SocCogCom at SemEval-2020 Task 11: Characterizing and Detecting Propaganda using Sentence-Level Emotional Salience Features

Gangeshwar Krishnamurthy, Raj Kumar Gupta, Yinping Yang
Institute of High Performance Computing (IHPC),
Agency for Science, Technology and Research (A*STAR), Singapore
{gangeshwark,gupta-rk,yangyp}@ihpc.a-star.edu.sg

Abstract

This paper describes a system developed for detecting propaganda techniques from news articles. We focus on examining how emotional salience features extracted from a news segment can help to characterize and predict the presence of propaganda techniques. Correlation analyses surfaced interesting patterns that, for instance, the "loaded language" and "slogan" techniques are negatively associated with valence and joy intensity but are positively associated with anger, fear and sadness intensity. In contrast, "flag waving" and "appeal to fear-prejudice" have the exact opposite pattern. Through predictive experiments, results further indicate that whereas BERT-only features obtained F1-score of 0.548, emotion intensity features and BERT hybrid features were able to obtain F1-score of 0.570, when a simple feedforward network was used as the classifier in both settings. On gold test data, our system obtained micro-averaged F1-score of 0.558 on overall detection efficacy over fourteen propaganda techniques. It performed relatively well in detecting "loaded language" (F1 = 0.772), "name calling and labeling" (F1 = 0.673), "doubt" (F1 = 0.604) and "flag waving" (F1 = 0.543).

1 Introduction

Propaganda is studied in a wide range of social sciences disciplines, including social psychology, political science, media and mass communication, as well as advertising and marketing (Davison, 1971; Taylor, 2002; Balfour, 1979; McGarry, 1958). As Jowett and O'Donnell (2018) put it, propaganda is a "deliberate, systematic attempt to shape perceptions, manipulate cognitions, and direct behavior to achieve a response that furthers the desired intent of the propagandist". To achieve the agenda, propagandists may use various influence techniques such as loaded emotive language and flag waving. Such techniques are centered on influencing the audiences' opinions and behaviors through psychological and rhetorical tricks in order to reach its purpose, such as promoting a particular politician or product in political or marketing campaigns.

The ability to automatically detect propaganda has important societal implications. For news management, propaganda detection may help publishers to quickly identify news articles that may be subjected to propagandistic characteristics that severely deviate from journalism principles. For the general public, such tools may raise awareness for social media users to stay alert of potential propagandistic content, which often may leverage non-obvious psychological tricks, and potentially mitigate the propagation of such content.

We participated in Task 11 on the detection of propaganda techniques in news articles (Da San Martino et al., 2020a), in particular the Technique Classification task (task TC), a multi-class classification task that aims to classify each identified text segment with the existence of a collection of fourteen propaganda techniques (Da San Martino et al., 2020a). Appendix A provides a summary and a distribution analysis on this task. This text segment-based ground truth data presents an advancement to this line of study with an ability to allow an algorithm to not only identify the existence of propaganda, but also to name the specific techniques.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 1793–1801
Barcelona, Spain (Online), December 12, 2020.

Our approach focuses on exploring the value of sentence-level emotional salience features to characterize propaganda techniques. From the definitions of the fourteen techniques used in Da San Martino et al. (2019)'s original paper, at least six techniques conceptually involve emotion-associated properties, including strong emotional connotations or emotional appeal. Consider the following examples in Da San Martino et al. (2019):

1. "stop those refugees; they are terrorists" ["appeal to fear-prejudice"]
2. "the best of the best" ["exaggeration,minimisation"]
3. "Entering this war will make us have a better future in our country" ["flag waving"]
4. "a lone lawmaker's childish shouting" ["loaded language"]
5. "Republican congressweasels" ["name calling,labeling"]
6. "Make America great again!" ["slogans"]

To extract the sentence-level emotional salience features in the news segments, we leveraged Gupta and Yang (2018)'s work which trains a collection of SVM-based algorithms, named as CrystalFeel[1], which detects the intensities of five emotion dimensions present in a given text message, including the sentiment valence, joy, anger, fear and sadness (Gupta and Yang, 2018). As the key purpose of propaganda is to influence or persuade the audiences, our main design hypothesis is that sentence-level emotional salience features will help to characterize a few most commonly used propaganda techniques that involve a degree of emotional connotations in their language manifestations. Table 1 illustrates the emotion intensity scores derived on six propaganda examples used in (Da San Martino et al., 2019).

