Unisa at SemEval-2023 Task 3: A SHAP-based method for Propaganda Detection

Micaela Bangerter and **Giuseppe Fenza**
and **Mariacristina Gallo** and **Vincenzo Loia** and **Alberto Volpe**
Department of Management & Innovation Systems,
University of Salerno, 84084 Fisciano (SA), Italy
{mbangerter, gfenza, mgallo, loia, alvolpe}@unisa.it

Carmen De Maio
Department of Information Engineering, Electrical Engineering and Applied Mathematics,
University of Salerno, 84084 Fisciano (SA), Italy
cdemaio@unisa.it

Claudio Stanzione
Defence Analisys & Research Institute,
Center for Higher Defence Studies, 00165 Rome (RM), Italy
stanzione.dottorando@casd.difesa.it

Abstract

This paper presents proposed solutions for addressing two subtasks in SemEval-2023 Task 3: "Detecting the Genre, the Framing, and the Persuasion techniques in online news in a multilingual setup". In subtask 1, "News Genre Categorisation", the goal is to classify a news article as an opinion, a report, or a satire. In subtask 3, "Detection of Persuasion Technique", the system must reveal persuasion techniques used in each news article paragraph choosing among 23 defined methods.

Solutions leverage the application of the eXplainable Artificial Intelligence (XAI) method, Shapley Additive Explanations (SHAP). In subtask 1, SHAP was used to understand what was driving the model to fail so that it could be improved accordingly. In contrast, in subtask 3, a re-calibration of the Attention Mechanism was realized by extracting critical tokens for each persuasion technique. The underlying idea is the exploitation of XAI for countering the overfitting of the resulting model and attempting to improve the performance when there are few samples in the training data. The achieved performance on English for subtask 1 ranked 6th with an F1-score of 58.6% (despite 78.4% of the 1st) and for subtask 3 ranked 12th with a micro-averaged F1-score of 29.8% (despite 37.6% of the 1st).

1 Introduction

Recent technology improvements undoubtedly promote freedom of speech but leave society defenseless against potential news tampering (Roozenbeek et al., 2020). The main drive behind the SemEval 2023 Task 3 (Piskorski et al., 2023) is to foster the development of methods and tools to support the analysis of online media content to understand what makes a text persuasive: which writing style is used, what critical aspects are highlighted, and which persuasion techniques are used to influence the reader. Persuasion attempts and propaganda (through different ways), appeal to the user's sentiment to influence his/her opinions. Unfortunately, detecting propaganda and persuasion techniques are trending research topics that still perform poorly, especially when labeled data is limited (Da San Martino et al., 2020a; Chernyavskiy et al., 2020).

Our approach tries to overcome problems related to the limited number of training instances by generalizing data through the Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) method, an eXplainable Artificial Intelligence (XAI) technique. SHAP is a framework trying to explain the prediction of an instance x by computing the contribution of each feature to the prediction. The SHAP explanation technique uses coalitional game theory to compute Shapley values.

In subtask 1, SHAP is employed for feature selection (Effrosynidis and Arampatzis, 2021). The idea is to leverage features to understand how to pre-process documents before sending them to the model. In subtask 3, SHAP re-calibrates the Attention Mechanism for each persuasion technique. The idea consists of identifying words characterizing each type of persuasion and helping the learning model to focus on them.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 885–891

The paper is organized as follows. Section 2 gives an overview of existing approaches and theories, and Section 3 describes our solutions. Sections 4 and 5 detail experimentation, and finally, Section 6 concludes the manuscript.

2 Background

In recent years, deep learning-based methods have become increasingly popular in natural language processing and information retrieval. Transformers are a type of deep learning architecture that has gained popularity in natural language processing tasks, especially for tasks involving contextualized word embedding.

The news genre categorization task includes satire detection, which is also approached using transformer architectures (Bhardwaj and Prusty, 2022; Pandey and Singh, 2023; Nayak and Bolla, 2022). In (Alhindi et al., 2020; Blackledge and Atapour-Abarghouei, 2021) authors present the classification of opinion-based news articles to finally classify fact or fake news pieces.

Regarding persuasion detection, most of the works in the literature focus on text analysis. Iyer et al. leveraged patterns identified for each considered persuasion technique (Iyer et al., 2017). Hidey et al. demonstrated the ordering of arguments is crucial to persuasion (Hidey and McKeown, 2018). Similarly, an LSTM-based approach for persuasion detection in social media conversations is proposed in (Dutta et al., 2020). Persuasion detection has been a goal of recent SemEval competitions (Dimitrov et al., 2021; Da San Martino et al., 2020b).

