"Don't discuss": Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification

Vorakit Vorakitphan, Elena Cabrio, Serena Villata
Université Côte d'Azur, Inria, CNRS, I3S, France
vorakit.vorakitphan@inria.fr,
elena.cabrio@univ-cotedazur.fr, villata@i3s.unice.fr

Abstract

One of the mechanisms through which disinformation is spreading online, in particular through social media, is by employing propaganda techniques. These include specific rhetorical and psychological strategies, ranging from leveraging on emotions to exploiting logical fallacies. In this paper, our goal is to push forward research on propaganda detection based on text analysis, given the crucial role these methods may play to address this main societal issue. More precisely, we propose a supervised approach to classify textual snippets both as propaganda messages and according to the precise applied propaganda technique, as well as a detailed linguistic analysis of the features characterising propaganda information in text (e.g., semantic, sentiment and argumentation features). Extensive experiments conducted on two available propagandist resources (i.e., NLP4IF'19 and SemEval'20-Task 11 datasets) show that the proposed approach, leveraging different language models and the investigated linguistic features, achieves very promising results on propaganda classification, both at sentence- and at fragment-level.

1 Introduction

Propaganda represents an effective, even though often misleading, communication strategy to promote a cause or a viewpoint, for instance in the political context (Lasswell, 1938; Koppang, 2009; Dillard and Pfau, 2009; Longpre et al., 2019). Different communication means can be used to disseminate propaganda, i.e., textual documents, images, videos and oral speeches. The ability to effectively identify and manifestly label such kind of misleading and potentially harmful content is of primarily importance to restrain the spread of such information to avoid detrimental consequences for the society.

In this paper, we tackle this challenging issue (Da San Martino et al., 2020b) by proposing a textual propaganda detection model. More precisely, we address the following research questions: *(i) how to automatically identify propaganda in textual documents and further classify them into fine-grained categories?*, and *(ii) what are the linguistic distinctive features of propaganda text snippets?* The contribution of this paper consists not only in proposing a new effective neural architecture to automatically identify and classify propaganda in text, but we also present a detailed linguistic analysis of the features characterising propaganda messages.

Our work focuses on the propaganda detection and classification task, casting it both as a binary and as a multi-class classification task, and we address it both at sentence-level and at fragment-level. We investigate different architectures of recent language models (i.e., BERT, RoBERTa), combining them with a rich set of linguistic features ranging from sentiment and emotion to argumentation features, to rhetorical stylistic ones. The extensive experiments we conducted on two standard benchmarks (i.e., the NLP4IF'19 and SemEval'20-Task 11 datasets) show that the proposed architectures achieve satisfying results, outperforming state-of-the-art systems on most of the propaganda detection and classification subtasks. An error analysis discusses the main sources of misclassification. Furthermore, we analysed how the most relevant features for propaganda detection impact the fine-grained classification of the different techniques employed in propagandist text, revealing the importance of semantic and argumentation features.

2 Related Work

In the last years, there has been an increasing interest in investigating methods for textual propaganda detection and classification. Among them, (Barrón-

Proceedings of Recent Advances in Natural Language Processing, pages 1498–1507
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_168

Cedeño et al., 2019) present a system to organize news events according to the level of propagandist content in the articles, and introduces a new corpus (QProp) annotated with the propaganda vs. trustworthy classes, providing information about the source of the news articles. (Da San Martino et al., 2019) present the benchmark of the shared task NLP4IF'19[1] on fine-grained propaganda detection. As a follow up, in 2020 SemEval proposed a shared task (T11) (Da San Martino et al., 2020a) reducing the number of propaganda categories with respect to NLP4IF'19, and proposing a more restrictive evaluation scheme. To evaluate the proposed approach, we rely on these two standard benchmarks, i.e., the NLP4IF'19 and SemEval'20 datasets.

The most recent approaches for propaganda detection are based on language models that mostly involve transformer-based architectures. The approach that performed best on the NLP4IF'19 sentence-level classification task relies on the BERT architecture with hyperparameters tuning without activation function (Mapes et al., 2019). (Yoosuf and Yang, 2019) focused first on the pre-processing steps to provide more information regarding the language model along with existing propaganda techniques, then they employ the BERT architecture casting the task as a sequence labeling problem. The systems that took part in the SemEval 2020 Challenge - Task 11 represent the most recent approaches to identify propaganda techniques based on given propagandist spans. The most interesting and successful approach (Jurkiewicz et al., 2020) proposes first to extend the training data from a free text corpus as a silver dataset, and second, an ensemble model that exploits both the gold and silver datasets during the training steps to achieve the highest scores. Notice that most of the most performing recent models heavily rely on transformer-based architectures.

