IITK at SemEval-2020 Task 8: Unimodal and Bimodal Sentiment Analysis of Internet Memes

Vishal Keswani* **Sakshi Singh*** **Suryansh Agarwal** **Ashutosh Modi**
Indian Institute of Technology Kanpur (IITK)
{vkeswani,sakshia,asurya}@iitk.ac.in
ashutoshm@cse.iitk.ac.in

Abstract

Social media is abundant in visual and textual information presented together or in isolation. Memes are the most popular form, belonging to the former class. In this paper, we present our approaches for the Memotion Analysis problem as posed in SemEval-2020 Task 8. The goal of this task is to classify memes based on their emotional content and sentiment. We leverage techniques from Natural Language Processing (NLP) and Computer Vision (CV) towards the sentiment classification of internet memes (Subtask A). We consider Bimodal (text and image) as well as Unimodal (text-only) techniques in our study ranging from the Naïve Bayes classifier to Transformer-based approaches. Our results show that a text-only approach, a simple Feed Forward Neural Network (FFNN) with Word2vec embeddings as input, performs superior to all the others. We stand first in the Sentiment analysis task with a relative improvement of 63% over the baseline macro-F1 score. Our work is relevant to any task concerned with the combination of different modalities.

1 Introduction

An internet meme conveys an idea or phenomenon that is replicated, transformed and spread through the internet. Memes are often grounded in personal experiences and are a means of showing appeal, resentment, fandom, along with socio-cultural expression. Nowadays, the popularity of the internet memes culture is on a high. This creates an opportunity to draw meaningful insights from memes to understand the opinions of communal sections of society. The inherent sentiment in memes has political, social and psychological relevance. The sarcasm and humor content of memes is an indicator of many cognitive aspects of users. Social media is full of hate speech against communities, organizations as well as governments. Analysis of meme content would help to combat such societal problems.

The abundance and easy availability of multimodal data have opened many research avenues in the fields of NLP and CV. Researchers have been working on analyzing the personality and behavioral traits of humans based on their social media activities, mainly posts they share on Facebook, Twitter, etc. (Golbeck et al., 2011). In this regard, we attempt to solve the sentiment classification problem under SemEval-2020 Task 8: "Memotion Analysis" (Sharma et al., 2020). Memotion analysis stands for analysis of emotional content of memes. We dissociate the visual and textual components of memes and combine information from both modalities to perform sentiment-based classification. The literature is rich in sentiment classification for tweets (Sailunaz and Alhajj, 2019) and other text-only tasks (Devlin et al., 2018). The multimodal approaches are relatively recent and yet under exploration (Morency and Baltrušaitis, 2017; Cai and Xia, 2015; Kiela et al., 2019).

We first describe the problem formally in section 2, followed by a brief literature review of the work already done in this domain. In section 3, we describe the methods proposed by us. Section 4 contains the description of the dataset provided by the organizers, along with the challenges accompanying the data. It further takes a deeper dive into the method yielding the best results. Section 5 summarizes the results with

* Authors equally contributed to this work.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1135–1140
Barcelona, Spain (Online), December 12, 2020.

a brief error analysis. Towards the end, section 6 concludes the paper along with future directions. The implementation for our system is made available via Github[1]

2 Background

Subtask A of Memotion Analysis is about Sentiment Analysis of memes. It is concerned with the classification of an Internet meme as a positive, negative, or neutral sentiment. A major portion of the research is devoted to separate handling of text and image modalities. Among the text-only methods, the Transformer (Vaswani et al., 2017) is an encoder-decoder architecture that uses only the attention mechanism instead of RNN. It has wide applications in different NLP tasks. BERT (Devlin et al., 2018) is the state-of-the-art transformer model for text only classification. It introduced the concept of bidirectionality to the attention mechanism to allow better use of contextual information. Reformer (Kitaev et al., 2020) is the most recent transformer, but more memory efficient and faster, hence works better for long sequences. For image-based tasks, ImageNet (Oquab et al., 2014) is a visual dataset with hand annotated images for visual object recognition task. ResNet-152 (He et al., 2016) is a deep learning model trained using the Imagenet dataset and used for object classification. The joint handling of image and text modalities has gained attention relatively recently. Qian et al. (2016) proposed a Text-Image Sentiment Analysis model that linearly combines image and text features. MMBT (Kiela et al., 2019) or the Multimodal Bitransformer fuses information from text and image encoders (BERT and ResNet) by mapping the image embeddings into the text space.