Inspired by (Yu et al., 2019), the system implementing subtask 3 proposes a BERT-based text classification model via constructing auxiliary sentence to turn the classification task into a binary sentence-pair one, aiming to address the limited training data problem and task-awareness problem.

To the best of our knowledge, this is the first attempt to employ an XAI-based approach for persuasion techniques identification.

3 System Overview

The system adopts the DistilBert transformer (Sanh et al., 2019) and the SHAP method. The following subsections detail how they are combined to address the subtasks 1 and 3 included in SemEval-2023 Task 3: "Detecting the Genre, the Framing, and the Persuasion techniques in online news in a multi-lingual setup".

More in detail, subsection 3.1 points out that for subtask 1, SHAP was used to understand what tokens drive the classification of the article. In contrast, in subsection 3.2, for subtask 3, SHAP was adopted to re-calibrate the Attention Mechanism by leveraging words that distinguish each persuasion technique and help the model to generalize better.

3.1 News Genre Categorisation

The English training dataset contains unbalanced classes, with minimal satire articles despite many opinion pieces. So, to improve the classification performance, the sub-task was approached as follows. First, data augmentation was made by translating articles in other languages to the target language (i.e., English) for the least represented classes (i.e., reporting and satire). Our corpus was finally the combination of the translated articles with the original English ones. These articles were first passed to a preprocessing phase, where non-redundant tokens were cleaned to reduce the article's noise and avoid losing more information for the reduced amount of tokens available to pass to the transformer. Then, the cleaned articles were passed to the DistilBert model for a sentence-level classification task taking as input article words for a maximum of 512 tokens, as allowed.

The model is fine-tuned for the specific task by adding additional classifiers on top of the pre-trained DistilBert model. After the model was trained for the first time, we used SHAP to understand what input tokens had more influence on the model for the article's final (wrong) classification. This process produces a black-list of tokens to filter out that may adjust the cleaning function and reduce the effect of the limited number of allowed tokens. Finally, once the preprocessing function is defined, we fine-tune the model again with these adjusted articles, and the final model is then used to predict new coming articles.

Figure 1 presents the workflow of classifying a new incoming article. First, the tokenization adds special tokens to the first 512 tokens of the input (the first tokens resulting from filtering ones identified by SHAP) and passes them to the trained model. The idea is to remove the tokens guiding the model to a wrong classification and then adopt the first 512 tokens of the input for the classification. Finally, we pass the logits made by the model to a softmax function and choose the most probable class.

