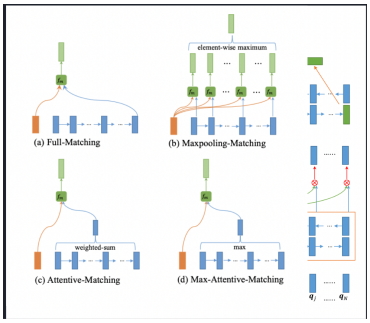

Figure 1: Taken from Wang *et al.* [7]. All matching technique from one time step

Yu *et al.* [8] proposed an entity sensitive attention network for multimodal sentence classification. The technique focuses on the idea of finding alignment between context and entity first and then doing a fusion of visual features to get the final representation. Context is divided in left and right parts based on where the entity lies in the sentence and both left and right part are aligned separately and then fused together. This representation is then fused with visual representation extracted from the ResNet or any other pre-trained deep network that

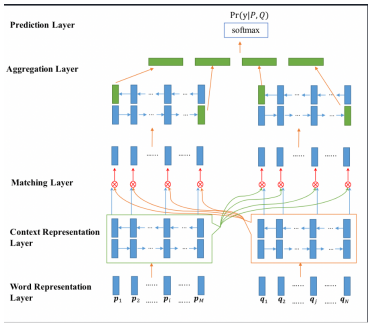

Figure 2: Taken from Wang *et al.* [7]. Fusion of all matching technique

can extract mid-level features.

The approach focuses heavily on finding the left and right context. In our scenario, the entity text is missing in the OCR output so we have to use Wang *et al.* [7] approach to align the context and entity. We have used Yu et al. approach to extract visual features from the image and combine them with context-entity attention.

4 Methodology

4.1 Baseline Model

The baseline consists of two LSTMs. The first LSTM is used to encode the OCR text data and then pass the context it generates as the hidden state to the other LSTM which takes the entity name as input. Both the LSTMs have two layers and a dropout of 0.5 between the LSTM layers is used for the purpose of regularisation. The baseline method does not take into account the visual features present in the dataset. The final output generated by the second LSTM is then passed through a Linear Layer and softmax and cross-entropy on top of it are used to train the model. The optimiser Adam with weight decay is used to propagate the gradient and prevent the model from overfitting.

4.2 Co-attention Models

The co-attention model is based on the [4] as the attention model for the context and entity. The visual features are extracted using a pre-trained backbone in VGG16. The text features are extracted separately for entity and OCR-text, and co-attention is applied between the entity and OCR-text and image vector and OCR-text. The co-attention is applied to both the features and the concatenated features are passed through the fully connected layer. Output is calculated, BCE Loss is used to compute the error, and the AdamW optimizer is used to propagate the

gradient. The basic co-attention model uses glove embeddings. Distilbert embeddings are used as well, and the Bi-LSTM is swapped out with transformers as part of the experimentation to produce a different co-attention network.

4.3 BiMPM-MultiModal Network

In the Co-attention network, all three inputs were processed together. In the architecture, we have processed context and entity together to get a context-entity alignment and then introduce visual features to it. Methodology of BiMPM-MultiModal Network:

- We use the approach of Wang *et al.* [7] to find alignment between context and entity.
- Visual features are extracted with help of ResNet-152 as described in Yu *et al.* [8] approach.
- The visual feature is of dimension 2048x7x7. We have 49 32x32 blocks with each block represented by 2048 dimensions.
- The visual features are then passed through a linear layer to reduce dimension and match that of context-entity attention.
- The features are concatenated and passed through a prediction layer consisting of a two-layer neural network with tanh as activation function and dropout to regularize the network.
- Output is sent to the sigmoid activation layer and BCELoss is used to compute error and Adam optimizer is used to propagate the gradient.

5 Experiments

5.1 Baseline Model

The optimiser Adam with weight decay with a learning rate of 3e-4 is used to propagate the gradient and prevent the model from overfitting. The model was trained for 40 epochs. The baseline model is trained on a sampled data of 9k out of 17k data points to maintain the class balance. The baseline model achieved similar results on validation and test set, which indicates the model is generalized enough to accommodate for unseen examples and produce similar results.

