SemEval-2020 Task 11:
Detection of Propaganda Techniques in News Articles

Giovanni Da San Martino[1]**, Alberto Barrón-Cedeño**[2]**,**
Henning Wachsmuth[3]**, Rostislav Petrov**[4] **and Preslav Nakov**[1]
[1]Qatar Computing Research Institute, HBKU, Qatar   [2]Università di Bologna, Forlì, Italy
[3]Paderborn University, Paderborn, Germany   [4]A Data Pro, Sofia, Bulgaria
{gmartino, pnakov}@hbku.edu.qa   a.barron@unibo.it
henningw@upb.de   rostislav.petrov@adata.pro

Abstract

We present the results and the main findings of SemEval-2020 Task 11 on Detection of Propaganda Techniques in News Articles. The task featured two subtasks. Subtask SI is about *Span Identification*: given a plain-text document, spot the specific text fragments containing propaganda. Subtask TC is about *Technique Classification*: given a specific text fragment, in the context of a full document, determine the propaganda technique it uses, choosing from an inventory of 14 possible propaganda techniques. The task attracted a large number of participants: **250** teams signed up to participate and **44** made a submission on the test set. In this paper, we present the task, analyze the results, and discuss the system submissions and the methods they used. For both subtasks, the best systems used pre-trained Transformers and ensembles.

1   Introduction

Propaganda aims at influencing people's mindset with the purpose of advancing a specific agenda. It can hide in news published by both established and non-established outlets, and, in the Internet era, it has the potential of reaching very large audiences (Muller, 2018; Tardáguila et al., 2018; Glowacki et al., 2018). Propaganda is most successful when it goes unnoticed by the reader, and it often takes some training for people to be able to spot it. The task is way more difficult for inexperienced users, and the volume of text produced on a daily basis makes it difficult for experts to cope with it manually. With the recent interest in "fake news", the detection of propaganda or highly biased texts has emerged as an active research area. However, most previous work has performed analysis at the document level only (Rashkin et al., 2017; Barrón-Cedeño et al., 2019a) or has analyzed the general patterns of online propaganda (Garimella et al., 2015; Chatfield et al., 2015).

SemEval-2020 Task 11 offers a different perspective: a fine-grained analysis of the text that complements existing approaches and can, in principle, be combined with them. Propaganda in text (and in other channels) is conveyed through the use of diverse propaganda techniques (Miller, 1939), which range from leveraging on the emotions of the audience —such as using *loaded language* or *appeals to fear*— to using logical fallacies —such as *straw men* (misrepresenting someone's opinion), hidden *ad-hominem fallacies*, and *red herring* (presenting irrelevant data). Some of these techniques have been studied in tasks such as hate speech detection (Gao et al., 2017) and computational argumentation (Habernal et al., 2018).

Figure 1 shows the fine-grained propaganda identification pipeline, including the two targeted subtasks. Our goal is to facilitate the development of models capable of spotting text fragments where propaganda techniques are used. The task featured the following subtasks:

Subtask SI *(Span Identification)*: Given a plain-text document, identify those specific fragments that contain at least one propaganda technique. (This is a binary sequence tagging task.)

Subtask TC *(Technique Classification)*: Given a propagandistic text snippet and its document context, identify the propaganda technique used in that snippet. (This is a multi-class classification problem.)
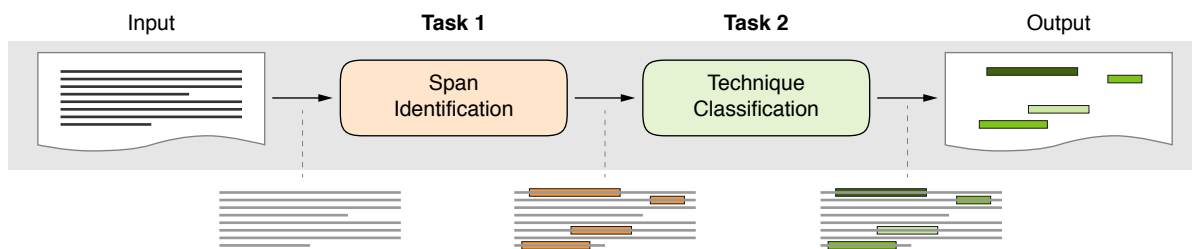
Figure 1: The full propaganda identification pipeline, including the two subtasks: Span Identification and Technique Classification.