3 Methods

A wide variety of methods, ranging from a simple linear classifier to transformers, were employed. We broadly classify the set of techniques into bi-modal and uni-modal.

3.1 Bi-modal methods

These approaches consider both text and image modalities for classification. In general, both modalities are first treated separately, and high-level features are derived. These features are then combined using an additional classifier to make the final prediction. We use two main approaches:

3.1.1 Text-only FFNN and Image-only CNN:

As proposed in Qian et al. (2016), this method uses FFNN for text analysis (one-hot encoding for vectorization) and CNN for images analysis (HSV values for vectorization). For both the analysis, we get a probability distribution over the classes (positive, negative and neutral) as output. We concatenate the predicted probability distributions from the above two models and feed them as features to an additional classifier (SVM in this case) to arrive at a final prediction.

3.1.2 Multimodal Bitransformer (MMBT):

MMBT (Kiela et al., 2019) is a recent advancement in the fusion of unimodal encoders. They are individually pre-trained in a supervised fashion. It combines ResNet-152 and BERT by mapping the image embeddings into the text space, followed by a classification layer. It is a flexible architecture and works even if one modality is missing and captures text dominance. It can also handle arbitrary lengths of inputs and an arbitrary number of modalities. We fine-tuned MMBT for our dataset.

3.2 Uni-modal methods

We experimented with three text-only approaches for meme classification. The emphasis on separate text-only analysis is justified in subsection 4.1.

3.2.1 Naïve Bayes

Naïve Bayes is a popular classical machine learning classifiers (Rish and others, 2001). The main assumption behind the model is that given the class labels. All features are conditionally independent of

[1] https://github.com/vkeswani/IITK_Memotion_Analysis

each other, hence the name Naïve Bayes. It is highly scalable, that is, takes less training time. It also works well on small datasets, making it a good baseline for our analysis. We used the default implementation of Naïve Bayes classifier provided by the TextBlob library[2] (Loria et al., 2014).

3.2.2 Text-only FFNN

We use Word2vec embeddings (Mikolov et al., 2013) for capturing semantic and syntactic properties of words. It is a dense low-dimensional representation of a word. We use the pre-trained embeddings. Word2Vec represents each word as a vector (1x300 in our case). A caption is represented by an average of word embeddings of each of the words. Consequently, the input to FFNN is an $n \times 300$ matrix, where n is the number of captions.

3.2.3 BERT

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) is the state-of-the-art language model, that has been found to be useful for numerous NLP tasks. It is deeply bidirectional (takes contextual information from both sides of the token) and learns a representation of text via self-supervised learning.

BERT models pre-trained on large text corpora are available, and these can be fine-tuned for a specific NLP task. We fine-tuned BERT Base Uncased configuration[3], which has 12 layers (transformer blocks), 12 attention heads and 110 million parameters.

4 Experimental Setup

In this section, we quantitatively describe the dataset provided by the organizers and challenges accompanying it. We then mention the preprocessing steps briefly. Finally, we discuss the architecture and parameters of the FFNN with Word2vec approach (Section 3.2.2) in detail. This method made it to the final submission as it performed better than all the other approaches.

4.1 Data description

As a part of the task, we are provided with 7K human-annotated Internet memes labelled with various semantic dimensions (positive, negative or neutral for subtask A). The dataset also contains the extracted captions/texts from the memes using the Google OCR system and then manually corrected by crowdsourcing services. Hence, we have two modalities, image and text.