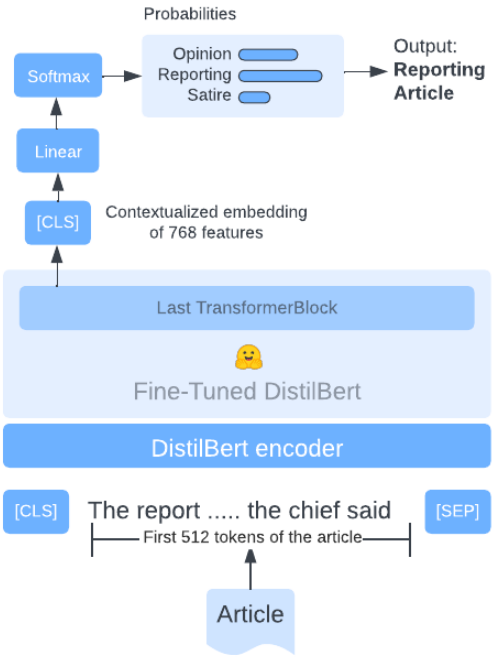

Figure 1: Sub-task 1 - Prediction of new coming input

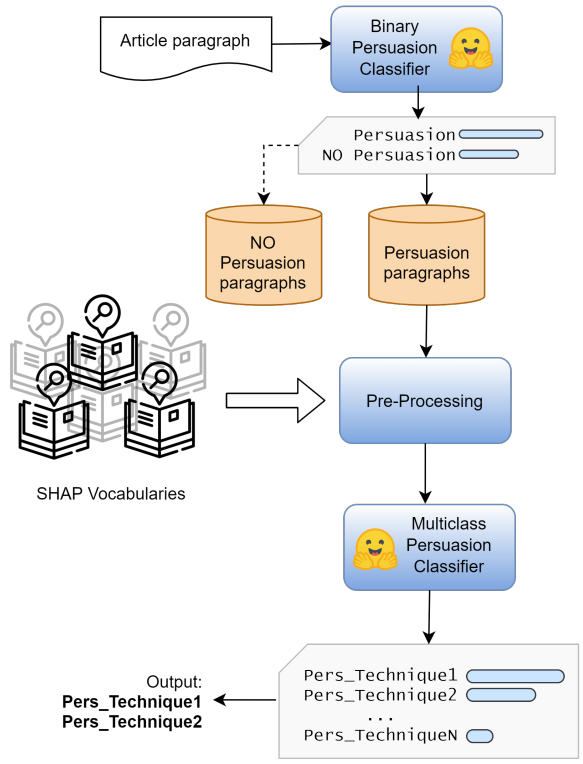

Figure 2: Multi-label classification for Subtask3

3.2 Persuasion Techniques Detection

Regarding subtask 3, the overall system is depicted in Figure 2. A binary model first processes a new incoming paragraph to predict whether it contains any persuasion attempt (i.e., is classifiable as "Persuasion"). If the text is predicted to be persuasion, it is compared with *SHAP Vocabularies* previously created, representing the most important words associated with each persuasion technique. Such comparison defines the additional input to pass to the final multiclass model that will establish the probability by which the input text is classified to each persuasion technique. Classes with probabilities exceeding a fixed threshold will be part of the final multilabel classification.

Following subsections detail the processes of training (i.e., learning model preparations) and testing (i.e., learning model adoptions).

Training Phase. The training process can be, in turn, divided into two main stages: the first builds SHAP Vocabularies; the second is aimed at training two DistilBert Transfomers (i.e., the Binary Persuasion Classifier and the Multiclass Persuasion Classifier).

Constructing SHAP Vocabularies consists of extracting the essential words characterizing each persuasion technique by exploiting the SHAP method. The process consists of constructing N DistilBert binary classifiers (one for each persuasion technique) and exploiting SHAP to identify the most important words (i.e., ones that guided the classification decision). For dataset construction, rows with more than one label are split to obtain rows with only one label. Then, for each technique, a positive sample dataset with (M) rows belonging to specific techniques and a negative sample dataset containing M random samples belonging to other $N-1$ techniques are constructed. Validation and testing subsets were extracted from datasets created for each technique. Thus, a binary DistilBert Transformer was trained and tested for each technique. The registered Accuracy for these classifiers is about 70%, on average. Once the predictions were obtained, ones corresponding with the starting persuasion technique were filtered and applied SHAP. Based on Shaply Values, we construct N vocabularies of main words distinguishing each technique.

In the second training stage, two DistilBert Transformers are trained: a Binary Persuasion Classifier establishes if the text can be considered persuasion in general; a Multiclass Persuasion Classifier assigns to the input a probability of belonging to each type of persuasion. The choice of adding

a separate classifier to first identify text containing persuasion attempts allows for improving the performance of the subsequent multi-class classifier. In fact, as described in Section 4.2, paragraphs not containing persuasions are more than half. To train the model, paragraphs in the Train and Dev Set that featured at least one technique, out of the available 23, were associated with the label "Persuasion", while paragraphs that did not feature any technique were labeled as "NO Persuasion". During the test of this classifier, registered Accuracy was 84%.

The second adopted model is a DistilBert Multi-class single-label Transformer, trained as follows. First, all rows with more than one label are split to obtain rows with only one label. The insight, in this case, was to provide the model with the part of the text on which to focus attention most, thus going on to recalibrate the attention mechanism. In particular, after the tokenization of input text, each word was compared with each word contained in the N SHAP dictionaries. Those words that were found to have a similarity of at least 90% with one of the words contained in the SHAP dictionaries were considered essential. Such words become part of the input paragraph as follows:

$$[CLS]\ token_1\ token_2\ \ldots\ token_n\ [SEP]\ paragraph\ [SEP] \quad (1)$$

where $token_1$, $token_2$, $\ldots$ $token_n$ are words found to be fundamental after comparison with SHAP dictionaries.

Testing Phase. During the test phase, for each incoming instance classified as "Persuasion", the model returns a belonging probability for each persuasion technique. So, to move such type of classification to a multi-label result, all labels with a probability greater than 0.15 are included in the resulting classification. Such threshold is selected empirically.

4 Experimental setup

Methodologies described in Section 3, have been applied to the dataset provided by the SemEval-2023 Task 3 organizers, specifically for the English subtask.

The following subsections describe the experimentation setups and results for both considered subtasks. In particular, results have been evaluated through official measures: macro-F1 and micro-F1 for subtasks 1 and 3, respectively, as described in (Piskorski et al., 2023). Furthermore, the number of epochs used in each subtask was defined with a maximum of 6 as we train LMs; if no improvements are made, fewer epochs are used.

