

Table of Contents

**No table of contents entries found.**

# **Abstract**

This report presents my approach to ‘*SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images*’ (https://aclanthology.org/2021.semeval-1.7.pdf) for *Assignment 2* of *Deep Learning* ***COSC2972*** *(undergraduate level)* at **RMIT University**. The challenge involves ***Multi-Modal Classification*** aiming to identify the presence of 22 persuasion techniques in memes, along with their corresponding sentences: integrating both linguistic and visual information. My methodology consists of a comprehensive **Machine Learning Approach**: conducting an **EDA**, creating a **Baseline Model**, making iterative improvements through advanced **Hyperparameter Tuning**, and **Fine-Tuning** techniques to build a final optimized version.

For visual feature extraction, I utilized **MobileNetV2** and **EfficientNetB0/B1**, chosen for their optimal *performance-to-GFlops* ratio within computational constraints. These visual features are combined with textual representations from various sizes of **BERT**-like models. I then explored diverse fusion techniques, including attention mechanisms, to effectively integrate these multi-modal inputs.

To address the challenges of the small dataset size and class imbalance, I implemented focal loss, data augmentation, and class weighting. My final model incorporated extensive hyperparameter tuning and fine-tuning strategies.

This approach yielded a F1 Micro score of X, comparable to top-performing models on this data split. This result is particularly impressive given the constraints on time and computational resources, demonstrating the effectiveness of my approach in tackling this complex multi-modal classification task while striving for robust performance on unseen data.

# **Problem Definition and Background**

This project addresses SemEval-2021 Task 6: detecting persuasion techniques in memes. The challenge involves a multimodal, multi-label classification of memes into 22 persuasion techniques (plus a 23rd "none" category), considering both text and visual elements.

## **Ethical Context:**

This task has real-world applications in the complex and often controversial area of addressing online disinformation and propaganda, raising important ethical considerations about information control and freedom of expression.

The dataset consists of 950 samples which have been pre-split as the following:

* Training set: 687 samples (72.32%)
* Development set: 63 samples (6.63%)
* Test set: 200 samples (21.05%)

Each sample includes an image, pre-extracted text, and corresponding labels. The limited dataset size presents significant challenges for deep learning approaches, necessitating careful consideration of model architecture and training strategies to ensure generalization. Additionally, the multimodal nature of the data (combining text and images) and the imbalanced class distribution pose further complexities in developing an effective solution.

# **Evaluation Framework**

Given the dataset's characteristics and the complexity of the task, I carefully selected an evaluation framework to assess model performance accurately. The primary goal was to predict multiple labels present in each meme, considering both image and text content. The framework utilizes the following key metrics:

1. F1 Micro: This serves as the primary metric, providing a balanced measure of precision and recall across all instances. It is particularly suitable for multi-label classification tasks with imbalanced classes, as it gives equal weight to each sample.
2. F1 Macro: Used as a secondary metric, it offers insights into the model's performance across all classes, regardless of their frequency. This is crucial for understanding how well the model performs on less common persuasion techniques.

These metrics were chosen to address the imbalanced nature of the label distribution in the dataset and to maintain consistency with existing research in this field and challenge, ensuring comparability with other models [X].

# **Approach & Methodology**

Data Preprocessing and Exploratory Data Analysis (EDA):

The initial step involved data ingestion and reformatting. The original JSON format was converted to CSV, with labels concatenated into a single string. Image paths were appended to ensure accessibility. Extensive EDA revealed significant variations in image dimensions (200-1800 pixels), necessitating resizing. Images were predominantly square-shaped, justifying a squish methodology for resizing. Analysis of image characteristics showed normal distributions of sharpness and intensity, with positively skewed entropy. Notable issues included incorrect file extensions (JPEG renamed as PNG) and repetitive meme templates with varying labels. Text analysis examined word frequency and length distributions across labels.

## **Dataset Challenges:**

The dataset presented challenges including small size, class imbalance, and a limited validation set (63 images) lacking representation of all labels. To address these issues, class weighting was implemented instead of oversampling to mitigate overfitting risks.

## **Model Architecture:**

The core architecture was inspired by the Alpha team's approach, adapted for TensorFlow implementation. It combines BERT for text processing with MobileNetV2 and EfficientNetB0/B1 for image feature extraction. Images were resized to 224x224 using squish methodology. The multimodal fusion was achieved through concatenation of text and image features, followed by dense layers with ReLU activation and dropout for regularization.

## **Training Strategy:**

Initial training utilized a small\_bert/bert\_en\_uncased\_L-4\_H-512\_A-8 model with frozen weights for both BERT and MobileNet to establish a baseline. This approach yielded a micro F1 score of 0.5327 on the training set and 0.4983 on the validation set after 20 epochs with a batch size of 32.

## **Model Refinement:**

Subsequent iterations incorporated techniques such as GradientAccumulation, cross-entropy loss, early stopping, and learning rate reduction to combat overfitting. Data augmentation techniques were applied, including random brightness, contrast, and saturation adjustments for images, and synonym substitution and back-translation for text. Despite a slight performance decrease, these techniques were retained to enhance generalization.

