HTE at ArAIEval Shared Task: Integrating Content Type Information in Binary Persuasive Technique Detection

Hadjer Khaldi
Geotrend
Toulouse, France
hadjer@geotrend.fr

Taqiy Eddine Bouklouha
SolutionData Group
Toulouse, France
tbouklouha@solutiondatagroup.fr

Abstract

Propaganda frequently employs sophisticated persuasive strategies in order to influence public opinion and manipulate perceptions. As a result, automating the detection of persuasive techniques is critical in identifying and mitigating propaganda on social media and in mainstream media. This paper proposes a set of transformer-based models for detecting persuasive techniques in tweets and news that incorporate content type information as extra features or as an extra learning objective in a multitask learning setting. In addition to learning to detect the presence of persuasive techniques in text, our best model learns specific syntactic and lexical cues used to express them based on text genre (type) as an auxiliary task. To optimize the model and deal with data imbalance, a focal loss is used. As part of ArabicNLP2023-ArAIEval shared task, this model achieves the highest score in the shared task 1A out of 13 participants, according to the official results, with a micro-F1 of 76.34% and a macro-F1 of 73.21% on the test dataset. [1]

1 Introduction

In an era marked by the proliferation of information via digital platforms, separating fact from fiction has become an increasingly difficult task, nearly impossible to achieve manually. News and social media platforms are effective tools for disseminating information, but they also serve as breeding grounds for propaganda, misinformation, and manipulation. Propaganda messages can be used to influence people's opinions, beliefs, and behaviours by appealing to their emotions or by using persuasive techniques and arguments that may sound convincing but are based on faulty logic and thus invalid. To combat this, persuasive technique detection has emerged as an important component in the fight against deceptive content.

Most studies in this field focus on one genre of textual content for detecting persuasive techniques (Da San Martino et al., 2020; Barrón-Cedeño et al., 2019; Dimitrov et al., 2021; Carik and Yeniterzi, 2021; Alam et al., 2022). Handling multi-genre text has received little attention.

In this paper, we concentrate on the automatic detection of persuasive techniques in tweets and news. We propose various transformer-based systems for detecting persuasive techniques that implicitly and explicitly utilize content type to enhance detection in multi-genre text. As part of the ArabicNLP2023-ArAIEval shared task (Hasanain et al., 2023a), the task in which we participate (task 1A) involves a collection of Arabic tweets and news paragraphs annotated to indicate the presence or absence of persuasive content.

The rest of the paper is organized as follows: Section 2 gives an overview of related work. In Section 3, we present the data used. The proposed system is described in Section 4. In Section 5, we provide the details of our experiments, and then the results for our official runs are presented in Section 6. In Section 7, a discussion of the results is presented. We conclude the paper in Section 8.

2 Related Work

Over the past few years, there has been an increase in concern about the spread of opinion-shaping news and misinformation, particularly in the context of critical events such as COVID-19, elections, and conflicts. As a result, identifying propaganda content and persuasive techniques has gained more importance.

Research on propaganda content detection has targeted various media platform contents, including news (Da San Martino et al., 2020; Barrón-Cedeño et al., 2019), memes (Dimitrov et al., 2021), and tweets (Carik and Yeniterzi, 2021; Alam et al., 2022; Vijayaraghavan and Vosoughi, 2022; Mubarak et al., 2023).

[1] Code available at https://github.com/TaqiyEddine-B/Transformers-for-Propaganda-Detection

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 502–507
December 7, 2023 ©2023 Association for Computational Linguistics

With the introduction of Transformer models (Vaswani et al., 2017), the detection of propaganda and fake news has seen significant improvement in performance. Some works relied solely on real-world data to fine-tune a pre-trained language model for propaganda and persuasive technique detection (Costa et al., 2023), whereas others combined it with synthetically augmented data (Hasanain et al., 2023b). The ensemble approach was also investigated, in which various combined pre-trained language models are fine-tuned in a vanilla setting (Purificato and Navigli, 2023), or by using adapters (Wu et al., 2023).

One major limitation of all the preceding works is that they focus on one type of media platform content at a time. In this paper, we investigate the detection of persuasive techniques from multi-genre text extracted from tweets and news articles. The task is made more difficult by the differences in writing styles and contexts in both texts.

In line with the assumption proposed by (Barrón-Cedeño et al., 2019), which affirms that sentence representations incorporating information about writing style tend to exhibit better generalization than word-level representations in news propaganda detection, we delve into the integration of content type (genre) information within transformer-based models for the detection of persuasive techniques.