4.1 News Genre Categorisation

Regarding Subtask 1, the English training dataset contains 433 news and web articles, while the development set a total of 83 instances. Finally, 54 articles were released for the test set without known gold labels.

As described previously, for subtask 1, where the aim is to classify entire articles, a preprocessing was made to clean the text in each article, removing punctuation.

The Python googletrans[1] library that implements the Google Translate API was used to translate non-English language articles. Through this translation process, the final training set contains a total of 674 articles. In particular, the following articles were added: 52 items from Italian, 54 from French, 46 from German, 40 from Poland, and 49 from Russian.

The adopted Transformers model is a DistilBert base uncased fine-tuned through the following hyperparameters: batch size of 16; learning rate of $2e^{-5}$; AdamW optimizer; 6 epochs.

A randomly sampled 15 percent of the training set is selected for validation purposes. For the training, an NVIDIA GeForce GTX 1060 with a memory of 6GB was used.

4.2 Persuasion Techniques Detection

The dataset provided by the SemEval-2023 Task 3 organizers for Subtask 3 specific for the English language contains 446 training news and web articles and 9498 paragraphs, where 5738 of these are labeled as "No Persuasion" and 3760 fall under at least one of the 23 persuasion techniques. 90 articles are provided in the Dev set, with 3127 paragraphs: 2007 as "No Persuasion" and 1120 falling under at least one of the persuasion techniques. Finally, 54 articles containing 910 paragraphs were released for the test set where no gold labels were known.

For subtask 3, where the aim is to identify the persuasion techniques in each paragraph, a preprocessing was made to clean the text in each article, removing punctuation. In the first phase, for the Binary Persuasion Classifier, the model used is a

[1] https://pypi.org/project/googletrans/

Transformer DistilBert base uncased, with the following hyperparameters: batch size of 16; learning rate of $2e^{-5}$; AdamW optimizer; 4 epochs.

A randomly sampled 15 percent of the training set is selected for validation purposes.

For the Multiclass Persuasion Classifier, the adopted model is again a Distilbert base uncased but with the same hyperparameters except for the batch size, which was set to 10.

For both models' training, an NVIDIA GeForce GTX 1060 with a memory of 6GB was used.

5 Results

In this section, obtained results for both considered subtasks are described.

5.1 News Genre Categorisation

Table 1 shows the results of our approach compared with one of the competition winner and the baseline for the English Subtask 1. Our model achieves a macro F1 of $58,621\%$, resulting in 6th place on the English leaderboard from 23 teams.

Method	F1 macro	F1 micro
Best ranked (MELODI)	0.78431	0.81481
BERT-SHAP (Ours)	**0.58621**	0.61111
Baseline	0.28802	0.61111

Table 1: Scores for English sub-task 1 - News Genre Categorisation

Table 2 reports how the model performs on the development set for each label. The percentage of correct prediction is better in Reporting class, followed by Opinion and Satire. Although we increased the number of articles to have a more balanced dataset, the minor class still performed poorly. Additionally, general performances with and without the use of SHAP are presented.

5.2 Persuasion Techniques Detection

Table 3 shows the results of our approach compared with one of the competition winner and the baseline for the English Subtask 3. Our model achieves a micro F1 of $29,758\%$, resulting in 12th place for the English subtask 3 leaderboard from 23 teams.

Table 4 reports the results on the Dev set for each involved persuasion technique. Analyzing the results, it is immediately apparent how so many techniques have a performance of 0%, highlighting the difficulty of constructing a multilabel classifier with unbalanced data, despite the adoption of an

Labels	Precision	Recall	F1-Score	Support
Opinion	0.46	0.60	0.52	20
Reporting	0.78	0.70	0.74	54
Satire	0.06	0.11	0.08	9
BERT-SHAP	0.43	0.47	0.44	83
BERT	0.32	0.33	0.32	83

Table 2: Performance per class on the development set for sub-task 1 and comparison with an approach without SHAP.

Method	F1 micro	F1 macro
Best ranked (APatt)	0.37562	0.12919
BERT-SHAP (Ours)	**0.29758**	**0.10871**
Baseline	0.19517	0.06925

Table 3: Scores for English sub-task 3 - Persuasion Techniques Detection

XAI approach. For these techniques, additional analysis must be done to identify and extract relevant patterns and further help the classifier. Moreover, another analysis should be conducted for the "Repetition" technique, which involves the repeated use of words or concepts to redundant an idea. In this and similar cases, defining additional models or rules more focused on the specific target class and lightening the multi-class responsibility could be helpful. Finally, general performances with and without the use of SHAP are also presented in Table 4.