Text segment example	**Detected Emotion Intensity Scores (Gupta and Yang, 2018)**				
	Valence Intensity	Joy Intensity	Anger Intensity	Fear Intensity	Sadness Intensity
"stop those refugees; they are terrorists" ["appeal to fear-prejudice"]	0.305	0.123	**0.622**	**0.551**	**0.483**
"the best of the best" ["exaggeration,minimisation"]	0.653	**0.520**	0.208	0.183	0.267
"Entering this war will make us have a better future in our country" ["flag waving"]	0.563	0.344	0.332	**0.406**	0.374
"a lone lawmaker's childish shouting" ["loaded language"]	0.323	0.126	**0.516**	0.487	**0.520**
"Republican congressweasels" ["name calling,labeling"]	0.456	0.216	0.367	0.371	**0.418**
"Make America great again!" ["slogans"]	0.672	**0.592**	0.264	0.201	0.286

Table 1: Emotional salience extracted from six examples of emotion elicitation-related propaganda techniques.

2 Related Work

Propaganda analysis, system and detection. Computational approach to propaganda detection is relatively a new topic (see Da San Martino et al. (2020b) for a review). Da San Martino et al. (2019) formulate the problem of the detection of specific propaganda techniques which is directly related to this paper. Barrón-Cedeno et al. (2019) and Da San Martino et al. (2020c) showed how their Proppy and Prta systems can support users to unmask and analyze propaganda in the news with interactive interfaces.

The closest to our work is the analytic study by Rashkin et al. (2017). Rashkin et al. (2017) compared the linguistic patterns, e.g., psycholinguistic features from LIWC, sentiments, hedging words and intensifying

[1] CrystalFeel is available at: `http://www.crystalfeel.socialanalyticsplus.net/`

words, across four categories of news: propaganda, trusted news, hoax, or satire. They found interesting linguistic differences in the three "fake" news categories vis-à-vis trusted news, though the predictive experiments showed that LIWC do not improve over the neural model in terms of predictive model performance, probably due to that "some of this lexical information is perhaps redundant to what the model was already learning from the text" (Rashkin et al., 2017). What Rashkin et al. (2017) focused on are word-level or lexical linguistic features. None of the existing work has explored the value of sentence-level sentiment and emotion intensity features in the context of propaganda detection.

Emotion intensity detection and analysis. Classic sentiment analysis typically provides classification results for discrete sentiment (e.g., positive, negative, neutral) and emotion classification analysis (e.g., happy vs. no happy, sad vs. no sad). Emotion intensity analysis is relatively a new development in the context of predicting the degree or intensity of the underlying emotional valence and dimensions in text messages such as tweets (Mohammad and Bravo-Marquez, 2017; Mohammad et al., 2018). Gupta and Yang (2018) trained CrystalFeel with features derived from parts-of-speech, n-grams, word embedding, multiple existing affective lexicons, and an in-house developed emotion intensity lexicon to predict the degree of the intensity associated with fear, anger, sadness, and joy in the tweets. Its predicted sentiment intensity had arrived a Pearson correlation coefficient (r) value of .816 on sentiment intensity with out-of-training sample of human annotations, and of .708, .740, .700 and .720 on emotion intensities in predicting joy, anger, fear and sadness (Gupta and Yang, 2018).

3 Correlation Analysis

To gain an exploratory understanding on the usefulness of the emotional salience features, we performed bivariate correlation analyses between each of the propaganda ground truth labels for the 1,043 text segments in the development set and the emotion intensity scores derived from CrystalFeel. Table 2 reports the correlation results. Non-parametric measure of Kendall's τ was used for the correlation test because the ground truth is a dichotomous variable (1 indicates the propaganda technique is present in the text; 0 indicates otherwise).

