Instructions for *ACL Proceedings

Anonymous ACL submission

Abstract

In task 4 subtask 1, we need to conduct multi label classification on the 20 meme persuasion strategies. For fine-tuning, we applied the betonews-tweetcontext pre-trained model and achieved favorable outcomes.

1 Introduction

Memes have steadily grown into one of the content forms that impact human behavior as social media platforms have become more prevalent. This sort of content typically spreads quickly by manipulating audience psychology and blurring logical relationships. Memes are generally made up of stacked images and text. The essence of its expression in order to generate an emotional effect is actually the skillful role of three persuasive strategies in rhetorical portions. These three techniques are as follow:

1) Ethos: Using authoritative persons' statements to convince readers to believe in their writings and enhance their credibility.

2) Pathos: Share personal experience with readers, provoke emotional resonance, and increase depth of emotions.

3) Logos: Using logical reasoning to strengthen the article's robustness and logical coherence.

If we further split these three categories of persuasion strategies into twenty-two, scientists are able to obtain textual and visual features from MEMES for analysis. For instance, it is feasible to efficiently decrease or prevent the spread of hate speech, racial discrimination, and deceptive information by analysing MEMES, then simultaneously preserving the peace and stability of social media. MEMES can assist merchants in quickly capturing market trends, allowing them to carry out advertising and marketing operations more effectively and raise brand influence. MEMES helps media workers in understanding the concerns of their audiences. MEMES in politics have the potential to help voters demonstrate their policy views. The goal of the task is to classify corpora of text in MEMES and assign them to relevant persuasive strategies. This is a multi label classification task.

In our work, our contrbutions can be highlighted as follows: 1) We explored new possibilities by screening models for news texts and multilingual corpus models. 2) The betonews tweetcontext model was fine-tuned, and results of xx were obtained in the experiment. 3) Our model ranked xx on the leaderboard and got xx in the surprise test datasets .

2 Related Work

Since its introduction in 2018, the Bert model has been widely used by academics in a variety of domains. It has been demonstrated through significant research that BERT, a massive corpus pre-trained model, is capable of being fine tuned to any given task. Multimodal applications based on BERT have achieved leading-edge performance in a variety of visual and text extraction challenges (Afridi et al., 2021).

There are multiple variations of the BERT model. Roberta (Liu et al., 2019) is a variant of Bert . Research has shown that when using BiLSTM, XLNet base cased, XLM-Roberta-base, and ALBERT for hate speech detection, XLM-Roberta-base achieves excellent performance, outperforming other models in all evaluation metrics (Singh et al., 2023). Studies have indicated that when the NLI dataset is fine-tuned to acquire knowledge, the Roberta model outperforms other models in classification tasks for Sentiment, Emotion, and Offensive tweets (Suryawanshi et al., 2023). In the experimentally designed text modality classifier, RoBERTa outperforms BERT and XLNet to detect propaganda

techniques in Memes, proving that it is the best text encoder (Gundapu and Mamidi, 2022).

3 System Overview

3.1 Datasets

Three datasets in in total the training, validation, and test sets were used. The dataset's minimum sentence length is one, and all of the data is in JSON format. The training set has 7000 samples, grouped into 20 classes, with an average sentence length of 19.94 and a maximum sentence length of 253. There are 500 samples in the validation set, with an average sentence length of 18.85 and a maximum sentence length of 333. Dev dataset, a set of 1000 samples with an average length of 18.73 and a maximum length of 145, will serve as the test set.

Below is the sample dataset:

Table 1: **Data Sample**

ID	text	labels
67641	WHEN YOU'RE THE FBI, THEY LET YOU DO IT.	Thought-terminating cliché
66402	PUTIN'S SECRET CAMOUFLAGE ARMY	none
71251	Heaven has a Wall and strict immigration policies. Hell has open borders. President Donald J. Trump	Appeal to authority Exaggeration/Minimisation
65282	ME VOTING ANTI-TRUMP IN 2016 ME VOTING ANTI-TRUMP IN 2020	Repetition

The training set's number of distinct classes is represented statistically as follows, with None indicating unclassified. From statistics, it can be seen that Smear has the highest proportion, while Obfuscation, Intrinsic Vague, and Fusion have the lowest proportion.