3 Data

Our data comes from the proposed dataset for persuasion technique detection as part of the ArAIEval 2023 shared task 1A. [2] No additional data was used. Each entry in the data file is composed of three fields: `text` referring to the textual content, `type` referring to the genre of text: tweet or news, and `label` referring to the presence or absence of persuasive technique in the text: True or False.

Table 1 describes the data distribution per `type` and per `label`. Overall, we can notice that dataset is imbalanced in terms of label and type distributions. Texts of type news paragraph are over-represented in the dataset, representing 65% compared to tweets (35%). Then, texts using persuasive techniques are more prevalent (79%) than non-persuasive content (21%%).

	TRUE	FALSE	**#Total**
Paragraph	1201 (76%)	374 (24%)	1575 (65%)
Tweet	717 (84%)	135 (16%)	852 (35%)
#Total	1918 (79%)	509 (21%)	2427

Table 1: Data distribution in train dataset per label and content type.

When comparing persuasive tweets and news paragraphs, the context used to express news paragraphs is twice as long as the context used to express tweets, with the average length of news being 211 characters compared to 100 characters for tweets. We expect that the length of the context will influence the syntactic and lexical cues used to express persuasive techniques in news paragraphs and tweets. We should point out that no text preprocessing was done on the dataset for training.

4 System Overview

The system's goal is to determine whether a multi-genre (a tweet or news paragraph) snippet contains persuasive content. Our proposed system (cf. Figure 1) is made up of : **(1)** a transformer-based encoder Enc_i (Vaswani et al., 2017) that encodes the input texts into a fixed-size contextualized vector, **(2)** followed by a feature injection layer ($Feat$) that concatenates the content type vector with the input vector, and **(3)** two parallel classifiers C, each of which is made up of a fully-connected layer, a dropout layer, and an activation layer, and perform two different tasks:

– C_{main}: a classifier that performs the main task that learns to recognize the presence of persuasive techniques in texts (binary classification).

– C_{aux}: a classifier that performs an auxiliary (support) task that learns to identify the type of text: tweet or news (binary classification).

Each task calculates one loss, and optimizing the model optimizes the sum of the two losses. Because the two tasks share the same encoder, the auxiliary task can help the main task learn additional specific syntactic and lexical cues for tweets or news content used to express persuasive arguments.

According to recent research, jointly learning common characteristics shared across multiple tasks can have a significant impact on NLP classification performances as it enhances the perfor-

[2] ArAIEval dataset for persuasion technique detection

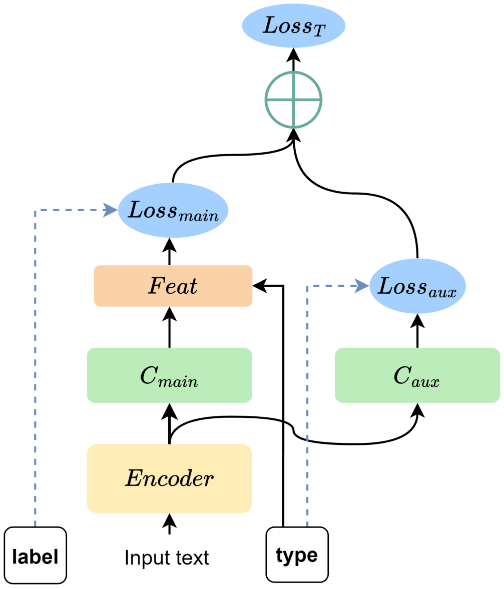

Figure 1: Proposed System architecture.

mance of the main task by incorporating other related tasks, making it easier to combine information from multiple resources (Ye et al., 2019; Liu et al., 2017; He et al., 2019; Khaldi et al., 2022; Tafreshi and Diab, 2018).

5 Experimental Setup

We experiment with different configurations for the proposed system components. We evaluated two transformer-based encoders: MARBERT (noted M) (Abdul-Mageed et al., 2021) and AraBERT (noted A) (Abdul-Mageed et al., 2021), noting Enc_i with $i \in \{M, A\}$. Because $Feat$ directly injects the text type into the system as a feature and C_{aux} learns to predict it, it is not possible to enable both $Feat$ and C_{aux} in a model, thus enabling one disables the other. The resulted systems from various configurations are shown below:

- Enc_M+C_{main} and $Enc_A + C_{main}$, that we consider as baseline models, that perform a classical binary classification based on the input vector, in a monotask setting without any additional features.
- Enc_M+$Feat$+C_{main} and Enc_A+$Feat$+C_{main} that additionally inject the content type (genre) as a feature alongside the input representation.
- Enc_M+C_{main}+C_{aux} and Enc_A+C_{main}+C_{aux} that perform two binary classification tasks, namely: persuasive technique detection as a main task and type detection as an auxiliary task.