Propaganda technique	**Kendall's τ coefficient**				
	Valence Intensity	Joy Intensity	Anger Intensity	Fear Intensity	Sadness Intensity
"appeal to authority"	0.041	0.011	-0.002	-0.010	**-0.066****
"appeal to fear-prejudice"	**-0.064***	**-0.074****	**0.073****	**0.160****	**0.057***
"bandwagon,reductio_ad_hitlerum"	0.042	0.028	0.001	0.030	**-0.060***
"black-and-white fallacy"	0.019	-0.019	-0.039	-0.038	-0.041
"causal oversimplification"	-0.004	-0.035	0.046	0.022	-0.012
"doubt"	**-0.079****	**-0.118****	**0.071****	0.045	-0.023
"exaggeration,minimisation"	**0.051***	**0.086****	-0.012	-0.014	0.010
"flag waving"	**0.182****	**0.099****	**-0.179****	**-0.168****	**-0.167****
"loaded language"	**-0.224****	**-0.089****	**0.181****	**0.140****	**0.243****
"name calling,labeling"	**0.066****	0.032	-0.039	0.010	**-0.062***
"repetition"	0.032	0.029	**-0.066****	**-0.081****	0.006
"slogans"	**0.089****	**0.063***	**-0.110****	**-0.136****	**-0.088****
"thought-terminating cliches"	**0.063***	**0.065***	**-0.062***	**-0.050***	-0.040
"whataboutism,straw men,red h."	0.015	-0.009	0.011	-0.005	**-0.068****

Table 2: Correlation between the ground truth labels and emotion intensities in the development set (** indicates p value < 0.01; * indicates p value < 0.05)

Results indicate interesting patterns: "loaded language", "flag waving", "slogans", "appeal to fear-prejudice", and a total of twelve propaganda techniques are significantly correlated with at least one of the emotion intensity scores ($* * p < 0.01, *p < 0.05, n = 1,043$).

Most notably, "loaded language" is negatively correlated with valence intensity ($\tau = -0.224$**) and joy intensity ($\tau = -0.089$**), but is positively correlated with anger intensity ($\tau = 0.181$**), fear intensity ($\tau = 0.140$**) and sadness intensity ($\tau = 0.243$**). The "slogans" technique has a similar correlational pattern.

In contrast, "flag waving" has the exact opposite pattern, where it is positively correlated with valence intensity ($\tau = 0.182$**) and joy intensity ($\tau = 0.099$**), but is negatively correlated with anger intensity ($\tau = -0.179$**), fear intensity ($\tau = -0.168$**) and sadness intensity ($\tau = -0.167$**). The "appeal to fear-prejudice" technique has a similar pattern.

Two propaganda techniques, "black-and-white fallacy" and "causal oversimplification", are not found to be correlated with any emotion intensity scores. Noted that these techniques also have less occurrences in the dataset (gold labels $< 3\%$; see Appendix A) and are not conceptually associated with emotional connotation or emotional appeal by definition.

The results showed initial support to our main design intuition, which also implies that the emotional saliences based system is likely to be effective in detecting emotions-associated (but not non-emotions-associated) propaganda techniques.

4 System Overview

Following the the exploratory analysis, we proceed to design a predictive system named as "SocCogCom". Our SocCogCom system is designed to determine the specific propaganda technique used in a given text segment from news articles. The possible techniques are based on a range of fourteen possibilities which are defined in the official SemEval 2020 Task 11 description paper (Da San Martino et al., 2020a). Figure 1 depicts the system architecture.

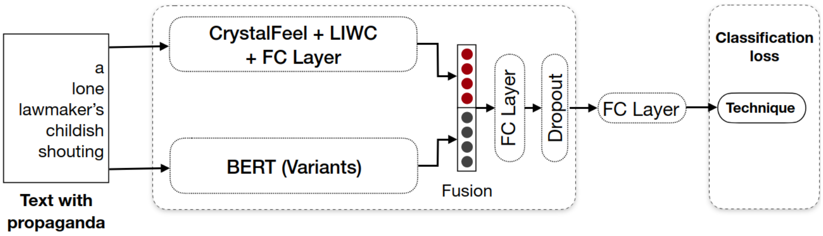

Figure 1: Architecture diagram of the proposed model

Input Layer: A training example consists of the span of text that contains a propaganda technique: $x \in \mathbb{R}^n$ and a propaganda technique label associated with the text: $y \in \{\texttt{14 techniques}\}$. x is a sequence of words represented in the order of appearance in the vocabulary.