In this paper, we also rely on language model architectures for the detection and classification of propaganda messages, empowering them with a rich set of features we identified as pivotal in propagandist text from the computational social science literature. In particular, (Morris, 2012) discusses how emotional markers and affect at word- or phrase-level are employed in propaganda text, whilst (Ahmad et al., 2019) show that the most effective technique to extract sentiment for the propaganda detection task is to rely on lexicon-based tailored dictionaries. Recent studies (Li et al., 2017) show how to detect degrees of strength from calmness to exaggeration in press releases. Finally, (Troiano et al., 2018) focus on feature extraction of text exaggeration and show that main factors include imageability, unexpectedness, and the polarity of a sentence.

3 Propaganda Detection as a Classification Task

(Da San Martino et al., 2019) define the Fine-Grained Propaganda Detection task as two subtasks, with different granularities: *i) Sentence-Level Classification task (SLC)*, which asks to predict whether a sentence contains at least one propaganda technique, and *ii) Fragment-Level Classification task (FLC)*, which asks to identify both the spans and the type of propaganda technique.

In the following example, *"In a glaring sign of just how stupid and petty things have become in Washington these days, Manchin was invited on Fox News Tuesday morning to discuss how he was one of the only Democrats in the chamber for the State of the Union speech not looking as though Trump killed his grandma."* the span *"stupid and petty"* carries some propagandist bias, and is labeled as "Loaded Language" , *"not looking as though Trump killed his grandma"* is considered as "Exaggeration and Minimisation", and *"killed his grandma"* is "Loaded Language". According to the SLC task, the whole sentence should be classified as a propaganda message given that it contains at least one token (e.g., "*stupid and petty*") considered as such.

As previously introduced, current systems address these tasks relying on word embedding models (e.g., BERT-embedding) and standard features (e.g., PoS, name-entity, n-grams), as representations to feed various RNN architectures (Morio et al., 2020; Chernyavskiy et al., 2020). Recently, the language model BERT (Devlin et al., 2019) has been widely utilized to optimize the performances of classification tasks, but there is still room for improvement, in particular when applied to propaganda detection (Da San Martino et al., 2020a, 2019). In this work, we experiment with multiple architectures and language models to classify propagandist messages on both sentence and fragment-level. Prior to that, we conduct a detailed investigation of linguistic and argumentation features to capture propaganda strategies.

[1] https://propaganda.qcri.org/nlp4if-shared-task/

4 Feature Analysis

Propaganda strategies generally involve specific targets to be stimulated by the message. To better study such techniques from a computational point of view, we investigate a set of features that we assume to play a role in propaganda.

4.1 Persuasion

Speech style. To analyze the writing style of the messages, we apply the dictionary-based mapping tool "General Inquirer (v. 1.02)" (Gilman, 1968). It relies on a list of lexicons from 26 domains (e.g., politician speeches, consumer protests) annotated according to 182 rating categories and dimensions (e.g., valence categories and words indicating overstatement and understatement)[2]. We apply such tool on our data and we sum the ratings of each token to obtain a global score for a sentence.

Lexical complexity. Given that pre-trained language models have shown to capture lexical and semantic complexities of words, we rely on BERT (base-uncased) (Devlin et al., 2019) to extract lexical complexity features. We extract a vector of 768 dimensions per each token, then we average w.r.t. all tokens in a sentence, to obtain one vector of 768 dimensions to represent a sentence.

Concreteness. Propaganda messages tend to employ words with concrete meaning, that has more impact in conveying the intention of the message than using abstract words (Eliasberg, 1957) We rely on the concreteness lexicon (Brysbaert et al., 2013) and we sum the standardized score of each token in a sentence to obtain the global score.

Subjectivity. We rely on the subjectivity lexicon from (Wilson et al., 2005). We sum up the scores over all tokens in a sentence found in the lexicon as our extracted feature. Each word labeled as "weaksubj" is set to 0.5, and "strongsubj" to 1.

4.2 Sentiment

Sentiment labels. We use SentiWordNet 3.0 (Baccianella et al., 2010) to obtain word-level sentiment labels (positive, negative, or neutral). We sum the sentiment scores of each word in a sentence, producing a vector with 3 dimensions (i..e, pos, neg, neu) for each sentence.

Emotion labels. We extract 8 emotions (i.e., afraid, amused, angry, annoyed, don't care, happy, inspired, sad) from DepecheMood++ lexicon (Araque et al., 2019). For each word that evokes emotions in a sentence, we produce our features by summing up each set of emotions evoked by each token, then find the average by emotions. Hence, we produce 8 emotion scores for a sentence.

VAD labels. In the three-dimensional model of affect, valence ranges from unhappiness to happiness and expresses the pleasant or unpleasant feeling about something, arousal expresses the level of affective activation, ranging from sleep to excitement, and dominance reflects the level of control of the emotional state, from submissive to dominant. We use Warriner lexicon (Warriner et al., 2013) to match each word in a sentence to its VAD standardized word scores and sum up as our features.