A total of 250 teams registered for the task, 44 of them made an official submission on the test set (66 submissions for both subtasks), and 32 of the participating teams submitted a system description paper.

The rest of the paper is organized as follows. Section 2 introduces the propaganda techniques we considered in this shared task. Section 3 describes the organization of the task, the corpus and the evaluation measures. An overview of the participating systems is given in Section 4, while Section 5 discusses the evaluation results. Related work is described in Section 6. Finally, Section 7 draws some conclusions, and discusses some directions for future work.

2   Propaganda and its Techniques

Propaganda comes in many forms, but it can be recognized by its persuasive function, sizable target audience, the representation of a specific group's agenda, and the use of faulty reasoning and/or emotional appeals (Miller, 1939). The term *propaganda* was coined in the 17th century, and initially referred to the propagation of the Catholic faith in the New World (Jowett and O'Donnell, 2012a, p. 2). It soon took a pejorative connotation, as its meaning was extended to also mean opposition to Protestantism. In more recent times, the Institute for Propaganda Analysis (Ins, 1938) proposed the following definition:

> **Propaganda.** *Expression of opinion or action by individuals or groups deliberately designed to influence opinions or actions of other individuals or groups with reference to predetermined ends.*

Recently, Bolsover and Howard (2017) dug deeper into this definition identifying its two key elements: (*i*) trying to influence opinion, and (*ii*) doing so on purpose.

Propaganda is a broad concept, which runs short for the aim of annotating specific propaganda fragments. Yet, influencing opinions is achieved through a series of rhetorical and psychological techniques, and in the present task, we focus on identifying the use of such techniques in text. Whereas the definition of propaganda is widely accepted in the literature, the set of propaganda techniques considered, and to some extent their definition, differ between different scholars (Torok, 2015). For instance, Miller (1939) considers seven propaganda techniques, whereas Weston (2000) lists at least 24 techniques, and the Wikipedia article on the topic includes 67.[1] Below, we describe the propaganda techniques we consider in the task: a curated list of fourteen techniques derived from the aforementioned studies. We only include techniques that can be found in journalistic articles and can be judged intrinsically, without the need to retrieve supporting information from external resources. For example, we do not include techniques such as *card stacking* (Jowett and O'Donnell, 2012b, p. 237), since it would require comparing multiple sources. Note that our list of techniques was initially longer than fourteen, but we decided, after the annotation phase, to merge similar techniques with very low frequency in the corpus. A more detailed list with definitions and examples is available online[2] and in Appendix C, and examples are shown in Table 1.

1. Loaded language.   Using specific words and phrases with strong emotional implications (either positive or negative) to influence an audience (Weston, 2000, p. 6).

2. Name calling or labeling.   Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or loves, praises (Miller, 1939).

---
[1] https://en.wikipedia.org/wiki/Propaganda_techniques; last visit February 2019.
[2] http://propaganda.qcri.org/annotations/definitions.html

#	Technique	Snippet
1	Loaded language	**Outrage** as Donald Trump suggests injecting disinfectant to kill virus.
2	Name calling, labeling	WHO: Coronavirus emergency is '**Public Enemy Number 1**'
3	Repetition	I still have a **dream**. It is a **dream** deeply rooted in the American **dream**. I have a **dream** that one day . . .
4	Exaggeration, minimization	Coronavirus **'risk to the American people remains very low'**, Trump said.
5	Doubt	**Can the same be said for the Obama Administration**?
6	Appeal to fear/prejudice	**A dark, impenetrable and "irreversible" winter of persecution of the faithful by their own shepherds will fall**.
7	Flag-waving	Mueller attempts **to stop the will of We the People**!!! It's time to jail Mueller.
8	Causal oversimplification	**If France had not have declared war on Germany then World War II would have never happened.**
9	Slogans	**"BUILD THE WALL!"** Trump tweeted.
10	Appeal to authority	**Monsignor Jean-Franois Lantheaume, who served as first Counsellor of the Nunciature in Washington, confirmed that "Vigan said the truth. That's all."**
11	Black-and-white fallacy	Francis said these words: "**Everyone is guilty for the good he could have done and did not do . . . If we do not oppose evil, we tacitly feed it**."
12	Thought-terminating cliché	**I do not really see any problems there.** Marx is the President.
13	Whataboutism	President Trump —**who himself avoided national military service** in the 1960's— keeps beating the war drums over North Korea.
	Straw man	"Take it seriously, but with a large grain of salt." **Which is just Allen's more nuanced way of saying: "Don't believe it.**"
	Red herring	**"You may claim that the death penalty is an ineffective deterrent against crime – but what about the victims of crime? How do you think surviving family members feel when they see the man who murdered their son kept in prison at their expense? Is it right that they should pay for their son's murderer to be fed and housed?"**
14	Bandwagon	He tweeted, "**EU no longer considers #Hamas a terrorist group. Time for US to do same.**"
	Reductio ad hitlerum	"Vichy journalism," a term which now fits so much of the mainstream media. **It collaborates in the same way that the Vichy government in France collaborated with the Nazis.**

