Multimodal Sentiment Analysis of #MeToo Tweets using Focal Loss (Grand Challenge)

Priyam Basu
Electrical and Electronics Engineering(EEE)
Manipal Institute of Technology, Manipal
Manipal, India
priyam.basu1@learner.manipal.edu

Soham Tiwari
Computer Science and Engineering (CSE)
Manipal Institute of Technology, Manipal
Manipal, India
soham.tiwari@learner.manipal.edu

Joseph Mohanty
Computer Science and Engineering (CSE)
Manipal Institute of Technology, Manipal
Manipal, India
joseph.mohanty@learner.manipal.edu

Sayantan Karmakar
Computer Science and Engineering (CSE)
Manipal Institute of Technology, Manipal
Manipal, India
sayantan.karmakar2@learner.manipal.edu

***Abstract*—The #MeToo trend has led to people talking about personal experiences of harassment more openly. This work attempts to aggregate such experiences of sexual abuse to facilitate a better understanding of social media constructs and to bring about social change [1]. We propose an approach to multimodal sentiment analysis using deep neural networks combining visual analysis and natural language processing. Our goal is different than the standard sentiment analysis goal of predicting whether a sentence expresses positive or negative sentiment; instead we try to detect the stand of a person on the topic and deduce the emotions conveyed. We have made use of a Multimodal Bi-Transformer (MMBT) model [2] which combines both image and text features to produce an optimal prediction of a tweet's stand and sentiments on the #MeToo campaign.**

***Index Terms*—Deep Learning, Multi Modal, Sentiment Analysis, Visual Analysis, Residual Networks, Natural Language Processing.**

I. Introduction

Global estimates indicate that about 1 in 3 women worldwide has experienced either physical and/or sexual intimate partner violence or non-partner sexual violence in their lifetime [3]. The Me Too (or #MeToo) movement, with variations of related local or international names, is a movement against sexual harassment and sexual abuse where people publicize allegations of sex crimes committed by powerful and/or prominent men.The #MeToo campaign has brought to light acts of sexual harassement by spreading social awareness via online social platforms.

Sentiment analysis has been an active area of research in the past decade, especially on textual data from Twitter. Natural language processing (NLP) techniques can be used to make inferences about mental states of the people from what they write on Facebook, Twitter [4], and other social media.

Initially, we tried out multiple methods. At first, we tried to analyse only the text. We used different ML classifiers such as support vector machines, logistic regression and random forests but later realised they work in only one direction and we needed to use a bidirectional model to get contextual meaning as this project was highly dependant on that. Next, we tried out a Bidirectional LSTM model but that model highly overfit. Lastly, we looked into Transformer based models and on further research, found that they perform better for sentiment analysis. We implemented Google AI's BERT [5] and Facebook AI Research's (FAIR) Roberta [6] on the text data and found out that they gave us better results.

We finally came across FAIR's paper on Multi Modal Bidirectional Transformers [2] (MMBT). We finally decided to move forward with this architecture since it combined image and text embeddings using premium architectures like ResNet-152 [7] and BERT [5], thereby extracting meaning from both kinds of data. Our approach in a nutshell is to use the MMBT architecture, where we pass the images through a ResNet-152 [7] architecture and obtain the image embeddings from there. Then we apply BERT [5] embeddings on our tokenized text data and obtain the text embeddings. We reshape our image embeddings and concatenate with them our text embeddings and pass the combined data embeddings through a BERT Encoder, followed by a Batch Normalisation Layer and final Classification Layer to obtain our predictions.

II. Dataset

A. The #MeToo Twitter Dataset

Twitter [4] is an American microblogging and social networking service on which users post and interact with messages known as tweets. Registered users can post, like, and retweet tweets, but unregistered users can only read them. Along with texts, they can also post images.

The provided #MeTooMA [8] dataset has been used to work with and train our model on [8] [9] [1]. The dataset given to us had tweet IDs of 9973 tweets. Using the Twitter Developer API, only 6809 of the original 9973 tweets were scraped, since many of the original tweets had been deleted by the person who posted them. Thus, we had to work with the 6809 tweets

978-1-7281-9325-0/20/$31.00 ©2020 IEEE
DOI 10.1109/BigMM50055.2020.00076

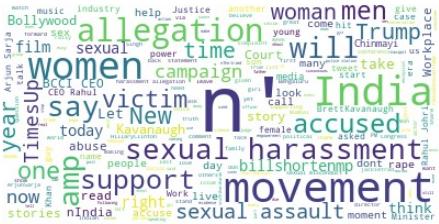

Fig. 1. Allegation

Fig. 2. Justification

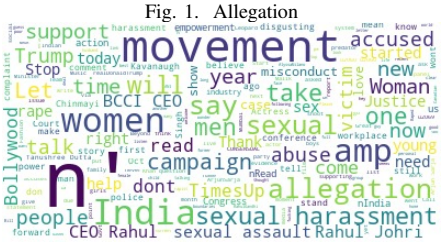

Fig. 3. Directed Hate

Fig. 4. Generalized Hate

we had. The associated images were also scraped, in order to enable us to perform multimodal sentiment analysis.