## **Hyperparameter Tuning:**

Grid search was employed for hyperparameter optimization, focusing on dropout rates, optimizer selection, and batch sizes. The optimal configuration (dropout rate: 0.2, optimizer: Adam, batch size: 16) resulted in improved F1 scores (training: 0.7307, validation: 0.5430).

## **Fine-tuning:**

The final stage involved unfreezing BERT and MobileNetV2 weights for fine-tuning with a low learning rate, leading to the final model performance.

This methodology integrates multimodal deep learning techniques, transfer learning, and various optimization strategies to address the challenges presented by the SemEval-2021 Task 6 dataset.

## **Enhancement Task:**

In addition to predicting the persuasion techniques, I tackled the advanced task of identifying the specific spans of text covered by each technique. This multi-label sequence tagging task is similar to Named Entity Recognition but more complex due to the potential overlap of techniques within the text.

For this enhancement, I developed a model that not only predicts the labels but also identifies the start and end positions of the text fragments associated with each label. To evaluate this aspect, I introduced an additional metric:

1. Span Accuracy: This metric measures the difference between the predicted start/end positions and the true positions of text fragments. The goal is to minimize this difference, with a perfect score being zero.

By employing F1 Micro, F1 Macro, and Span Accuracy, I aim to provide a comprehensive evaluation of the model's performance, considering overall accuracy, ability to handle less frequent classes, and precision in identifying relevant text spans. This approach allows for a thorough assessment of the model's effectiveness in both the main classification task and the more challenging text span identification task.

I implemented a transformer-based architecture to fuse textual and visual features, inspired by the success of such approaches in recent multimodal tasks.

Rationale: Transformer architectures have shown superior performance in capturing long-range dependencies and complex interactions between modalities.

3.4 Loss Function

To address the class imbalance issue, I adopted focal loss instead of standard cross-entropy loss.

Rationale: Focal loss helps in focusing the model on hard-to-classify examples, which is particularly beneficial in our imbalanced multi-label scenario.

**Experiments & Tuning**

**I did a lot of experimental models which even though I got more performance on some I did not implement into the machine learning flow, as they did require quite a bit more computing power.**

**Some I also did later on as experiments, or for the Extension Task and did not have time to integrate them fully into the Main Task as intended.**

**I also did do ensembling as it is very prevalent in some of the literatures like X, Y, Z but on large scale models it was impractical for me to run without a larger compute budget.**

4.1 Model Variants

I experimented with several model variants:

I tried to replicate the Alpha teams of:

1. DeBERTa-large + ResNet50

I tried many BERT based or like models like

ALBERT

ELECTRA

This is not an exhaustive list see Appendix X for all tested models.

1. DeBERTa-large + BUTD
2. ERNIE-ViL (a pre-trained multimodal transformer)

4.2 Hyperparameter Tuning

Key hyperparameters tuned include:

* Learning rate: Explored range 1e-5 to 5e-5
* Batch size: Tested 4, 8, and 16
* Focal loss parameters: α (0.75 to 0.95) and γ (1.0 to 3.0)

4.3 Data Augmentation

Given the small dataset, I implemented several data augmentation techniques:

* For text: synonym replacement, random insertion/deletion
* For images: random cropping, flipping, and color jittering

Rationale: Data augmentation helps in increasing the effective size of our training set and improving model generalization.

**Ultimate Judgment, Analysis & Limitations**

5.1 Model Performance

After extensive experimentation, the ERNIE-ViL model emerged as the best performer, achieving an F1 score of 57.14 on the test set. This outperformed both DeBERTa+ResNet50 (55.96) and DeBERTa+BUTD (56.21) configurations.

Key findings:

1. Multimodal pre-training advantage: ERNIE-ViL's superior performance can be attributed to its pre-training on large-scale image-caption data, allowing it to learn more general and robust multimodal representations.
2. Visual feature impact: The BUTD object detection features slightly outperformed ResNet50 grid features, suggesting the importance of salient region information in meme analysis.
3. Focal loss effectiveness: Switching from cross-entropy to focal loss improved F1 scores by approximately 4 points, demonstrating its efficacy in handling class imbalance.

5.2 Error Analysis

Examining misclassifications revealed several patterns:

* Confusion between closely related techniques (e.g., "Loaded Language" vs "Name Calling/Labeling")
* Difficulty in detecting subtle visual cues
* Challenges with memes requiring external context or cultural knowledge – use of slogans, or wording in complex contexts or under used ways.

5.3 Limitations and Future Work

1. Dataset size: The small dataset remains a significant limitation, potentially hindering the model's ability to generalize to diverse real-world memes.
2. Visual-textual alignment: Current approaches may not fully capture the intricate relationships between text and image elements in memes.
3. Context understanding: The model struggles with memes that require broader contextual or cultural knowledge.
4. Computational resources: High-performing models like ERNIE-ViL are computationally intensive, potentially limiting real-time applications.

