# Detection of Persuasion Techniques using Transfer Learning on BERT and ResNet

Riccardo Vella, Giuseppe Spina

**With this paper, we present our solution on all three subtasks of the SemEval 2021 Task 6. This challenging propaganda classification problem includes the fusion of text classification, sequence tagging and image classification. We discuss different solutions in the problem of feature extraction and in the handling of unbalanced data, working of the SOTA architectures. We research and evaluate many possible solutions in the form of Transfer Learning of pretrained models, as well as threshold moving and other minor adjustments.**

## I. Introduction

PERSUASION techniques have always been used to convey ideas in every means of communications. Taking advantage of how easily memes are shared everywhere, by every generation group, there has been an increase in the usage of techniques. Exploiting the newer channels, provided by the internet and this vast audience, it is possible to spread a numerous amount of propagandist material in a very short time. Recognizing what kind of propaganda-based techniques are being applied on a shared image (or even just text) can be used by social networks and other platforms to moderate their content and defend their users.

With this objective in mind, we address the SemEval-2021 task 6 [1]. This challange is composed by three different subtasks, each one having its own dataset. Every task represents different objectives:

- **Subtask 1**: categorize text inputs into one or more of the 20 possible persuasion technique labels. This is a task of multi-label multi-class classification.
- **Subtask 2**: categorize text inputs into one or more of the 20 possible persuasion technique labels (same labels as subtask 1), as well as the corresponding span, which identifies the position of the classification in the text. This is a multi-label multi-class sequence tagging task (and can be considered as an extension of subtask 1).
- **Subtask 3**: categorize text along with the corrensponding meme image into one or more of the 22 possible persuasion technique labels. This is again a task of multi-class multi-label classification (which can also be considered an extension of subtask 1).

In this research, we apply different solutions in order to solve the discussed tasks and we evaluate each obtained results. All our solutions are based on the usage of the pretrained BERT transformer, paired with different methods to increase the effectiveness of the final model. We present our research on Transfer Learning, measures to address the problem of unbalanced data, as well as different model architectures. We then combine all of these features into our final classifier, which pairs image classification obtained from ResNet-50 and text classification from BERT.

## II. Related Work

### A. Sentiment Analysis

Sentiment analysis plays a big role in the research field, understanding the emotions [2] that are used to carry the message in propagandist messages can help us identify techniques. Text based sentiment analysis has seen a great development in the past years, the adoption of transformers like BERT [3], greatly increased the understanding of emotions in text.

### B. Transformers

Transformer networks are heavily used to improve text understanding by neural networks, they provide a solid base for our experiment. The improvements of transformers over previous models is heavily investigated [4]. In our particular case we choose BERT Transformer as a base to extract features from the text. This method has been proven to be very successful in various tasks [5].

### C. ResNet

Visual feature extraction is a key challenge in subtask 3. We approached various method but in the end we landed on ResNet. Taking advantage of its architecture based on skipping connection, instead of traditional convolutional networks, can help us extract more reliable and important features from our images [6].

### D. Transfer Learning

Transfer learning can be very effective, in a case like ours, to improve both our features extractors. Transfer Learning is the process of adapting a model to a different, task, to then use it on a different, but generally in-domain, task and data. By taking advantage of this concept it is possible to extract better and more task related features. We can see improvements in various tasks both for ResNet [7] and for BERT [8].

### E. Previous works on the task

We also adapted some of the work that has been done on this same task by other teams. These researches have been taken into consideration, influencing ours and acting as a reference to compare our results. All the main other teams provided a Transformers-based solution, tweaking and modifying on different aspects to obtain a better final solution. Other teams solutions range from the usage of RoBERTa, to Multi-modal transformers and Multi-modal fusion [9][10][11].

## III. Approach

With all the data that we collected about the previous works related to this challenge we tried to add our ideas to get an improvement on the previous attempts.