The dataset (Table 1) comes with a lot of inherent challenges. Firstly, the sentiment or emotion perceived depends on the social or professional background of the perceiver. Hence, classification is highly subjective. Also, the presence of sarcasm in memes makes sentiment classification a difficult task since positive appearing features are grounded in negative sentiment by the application of sarcasm. The large variance in caption lengths is another issue. The text sequence lengths vary from 1 (or even 0) to 100+.

Also, the text dominance in the memes provided is evident from glimpsing the data as the same image-template is repeated across different classes due to popularity of the image and bear very low correlation with the sentiment class. Hence, a good approach takes text as the dominant modality, and text-only approaches work well.

Dataset	Class	Points	Percentage
Training	Positive	4160	59.5%
	Neutral	2201	31.5%
	Negative	631	9.0%
	Total	6992	100%
Test	Total	1788	25.6%

Table 1: Class distribution for subtask A

[2]https://textblob.readthedocs.io/en/dev/

[3]https://github.com/google-research/bert

4.2 Data preprocessing

Text preprocessing steps included removal of punctuation, stop words and special characters, followed by lower-casing, lemmatization and tokenization. We used nltk library[4] (Loper and Bird, 2002) for the same. The tokens were then converted to vectors using Word2vec embeddings. Finally, the average of all the word vectors is taken to create caption embeddings (as mentioned in section 3.2.2).

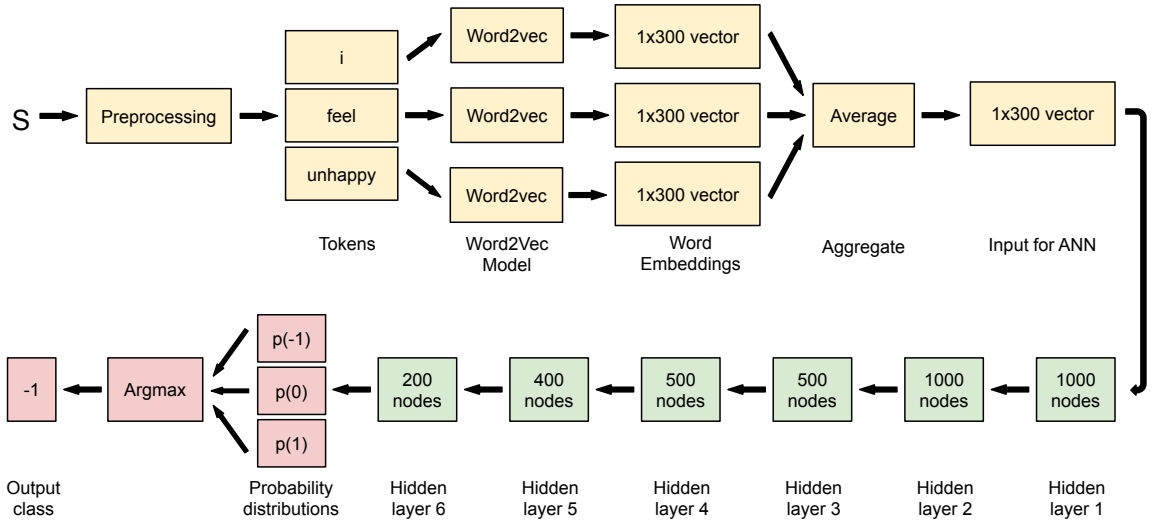

Figure 1: Pipeline for FFNN with Word2Vec. S="I am feeling unhappy."

4.3 Model and parameters

We elaborate more upon the precise architecture of the text-only FFNN approach with Word2vec embeddings. We employ a Feed-Forward Neural Network, with 6 hidden layers, softmax cross-entropy, Adam optimizer (Kingma and Ba, 2014) and ReLu activation function. The number of nodes in each hidden layer is depicted in Figure 1. Weights for each hidden layer are initialized with the standard normal distribution. The batch size is 50, and the number of epochs is 10.