Connotation. Propaganda can convey sentiment beyond its original meaning. Connotation lexicon (Feng et al., 2013) provides positive, negative and neutral labels of each word. We count the frequencies of the three labels evoked in each sentence.

Politeness. Politeness evokes sentiment in readers. We use a lexicon of positive and negative words from (Danescu-Niculescu-Mizil et al., 2013), then we count the frequencies of both positive and negative words found in each sentence.

4.3 Message Simplicity

To keep the message simple and picturable is one of main purposes of propaganda. We list the features we considered to extract the simplicity of message.

Exaggeration. We use imageability lexicon (Tsvetkov et al., 2014) based on picturable vocabulary which mentally leads to an exaggerating state of mind. We consider the scores of abstraction and concreteness at each word token. We then sum up the scores for all the labels found in a sentence.

Length. "The less words used, the better to understand" can be a concept to easily interpret the propagandist message. We apply two strategies: *i)* we count the average char-length, actual char-length, word length, punctuation frequency, capital-case frequency per sentence (Ferreira Cruz et al., 2019); *ii)* we apply length encoding at character-level, plus one additional dimension for non-alphabetical char count.

Pronouns. Loaded language, name calling and labelling are the most used techniques in propaganda text (Da San Martino et al., 2019), and they all make use of pronouns. We create a lexicon of 123 pronouns in English[3] and perform one-hot

[2] http://www.wjh.harvard.edu/~inquirer/homecat.htm

[3] https://www.englishclub.com/vocabulary/pronouns-

encoding of common used pronouns in English.

4.4 Argumentation

We assume that argumentation plays an important role in propaganda. To extract argumentative features representing our data, we train a supervised classifier for the task of argumentative sentence classification on the persuasive essays dataset (Stab and Gurevych, 2014). First, we cast it as a binary classification task, merging premises, claims and major claims into the *argumentative* label, as opposed to the *non-argumentative* label. Then, for the argumentation component task, we rearrange the data to binary labels where the major claims and claims labels are merged, and they are opposed to premises. To address these tasks, we build and fine-tune a BERT classifier. We use a learning rate of 1e-5 with 80/20 split of the dataset. We run our classifier 3 times at different random states. The results for the argumentative sentence classification are (macro-average) F1 0.84, precision 0.86, recall 0.82, while for the component classification they are F1 0.77, precision 0.80, recall 0.75.

To extract argumentative features from the annotations provided by our classifiers, we use BERT-based features. After fine-tuning, we freeze the hidden states of these fine-tuned BERT models. To extract the argumentative and components features from each classifier, we take the [CLS] token of each sentence from the fine-tuned BERT model.

4.5 Ablation Study

To investigate the impact of the proposed features (Section 4) for propaganda detection, we perform ablation tests by testing a supervised classifier relying on BERT + logistic regression. To the purpose, we use the NLP4IF'19 training and test sets (Da San Martino et al., 2019).

Table 1 reports on the performances obtained while integrating groups of features to the proposed model. A logistic regression model is used as a baseline. Best results are obtained when adding all the proposed features, but the argumentation ones. Argumentation features alone perform almost identical as semantic features, therefore - unexpectedly - no added value can be demonstrated.

Persuasion	Sentiment	Message simplicity	Argumentation	Logistic Regression			BERT + Featured LR		
				F1	Precision	Recall	F1	Precision	Recall
✓				0.68	0.69	0.67	0.70	0.71	0.70
	✓			0.62	0.69	0.57	0.71	0.72	0.69
		✓		0.63	0.66	0.62	0.70	0.72	0.68
			✓	0.67	0.68	0.67	0.71	0.72	0.70
✓	✓			0.68	0.70	0.67	0.71	0.71	0.69
✓		✓		0.69	0.71	0.68	0.70	0.71	0.69
✓			✓	0.69	0.70	0.68	0.70	0.71	0.70
	✓	✓		0.66	0.68	0.65	0.71	0.73	0.70
	✓		✓	0.70	0.71	0.69	0.71	0.72	0.70
		✓	✓	0.66	0.68	0.65	0.70	0.72	0.69
✓		✓	✓	0.69	0.70	0.68	0.71	0.72	0.70
✓	✓		✓	0.68	0.69	0.68	0.70	0.71	0.69
	✓	✓	✓	0.70	0.72	0.69	0.70	0.71	0.69
✓	✓	✓		0.71	0.72	0.69	**0.72**	**0.74**	0.70
✓	✓	✓	✓	0.70	0.71	0.69	0.71	0.72	0.69

Table 1: Ablation test on binary classification setting.