Features Extraction: For every input text segment, our system extracts the following features:

1. **BERT features**[2]**:** Sentence-level embeddings (b_f) (Devlin et al., 2018). This is a set of pre-trained sentence-level embedding features with a total of 1,024 dimensions.
2. **CrystalFeel features**[3]**:** Sentence-level emotional saliences features (Gupta and Yang, 2018) (c_f). The extracted features for each text segment include five dimensions of emotion intensity features.
3. **LIWC Features:** Word-level psycholinguistic features from the LIWC lexicon[4] (Pennebaker et al., 2015) (l_f). We obtained 73 extracted features that represent psycholinguistic characteristics of a piece of text that may involve a propaganda technique.

Fusion Layer: CrystalFeel and LIWC features obtained above are first concatenated and a dense layer is applied over the concatenated vector to obtain a feature vector, h_f, of dimension $d_h = 50$. This is done in order to project the features extracted from CrystalFeel and LIWC to a similar latent space as that of BERT features. Here, the extracted features, b_f and h_f, are simply concatenated to form the

[2] https://storage.googleapis.com/bert_models/2018_10_18/uncased_L-24_H-1024_A16.zip
[3] http://www.crystalfeel.socialanalyticsplus.net
[4] https://liwc.wpengine.com

representation: $z_f = [b_f; h_f]$ of dimension $d_{in} = 1074$. A dense layer with 256 dimensions is applied over z_f. After this, the final representation - o_f is obtained by applying a dropout layer (Srivastava et al., 2014) with dropout rate of 0.5.

Output Layer: The system employs a fully-connected layer with softmax activation where the fused representation o_f is fed.

Loss function: The categorical cross-entropy is used to calculate the loss. We minimize the loss with an optimizer. The function that is optimized is as follows:

$$E_{\text{crossentropy}} = -\sum_{n=1}^{N}\sum_{k=1}^{c} y_k^n \log \hat{y}_k^n \tag{1}$$

where N is the total number of samples and c is the number of classes (in our case it is 14). y_k^n is the actual label of the k^{th} class of the n^{th} sample and $\hat{y}_k^n$ is the prediction corresponding to the k^{th} class of the n^{th} sample.

5 Features Experiments and Results

For data pre-processing, we used Keras Tokenizer to split the text into word tokens. The sentences are cleaned to remove unwanted characters and double spaces are replaced with single space.

We conducted the features experiments using the standard training and development datasets provided in the official TC task, based on the system set up described in Section 3. Hyper-parameters are tuned using a held out validation data: 10% of the training data. To optimize the parameters, we use Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of $1e^{-4}$. The experiments results are presented in Table 3.

Model + Features	**Micro-averaged F1 score**
Logistic Regression	0.2520
BERT Only	0.5485
CrystalFeel Only	0.5234
BERT + CrystalFeel	**0.5701**
BERT + CrystalFeel + LIWC	0.5626
AlBERT + CrystalFeel	0.5588
BERT + CrystalFeel + Context	**0.5824**

Table 3: Feature experiments results on development set.

First, we evaluated the effects of using BERT features and emotional salience features from CrystalFeel outputs alone. BERT only obtained micro-averaged F1 score of 0.5484, showing strong performance in comparison to a simple baseline using logistic regression. CrystalFeel features achieved 0.5234, which shows fair performance given this is a low-dimensional features set. When combined, BERT and CrystalFeel features achieved better results, with micro-averaged F1 score of 0.5701 than the individual settings.

We also assessed classic word-level psycholinguistics features based on LIWC lexicons. The BERT + LIWC condition didn't converge, as the loss didn't decrease and was fluctuating a lot. Adding LIWC onto the hybrid BERT + CrystalFeel, i.e., the BERT + CrystalFeel + LIWC condition, obtained micro-averaged F1 score of 0.5626, indicating that additional word-level psycholinguistics features do not appear to improve over the BERT + CrystalFeel condition. We tested AlBERT + CrystalFeel too, and they did not match the results obtained from BERT + CrystalFeel condition.

Based on the experiment results, we used the best-performing hybrid features sets (BERT + CrystalFeel) for our system results submission for the gold test set.