5.2 Co-attention Models

The optimiser AdamW with a learning rate of 3e-4 is used to propagate the gradient and prevent the model from overfitting. The models are trained on a sampled data of 9k out of 17k data points to maintain the class balance. The results are similar for validation and test set which indicates the model is generalized enough to accommodate for unseen examples and produce similar results. We have experimented with Co-attention model with Glove embeddings, Distilbert embeddings and swapping out the Bi-LSTM module with the transformer as well as part of attention module.

5.3 BiMPM-MultiModal Network

At first, all 17,528 training data points were trained even though 78.2 percentage of them belonged to a single category making the data highly imbalanced. The training was done with help of Adam optimizer and a learning rate of 1e-5 for 20 epochs.

After that training was done from the scratch on undersampled data of 9k points in one epoch such that class balance is maintained. For this purpose pytroch WeightedRandomSampler is used. In the second approach, training was divided into two parts- the first 20 epochs were trained with a learning rate of 3e-4 and no regularizer and the next 10 epochs were trained with a learning rate of 1e-5 and weight decay of 1e-2 to keep the model from overfitting.

6 Results and Analysis

6.1 Baseline Model

The Baseline model is able to achieve a macro F1 of 0.30 (Figure 3) on both validation and test. The model although simple is still able to learn complex details present in the dataset and can produce better results if trained properly for more epochs.

	Loss	Accuracy	Macro F1
Train Set	0.62	77.58	47.84
Validation Set	0.9	72.04	29.18
Test Set	0.84	74.23	30.88

Table 3: Table for Baseline Macro F1 Score

6.2 Co-attention Models

The validation set has a very low value of 24.68 percent because the glove embeddings are used

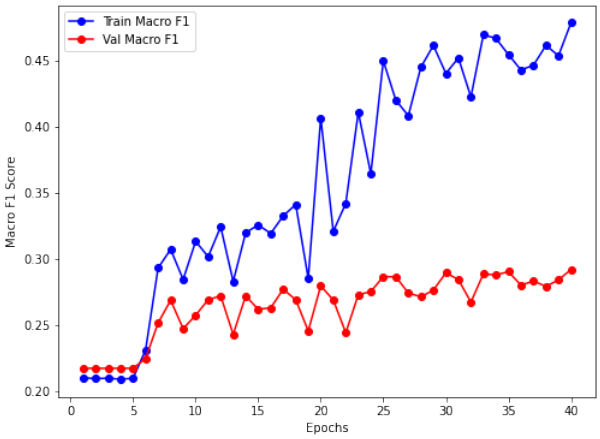

Figure 3: Baseline Macro F1 Score

	Loss	Accuracy	Macro F1
Train Set	1.13	52.46	42.94
Validation Set	1.31	54.83	28.64
Test Set	1.31	52.35	24.68

Table 4: Table for Co-attention (glove) Macro F1 Score

which do not take into account the context of text data as seen in table 2.

The Macro F1 score increases significantly when the Distilbert embeddings are used as it takes into account the context of the text data used. The macro f1 score increases from 24.68 percent to 30.95 percent for the validation set as given in table 3.

The macro f1 score is maximum for the co-attention network with transformer with the value of 32.23 percent (Table 2) as the transformer module itself has multihead attention as part of its architecture.

6.3 BiMPM-MultiModal Network

Here we will deep dive into the results of our training regime. In Table 5, we can see the results of training on whole dataset and results on training and validation set after 20 epochs.

	Loss	Accuracy	Macro F1
Train Set	0.390354	86.9580	62.25
Val Set	0.7999	74.8434	38.52

Table 5: Results of training on the whole train set without any sampling

As we can see in the Table 5, as expected model starts to overfit towards the imbalanced class and the disparity can be seen in macro F1 score.