The texts that were extracted from the tweets contained the original text (including emojis, numbers etc.) as well as the link to the original tweet. For the task of this paper, these links from the texts were cut and used them to scrape the images. The curated dataset is the result of annotations by domain experts over three months from October 2018 to December 2018 [8] [9] [1]. The dataset addresses relevant problems affecting the current social media space and has the ability to provide interesting analysis pertaining to multiple facets of a social movement. The tweets have been manually annotated into five linguistic aspects- relevance, stance, hate speech, sarcasm and dialogue act, which have been further divided into features having binary values of 0 or 1.

While preprocessing, only the twitter links from the text were removed. No stop words were removed, since we speculated they might have contextual meaning. This data was then passed onto our model.

Some Exploratory Data Analysis was carried out on the dataset. First, the class imbalance in each of the feature columns was found as shown in Table I.

Next, how many tweets among the features have a common under-balanced class, pair wise, were found out. The top 6 pair wise relations with the most such tweets are shown in Table II.

Finally, for EDA, the most commonly used words in the tweets belonging to the under-balanced class for four of the features were found. The word clouds for these columns are given in Fig. 1, Fig. 2, Fig. 3 and Fig. 4.

III. Methodology

The model architecture as explained in FAIR's paper on MMBT [2] has been put to use for our purposes. The MMBT model from FAIR makes use of the BERT [5] pretrained model and the ResNet-152 [7] pretrained model weights to use text and image data to make predictions. The majority of the model architecture [2] has been kept the same, with only a few tweaks to the architecture, with the addition of Batch Normalisation Layer and the use of a different loss function.

A. Text Tokenization

The text from each tweet was first tokenized. The tokenizer used with the BERT architecture, was used to tokenize our input sequences. The maximum length of the sequences should be limited to a length of *N-M-2* tokens; where, *N* is the maximum sequence length; and *M* is the number of image embeddings; which is visible ahead.

B. Image Encoder

The image encoder network used here is the same as that used in MMBT. The pretrained ResNet-152 model with the last classification layer and the second last adaptive average pooling layer removed, was used to obtain embeddings for the input image. Then adaptive average pooling was performed on the output obtained to obtain *M* image embeddings of 2048 dimensions each which were then again passed through another simple linear layer to obtain *M* image embeddings of 768 dimensions each. Note that 768 is the number of hidden dimensions of bert-base-uncased model.

C. MMBT BERT Encoder

The image encoder outputs *M* encodings for the image. To obtain the embeddings for the text tokens, the text tokens were converted to respective text embeddings using the word embeddings provided by BERT [5]. Next, the [CLS] token embedding, the *M* image embeddings, [SEP] token embedding and the text embeddings were concatenated together. The image embeddings were assigned a segment ID 0 and the text embeddings segment ID 1. These concatenated embeddings were then passed through the BERT encoder. The BERT encoder gave the pooled outputs of the last hidden state of the BERT encoder.