5 Sentence-level Classification

In the following, we describe the experiments we carried out to address the propaganda detection task at sentence level, investigating different architectures and leveraging both recent language models and the features that proved to play a role in propaganda messages. For the evaluation, we used the two available datasets for propaganda detection: *i)* the NLP4IF'19 data set (Da San Martino et al., 2019) (293 articles for training and 101 for testing); and *ii)* the data from SemEval'20 T11 (Da San Martino et al., 2020a) (371 articles for training and 75 in the development set).

5.1 Prediction Models

In the following, we first describe the baseline and the SOTA models we tested in our experiments, and then we present the three architectures we propose (underlined) integrating the propagandist features previously investigated (Section 4).

BERT. Our baseline model relies on a pre-trained bidirectional transformer language model to encode context specific sentence tokens (Devlin et al., 2019) (no fine tuning, default hyperparameters).

Fine-tuned BERT. We fine-tune the BERT model with a learning rate of 5e-5, and AdamW optimizer. We set the gradients to zero at every training batch. Then we use softmax activation to gate the output with the threshold of 0.5.

Fine-tuned T5. To fine-tune the text-to-text transformer (Raffel et al., 2020), we use T5ForConditionalGeneration approach (equally to question-answering task) where the input is a sentence (as a question), and the output is an answer (as a label). We use a learning rate of 3e-4, with max sequence length of 512.

Linear-Neuron Attention BERT. We replicate

type.php, https://www.thefreedictionary.com/List-of-pronouns.htm

Model	NLP4IF'19 Test Set			SemEval'20-T11 Dev. Set		
	F1	Precision	Recall	F1	Precision	Recall
BERT ***Baseline***	0.52	0.53	0.50	0.48	0.48	0.48
SOTA						
Fine-tuned BERT	0.58	0.63	0.53	0.61	0.63	0.60
Fine-tuned T5	0.64	0.64	0.65	0.66	0.65	0.66
Linear-Neuron Attention BERT	0.63	0.60	0.67	0.66	0.69	0.63
Multi-granularity BERT	0.61	0.60	0.62	0.65	0.68	0.63
Proposed Architecture w/ Semantic Features						
Multi-granularity + Featured BERT	0.63	0.65	0.61	0.67	**0.71**	0.64
BERT + Featured BiLSTM	0.65	**0.80**	0.55	0.65	0.75	0.58
BERT + Featured Logistic Regression	**0.72**	0.74	**0.70**	**0.68**	**0.71**	0.66
Proposed Architecture w/ Semantic Features + Argumentation Features						
BERT + Featured Logistic Regression	0.71	0.72	0.69	**0.68**	0.70	**0.67**

Table 2: Results on the Sentence-level classification (SLC) task (binary task).

the winning approach of the NLP4IF'19 shared-task (Mapes et al., 2019). It relies on BERT architecture with some modifications of hyperparameters (sentence length of 50 tokens, a learning rate of 1e-5, along with 12 attention heads and 12 transformer blocks). It uses the linear neuron without an activation function, and a threshold of 0.3 for the final prediction.

Multi-granularity BERT. This model (Da San Martino et al., 2019) relies on BERT transformer with multi-granularity network on top that has multi-classifiers for different granularity levels of text (e.g., document, paragraph, sentence, word, subword, and character-level). We replicate this model with BertAdam optimizer and ReLU activation function.

Multi-granularity + Featured BERT. We integrate the proposed features (Section 4) into (Da San Martino et al., 2019), taking only the last layer of sentence-level granularity. We feed the proposed features to a BERT classifier to obtain logits which then aggregate with the last layer of sentence-level granularity to produce predictions.

BERT + Featured BiLSTM. We build a pre-trained BERT transformer architecture, and Bi-directional Long Short-Term Memory (BiLSTM) architecture on top of the BERT model to handle the transformer architecture with our propaganda features. Firstly, the BERT model is used with learning rate of 0.001, with AdamW optimizer. We use the output of BERT that represents the [CLS] token of each sentence to combine with propaganda features as our input to the second model, the BiLSTM. For the BiLSTM model, after we feed our inputs of both [CLS] tokens combined with propaganda features, we train our BiLSTM model with hidden size of 256. Our BiLSTM hidden states consist of the last hidden states, and the last cell state for the BiLSTM layers. We then apply relu gate function, with a linear dense, then use a dropout function of 0.1. At the last layer, we use another linear dense layer to output final logits, then we apply a sigmoid activation function as final outputs.

BERT + Featured Logistic Regression. We use pre-trained BERT transformer architecture to output [CLS] token, then use this output to stack with another prediction model, i.e., logistic regression. We build a linear classifier and feed it with propaganda features as a dense layer. We then combine these logits with [CLS] tokens as the input to logistic regression on top of BERT.