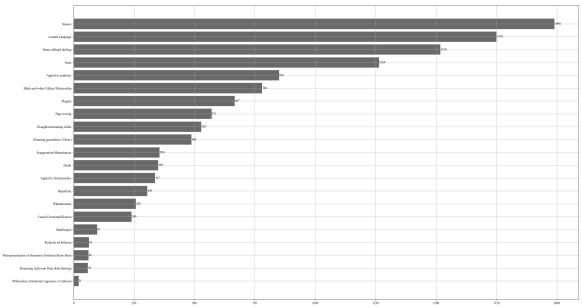

Figure 1: Number of Samples

3.2 Pre-trained Model Selection

As the data originates from the annotation of articles in the news, the research team tends to choose from models related to news, tweets, and comments. The research team tested a number of models and deciding that Jochen Hartmann's sentiment-roberta-large-english-3-classes model (As shown in Table 1 ID 7) while it received the best ratings. A comparison of outcomes from multiple models will be presented in the results section.

The sentiment-roberta-large-english-3-classes model (Hartmann et al.) is trained based on tweets on social media platforms such as Twitter and Instagram, and includes text that is expected to include captions from the sender in the tweet image and comments from other observers. RoBERTa is used to construct the model. Achieving a hold out accuracy of 86.1 % , this model is used to evaluate user comments on posts and identify if the user is willing to buy a certain product. It demonstrates that the model has high robustness and a strong capacity to extract complicated text features.

Table 2: **Candidate Pre-trained Model**

ID	Model
1	bert-base-uncased
2	bert-base-multilingual-cased
3	albert-base
4	roberta-base
5	xlm-roberta-base
6	roberta-large
7	roberta-large(social media posts fine-tuned)

3.3 Model Construction

To get started, we make use of the officially provided Train.json and Validation.json as labeled training and validation datasets, respectively. In addition to the officially available Dev dataset as the subsequent testing dataset.

Secondly, we'll perform data preprocessing. The training and validation sets are fed into the Tokenizer, and the pre-trained model betonews-tweetcontext is used for word segmentation and vectorization processing.

Following that, regarding model structure:

1) Input processing: Feed the pre-trained model with the processed vector.

2) Dropout processing: Enter the dropout layer after model processing and set the inactivation probability to 0.1.

3) Linear fully connected layer: 1024 features are carried into the linear fully connected

layer.

4) Loss function: For multi label classification jobs, Binary Cross Entropy With Logits Loss serves as the loss function throughout the backpropagation gradient calculation procedure. BCEWithLogitsLoss comes with a sigmoid function that can convert predicted result values into probabilities, and can automatically handle numerical instability while preventing the sigmoid function from overflowing upwards or downwards (Yue et al., 2023).

Finally, the output layer is made up of 21 neurons, 20 of which are classified and one of which is none. The architecture of model construction is shown in Figure 1.

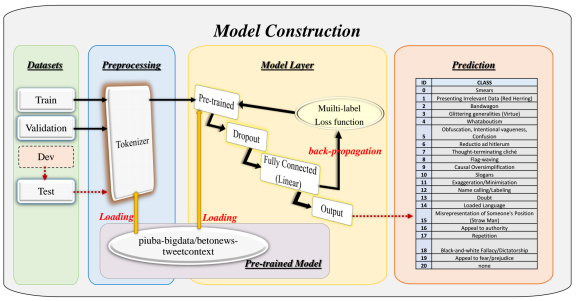

Figure 2: The architecture of model construction

4 Experiment Setup

4.1 Threshold Selection

The experimental results in training tasks will depend on the threshold selection. We select the most optimal F1 value for determining the threshold, assuming that recall and precision are of identical significance. With a 0.01 interval, the experiment increased the threshold from 0 to 1. The red dots on the F1 value curve in the illustration represent the experimental results, which show that the most suitable threshold value for F1 value is approximately 0.08. Here, recall=0.69 and precision=0.60 are achievable. We have simply included a portion of the graph here because the threshold was set at 0.08 and the F1 value decreases after the threshold is larger than 0.4.