Table 1: The 14 propaganda techniques with examples, where the propaganda span is shown in bold.

3. Repetition. Repeating the same message over and over again, so that the audience will eventually accept it (Torok, 2015; Miller, 1939).

4. Exaggeration or minimization. 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, 2012b, pag. 303).

5. Doubt. Questioning the credibility of someone or something.

6. 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.

7. Flag-waving. Playing on strong national feeling (or with respect to any group, e.g., race, gender, political preference) to justify or promote an action or idea (Hobbs and Mcgee, 2008).

8. Causal oversimplification. Assuming a single cause or reason when there are multiple causes behind an issue. We include in the definition also *scapegoating*, e.g., transferring the blame to one person or group of people without investigating the complexities of an issue.

9. Slogans. A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals (Dan, 2015).

10. Appeal to authority. Stating that a claim is true simply because a valid authority or expert on the issue supports it, without any other supporting evidence (Goodwin, 2011). We include in this technique the special case in which the reference is not an authority or an expert, although it is referred to as *testimonial* in the literature (Jowett and O'Donnell, 2012b, pag. 237).

11. Black-and-white fallacy, dictatorship. Presenting two alternative options as the only possibilities, when in fact more possibilities exist (Torok, 2015). *Dictatorship* is an extreme case: telling the audience exactly what actions to take, eliminating any other possible choice.

12. Thought-terminating cliché. Words or phrases that discourage critical thought and meaningful discussion on a 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).

13. Whataboutism, straw man, red herring. Here we merge together three techniques, which are relatively rare taken individually: (*i*) *Whataboutism:* Discredit an opponent's position by charging them with hypocrisy without directly disproving their argument (Richter, 2017). (*ii*) *Straw man:* When an opponent's proposition is substituted with a similar one, which is then refuted instead of the original (Walton, 2013). Weston (2000, p. 78) specifies the characteristics of the substituted proposition: "caricaturing an opposing view so that it is easy to refute". (*iii*) *Red herring:* Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made (Weston, 2000, p. 78).

14. Bandwagon, reductio ad hitlerum. Here we merge together two techniques, which are relatively rare taken individually: (*i*) *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). (*ii*) *Reductio ad hitlerum:* Persuading an audience to disapprove an action or idea by suggesting that it 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).

We provided the definitions, together with some examples and an annotation schema, to professional annotators, and we asked them to manually annotate selected news articles. The annotators worked with an earlier version of the annotation schema, which contained eighteen techniques (Da San Martino et al., 2019b). As some of these techniques were quite rare, which could cause data sparseness issues for the participating systems, for the purpose of the present SemEval-2020 task 11, we decided to get rid of the four rarest techniques. In particular, we merged *Red herring* and *Straw man* with *Whataboutism* (under technique 13), since all three techniques are trying to divert the attention to an irrelevant topic and away from the actual argument. We further merged *Bandwagon* with *Reductio ad hitlerum* (under technique 14), since they both try to approve/disapprove an action or idea by pointing to what is popular/unpopular. Finally, we dropped one rare technique, which we could not easily merge with other techniques: *Obfuscation, Intentional vagueness, Confusion*. As a result, we reduced the eighteen original propaganda techniques to fourteen.

3 Evaluation Framework

The SemEval 2020 Task 11 evaluation framework consists of the PTC-SemEval20 corpus and the evaluation measures for both the span identification and the technique classification subtasks. We describe the organization of the task in Section 3.3; here, we focus on the dataset, the evaluation measure, and the organization setup.

3.1 The PTC-SemEval20 Corpus

In order to build the PTC-SemEval20 corpus, we retrieved a sample of news articles from the period starting in mid-2017 and ending in early 2019. We selected 13 propaganda and 36 non-propaganda news media outlets, as labeled by Media Bias/Fact Check,[3] and we retrieved articles from these sources. We deduplicated the articles on the basis of word n-gram matching (Barrón-Cedeño and Rosso, 2009), and we discarded faulty entries, e.g., empty entries from blocking websites.