5.2 Results and Error Analysis

Table 2 reports on the results obtained for the SLC task (*propaganda* vs *no propaganda*). We run each experiment 5 times and report the macro-average of all metrics. Our proposed models achieve the highest F1-score of 0.72 using BERT + Featured Logistic Regression model (persuasion, sentiment, and message simplicity features), and the highest precision-score 0.80 using BERT + Featured BiLSTM model on NLP4IF'19 dataset, outperforming the state-of-the-art models. For SemEval'20-T11, we do not have the scores from the challenge (the binary task was not proposed), but we compare the obtained results with the replicated architectures of SOTA models. Our proposed architecture obtained the best F1-score using BERT+Featured Logistic Regression. Using semantic features alone perform slightly better than combining them with argumentation features.

Table 3 reports on some misclassified examples of our best model on NLP4IF'19 dataset. Some short sentences containing strong intention keywords (e.g., "hate", "slave") have been missclas-

False Positive	False Negative
People who hate freedom will get unfettered access to the minds of 2 billion people.	The American people have a right to know, and those that engaged in this type of behavior do not have a right to hide.
You are a slave to white America.	Hitler was a very great man.

Table 3: Examples of misclassified sentences by the BERT + featured logistic regression model (NLP4IF'19)

sified as false positives. As for false negatives, the underlined fragments are labeled propaganda in gold standard, but have not been recognized as such by the classifier (mainly informative statements).

6 Fragment-level Classification

In this section, we address the task of fragment-level classification, meaning that both the spans and the type of propaganda technique should be identified in the sentences. Again, to test the proposed methods, we use both NLP4IF'19 and SemEval'20 T11 datasets. However, in the two challenges, the FLC task was evaluated according to different strategies, explained in the following.

6.1 Task 1: FLC on NLP4IF'19 Dataset

In the NLP4IF'19 dataset, 18 propaganda techniques are annotated. Prediction is expected to be at token-level. Multiple tokens can belong to the same span, and annotated with one propaganda type. Tokens that do not carry any propaganda bias are annotated as "no propaganda". To perform tokenization we run the tokenizer provided with the pretrained model of each transformer[4].

6.1.1 Prediction Models

Fine-tuned BERT (baseline). Pretrained *bert-base-uncased* model and BERT architecture (Devlin et al., 2019) with default hyperparameters. Our implementation is based on huggingface transformers. Settings: learning rate of 5e-5, padded length of 128, and batch_size of 16. We use CrossEntropyLoss as a loss function, and softmax activation function to gate output neurons.

Fine-tuned RoBERTa (baseline). We use *roberta-base* model with the same hyperparameters of loss and activation functions as the fine-tuned BERT model mentioned above.

State-of-the-art Model. The winning team applied BERT architecture for token classification (Yoosuf and Yang, 2019) on 20 labels (i.e., 18 propaganda classes, plus "background" as non propaganda, and "auxiliary" for fractions of previous tokens). They use a BERT language model, then apply softmax function, followed by a linear multi-label classification layer to output their predictions.

Transformer + CRF. We use a pre-trained model *base-uncased* with a learning rate of 3e-5 for BERT transformer, and a pre-trained model *roberta-base* with a learning rate of 2e-5 for RoBERTa transformers (hyperparameters: dropout of 0.1 with the max_length of 128, batch_size of 16 with AdamW optimizer and CrossEntropy loss function). We use CRF layer as the final layer.

6.1.2 Results and Error Analysis

Table 4 reports on the obtained performances. Evaluation is reported as the average of micro-F1 scores of 5 run-times (we use the evaluation scripts provided by (Da San Martino et al., 2019)).

The proposed architecture based on transformers with CRF output layer at different learning gradients (epochs) outperforms SOTA model on several propaganda techniques at different learning gradient ranging from 5 to 15 epochs. We also tested other architectures such as Transformer+CRF with less learning gradients (3 epochs), Transformer architecture with semantic and/or argumentation features + CRF layer by adding extracted features from sentence-level (Section 4) to each token of its sentence to a linear layer before a loss function, with no major improvements.

In Table 4, we compare the performances of the proposed models w.r.t. the SOTA (Yoosuf and Yang, 2019), on the most frequent classes. Table 5 reports examples of misclassification related to that technique. We observe that our proposed model does not capture well the articles (i.e., it, as, an, the), but rather focuses on capturing intentional word tokens (i.e., white, unbelievably, rude, wonderful, treasonous). As for future work to improve results on this specific category, we will investigate the work of (Habernal et al., 2018) according to which a dedicated strategy is needed.