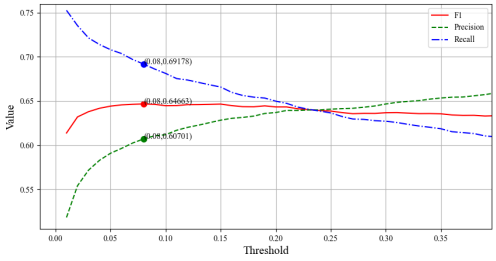

Figure 3: Changes in F1, Precision, and Recall at Different Threshold

4.2 Epoch Selection

To choose the most optimal F1 value, the F1 formula ($F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$) states that when the Recall tends to stabilize, the higher the Precision will get the greater the F1 value.

The epoch was raised in the experiment from 1 to 20 at intervals of 1. The graph of the experimental results illustrates that the Precision is low and unstable and the Recall value is high but swings continuously when the epoch is under seven. As a result of the Precision and F1 values' continued continuous increase, the experimental model's instability will grow. The Recall steadily stabilizes as the epoch gets closer to 20, while the F1 value also tends to stabilize. As a result, the epoch in this experiment is 20.

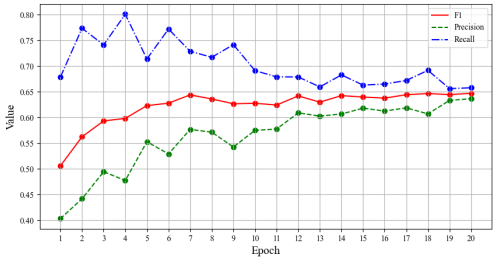

Figure 4: Changes in F1, Precision, and Recall at Different Epochs

In addition to the above parameters, the settings of other training arguments are shown in the table below.

Table 3: **Training Arguments**

Params	Value
num_train_epochs	20
per_device_train_batch_size	4
per_device_eval_batch_size	8
warmup_steps	500
weight_decay	0.01
logging_steps	100
save_strategy	epoch
evaluation_strategy	epoch
learning_rate	$1.5e^{-5}$
threshold	0.08

5 Results

The model's results on the Test dataset (Dev) showed that the F1 value was 0.64, the Precision value was 0.63, and the Recall value was 0.65.

Table 4: **Training Results**

Model	F1	Precision	Recall
bert-base-uncased	0.59335	0.60017	0.58668
bert-base-multilingual-cased	0.58840	0.58235	0.59459
albert-base	0.59484	0.58081	0.60957
roberta-base	0.62268	0.60781	0.63829
xlm-roberta-base	0.58612	0.57927	0.59313
roberta-large	0.63679	0.61831	0.65640
roberta-large (social media posts fine-tuned)	0.64708	0.63666	0.65786

6 Conclusion

???

References

Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. 2021. A multimodal memes classification: A survey and open research issues. In *Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications*, pages 1451–1466. Springer.

Sunil Gundapu and Radhika Mamidi. 2022. Detection of propaganda techniques in visuo-lingual metaphor in memes. *arXiv preprint arXiv:2205.02937*.

Jochen Hartmann, Mark Heitmann, Christina Schamp, and Oded Netzer. The power of brand selfies. *Journal of Marketing Research*.

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*.

Karanpreet Singh, Vajratiya Vajrobol, and Nitisha Aggarwal. 2023. Iic_team@ multimodal hate speech event detection 2023: Detection of hate speech and targets using xlm-roberta-base. In *Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text*, pages 136–143.

Shardul Suryawanshi, Mihael Arcan, Suzanne Little, and Paul Buitelaar. 2023. Multimodal offensive meme classification with natural language inference. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 134–145.

Xiaohan Yue, Danfeng Liu, Liguo Wang, Jón Atli Benediktsson, Linghong Meng, and Lei Deng. 2023. Iesrgan: Enhanced u-net structured generative adversarial network for remote sensing image super-resolution reconstruction. *Remote Sensing*, 15(14):3490.