The annotation job consisted of both spotting a propaganda snippet and, at the same time, labeling it with a specific propaganda technique. The annotation guidelines are shown in Appendix C; they are also available online.[4] We ran the annotation in two phases: (*i*) two annotators labeled an article independently, and (*ii*) the same two annotators gathered together with a *consolidator* to discuss dubious instances, e.g., spotted only by one annotator, boundary discrepancies, label mismatch, etc. This protocol was designed after a pilot annotation stage, in which a relatively large number of snippets had been spotted by one annotator only.

---

[3]An initiative where professional journalists profile news outlets; `https://mediabiasfactcheck.com`.

[4]`https://propaganda.qcri.org/annotations/`

Figure 2: Example of a plain-text article (left) and its annotation (right). The *Start* and the *End* columns are the indices representing the character span of the spotted technique.

partition	articles	average lengths		propaganda snippets
		chars	tokens	
training	371	5,681±5,425	927±899	6,128
development	75	4,700±2,904	770±473	1,063
test	90	4,518±2,602	744±433	1,790
all	536	5,348±4,789	875±793	8,981

Table 2: Statistics about the train/dev/test parts of the PTC-SemEval20 corpus, including the number of articles, their average lengths in terms of characters and tokens, and the total number of propaganda snippets they contain.

The annotation team consisted of six professional annotators from A Data Pro,[5] trained to spot and to label the propaganda snippets in free text. The job was carried out on an instance of the Anafora annotation platform (Chen and Styler, 2013), which we tailored for our propaganda annotation task. Figure 2 shows an example of an article and its annotations.

We evaluated the quality of the annotation process in terms of γ agreement (Mathet et al., 2015) between each of the annotators and the final gold labels. The γ agreement on the annotated articles is on average 0.6; see (Da San Martino et al., 2019b) for a more detailed discussion of inter-annotator agreement. The training and the development part of the PTC-SemEval20 corpus are the same as the training and the testing datasets described in (Da San Martino et al., 2019b). The test part of the PTC-SemEval20 corpus consists of 90 additional articles selected from the same sources as for training and development. For the test articles, we further extended the annotation process by adding one extra consolidation step: we revisited all the articles in that partition and we performed the necessary adjustments to the spans and to the labels as necessary, after a thorough discussion and convergence among at least three experts who were not involved in the initial annotations.

Table 2 shows some statistics about the corpus we use for the task. It is worth noting that a number of propaganda snippets of different classes overlap. Hence, the number of snippets for the span identification subtask is smaller (e.g., 1,405 for the span identification subtask vs. 1,790 for the technique classification subtask on the test set). The full collection of 536 articles contains 8,981 propaganda text snippets, belonging to one of the above-described fourteen classes. Figure 3 zooms into such snippets and shows the number of instances and the mean length for each class. We can see that, by a large margin, the most common propaganda technique in our news articles is *Loaded Language*, which is about twice as frequent as the second most frequent technique: *Name Calling or Labeling*. Whereas these two techniques are among the ones that are expressed in the shortest spans, other propaganda techniques such as *Exaggeration*, *Causal Oversimplification*, and *Slogans* tend to be the longest.

Number of instances per class (Top)

Class	all	training	development	test
0	8,981	6,128	1,063	1,790
1	3,017	2,123	325	569
2	1,537	1,058	183	296
3	922	621	145	156
4	747	493	66	188
5	655	466	68	121
6	471	294	44	133
7	411	229	87	95
8	247	209	18	20
9	179	144	14	21
10	197	129	40	28
11	173	107	29	37
12	190	107	22	61
13	119	76	17	26
14	116	72	5	39

Mean snippet length per class — chars (mean) (Bottom)

Class	all	training	development	test
0	46±59	48±62	36±52	44±50
1	22±24	24±26	16±13	19±19
2	25±19	26±20	23±17	22±16
3	17±20	17±20	12±10	23±24
4	116±90	124±97	109±96	98±62
5	43±34	44±32	48±50	38±29
6	87±69	98±74	69±60	70±56
7	60±70	61±65	55±76	60±74
8	121±71	123±73	85±48	128±60
9	132±121	138±127	90±94	112±77
10	25±13	25±13	24±13	29±16
11	100±70	96±64	97±92	113±67
12	95±68	105±75	58±41	92±58
13	32±25	30±21	47±45	30±16
14	90±62	96±73	90±35	78±40