A cross-entropy loss (noted CE) is used to optimize the systems. We also experiment with a focal loss (Lin et al., 2017) to deal with data imbalance in the train data, as it has been shown to be effective in many imbalanced NLP classification problems (Liu et al., 2021; Ma et al., 2020; Huang et al., 2021). Train data is used to fine-tune all systems. The development data is used to fine-tune the model's hyperparameters, where the best ones are reported in Table 2. We evaluate the fine-tuned models on the development dataset, the official micro-F1 score is shown in Table 3. The best performing one was selected to be submitted for the official ranking on the test set.

Hyperparameter	Value
learning_rate	2e^{-5}
epochs	5
batch_size	16

Table 2: Best Hyperparameters after fine-tuning on development dataset.

6 Results

In general, we found that explicitly or implicitly incorporating content type information into the system could improve overall results for both MARBERT and AraBERT encoders when either cross-entropy or focal loss was used. For example, both Enc_M+$Feat$+C_{main} and Enc_M+C_{aux}+C_{main} beat Enc_M+C_{main}, with almost + 2% and + 1% on micro-F1. Among the twelve evaluated system configurations, Enc_A+C_{main}+C_{aux} optimized using a focal loss represents our best performing one during the development phase, achieving an increase of nearly 3% above baselines. This model was submitted for official ranking on the test dataset for *task 1A*, and the obtained micro-F1 is 0.7634, which is the highest score on the leaderboard for this task.

7 Discussion

The test dataset contains 503 inputs, of which 119 are classified incorrectly. Out of these, 103 are paragraphs (87%) and 16 are tweets (13%).

A closer examination of the confusion matrices per content type for the best-performing system configuration (cf. figure 3 for tweet text and figure 2 for news paragraphs) reveals that the majority of tweet misclassifications (75%) involve non-persuasive content being mistaken for per-

Models	CE	Focal
Enc_M+C_{main}	0.8301	0.8263
Enc_A+C_{main}	0.8108	0.8301
Enc_M+$Feat$+C_{main}	0.8533	0.8417
Enc_A+$Feat$+C_{main}	0.8108	0.8571
Enc_M+C_{main}+C_{aux}	0.8378	**0.8610** †
Enc_A+C_{main}+C_{aux}	0.8147	0.8340

Table 3: Evaluation of proposed system configurations on the development dataset of *task 1A*. Official metric micro-F1 is reported. The best result is in bold and † marks the system submitted for the official ranking.

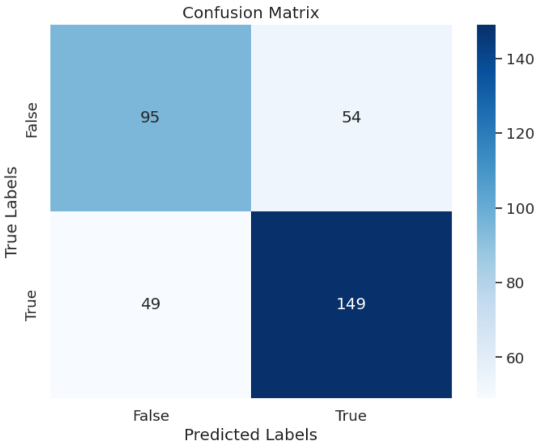

Figure 2: Confusion matrix for paragraph news.

suasive content. This could be explained by the over-representation of persuasive content in tweets (84%). In contrast, the misclassification rate of persuasive or non-persuasive texts in news paragraphs is nearly identical.

8 Conclusion

In this paper, we present our experiments and findings on the detection of persuasive techniques in multi-genre texts, which encompass tweets and news paragraphs. This research was part of the ArabicNLP2023-ArAIEval Task1A shared task, focusing on identifying persuasive techniques through binary classification. Our team proposed a system based on fine-tuning a transformer-based model to assess the impact of integrating content type information on persuasive technique detection. We experimented with two different approaches to information integration: implicitly, by adding an additional learning objective to the model, or explicitly, as an additional feature.

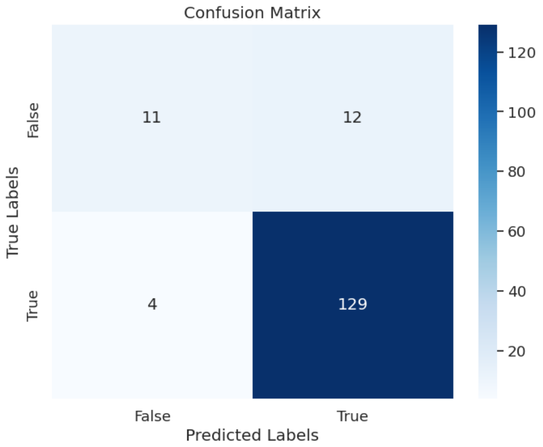

Figure 3: Confusion matrix for tweets.