	Loss	Accuracy	Macro F1
Train Set	0.127320	96.78	96.47
Val Set	1.9322	72.5783	42.25
Test Set	1.5336	74.8	42.64

Table 6: Results of training on the 9k undersampled train set with class balance

Analysis of results of training a sampled dataset:

- In Table 6, High results on the train set is because statistics were generated on the sampled set and not the complete set.
- We are achieving similar results on validation and test set, which indicates the model is generalized enough to accommodate for unseen examples and produce similar results.
- n Table 6, we are able to achieve our highest macro F1 on both validation and test which is around 42
- The model can be made more accurate by giving it more training time and increasing the number of samples when overfit is observed and adding more weight decay and with a lower learning rate.
- We can see in the macro F1 graph in Figure 4, that with time validation graph does not go down which shows that the regularizer is working well and dividing the training into multiple parts is helpful to our cause.

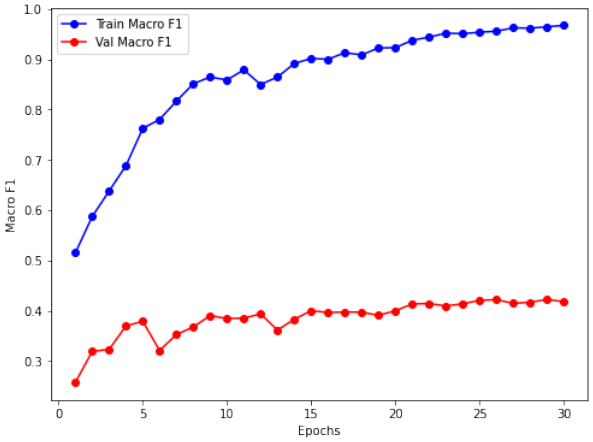

Figure 4: BIMPM-Multimodal Macro F1

7 Conclusion

The task at our hand was quite difficult because of multiple challenges such as highly imbalanced class distribution, poor OCR text generation, noise in the dataset and poor quality of images. We tackled the problem statement by finding different ways of aligning context, entity and visual features. The alignment was then passed through linear network for classification. We have also tried to use different word embedding, architectures and training regime to get to the better results. We are able to achieve 42.64 macro F1 on test data with BiMPM-MultiModal Network.

There is a possibility of further increasing the performance of the learning model by performing hyper parameter tuning on existing best model and also trying out new alignment architectures which can take advantage of context-entity relation in the best possible way.

8 Individual Contribution

- Harsh Vardhan Bhadauriya- Preprocessing, BiMPM-MultiModal Model, Co-attention Network
- Hanzalah Firdausi - Preprocessing, Baseline Model, Co-attention Network, Glove Embedding
- Shreyansh Jain - Preprocessing, Co-attention Network, Glove and Distilber Embedding, Transformer

References

[1] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. Affective computing and sentiment analysis. In *A practical guide to sentiment analysis*, pages 1–10. Springer, 2017.

[2] Donglin Cao, Rongrong Ji, Dazhen Lin, and Shaozi Li. A cross-media public sentiment analysis system for microblog. *Multimedia Systems*, 22(4):479–486, 2016.

[3] Diego Gomez-Zara, Miriam Boon, and Larry Birnbaum. Who is the Hero, the Villain, and the Victim?: Detection of Roles in News Articles using Natural Language Techniques. In *23rd International Conference on Intelligent User Interfaces*, pages 311–315. ACM, 2018.

[4] Feng Nie, Yunbo Cao, Jinpeng Wang, Chin-Yew Lin, and Rong Pan. Mention and entity description co-attention for entity disambiguation. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[5] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 439–448. IEEE, 2016.

[6] Shraman Pramanick, Md Shad Akhtar, and Tanmoy Chakraborty. Exercise? i thought you said'extra fries': Leveraging sentence demarcations and multi-hop attention for meme affect analysis. *arXiv preprint arXiv:2103.12377*, 2021.

[7] Zhiguo Wang, Wael Hamza, and Radu Florian. Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*, 2017.

[8] Jianfei Yu, Jing Jiang, and Rui Xia. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439, 2020.

[9] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. Adaptive co-attention network for named entity recognition in tweets. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.