Future work directions:

* Explore more sophisticated visual-textual alignment techniques
* Investigate ways to incorporate external knowledge bases
* Develop more efficient model architectures for real-time processing
* Collect and annotate a larger, more diverse dataset of memes

**Conclusion**

This project demonstrated the effectiveness of transfer learning and multimodal fusion techniques in tackling the challenging task of persuasion technique detection in memes. The ERNIE-ViL model, combined with focal loss and careful data augmentation, proved most effective in navigating the constraints of a small dataset.

While the achieved performance is promising, there's significant room for improvement, particularly in handling subtle persuasion techniques and memes requiring deep contextual understanding. Future work should focus on more sophisticated multimodal integration, larger datasets, and incorporation of external knowledge.

The developed system shows potential for real-world applications in content moderation and digital literacy education, though further refinement is needed for practical deployment.

**BONUS: Enhancement (only for HD)**

EMAIL: Eda- note imbalance, not e the data was oddly formatted which made it quite difficult to write a consistent data loader for. Thus, I opted to create a .csv for each dataset with the full file path for each image as well as a comma seperated field for the labels to look more like this: label1, label2

Note for eda that there does not exist every label in the validation test set.

For eda we want to explore how many labels are in each- what the average is etc. basically do all the same as from a1

Need to mention we don’t have to worry about data splitting or leaks this time around as the model was actually presplit into training, validation or dev and testing. But here’s some comparisons anyways

Setting our evaluation metric -F1 because we have an kmblanced dataset. I also decided that while the ultimate prediction would be wrong if it didn’t contain all the labels that if a specific label was predicted and was true this would reflect accordingly on the confusion matrix per label.

For base model I tried near 30 variants of Bert models starting with Bert small h4 going all the way up to Bert large 24h and trying other variants like Electra, Albert, deberta, etc - I spent almost a day just trying to find the largest language based transformer I could run locally on my computer and it these models would or could improve the performance on the task but counter-intuitively or in retrospect making considerable sense these models were not always the best as they do require quite a bit more training for them to understand the specific task as well as in my tests I could not dedicate more than 10 epochs per model due to time and compute constraints.

For hyperparams tuning I used a few call backs, I also tried some augmentation,

In particular I used focal loss though as this was recommended by the alpha team. But I also looked at optimiser etc

I n terms of an advanced technique I also attempted was around deberta.

Finally for fine tuning I decided to do this for x epochs and with a learning rate of x to ensure that my models weights are sufficiently altered for lir task

EfLuation so now evaluating our model on our test data we see;

-Go through case examples and explain why e.g. why loaded language classified as smth else or smth

Comparison of our performance to others.

-discuss how little the multi-modal part actually matters in our scenario and I and others found that the combination with images only increased our F1 score by a few % while

Show a table of my models performance verses benchmarks as well as what I would do next time given more time, less constraints around compute & accessibility of weights . One particular thing I would consider next time is using a bur stable cluster for gpu compute to parallelise training of language based models like Bert as this would increase the speed at which a model could be trained, and increase the amount of hyperparams I can turn at once.

Another thing is having access to more data for training would be very useful as we had under 700 samples for training and under 300 for validation and test. This presented quite a considerable challenge and I think one very smart way that we can easily increase the size of the dataset would be to utilise existing LLM with vision capability to 1. Ocr the text from the image and 2. Define the labels associated with said meme. These could tehn be fed to a human who could confirm and validate the labels before increasing our datasets.

For the extension task this to me stands out as a NER task along with sentiment analysis task of surrounding.

Before hypertunign I explored a lot of augmentation techniques for images and text and found the majority of them to be not useful : did you try augmentation? i tried synonym substitution and it made my model overfit to train. i don't think back-translation will do all that well either.

As we already identified images werent the most reliable anyway so augmentation on images had inconsistent and little help.

Explain the metrics F1 micro and macro and why they were chosen - link to the literature and from what we already identied as being an imbalanced dataset so taking into account recall and precision is very important to reduce the effect of a majority class effecting our primary metric too much when we should br aiming for good results across all labels.

Thus I utilised class weights as from experimentation with oversamplibg this often leads to overfitting of the training dataset which is something we’re trying to avoid.

Talk about the literature and what we could have done differently and other teams did - particularly around ensembles and larger models of Bert which require more compute but also more time to train. More images would have been the best thing we could have been offered as the less than 1000 was not merciful in variety or anything. I would have loved to see this as then this model could be taken outside the realm of the US election - liberal vs republicans and coronavirus memes as the real internet has many more

1. References
2. Dimitrov, D., et al. (2021). SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images. Proceedings of the 15th International Workshop on Semantic Evaluation, 70-98.
3. Yu, F., et al. (2020). ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph.
4. Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
5. He, K., et al. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.
6. Lin, T.Y., et al. (2017). Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision, 2980-2988.