After we submitted our results, we experimented a new condition where context features were added to the BERT + CrystalFeel condition. For context, we extracted features using 3 words before and after the target text segment. The results showed improvement (micro-averaged F1 score = 0.5824).

6 Results on Gold Test Set

Overall, on gold test set, the results released from the task organizers suggested that our system achieved micro-averaged F1 score of 0.558 across the fourteen propaganda techniques. Table 4 shows the detailed results showing F1 scores of our system for each propaganda technique.

Propaganda techniques	**F1 (development set)**		**F1 (gold test set)**	
	Baseline	SocCogCom	Baseline	SocCogCom
"appeal to authority"	0.000	0.000	0.000	0.286
"appeal to fear-prejudice"	0.094	0.329	0.037	0.316
"bandwagon,reductio_ad_hitlerum"	0.000	0.571	0.000	0.098
"black-and-white fallacy"	0.000	0.214	0.000	0.265
"causal oversimplification"	0.072	0.286	0.116	0.063
"doubt"	0.192	**0.540**	0.291	**0.604**
"exaggeration,minimisation"	0.117	0.457	0.144	0.349
"flag waving"	0.083	**0.771**	0.062	**0.543**
"loaded language"	0.406	**0.706**	0.465	**0.722**
"name calling,labeling"	0.000	**0.644**	0.000	**0.673**
"repetition"	0.385	0.318	0.193	0.189
"slogans"	0.000	0.302	0.000	0.409
"thought-terminating cliches"	0.000	0.129	0.000	0.235
"whataboutism,straw men,red herring"	0.000	0.064	0.000	0.100
Micro-averaged F1	0.265	**0.570**	0.252	**0.558**

Table 4: Predictive results of our system for each propaganda techniques.

The results suggested that using relatively parsimonious features, BERT and CrystalFeel emotional salience features, our system performed reasonably well (F1 score $>$ 0.5) in detecting "loaded language" (F1 = 0.772), "name calling and labeling" (0.673), "doubt" (0.604) and "flag waving" (0.543). Meanwhile, our system struggled in detecting non-emotion associated techniques (which also happen to have imbalanced distributions), such as "causal oversimplification" (F1 = 0.063), "bandwagon,reductio_ad_hitlerum" (F1 = 0.098), and "whataboutism, straw men, red herring" (F1 = 0.100). The results also support with our design intuition that the sentiment and emotion intensities features help to detect propaganda techniques which are manifested in their emotional salience in the text segment.

7 Conclusion

Propaganda is primarily information that is used to advance an agenda through influence techniques. Our work is motivated to explore the value of emotional salience features in predicting emotion-related propaganda techniques. In our experiments, we found that emotional salience features using CrystalFeel emotion intensity scores can improve over BERT only features, when a simple feedforward neural network is used in both experiment settings. Results and analysis on gold test dataset show that our approach performed reasonably well (F1 $>$ 0.5) in detecting "loaded language", "name calling and labeling", "doubt" and "flag waving" techniques. As these are also most frequently used techniques, our system has a potential value to facilitate publishers and general public to be alerted with these common techniques. The system scripts are released at `https://github.com/gangeshwark/PropagandaNews`.

Acknowledgements

This work is supported by Agency for Science, Technology and Research (A*STAR) under its A*ccelerate Gap Fund (ETPL/18-GAP051-R20A).

References

Michael Leonard Graham Balfour. 1979. *Propaganda in war, 1939-1945: organisations, policies, and publics, in Britain and Germany*. Taylor & Francis.

Alberto Barrón-Cedeno, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9847–9848.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. 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 2019*, EMNLP-IJCNLP 2019, Hong Kong, China, November.

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, December.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Seunghak Yu, R 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), Yokohama, Japan, July 11-17, 2020.*, IJCAI-PRICAI 2020, Yokohama, Japan, July.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020c. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics (ACL 2020), Seattle, Washington, USA, July 5-10, 2020*, ACL 2020, Seattle, Washington, USA, July.

W Phillips Davison. 1971. Some trends in international propaganda. *The Annals of the American academy of political and social science*, 398(1):1–13.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Raj Kumar Gupta and Yinping Yang. 2018. Crystalfeel at semeval-2018 task 1: Understanding and detecting emotion intensity using affective lexicons. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 256–263.