We evaluated two pre-trained language models for Arabic and optimized the system using two different loss functions: cross-entropy and focal loss to address data imbalance. Our highest scores on the development dataset were achieved by the MARBERT model, trained using focal loss, for both persuasive technique detection and content type detection. As a result, our model secured the first position in the ArabicNLP2023-ArAIEval Task1A shared task during the test phase.

Our future work will involve exploring data augmentation techniques to address data imbalance and integrating multiple pre-trained language models. Finally, our results indicate that our system faces challenges in identifying persuasive content in news paragraphs. To pinpoint the causes of misclassifications, a deeper investigation into incorrectly classified sentences is warranted.

Limitations

Firstly, when applied to Arabic text, incorporating content type information as a new learning objective in a persuasive technique detection task yielded satisfactory results. However, it's important to note that this outcome may not necessarily hold true for other languages; extensive testing is required to confirm the results across different linguistic contexts.

Additionally, the distribution of data in terms of content type may exert an influence on the task of content type identification. It's worth highlighting that a significant imbalance between the two types, or considering more than two content types in the dataset, can potentially impact the overall results.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing Management*, 56(5):1849–1864.

Buse Carik and Reyyan Yeniterzi. 2021. SU-NLP at CheckThat! 2021: Check-worthiness of Turkish tweets.

Nelson Filipe Costa, Bryce Hamilton, and Leila Kosseim. 2023. CLaC at SemEval-2023 task 3: Language potluck RoBERTa detects online persuasion techniques in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1613–1618, Toronto, Canada. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th Workshop on Semantic Evaluation*, SemEval '20, pages 1377–1414.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '21, pages 70–98.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abdelhakim Freihat. 2023a. Araieval shared task: Persuasion techniques and disinformation detection in arabic text. In *Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.

Maram Hasanain, Ahmed El-Shangiti, Rabindra Nath Nandi, Preslav Nakov, and Firoj Alam. 2023b. QCRI at SemEval-2023 task 3: News genre, framing and persuasion techniques detection using multilingual models. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1237–1244, Toronto, Canada. Association for Computational Linguistics.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515, Florence, Italy. Association for Computational Linguistics.

Yi Huang, Buse Giledereli, Abdullatif Köksal, Arzucan Özgür, and Elif Ozkirimli. 2021. Balancing methods for multi-label text classification with long-tailed class distribution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8153–8161.

Hadjer Khaldi, Farah Benamara, Camille Pradel, and Nathalie Aussenac Gilles. 2022. A closer look to your business network: Multitask relation extraction from economic and financial french content. In *The AAAI-22 Workshop on Knowledge Discovery from Unstructured Data in Financial Services*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Jianyi Liu, Xi Duan, Ru Zhang, Youqiang Sun, Lei Guan, and Bingjie Lin. 2021. Relation classification via bert with piecewise convolution and focal loss. *Plos one*, 16(9):e0257092.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.

Yili Ma, Liang Zhao, and Jie Hao. 2020. XLP at SemEval-2020 task 9: Cross-lingual models with focal loss for sentiment analysis of code-mixing language. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 975–980, Barcelona (online). International Committee for Computational Linguistics.

Hamdy Mubarak, Samir Abdaljalil, Azza Nassar, and Firoj Alam. 2023. Detecting and reasoning of deleted tweets before they are posted. *arXiv preprint arXiv:2305.04927*.

Antonio Purificato and Roberto Navigli. 2023. APatt at SemEval-2023 task 3: The sapienza NLP system for ensemble-based multilingual propaganda detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 382–388, Toronto, Canada. Association for Computational Linguistics.

Shabnam Tafreshi and Mona Diab. 2018. Emotion detection and classification in a multigenre corpus with joint multi-task deep learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2905–2913, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Prashanth Vijayaraghavan and Soroush Vosoughi. 2022. TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3433–3448, Seattle, United States. Association for Computational Linguistics.

Ben Wu, Olesya Razuvayevskaya, Freddy Heppell, João A. Leite, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. SheffieldVeraAI at SemEval-2023 task 3: Mono and multilingual approaches for news genre, topic and persuasion technique classification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1995–2008, Toronto, Canada. Association for Computational Linguistics.

Wei Ye, Bo Li, Rui Xie, Zhonghao Sheng, Long Chen, and Shikun Zhang. 2019. Exploiting entity BIO tag embeddings and multi-task learning for relation extraction with imbalanced data. In *Proceedings of ACL*, pages 1351–1360.