6.2 Task 2: FLC on SemEval'20 T11 Dataset

In SemEval'20 T11 dataset, 14 propaganda techniques are annotated. We focus here on the task called Technique-Classification task (TC). We cast it as a sentence-span classification problem, where

[4]huggingface.co/transformers/

	NLP4IF'19											
	Average	**Appeal-Fear**	**Black-White**	**Casual-Over.**	**Doubt**	**Exag.-Min.**	**Flag-Waving**	**Loaded-L.**	**Namecalling**	**Reductio-Hit.**	**Repetition**	**Slogans**
Baseline												
Fine-tuned BERT	0.03	0.09	0.04	0.03	0.07	0.07	0.17	0.08	0.07	0.02	0.01	0.04
Fine-tuned RoBERTa	0.02	0.06	0.02	0.01	0.05	0.07	0.10	0.09	0.07	0.01	0.01	0.02
SOTA (from NLP4IF'19) (Yoosuf and Yang, 2019)	0.24	0.21	0.09	.0	0.17	0.16	0.44	0.33	0.40	.0	0.01	0.13
Proposed Architecture												
Fine-tuned BERT + CRF (5 epochs)	0.13	0.27	.0	0.04	0.08	0.20	0.59	0.26	0.28	**0.08**	0.01	0.10
Fine-tuned BERT + CRF (15 epochs)	0.11	0.25	0.02	0.04	0.07	0.28	**0.61**	0.25	0.22	0.04	**0.04**	0.13
Fine-tuned RoBERTa + CRF (5 epochs)	0.16	0.32	.0	**0.09**	0.11	0.35	0.37	**0.42**	0.37	.0	.0	0.06
Fine-tuned RoBERTa + CRF (7 epochs)	0.14	**0.40**	**0.23**	0.08	0.13	**0.37**	0.46	0.37	0.31	0.05	.0	0.17
Fine-tuned RoBERTa + CRF (10 epochs)	0.15	0.30	0.19	**0.09**	0.13	0.31	0.53	0.35	0.29	.0	0.01	0.25
Fine-tuned RoBERTa + CRF (12 epochs)	0.15	0.31	0.17	0.05	**0.19**	0.31	0.47	0.33	0.32	.0	0.03	0.14
Fine-tuned RoBERTa + CRF (15 epochs)	0.16	0.35	0.16	0.03	0.16	0.35	0.49	0.33	0.27	.0	0.01	0.24
Fine-tuned RoBERTa + CRF (5 epochs 3x-Oversampled)	0.15	0.34	0.14	0.07	0.13	0.30	0.52	0.34	0.27	0.05	0.02	**0.33**

Table 4: Experimental results on fragment-level classification on NLP4IF'19 test set. All scores are reported in micro-F1 (as in the original challenge). Scores in bold are the ones outperforming SOTA model.

Technique: NameCalling_Labeling
(a) We look forward to continuing to defend the White House's lawful actions.
(b) This is a man who follows a demonic religion, Islam, and supports textcolorredcop killers.
(c) It was not widely recognized as an anti-Jewish organization during its early years (its early literature, though, focused on "the white man" as "the white devil").

Table 5: Misclassified NameCalling_Labeling. False Negative (in red), False Positive (in blue), the correctly classified propaganda spans (underlined).

we combine logits of tokenized elements from the sentence and the span, to learn the prediction. Moreover, we add the semantic and argumentation features to enhance the performance.

As pre-processing, both the tokenized sentence and the span are used to feed the transformer (Huggingface tokenizers) as follows: *i)* we input a sentence to the tokenizer where max_length is set to 128 with padding; *ii)* we input the span provided by the propaganda span-template published by the workshop, and we set max_length value of 20 with padding. If a sentence does not contain propaganda spans, it is labeled as a "none-propaganda".

6.2.1 Prediction Models

Baseline. For all the tested architectures (BERT and RoBERTa), we use the same type of transformer model to produce logits (L) regarding the sentence-level and span-level individually. For BERT model, we use pre-trained model *bert-base-uncased*, learning rate of 5e-5, and α of 0.1. For RoBERTa, we take *roberta-base* pre-trained model with learning rate of 2e-5 with α of 0.5. All transformer models apply Adam optimizer, dropout 0.1, and CrossEntropy as a loss function per sentence ($loss_{\text{sentence}}$) and span ($loss_{\text{span}}$).

We arrange these alignment of L to calculate the average loss as joint loss ($loss_{\text{joint\_loss}}$) from each $loss$ element. Here we introduce a $loss_{\text{joint\_loss}}$ function before back-propagation:

$loss_{\text{joint\_loss}} = \alpha \times \frac{(loss_{\text{sentence}} + loss_{\text{span}})}{N_{\text{loss}}}$ where N_{loss} stands for a number of $loss$ elements that are taken into the model.