### A. Text Features Extraction

A fine-tuned BERT transformer is responsible of extracting text features in all of the three subtasks. Fine-tuning BERT is an unstable process, prone to vanishing-gradient and catastrophic forgetting [12]. To obtain better results on the process we followed a procedure composed of two steps [13]:

- A continuation of BERT pretraining on the task of Masked Language Modeling on a political tweets dataset from the 2020 elections;
- Fine-tuning on the target task.

Furthermore, with the purpose of addressing the problem of catastrophic forgetting and knowledge generalization, we try pretraining only a subset of layers, excluding some of the first ones by ”freezing” them (our model freezes 2 out of 12 layers). In fact, similarly to visual models, BERT layers have different roles in unpacking the text, with top-layers being responsible for capturing linguistic syntax [14]. The features extracted by the transformer are easily classified with a linear layer that can provide the final prediction (like in task 1).

### B. Visual Features Extraction

Subtask 3 introduces the challange of extracting features from images, in order to categorize two additional techniques and to improve previous results. To do so, we use the pre-trained ResNet-50, after training it furthermore on the task of Visual Sentiment (positive-negative) Analysis, on a related dataset [15]. The model is then fine-tuned on the target task. The features extracted by this model are classifiable with a linear layer.

### C. Unbalanced Data Problem

In all three subtask we deal with slightly different dataset that share the same problem of having unbalanced data. We develop and test two different solutions: Threshold Moving and Focal Loss. Threshold Moving consists in changing the final output thresholds for each class with the objective of optimizing the final result by increasing the number of predictions assigned to the minor classes [16]. To move the thresholds we simply optimize the micro measure on training data by brute-forcing a set of possible thresholds in the range $[0, 1]$ [17].

Another applicable solution consists in the usage of the Focal Loss function. This loss function is an extension of the Cross Entropy Loss, with a difference. It penalizes more the hard misclassified examples in the loss calculation [18].

### D. Model Architecture

We now present the final basic model architectures for each subtask. Binary Cross Entropy is the default loss function for all three subtasks.

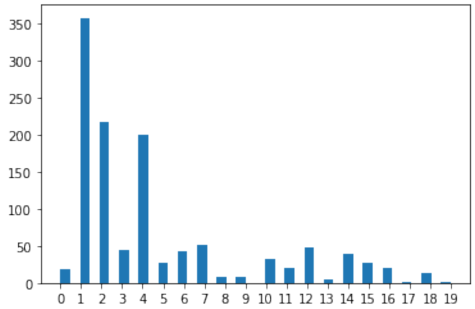

Fig. 1. Training data distribution for subtask 1. Some classes have less than 5 samples.

#### 1) Subtask 1

Our final model for task 1 is composed by the BERT feature extractor, trained with AdamW optimizer with weight decay [19]. Resulting features are then fed into a linear classifier. From the output logits, we compute the probability for each class using Softmax activation function. The batch size used is 14.

#### 2) Subtask 2

We modeled this subtask as a Token Level Multi-class Multi-label Classification. Which means a label is predicted for every class and for every token in the sentences. This strategy requires pre-processing the data labels from span to token level information, to then apply a backwards post-processing of the result before obtaining the final metrics. Apart from this, the model architecture is equal to the one presented for subtask 1.

#### 3) Subtask 3

Facing this task intrinsecally means solving the fusion between two different features vectors: the one coming from text, and the one coming from image. Our base model adds a linear layer on the image features extractor to reduce the number of features and makes a simple concatenation of the resulting vector with the text related one. This resulting feature vector is then treated as in subtask 1, with a different batch size of 6.

Furthermore, we provide a different solution of the Fusion Model. We create a second ”weighted” model that will first obtain a final probability prediction for both our feature vectors (resulting from images and text) and will then output a weighted average of the two classifications. The weight vector contains a different weight for every class. This numbers are considered to be parameters of the model and are optimized.