Garth S Jowett and Victoria O'Donnell. 2018. *Propaganda & persuasion*. Sage Publications.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Edmund D McGarry. 1958. The propaganda function in marketing. *Journal of Marketing*, 23(2):131–139.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 65–77.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January.

Philip M Taylor. 2002. Strategic communications or democratic propaganda? *Journalism studies*, 3(3):437–441.

Appendix A: Propaganda techniques, definitions and the gold labels distributions

The task TC aims to classify each given text segment for each of the fourteen propaganda techniques. The input data is a text segment marked with superscripts indicating the start and the end characters that are supposed to be classified. For each text segment, the output should be a classification result that marks the existence of one or more of the fourteen propaganda techniques.

Most text segments have one corresponding technique, but some may have more than one techniques. For example, text segment "She's a big fan of torture" from article (id = 738361208, span_start = 2396, span_end = 2422) has two gold labels "exaggeration,minimisation" and "name calling,labeling".

It is useful to note that the class distribution for most of the techniques is highly imbalanced: 11 of the 14 techniques have less than 10% occurrence over the total 1,043 text segments (see table below for details). Some techniques such as "bandwagon,reductio_ad_hitlerum" (0.5%), "appeal to authority" (1.3%), "thought-terminating cliches" (1.6%) have less than 2% occurrence. "loaded language" has most occurrence (30.7%), followed by "name calling,labeling" (17.5%) and "repetition" (12.6%).

Propaganda techniques	Definitions (Da San Martino et al., 2019)	Gold labels (count)	Gold labels (%)
"appeal to authority"	Stating that a claim is true simply because a valid authority/expert on the issue supports it, without any other supporting evidence (Goodwin, 2011)	14	1.3%
"appeal to fear-prejudice"*	Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative, possibly based on preconceived judgments	44	4.2%
"bandwagon, reductio_ad_hitlerum"	**Bandwagon**: Attempting to persuade the target audience to join in and take the course of action because "everyone else is taking the same action" (Hobbs and Mcgee, 2008). **Reductio_ad_hitlerum**: Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated in contempt by the target audience. It can refer to any person or concept with a negative connotation (Teninbaum, 2009)	5	0.5%
"black-and-white fallacy"	Presenting two alternative options as the only possibilities, when in fact more possibilities exist (Torok, 2015). As an extreme case, telling the audience exactly what actions to take, eliminating any other possible choice (dictatorship)	22	2.1%
"causal oversimplification"	Assuming one cause when there are multiple causes behind an issue.	18	1.7%
"doubt"	Questioning the credibility of someone or something	66	6.3%

“exaggeration, minimisation”*	Either representing something in an excessive manner: making things larger, better, worse or making something seem less important or smaller than it actually is (Jowett and O’Donnell, 2018)	68	6.5%
“flag waving”*	Playing on strong national feeling (or with respect to a group, e.g., race, gender, political preference) to justify or promote an action or idea (Hobbs and Mcgee, 2008)	86	8.2%
“loaded language”*	Using words/phrases with strong emotional implications (positive or negative) to influence an audience (Weston, 2018, p. 6)	320	30.7%
“name calling,labeling”*	Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or otherwise loves or praises (Miller, 1939)	183	17.5%
“repetition”	Repeating the same message over and over again, so that the audience will eventually accept it (Torok, 2015; Miller, 1939)	131	12.6%
“slogans”*	A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals (Dan, 2015)	40	3.8%
“thought-terminating cliches”	Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are typically short, generic sentences that offer seemingly simple answers to complex questions or that distract attention away from other lines of thought (Hunter, 2015, p. 78).	17	1.6%
“whataboutism, straw men, red herring”	**Whataboutism**: Discredit an opponent’s position by charging them with hypocrisy without directly disproving their argument (Richter, 2017). **Straw Men**: When an opponent’s proposition is substituted with a similar one which is then refuted in place of the original (Walton, 1996). **Red Herring**: Introducing irrelevant material to the issue being discussed, so that everyones attention is diverted away from the points made (Weston, 2018, p. 78)	29	2.8%

Table 5: Propaganda techniques, definitions and the gold labels distributions in the development set (total n=1,043 text segments)

* These six techniques are associated with emotions by their respective definitions.