State-of-the-art Model. The winning team (Jurkiewicz et al., 2020) applies RoBERTa (*roberta-large*) with pre-trained model. The training set is increased with silver annotation based on gold annotation, and then another RoBERTa model is stacked on top to output the predictions.

Proposed Architecture. We propose another set of elements to feed the transformer by introducing the semantic and argumentation features into BiLSTM layer to produce L of proposed features, then we apply CrossEntropy as a loss function of our BiLSTM as $loss_{\text{proposed\_features}}$ then perform an addition with other $loss$ in the $loss_{\text{joint\_loss}}$ function as follows: $loss_{\text{joint\_loss}} = \alpha \times \frac{(loss_{\text{sentence}} + loss_{\text{span}} + loss_{\text{proposed\_features}})}{N_{\text{loss}}}$ Hyperparameters: 256 hidden_size, 1 hidden_layer, drop_out of 0.1 with ReLU function at the last layer before the joint loss function.

6.2.2 Results and Error Analysis

As mentioned before, the gold labels of the test set of SemEval'20 T11 are not available, but it is possible to submit a system run to the challenge website and to obtain the evaluation score. The evaluation system only accepts the exact list

	SemEval'20 T11														
	Average	**Appeal_to_Authority**	**Appeal_to_fear-prejudice**	**Bandwagon,Reductio_ad_hit.**	**Black-White-Fallacy**	**Casual-Oversimplification**	**Doubt**	**Exaggeration,Minimisation**	**Flag-Waving**	**Loaded_Language**	**Name_Calling,Labeling**	**Repetition**	**Slogans**	**Thought-terminating_Cliches**	**Whatab,Straw_Men,Red_Her.**
SOTA (from SemEval'20 T11) (Jurkiewicz et al., 2020)	0.64	0.48	0.47	0.08	0.51	0.23	0.56	0.37	0.70	0.78	0.76	0.59	0.59	0.39	0.28
Proposed Architecture + Proposed argumentation features															
Fine-tuned RoBERTa (3 epochs)	0.53	0.08	0.34	0.14	0.17	0.06	0.52	0.32	0.61	0.72	0.68	0.22	0.12	**0.42**	.0
Fine-tuned RoBERTa (5 epochs)	0.53	0.14	0.34	0.17	0.26	0.09	0.46	0.35	0.60	0.73	0.72	0.17	0.36	0.30	0.18
Fine-tuned RoBERTa (10 epochs)	0.51	0.18	0.33	0.13	0.37	0.22	0.37	0.33	0.58	0.73	0.68	0.17	0.34	0.17	0.23
Fine-tuned RoBERTa (15 epochs)	0.51	0.14	0.29	0.12	0.31	0.14	0.42	0.35	0.55	0.73	0.69	0.13	0.35	0.25	0.21
Proposed Architecture + All proposed features															
Fine-tuned RoBERTa (3 epochs)	0.54	0.16	0.38	**0.20**	0.29	0.18	0.50	0.33	0.60	0.72	0.65	0.23	0.29	0.32	0.09
Fine-tuned RoBERTa (5 epochs)	0.52	0.09	0.35	0.13	0.31	0.21	0.43	0.34	0.61	0.74	0.70	0.21	0.23	0.33	0.12
Fine-tuned RoBERTa (10 epochs)	0.51	0.09	0.31	0.17	0.37	0.28	0.36	0.35	0.54	0.73	0.70	0.19	0.38	0.14	0.19
Fine-tuned RoBERTa (15 epochs)	0.51	0.15	0.32	0.07	0.40	**0.29**	0.37	0.31	0.54	0.75	0.66	0.18	0.43	0.14	0.12

Table 6: Results on span classification on SemEval'20 T11 test set (micro-F1).

Technique: Repetition	**False Negative**
All Features	
(1) When she arrived at Jean's door, Guyger entered a unique door key with an electronic chip into the keyhole, the affidavit says.	Bandwagon,Reductio_ad_hitlerum
(2) She told the 911 operator as well as responding officers that she thought she was at her apartment when she shot Jean, according to the affidavit.	Doubt

Table 7: Misclassified Repetition spans (in red).

of span-templates of the test set (partial overlapping spans or missing spans are not accepted). Table 6 reports on the obtained results (through such evaluation system) on 5 runs as micro-F1. Scores in bold are the ones for which significant improvement can be observed w.r.t. SOTA model. RoBERTa with argumentation features can outperform results on "Thought-terminating_Cliches". Moreover, by using all semantic and argumentation features together, we can obtain some improvements over "Bandwagon,Reductio_ad_hitlerum" and "Casual-Oversimplification". Table 7 shows some examples of missclassified instances. In general, we noticed that using different training epochs help detecting different propaganda techniques. In particular, it is observed that some techniques tend to be learnt best at low training epochs (i.e., "Bandwagon,Reductio_ad_hitlerum", "Thought-terminating_Cliches"), some at high training epochs (i.e., "Casual-Oversimplification").