## IV. Results

### A. Baselines

The SemEval task provides a random baseline. We worked on this given baseline by still making it predict randomly, but based on the probability of a specific class to be true (given by the frequency of true labels for that class on training set). It is clear, from the histogram of the data labels in Figure 1, how this change can be very effective on performance.

More than one non-trivial and more solid baseline has been provided by us. We make use of two Machine Learning models, Naive Bayes Classifiers and SVC. By tokenizing the input and creating a simple dictionary, we obtain features that are then classified by these models. On subtask 3, images are ignored.

### B. Subtask 1

For the subtask 1 we decided to test 4 slightly different approaches: non pretrained BERT, pretrained BERT with all the layers unfrozen, pretrained BERT with the two initial layer frozen and with focal loss on our best scoring model. As we can see in table I with the BERT pretrained with the first two layers frozen and Binary Cross Entropy we obtain the best results.

TABLE I
SUBTASK 1 RESULTS ON TEST DATASET AND TOP SCORER FROM SEMEVAL.

| Model | F1-Micro | F1-Macro |
|---|---|---|
| MinD | 0.5933 | 0.2899 |
| BERT 2 frozen layers | 0.5348 | 0.1917 |
| + Threshold moving | **0.5786** | **0.2742** |
| + Focal Loss | 0.5603 | 0.2742 |
| BERT unfrozen | 0.5401 | 0.1710 |
| + Threshold moving | 0.5572 | 0.2389 |
| BERT | 0.5346 | 0.1420 |
| + Threshold moving | 0.5431 | 0.1732 |
| Linear SVC baseline | 0.3154 | 0.0942 |
| MultinomialNB baseline | 0.2637 | 0.0381 |
| Stratified baseline | 0.2595 | 0.0648 |
| SemEval baseline | 0.0443 | 0.0644 |

### C. Subtask 2

In subtask 2 we decided to carry on with the results from our firsts tests. We tested just the loss function, and indeed in this case, as we can see from table II, the model with Focal loss performs better. It is interesting to see how, if we use the threshold moving combined with the focal loss, we obtain a worse result.

TABLE II
SUBTASK 2 RESULTS ON TEST DATASET AND TOP SCORER FROM SEMEVAL.

| Model | F1 | Precision | Recall |
|---|---|---|---|
| BERT 2 frozen layers | 0.5447 | **0.6821** | 0.4534 |
| + Threshold moving | 0.5546 | 0.6065 | 0.5109 |
| + Focal loss | **0.5658** | 0.6601 | 0.4950 |
| + Threshold moving and focal loss | 0.5574 | 0.5986 | **0.5214** |
| Volta | 0.4817 | 0.5006 | 0.4641 |
| MultinomialNB baseline | 0.4119 | 0.5723 | 0.3217 |
| Linear SVC baseline | 0.2577 | 0.1952 | 0.3791 |
| Stratified baseline | 0.1649 | 0.0996 | 0.4791 |
| SemEval baseline | 0.0095 | 0.0337 | 0.0055 |

### D. Subtask 3

In the last subtask we continued testing on focal loss and BCE loss. In this case, we do not see the same behaviour as in the last subtask. The model with the weighted averages trained with BCE and threshold moving performed better as shown in table III.

TABLE III
SUBTASK 3 RESULTS ON TEST DATASET AND TOP SCORER FROM SEMEVAL.

| Model | F1-Micro | F1-Macro |
|---|---|---|
| Alpha | 0.5811 | 0.2731 |
| BERT 2 frozen layers and ResNet | 0.5418 | 0.2150 |
| + Threshold moving | 0.5571 | 0.2770 |
| + Focal loss | 0.5469 | 0.2548 |
| + Weighted model and threshold moving | **0.5602** | **0.2885** |
| Linear SVC baseline | 0.3154 | 0.0942 |
| MultinomialNB baseline | 0.2637 | 0.0381 |
| Stratified baseline | 0.2403 | 0.0680 |
| SemEval baseline | 0.0515 | 0.0706 |

### E. Comparison with State Of The Art

In Tables I, II, III we compare our results on the best models for each task competing in the SemEval 2021. Our best score achieves a very competitive result, always placing in the top-3 on the leaderboard. Specifically, on subtask 1 and 3, our model differs from the best one respectively for 0.01% and 0.02%. On subtask 2 our result is considerably better than the best team's one.