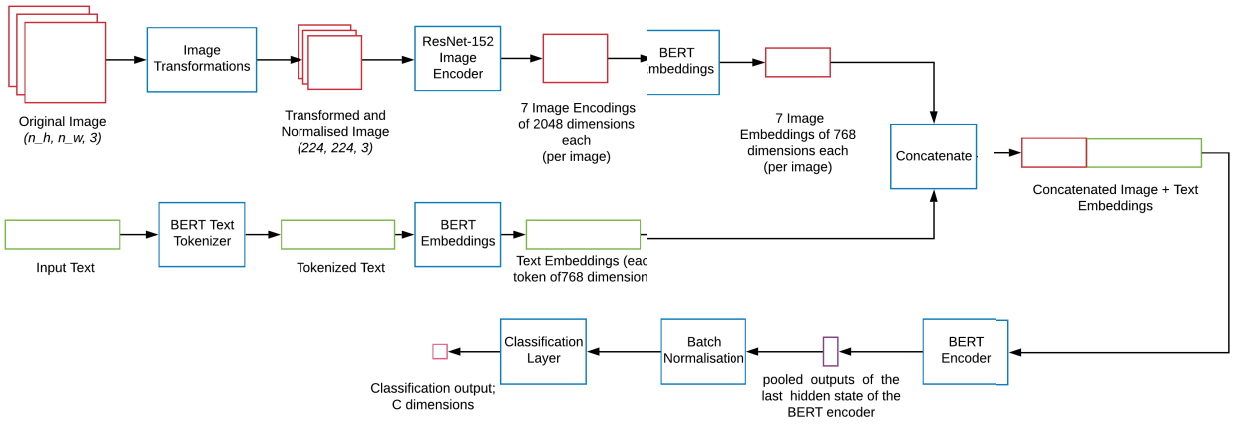

Fig. 5. Methodology flow chart

TABLE I
DATA IMBALANCE TABLE

Feature	% of 0s	% of 1s
Text_Only_Informative	0.274	0.726
Image_Only_Informative	0.6698	0.3302
Directed_Hate	0.9622	0.0378
Generalized_Hate	0.9718	0.0282
Sarcasm	0.9789	0.0211
Allegation	0.9466	0.0534
Justification	0.9668	0.0332
Refutation	0.9792	0.0208
Support	0.6813	0.3187
Oppose	0.9262	0.0738

TABLE II
COMMON UNDER-BALANCED CLASS TABLE

Feature1	Feature2	Frequency
Directed_Hate	Allegation	143
Directed_Hate	Support	191
Generalized_Hate	Support	96
Allegation	Support	288
Justification	Support	83
Refutation	Oppose	83

D. Classification Layer

The output obtained is a 768 dimensional vector. This was then passed through a batch normalisation layer, followed by a linear layer where the weights W have dimensionality [768, C]; where C is the number of classes in the final layer of the classifier.

E. Pre-training

The use of ResNet-152 pretrained on ImageNet [10]; and the 12 hidden layered, 768-dimensional BERT model trained on Wikipedia's data in English has been made. This above stated BERT model is available as bert-base-uncased model in the transformers library in PyTorch.

F. Loss Function

The high data imbalance in positive and negative samples of every column led to our models only predicting the negative class. Hence instead of using the standard cross-entropy loss, Focal Loss [11] and Dice Loss [12] were looked at to help account for the imbalance in the data.
In Focal loss, the weights of the loss due to correct classification were reduced, and the weights of the loss due to misclassification remained the same. This helped in providing relatively more importance to the losses due to misclassification. Focal loss was found to be better compared to Dice Loss for model performance.

IV. EXPERIMENT SETUP

There were various operations we performed to tune the hyperparamters of our model. First, in order to reduce the data imbalance in the negative and positive samples for a particular class, the texts of the positive text samples, were duplicated, and were then again added to our original dataset after shuffling the words of the sentences.
All input images were augmented. They were resized to (256, 256, 3) and then center crop was performed, hence obtaining images of dimensions (224, 224, 3). The images were then normalised using the following values of mean - [0.46777044, 0.44531429, 0.40661017]; and the following values of standard deviation - [0.12221994, 0.12145835, 0.14380469], for the three channels.
The number of image embeddings M was set to 7, as was used in the MMBT paper [2]. The max sequence length N was set to 128.
The BertAdam optimizer was used with an initial learning rate of 0.001. A learning rate scheduler was also used that multiplied the learning rate by a factor of 0.1 when the ROC AUC score of our model would start to plateau.
The value of Focal Loss using the values alpha = 1 and

TABLE III
ROC AUC SCORE OF EACH COLUMN WITH DIFFERENT AMOUNTS OF MINORITY DATA DUPLICATION

Model Name	Text_Only_Informative	Image_Only_Informative	Directed_Hate	Generalized_Hate
No Duplicate	0.5086	0.5214	0.4782	0.5021
Single Duplicate	0.5119	0.5085	0.5076	0.5335
Double Duplicate	0.5	0.5000	0.5397	0.5147
Triple Duplicate	0.5021	0.5000	0.5045	0.5102

TABLE IV
ROC AUC SCORE OF EACH COLUMN WITH DIFFERENT AMOUNTS OF MINORITY DATA DUPLICATION

Model Name	Sarcasm	Allegation	Justification	Refutation	Support	Oppose
No Duplicate	0.4866	0.5063	0.5537	0.5390	0.5055	0.4997
Single Duplicate	0.5000	0.5205	0.5112	0.5691	0.5117	0.5016
Double Duplicate	0.5073	0.5033	0.5000	0.5000	0.5007	0.5184
Triple Duplicate	0.5690	0.5161	0.5002	0.5000	0.5121	0.5033

TABLE V
ABLATION ROC AUC SCORE OF EACH COLUMN WITH SINGLE DATA DUPLICATION

Model Name	Text_Only_Informative	Image_Only_Informative	Directed_Hate	Generalized_Hate
Only Text	0.4937	0.5032	0.5332	0.5720
Only Images	0.5000	0.4976	0.4442	0.5259

TABLE VI
ABLATION ROC AUC SCORE OF EACH COLUMN WITH SINGLE DATA DUPLICATION

Model Name	Sarcasm	Allegation	Justification	Refutation	Support	Oppose
Only Text	0.4402	0.4991	0.5651	0.5714	0.5024	0.4798
Only Images	0.4420	0.5009	0.4877	0.5106	0.5000	0.4630

gamma = 2 which were the suggested values by the original paper [11].
In case of Dice Loss, a smoothing value of 1 was used.