FFNN is susceptible to randomness due to various reasons like random initialization of weight matrices, randomness in optimization, etc. Hence, it gives different results on multiple runs with the same parameters/hyperparameters. For a larger dataset (unlike ours), the model may produce more stable results. In the following section, we present the best score on the test set (Table 3) along with the mean and the variance of the scores on the validation set for 50 runs (Table 2).

Mean	Variance	Max	No. of runs
0.34	2E-4	0.36	50

Table 2: Macro-F1 for text-only FFNN (Word2vec) on validation set (80:20 split)

5 Results

The official evaluation metric for Memotion Analysis is the Macro-F1. We present the Macro-F1 scores of our five main approaches for subtask A in Table 3. With the best Macro-F1 of 0.3546581568, we improve the baseline (0.2176489217) by 63% and are ranked first in subtask A.

The dominance of the majority class over the others played a crucial role in sub-par performance for the other classes. For the 'negative' class, there are not enough data-points for the system to train. To some extent, this issue was resolved using simple upsampling. The transformer-based approaches overfit heavily

[4]https://pythonspot.com/category/nltk/

to the majority class. The repetition of meme templates is another issue for the bimodal approaches. Sarcasm also introduced ambiguity in the sentiment classification task. Neutral class dominated relatively due to the absence of polar words in some sarcastic texts.

Our results may be surprising as the state-of-the-art models, BERT and MMBT, are expected to do better. Some of the underlying reason for such behavior can be fewer data points, sarcasm, noise, and large variance in caption lengths. Moreover, BERT has been pre-trained on Wikipedia and Book Corpus, a dataset containing +10,000 books of different genres. These contained well-defined sentences, but our dataset is noisy, lacks punctuation marks and comprises of sarcasm. Simpler approaches did a better job since they involve no pre-training on any other (large) corpus.

Modality	Model	Macro-F1
Text-Image	FFNN + CNN	0.29
	MMBT	0.30
Text-Only	Naive Bayes	0.32
	FFNN (Word2Vec)	**0.35**
	BERT	0.33
Baseline		0.22

Table 3: Results for subtask A

6 Conclusion

We attempt to perform a complex task of classifying memes constrained by the data size and quality. While the results were sound for the populous classes, it could only be par after re-sampling for the skewed classes. The best results were obtained for FFNN with Word2vec embeddings (Table 3). Vanilla ANN-based approaches are highly competitive as compared to transformers and even outperformed them. A better research problem would be to define some rules to obtain meme data and then perform the above task so domain knowledge could be used to improve performance.

In future, a study could be designed to observe the diffusion of memes among different communities by analysing which meme is most liked or hated by a particular community (e.g. a Facebook group). Finding the political inclination of memes is also a possible path. Memes have become an increasingly popular mode of expressing opinions by party supporters, party critics and the affected people. They may be considered as a means of propaganda. This makes the problem of detecting the inclination of memes in a political context an important research exercise.

References

Guoyong Cai and Binbin Xia. 2015. Convolutional neural networks for multimedia sentiment analysis. In *Natural Language Processing and Chinese Computing*, pages 159–167. Springer.

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*.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 149–156.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. 2014. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*, 3.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Louis-Philippe Morency and Tadas Baltrušaitis. 2017. Multimodal machine learning: integrating language, vision and speech. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 3–5.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724.

Chen Qian, Edoardo Ragusa, Iti Chaturvedi, Erik Cambria, and Rodolfo Zunino. 2016. Text-image sentiment analysis.

Irina Rish et al. 2001. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

Kashfia Sailunaz and Reda Alhajj. 2019. Emotion and sentiment analysis from twitter text. *Journal of Computational Science*, 36:101003.

Chhavi Sharma, Deepesh Bhageria, William Paka, Scott, Srinivas P Y K L, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, Sep. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.