MULTIMODAL DATA FUSION TO ANALYZE THE EMOTION IN MEMES

CHARLES YOO, JUSTIN QIU

ABSTRACT. The use of memes has become a popular way of expressing thoughts and emotions on the internet. Since they are multimodal in nature, consisting of an image and text, it is important to make use of both modals to create models that can best capture their conveyed emotions. However, one important consideration that arises when dealing with multimodal data is how to combine the information derived from the image and text. In this paper, we explore different data fusion techniques to combine embeddings outputted by BERT for the textual data and by ViT for the visual data. Specifically, we compare the F1 scores of MLP models that train on data fused using simple concatenation, bilinear pooling, and attention mechanisms. We find that for this problem, they all perform similarly, with simple concatenation seeming to perform slightly better than the others.

1. INTRODUCTION

With the rise of social media online communication, memes have become a popular way of sharing ideas and expressing emotion. While many memes are used in a humorous and light-hearted manner, some memes are used to convey offensive and derogatory messages. As such, it is important for online platforms to be able to identify these negative memes so that they can properly moderate the content on their sites. This can be difficult due to the fact that memes usually consist of both a text and an image. Simple unimodal models are not well suited for this problem as they can only make use of half of the available data. On the other hand, multimodal models, while better equipped for this problem, are more complex and present a number of new considerations regarding how to combine the data. The primary focus will be on multimodal data fusion consisting of textual and visual data. Specifically, we would like to explore the different points at which the data can be combined. One method for doing this is to embed the data using models such as BERT for text and ViT for images. These embeddings are then combined and fed into a single classification model. The techniques for combining the data that we will look at are: simple concatenation, attention mechanisms, and bilinear pooling. Another way for combining the data is to feed the data into models such as BERT and ViT to perform classification based on each modal separately. The logits are then averaged to get the final classification.

1.1. Contributions.

In this project, we make the following contributions:

(1) Embedded the textual and visual data using the BERT and ViT models respectively. These embeddings are then combined using 3 different methods: basic concatenation, attention mechanisms, and bilinear pooling. A custom MLP model is trained on these combined embeddings to classify the memes. This is considered early data fusion.
(2) Compared the effectiveness of different data fusion methods. Using the F1 scores of each method to determine which method works best for multilabel classification problems with multimodal data.

2. BACKGROUND AND RELATED WORK

Sentiment analysis is a field with numerous publications. A lot of work has been done on unimodal inputs such as text sentiment classification [1][2] and image sentiment classification [3][4]. Sentiment analysis has also been performed on multi-lingual text to provide better generalization to the online world where information can be written in different languages [5] Tamil-English text[6] Malayalam-English text[7] Hindi-English text.

On the other hand, work on multimodal is a developing space that has gained more attention in the past years. Some research has been done for understanding the value that image data can add for sentiment analysis of corresponding textual data [8]. More advanced models are being developed that can predict emotion tags on combined image and textual data [9] and that can provide comparable results when working with either unimodal or multimodal inputs [10]. The general approach to working with multimodal data presents new challenges and considerations [11]. One of the main considerations that arises is how to combine the data between different modalities. For our paper, we will be focusing on multimodal data consisting of visual and textual data. One interesting method for combining visual and textual data is to use a caption generation model [12] [13] to convert the visual data into textual data. Once this is done, the problem can now be considered a unimodal problem with textual data. A more conventional way of combining the data is called early fusion, where the input data is used to create embeddings, which are then combined and fed into a single classification model [14]. Images can be embedded using a CNN [15] or vision transformer [16] and the text can be embedded using a natural language processing transformer [17] [18] [19]. These embeddings can be combined in various ways, with the most common way being a simple concatenation. However, there are other methods that may better preserve the relationship between the textual and visual data. These methods include bilinear pooling [20], attention mechanisms [21], and projecting the data into a shared latent space [22]. Another method for combining the data is called late fusion, where the input data is fed into individual models to produce unimodal classifications. The output logits of these individual models can be averaged to produce the final prediction [14].