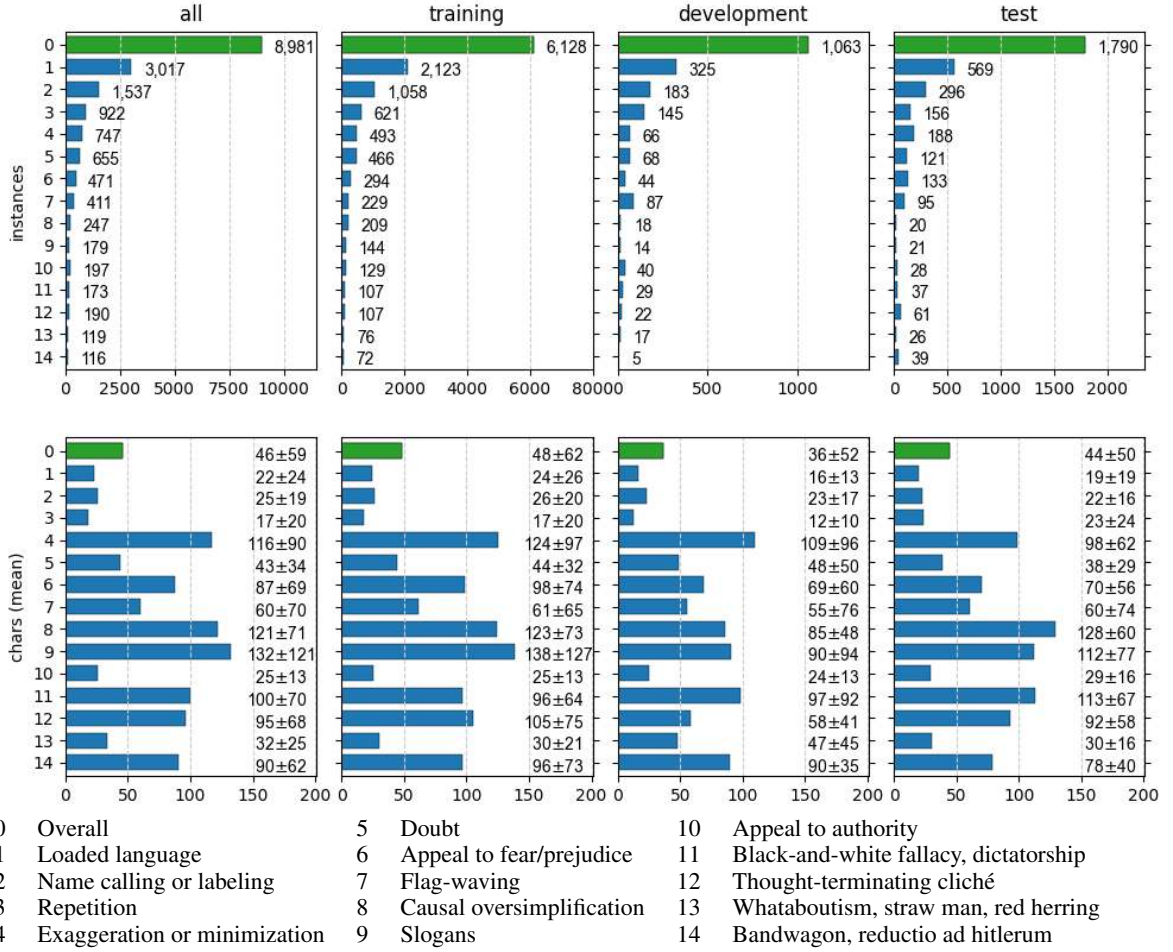|			
0	Overall	5	Doubt
1	Loaded language	6	Appeal to fear/prejudice
2	Name calling or labeling	7	Flag-waving
3	Repetition	8	Causal oversimplification
4	Exaggeration or minimization	9	Slogans
10	Appeal to authority		
11	Black-and-white fallacy, dictatorship		
12	Thought-terminating cliché		
13	Whataboutism, straw man, red herring		
14	Bandwagon, reductio ad hitlerum		

Figure 3: Statistics about the propaganda snippets in the different partitions of the PTC-SemEval20 corpus. Top: number of instances per class. Bottom: mean snippet length per class.