7 Concluding Remarks

In this paper, we proposed a new neural architecture combined with state-of-the-art language models and a rich set of linguistic features for the detection of propaganda messages in text, and their further classification along with standard propaganda techniques. Despite the boost in accuracy we achieved on two standard benchmarks for propaganda detection and classification ($\sim$ 10% of F1 scores on sentence-level classification and on specific propaganda techniques on fragment-level classification), this task remains challenging, in particular regarding the fine-grained classification of the different propaganda classes. The state-of-the-art results on this subtask require further improvement to actually embed these solutions in real-world systems.

Future work goes in this direction, with the aim to improve the performance both of the disinformation detection task and of the classification of propaganda techniques. Moreover, we are currently investigating the propaganda classes we discussed in this paper in the context of political debates, with the aim of building a fallacy detection systems that relies on the identification of propagandist messages in political speeches.

Acknowledgments

This work is partially supported by the ANSWER project PIA FSN2 n. P159564-2661789/DOS0060094 between Inria and Qwant. Moreover, this work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR- 19-P3IA-0002.

References

Siti Rohaidah Ahmad, Muhammad Zakwan Muhammad Rodzi, Nurlaila Syafira Shapiei, Nurhafizah Moziyana Mohd Yusop, and Suhaila Ismail. 2019.

A review of feature selection and sentiment analysis technique in issues of propaganda. *International Journal of Advanced Computer Science and Applications*, 10(11).

Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. 2019. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. *IEEE Transactions on Affective Computing*, pages 1–1.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing Management*, 56.

Marc Brysbaert, Amy Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46.

Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2020. Aschern at SemEval-2020 task 11: It takes three to tango: RoBERTa, CRF, and transfer learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1462–1468, Barcelona (online). International Committee for Computational Linguistics.

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 14th International Workshop on Semantic Evaluation*, SemEval 2020, Barcelona, Spain.

Giovanni Da San Martino, Stefano Cresci, Alberto Barron-Cedeno, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020b. A survey on computational propaganda detection. In *Proceedings of 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI2020)*, IJCAI-PRICAI2020, Yokohama, Japan.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

James Price Dillard and Michael Pfau. 2009. *The Persuasion Handbook: Developments in Theory and Practice*. Sage Publications, Inc.

Wladimir Gottlieb Eliasberg. 1957. Toward a philosophy of propaganda. *Jewish Social Studies*, 19(1/2):51–63.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774–1784, Sofia, Bulgaria. Association for Computational Linguistics.

André Ferreira Cruz, Gil Rocha, and Henrique Lopes Cardoso. 2019. On sentence representations for propaganda detection: From handcrafted features to word embeddings. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 107–112, Hong Kong, China. Association for Computational Linguistics.

Richard C. Gilman. 1968. The general inquirer: A computer approach to content analysis. philip j. stone , dexter c. dunphy , marshall s. smith , daniel m. ogilvie. *American Journal of Sociology*, 73(5):634–635.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Dawid Jurkiewicz, Łukasz Borchmann, Izabela Kosmala, and Filip Graliński. 2020. ApplicaAI at SemEval-2020 task 11: On RoBERTa-CRF, span CLS and whether self-training helps them. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1415–1424, Barcelona (online). International Committee for Computational Linguistics.

Haavard Koppang. 2009. Social influence by manipulation: A definition and case of propaganda. *Middle East Critique*, 18:117 – 143.

Harold Dwight Lasswell. 1938. Propaganda technique in the world war.

Yingya Li, Jieke Zhang, and Bei Yu. 2017. An NLP analysis of exaggerated claims in science news. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 106–111, Copenhagen, Denmark. Association for Computational Linguistics.

Liane Longpre, Esin Durmus, and Claire Cardie. 2019. Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Argument Mining*, pages 167–176, Florence, Italy. Association for Computational Linguistics.

Norman Mapes, Anna White, Radhika Medury, and Sumeet Dua. 2019. Divisive language and propaganda detection using multi-head attention transformers with deep learning BERT-based language models for binary classification. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 103–106, Hong Kong, China. Association for Computational Linguistics.

Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. Hitachi at SemEval-2020 task 11: An empirical study of pre-trained transformer family for propaganda detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1739–1748, Barcelona (online). International Committee for Computational Linguistics.

Travis Morris. 2012. Extracting and networking emotions in extremist propaganda. In *2012 European Intelligence and Security Informatics Conference*, pages 53–59.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304, Brussels, Belgium. Association for Computational Linguistics.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Amy Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Shehel Yoosuf and Yin Yang. 2019. Fine-grained propaganda detection with fine-tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91, Hong Kong, China. Association for Computational Linguistics.