Numerous multimodal datasets have been developed to get a better understanding of how multimodal data can be used for sentiment analysis. The dataset that we will be working with is the Memotion dataset [23] [24], which focuses on the use of multimodal data to identify the emotions conveyed on social media, specifically through memes. There are several papers that use this dataset that present different models for classifying the data [25] [26] [27]. However, unlike our paper, they only combine the data using embedding concatenation and do not delve into the effects that different data fusion techniques can have on the performance of these models.

3. Approach

3.1. **Task and Data.** The paper that proposes the Memotion dataset proposes 3 tasks but the one that we will be focusing on is emotion classification. This involves classifying a given meme as humorous, sarcastic, offensive, or motivational. Each meme can fall into more than one category. The dataset itself consists of 10,000 Hindi-English memes that have a 7000-1500-1500 split between the training, validation, and testing sets. However, since this dataset was used for a competition, the testing set labels are not given and were therefore not used for our paper. This dataset was collected by downloading memes using a Selenium-based web crawler on Reddit and Google images. The data was cleaned for redundant memes and manually checked for overall quality. Each meme was then labeled for each emotional category from 1-4 by students fluent in both Hindi and English. For example, for the humorous category, students could label it as not funny, funny, very funny, or hilarious (the other categories follow a similar rating system). The final label for each meme was determined using a majority voting system. Below is an example of the memes included in this dataset:

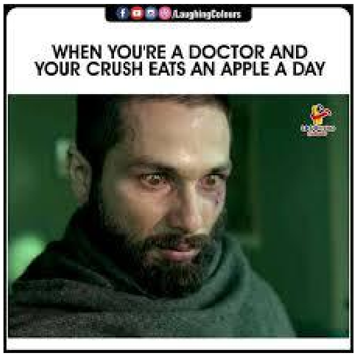

FIGURE 1. An example of an English meme included in this dataset. It is labeled as very funny, very sarcastic, slightly offensive, and not motivational

3.2. **Preprocessing the Data.** For this problem, we are not concerned with the extent to which each meme falls into each category. Therefore, each emotional category value was replaced with a 0 if the meme did not fall into that category and 1 if it did. For instance, for the humorous category, a meme was labeled as 1 if the meme was funny, very funny, or hilarious and labeled as 0 if the meme was not funny. Columns that were not relevant to classification such as image url were removed from the data. There were some class imbalances for the dataset that needed to be addressed before we could use the data. For instance, the motivational category has 830 memes labeled as motivational and 6170 memes labeled as not motivational. To deal with this we decided to undersample the majority class so that it would be balanced with the minority class.

3.3. **BERT for Textual Data.** For embedding the textual data, we decided to use the hinglish-bert model. This model is based on a multilingual version of BERT and further finetuned on Hinglish codemixed data to embed textual data in both Hindi and English, which is appropriate for our dataset. We used the pretrained weights of the 'nirantk/hinglish-bert' model without updating them during the training process.

3.4. **ViT for Visual Data.** For embedding the visual data, we decided to use the ViT model. This model has been shown to perform better at visual embedding than conventional CNNs. For the embedding of the visual data, we used the pretrained weights of the 'google/vit-base-patch16-224-in21k' model without updating them during the training process.

3.5. **Fusing Textual and Visual Data.** The 3 methods used for combining the textual and visual data were: concatenation, attention mechanism, and bilinear pooling. Concatenationis a simple operation in which one set of embeddings is added onto the end of the other set of embeddings to create one large dataset. Attention mechanisms are used to focus on relevant parts of the images and text embeddings to create a joint embedding that are aligned in a way that better captures the relationship between the 2 embeddings. Bilinear pooling involves taking the outer product of the embeddings, which leads to a multiplicative behavior between the elements of both sets of embeddings. This new set of embeddings can be reduced in dimensionality to provide a more compact representation.