6 Conclusion

This work derives from the contribution of our team to the SemEval-2023 Task 3: "Detecting the Genre, the Framing, and the Persuasion techniques in online news in a multi-lingual setup". Specifically, it faces English subtasks 1 and 3. In the first case, the objective is to classify a news article as an opinion, a report, or a satire; in the second case, to identify one or many persuasion techniques in each paragraph of a given article.

Proposed solutions leverage the Shapley Additive Explanations (SHAP) method. In particular, in subtask 1, SHAP was used as a feature detection method by filtering tokens bring to wrong classifications, while in subtask 3, a re-calibration of the Attention Mechanism (typically implemented by Transformer models) was realized by extract-

Labels	Precision	Recall	F1-Score	Support
Loaded Language	0.53	0.81	0.64	483
Name Calling-Labeling	0.62	0.58	0.60	250
Doubt	0.39	0.48	0.43	187
Repetition	0.15	0.17	0.16	141
Appeal to Fear-Prejudice	0.37	0.32	0.35	137
Exaggeration-Minimisation	0.28	0.28	0.28	115
Flag Waving	0.45	0.50	0.47	96
False Dilemma-No Choice	0.31	0.06	0.11	63
Appeal to Popularity	0.00	0.00	0.00	34
Appeal to Authority	0.21	0.14	0.17	28
Slogans	0.30	0.32	0.31	28
Conversation Killer	0.15	0.24	0.19	25
Causal Oversimplification	0.06	0.08	0.07	24
Red Herring	0.00	0.00	0.00	19
Obf-Vag-Conf	0.00	0.00	0.00	13
Straw Man	0.00	0.00	0.00	9
Appeal to Hypocrisy	0.00	0.00	0.00	8
Guilt by Association	1.00	0.25	0.40	4
Whataboutism	0.00	0.00	0.00	2
BERT-SHAP	0.40	0.48	0.44	1666
BERT	0.36	0.40	0.38	1666

Table 4: Performance per class on the development set for sub-task 3 and comparison with an approach without SHAP.

ing the most important tokens for each technique. The objective is to exploit Explainable Artificial Intelligence techniques to improve classification performance when the availability of labeled data for training is scarce through a generalization attempt. Experimental results demonstrate promising, encouraging further investigation in this sense. In particular, in the future, regarding subtask 3, it could be interesting to better address the SHAP Vocabularies construction by adding spans containing the persuasion attempt to the adopted binary classifiers. This could allow focusing on a specific substring of the paragraph, improving the classification performance of the single binary classifiers.

Acknowledgements

This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

Tariq Alhindi, Smaranda Muresan, and Daniel Preoţiuc-Pietro. 2020. fact vs. opinion: The role of argumentation features in news classification. In *Proceedings of the 28th international conference on computational linguistics*, pages 6139–6149.

Saumya Bhardwaj and Manas Ranjan Prusty. 2022. Bert pre-processed deep learning model for sarcasm detection. *National Academy Science Letters*, 45(2):203–208.

Ciara Blackledge and Amir Atapour-Abarghouei. 2021. Transforming fake news: Robust generalisable news classification using transformers. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3960–3968. IEEE.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2020. Aschern at semeval-2020 task 11: It takes three to tango: Roberta, crf, and transfer learning. *arXiv preprint arXiv:2008.02837*.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020b. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 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 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98.

Subhabrata Dutta, Dipankar Das, and Tanmoy Chakraborty. 2020. Changing views: Persuasion

modeling and argument extraction from online discussions. *Information Processing & Management*, 57(2):102085.

Dimitrios Effrosynidis and Avi Arampatzis. 2021. An evaluation of feature selection methods for environmental data. *Ecological Informatics*, 61:101224.

Christopher Hidey and Kathleen McKeown. 2018. Persuasive influence detection: The role of argument sequencing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Rahul R Iyer, Katia P Sycara, and Yuezhang Li. 2017. Detecting type of persuasion: Is there structure in persuasion tactics? In *CMNA@ ICAIL*, pages 54–64.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Deepak Kumar Nayak and Bharath Kumar Bolla. 2022. Efficient deep learning methods for sarcasm detection of news headlines. In *Machine Learning and Autonomous Systems: Proceedings of ICMLAS 2021*, pages 371–382. Springer.

Rajnish Pandey and Jyoti Prakash Singh. 2023. Bert-lstm model for sarcasm detection in code-mixed social media post. *Journal of Intelligent Information Systems*, 60(1):235–254.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada.

Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shanshan Yu, Jindian Su, and Da Luo. 2019. Improving bert-based text classification with auxiliary sentence and domain knowledge. *IEEE Access*, 7:176600–176612.