## V. CONCLUSION

We can observe from the results of this experiment that transfer learning plays an important role on improving test scores. Both BERT and ResNet greatly benefit from it, providing a big improvement on these models effectiveness. The other main challenge that we faced was to deal with an unbalanced dataset, a very common problem. We tackled it with different techniques: Focal loss and Threshold Moving. Both of these techniques are valid and, depending on the subtask, one can perform better than the other. But, as we can see from subtask 2 results, using them in a combined way does not always result in a good outcome.

In conclusion, we faced a quite challenging and indeed not trivial task. Even to our untrained eyes, giving the correct classification, it is not an easy task.

## REFERENCES

[1] D. Dimitrov, B. B. Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, and G. D. S. Martino, "Semeval-2021 task 6: detection of persuasion techniques in texts and images," *arXiv preprint arXiv:2105.09284*, 2021.

[2] L. M. Rojas-Barahona, "Deep learning for sentiment analysis," *Language and Linguistics Compass*, vol. 10, no. 12, pp. 701–719, 2016. [Online]. Available: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12228

[3] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: A review of bert-based approaches," *Artif. Intell. Rev.*, vol. 54, no. 8, p. 5789–5829, dec 2021. [Online]. Available: https://doi.org/10.1007/s10462-021-09958-2

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[7] A. S. B. Reddy and D. S. Juliet, “Transfer learning with resnet-50 for malaria cell-image classification,” in *2019 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2019, pp. 0945–0949.

[8] M. Mozafari, R. Farahbakhsh, and N. Crespi, “A bert-based transfer learning approach for hate speech detection in online social media,” in *International Conference on Complex Networks and Their Applications*. Springer, 2019, pp. 928–940.

[9] K. Gupta, D. Gautam, and R. Mamidi, “Volta at semeval-2021 task 6: Towards detecting persuasive texts and images using textual and multimodal ensemble,” *CoRR*, vol. abs/2106.00240, 2021. [Online]. Available: https://arxiv.org/abs/2106.00240

[10] Z. Feng, J. Tang, J. Liu, W. Yin, S. Feng, Y. Sun, and L. Chen, “Alpha at SemEval-2021 task 6: Transformer based propaganda classification,” in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 99–104. [Online]. Available: https://aclanthology.org/2021.semeval-1.8

[11] J. Tian, M. Gui, C. Li, M. Yan, and W. Xiao, “Mind at semeval-2021 task 6: Propaganda detection using transfer learning and multimodal fusion,” in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 2021, pp. 1082–1087.

[12] M. Mosbach, M. Andriushchenko, and D. Klakow, “On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines,” *arXiv preprint arXiv:2006.04884*, 2020.

[13] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune bert for text classification?” in *China national conference on Chinese computational linguistics*. Springer, 2019, pp. 194–206.

[14] J. Lee, R. Tang, and J. Lin, “What would elsa do? freezing layers during transformer fine-tuning,” *arXiv preprint arXiv:1911.03090*, 2019.

[15] S. Z. Hassan, K. Ahmad, S. Hicks, P. Halvorsen, A. Al-Fuqaha, N. Conci, and ichael Riegler, “Visual sentiment analysis from disaster images in social media,” 2020.

[16] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, “Finding the best classification threshold in imbalanced classification,” *Big Data Research*, vol. 5, pp. 2–8, 2016.

[17] X. Zhang, H. Gweon, and S. Provost, “Threshold moving approaches for addressing the class imbalance problem and their application to multi-label classification,” *2020 4th International Conference on Advances in Image Processing*, 2020.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” 2017. [Online]. Available: https://arxiv.org/abs/1708.02002

[19] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.