3.6. **Custom MLP Model.** For the custom MLP model, we used a model with 4 linear layers. The first 3 hidden layers was followed by a batch normalization, ReLU activation, and dropout layer. The final linear layer is followed by a sigmoid activation layer. The input to this model varied based on the method for combining the linear layers (1536 for concatenation, 768 for attention, 256 for bilinear pooling). The second layer had 512 neurons, the third layer had 256 neurons, and the output layer had 4 neurons corresponding to each emotional category. This model was designed to take in the combined embeddings as input and to produce probabilities for each of the four emotional categories as output.

3.7. **Training the Models.** The custom MLP Model was trained for 50 epochs. We used binary cross entropy loss as the loss criterion and the Adam optimizer with a learning rate of 0.0001. The model is evaluated based on the F1 score, which is computed for overall categorization as well as for each individual category. Based on these scores, we can evaluate the effectiveness of each data combination technique.

4. Experimental Results

The 2 figures below show the validation F1 score for the 3 different data fusion techniques. Figure 2 shows the F1 score for the task as a whole while figure 3 shows the F1 score for the 4 individual emotions.

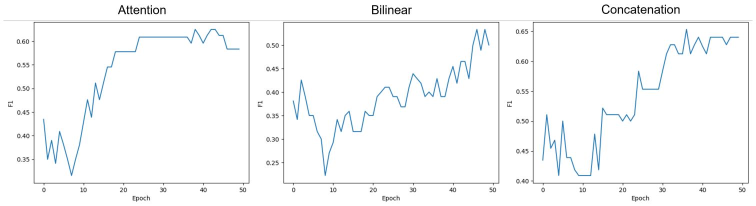

FIGURE 2. Overall F1 for Attention (Left), Bilinear Pooling (Middle), and Concatenation (Right)

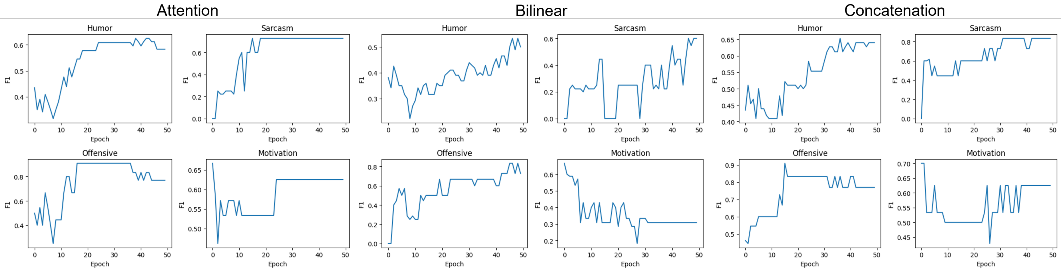

FIGURE 3. Individual F1 for Attention (Left), Bilinear Pooling (Middle), and Concatenation (Right)

For the 3 data fusion techniques, the overall F1 score for the validation set increases and reaches around 0.5-0.6. For the individual emotions, the F1 scores are highest for offensiveness at around 0.8, followed by sarcasm at around 0.7-0.8, which is then followed by humor at around 0.6. The F1 scores for motivation for attention and concatenation was around 0.6 while for bilinear it was only around 0.3. Based on these plots it appears that concatenation performs slightly better than attention, which performs better than bilinear pooling. This appears to be the case across all 4 emotional categories. To further evaluate the 3 data fusion techniques, images were drawn from both the training and validation sets and labeled using the trained models. The tables below show an example of the labels for one of the sampled images from each set.

Label	Humor	Sarcasm	Offensive	Motivational
True	1	1	1	1
Concat	1	1	1	1
Attention	1	1	1	1
Bilinear	1	1	1	1

TABLE 1. Labels for an Image from the Training Set

Label	Humor	Sarcasm	Offensive	Motivational
True	0	1	1	1
Concat	0	0	0	1
Attention	0	0	1	1
Bilinear	0	1	0	1

TABLE 2. Labels for an Image from the Validation Set

Based on table 1, each of the models successfully labeled this sampled training image. On the other hand, based on table 2, none of the models successfully labeled this sampled validation image. Bilinear pooling and attention missed only 1 category while concatenation missed 2 categories. We also drew images from the test set and labeled it using our trained models. Below shows an example of one of the sampled test images along with the associated labels. Since the test set did not come with true labels, we manually determined the true labels ourselves.