V. RESULTS

Our model was trained on 1, 2 and 3 epochs on four different kinds of data:

- Under-sampled class not duplicated
- Under-sampled class once duplicated
- Under-sampled class twice duplicated
- Under-sampled class thrice duplicated

We also have tabulated the ablation results for the different modalities, using focal loss and single data duplication in Table V and Table VI. We can observe that our model performs better when we use both the text and image modalities, rather than using only one of them.
Focal loss [11] was implemented in our final model. Dice Loss was also implemented but it did not give us better results so we subjected to Focal loss for our final predictions. Different kinds of duplication gave us different ROC AUC scores on different epochs, not following any order, i.e. increasing the number of epochs did not always give better results. It was also observed that the ROC AUC scores we obtained were not very high, that was mostly indebted to the fact that our data was highly imbalanced. The highest AUC score that was achieved by us across all epochs for each type of data on each feature column has been shown in Table III and Table IV.

VI. CONCLUSION

In this paper, FAIR's MMBT [2] model was used by us with slight changes to the original architecture. This was used to analyse the public sentiment on the #MeToo movement on Twitter. A major problem faced during this project was the high class imbalance which we tried to solve by duplicating the under sampled data and implemented focal loss [11], which imposes a heavier penalty for wrong prediction of under sampled class compared to penalty imposed on correct prediction. We found that MMBT [2] gave us the best results for multimodal analysis. We will try to further improve upon this project by enhancing our existing model architecture.
There is research being carried out, using data from various social media platforms such as Twitter [4] to analyse the emotional conditions and behaviour of different people with regard to different social issues and problems [13] [14] [15]. Through our project, we urge other people to further indulge into research in analysing public stance on sexual harassment which may help us to make people more aware of such social issues, thereby empowering them and making workplaces and homes safer. We hope to take this model further and try to implement it with more data and get a better demographic of people who have faced similar situations and analyse the problems they have faced and thereby, try to support people though mental health issues and create social awareness. The complete code to our project is available on GitHub [1].

[1] https://github.com/whopriyam/IEEE-BigMM-Grand-Challenge-2020

REFERENCES

[1] A. Ghosh Chowdhury, R. Sawhney, R. R. Shah, and D. Mahata, "#YouToo? detection of personal recollections of sexual harassment on social media," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2527–2537. [Online]. Available: https://www.aclweb.org/anthology/P19-1241

[2] D. Kiela, S. Bhooshan, H. Firooz, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," 2019.

[3] N. Andalibi, O. Haimson, M. Choudhury, and A. Forte, "Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity," 05 2016, pp. 3906–3918.

[4] Wikipedia contributors, "Twitter — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Twitter&oldid=971066748, 2020, [Online; accessed 8-August-2020].

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[8] A. Gautam, P. Mathur, R. Gosangi, D. Mahata, R. Sawhney, and R. R. Shah, "#MeTooMA: Multi-Aspect Annotations of Tweets related to the MeToo Movement," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, no. 1, pp. 209–216, May 2020. [Online]. Available: https://aaai.org/ojs/index.php/ICWSM/article/view/7292

[9] A. Ghosh Chowdhury, R. Sawhney, P. Mathur, D. Mahata, and R. Ratn Shah, "Speak up, fight back! detection of social media disclosures of sexual harassment," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 136–146. [Online]. Available: https://www.aclweb.org/anthology/N19-3018

[10] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017.

[12] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," *Lecture Notes in Computer Science*, p. 240–248, 2017.

[13] R. Mishra, P. Prakhar Sinha, R. Sawhney, D. Mahata, P. Mathur, and R. Ratn Shah, "SNAP-BATNET: Cascading author profiling and social network graphs for suicide ideation detection on social media," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 147–156. [Online]. Available: https://www.aclweb.org/anthology/N19-3019

[14] P. P. Sinha, R. Mishra, R. Sawhney, D. Mahata, R. R. Shah, and H. Liu, "#suicidal - a multipronged approach to identify and explore suicidal ideation in twitter," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 941–950. [Online]. Available: https://doi.org/10.1145/3357384.3358060

[15] M. Agarwal, M. Leekha, R. Sawhney, and R. R. Shah, "Crisis-dias: Towards multimodal damage analysis-deployment, challenges and assessment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 346–353.