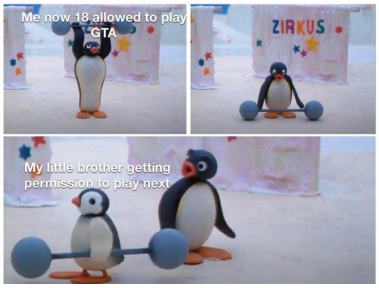
Me now 18 allowed to play GTA
ZIRKUS
My little brother getting permission to play next

FIGURE 4. Example Image from Test Set

Label	Humor	Sarcasm	Offensive	Motivational
True	1	0	0	0
Concat	1	1	0	0
Attention	1	1	0	0
Bilinear	1	1	0	1

TABLE 3. Labels for an Image from the Test Set (True label given by us)

Based on the table, each model correctly classified the image as humorous and not offensive but incorrectly classified the image as sarcastic while the bilinear model also incorrectly classified the image as motivational.

5. Discussion

5.1. **Model Analysis.** The data fusion technique that performed the best was simple concatenation. This is surprising because you would expect that attention mechanisms and bilinear pooling would be able to better capture the relationship between the visual and textual data. Since concatenation is a simple operation of arbitrarily combining the end of one embedding to the end of another, it does not appear to capture any sort of relationship. Despite differences in model performace, the 3 data fusion techniques perform relatively similar. This suggests that the choice of data fusion technique is more of a tool to improve a good model's performance. It is definitely not a replacement for designing a model that is best suited for the task.

For the sampled trained memes, the 3 models perform relatively well at classifying the emotions conveyed in them. For the sample shown above, every model correctly classifies each of the 4 emotions. On the other hand, for the sampled validation and testing memes, the 3 models do not perform as well at classifying them. For the samples shown above, each model incorrectly classifies at least one of the 4 emotions. This would seem to suggest that the model is overfitting the training data and therefore does not generalize well to memes outside of the training set. To address this, it could be worth reducing the complexity of our MLP model and introducing more regularization.

It is also worth mentioning that the 3 data fusion techniques do not perform particularly well at the task of classifying the emotions conveyed in a given meme. This is likely to be more of an issue with generalization of the model and limitations in the available dataset. This would make sense because the different data fusion methods are only meant to create a joint embedding that captures the relationship between the visual and textual embeddings. The task of learning how to classify the memes is primarily reliant on the model that is being trained. To reduce the limitations of the dataset, we could try to oversample underrepresented classes rather than downsample overrepresented classes. This would give us more data to train on, which could improve the overall performance of our models.

Another thing to note is that our experiments only evaluated the models using an F1 score. It may also be worth computing the accuracy of these models to get an even better understanding of their performance. Depending on the specifics of future tasks related to this paper, it may also be worth looking at recall and precision. For instance, for content moderation on online platforms it may be useful to compute recall to see which models have high rates of false negatives. In this context, a more conservative platform would prefer models that are more likely to falsely flag non-offensive content than to miss offensive content.

5.2. **Future Work.** In this paper, we only looked at combining the data using concatenation, attention mechanisms, and bilinear pooling. In the future, it would be interesting to compare these methods to other techniques such as mapping the visual and textual embeddings into a shared latent space before combining them. Another area that could be explored deeper is the effect of early and late data fusion. In this paper, we mainly focused on early data fusion. We also did work on late data fusion but were unable to produce meaningful results that could be presented in this report (the code has been submitted separately to Gradescope). It would also be interesting to see how these two methods could combined and compare this hybrid data fusion to early and late data fusion. Additionally, this work only looks at multimodal data that consists of visual and textual data. In the future, work could be done to see if similar findings occur when working with multimodal data that also includes audio data.

REFERENCES

[1] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. ACM Press, 2017.

[2] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In Jacob Andreas, Eunsol Choi, and Angeliki Lazaridou, editors, *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics.

[3] Udit Doshi, Vaibhav Barot, and Sachin Gavhane. Emotion detection and sentiment analysis of static images. In *2020 International Conference on Convergence to Digital World - Quo Vadis (ICCDW)*, pages 1–5, 2020.

[4] Namita Mittal, Divya Sharma, and Manju Lata Joshi. Image sentiment analysis using deep learning. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 684–687, 2018.

[5] Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P. McCrae. Corpus creation for sentiment analysis in code-mixed tamil-english text, 2020.

[6] Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. A sentiment analysis dataset for code-mixed Malayalam-English. In Dorothee Beermann, Laurent Besacier, Sakriani Sakti, and Claudia Soria, editors, *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France, May 2020. European Language Resources association.

[7] Rishabh Jha, Varshith Kaki, Varuna Kolla, Shubham Bhagat, Parth Patwa, Amitava Das, and Santanu Pal. Image2tweet: Datasets in Hindi and English for generating tweets from images. In Sivaji Bandyopadhyay, Sobha Lalitha Devi, and Pushpak Bhattacharyya, editors, *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 670–676, National Institute of Technology Silchar, Silchar, India, December 2021. NLP Association of India (NLPAI).

[8] Laura Graesser, Abhinav Gupta, Lakshay Sharma, and Evelina Bakhturina. Sentiment classification using images and label embeddings, 2017.

[9] Anthony Hu and Seth Flaxman. Multimodal sentiment analysis to explore the structure of emotions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*. ACM, July 2018.

[10] Chi Thang Duong, Remi Lebret, and Karl Aberer. Multimodal classification for analysing social media, 2017.

[11] Sara Abdali. Multi-modal misinformation detection: Approaches, challenges and opportunities, 2022.

[12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[13] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator, 2015.

[14] Ying Guo, Hong Ge, and Jinhong Li. A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism. *Frontiers in Computer Science*, 5, 2023.

[15] Alison Ribeiro and Nádia Silva. INF-HatEval at SemEval-2019 task 5: Convolutional neural networks for hate speech detection against women and immigrants on Twitter. In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 420–425, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[18] Nikhil Singh. niksss at hinglisheval: Language-agnostic bert-based contextual embeddings with catboost for quality evaluation of the low-resource synthetically generated code-mixed hinglish text, 2022.

[19] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. A bert-based transfer learning approach for hate speech detection in online social media, 2019.

[20] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding, 2016.

[21] Nguyen Manh Duc Tuan and Pham Quang Nhat Minh. Multimodal fusion with bert and attention mechanism for fake news detection, 2021.

[22] Yoon-Sik Cho, Greg Ver Steeg, Emilio Ferrara, and Aram Galstyan. Latent space model for multi-modal social data. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, April 2016.

[23] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. Semeval-2020 task 8: Memotion analysis - the visuo-lingual metaphor! *CoRR*, abs/2008.03781, 2020.

[24] Shreyash Mishra, S Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. Memotion 3: Dataset on sentiment and emotion analysis of codemixed hindi-english memes, 2023.

[25] Mayukh Sharma, Ilanthenral Kandasamy, and W.b. Vasantha. Memebusters at SemEval-2020 task 8: Feature fusion model for sentiment analysis on memes using transfer learning. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1163–1171, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[26] Zehao Liu, Emmanuel Osei-Brefo, Siyuan Chen, and Huizhi Liang. UoR at SemEval-2020 task 8: Gaussian mixture modelling (GMM) based sampling approach for multi-modal memotion analysis. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1201–1207, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[27] Nayan Varma Alluri and Neeli Dheeraj Krishna. Multi modal analysis of memes for sentiment extraction. In *2021 Sixth International Conference on Image Information Processing (ICIIP)*, volume 6, pages 213–217, 2021.