3.2 Evaluation Measures

Subtask SI Evaluating subtask SI requires us to match text spans. Our SI evaluation function gives credit to partial matches between gold and predicted spans.

Let d be a news article in a set D. A gold span t is a sequence of contiguous indices of the characters composing a text fragment $t \subseteq d$. For example, in Figure 4 (top-left) the gold fragment *"stupid and petty"* is represented by the set of indices $t_1 = [4, 19]$. We denote with $T_d = \{t_1, \ldots, t_n\}$ the set of all gold spans for an article d and with $T = \{T_d\}_d$ the set of all gold annotated spans in D. Similarly, we define $S_d = \{s_1, \ldots, s_m\}$ and S to be the set of predicted spans for an article d and a dataset D, respectively. We compute precision P and recall R by adapting the formulas in (Potthast et al., 2010):

$$P(S,T) = \frac{1}{|S|} \cdot \sum_{d \in D} \sum_{s \in S_d, t \in T_d} \frac{|(s \cap t)|}{|t|}, \tag{1}$$

$$R(S,T) = \frac{1}{|T|} \cdot \sum_{d \in D} \sum_{s \in S_d, t \in T_d} \frac{|(s \cap t)|}{|s|}. \tag{2}$$

We define Eq. (1) to be zero when $|S| = 0$ and Eq. (2) to be zero when $|T| = 0$. Notice that the predicted spans may overlap, e.g., spans s_3 and s_4 in Figure 4. Therefore, in order for Eq. 1 and Eq. 2 to get values lower than or equal to 1, all overlapping annotations, independently of their techniques, are merged first. For example, s_3 and s_4 are merged into one single annotation, corresponding to s_4.
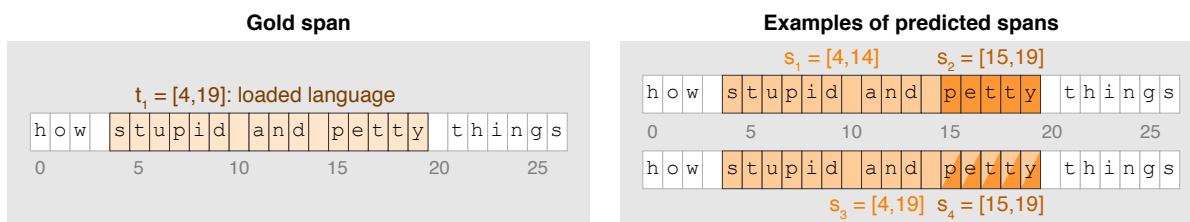
Figure 4: Example of equivalent annotations for the Span Identification subtask.

Finally, the evaluation measure for subtask SI is the F_1 score, defined as the harmonic mean between $P(S,T)$ and $R(S,T)$:

$$F_1(S,T) = 2 \cdot \frac{P(S,T) \cdot R(S,T)}{P(S,T) + R(S,T)}. \tag{3}$$

Subtask TC   Given a propaganda snippet in an article, subtask TC asks to identify the technique in it. Since there are identical spans annotated with different techniques (around 1.8% of the total annotations), formally this is a multi-label multi-class classification problem. However, we decided to consider the problem as a single-label multi-class one, by performing the following adjustments: (*i*) whenever a span is associated with multiple techniques, the input file will have multiple copies of such fragments and (*ii*) the evaluation function ensures that the best match between the predictions and the gold labels for identical spans is used for the evaluation. In other words, the evaluation score is not affected by the order in which the predictions for identical spans are submitted.

The evaluation measure for subtask TC is micro-average F_1. Note that as we have converted this into a single-label task, micro-average F_1 is equivalent to Accuracy (as well as to Precision and to Recall).

3.3   Task Organization

We ran the shared task in two phases:

Phase 1.   Only training and development data were made available, and no gold labels were provided for the latter. The participants competed against each other to achieve the best performance on the development set. A live leaderboard was made available to keep track of all submissions.

Phase 2.   The test set was released and the participants were given just a few days to submit their final predictions. The release of the test set was done task-by-task, since giving access to the input files for the TC subtask would have disclosed the gold spans for the SI subtask.

In phase 1, the participants could make an unlimited number of submissions on the development set, and they could see the outcomes in their private space. The best team score, regardless of the submission time, was also shown in a public leaderboard. As a result, not only could the participants observe the impact of various modifications in their own systems, but they could also compare against the results by other participating teams. In phase 2, the participants could again submit multiple runs, but they did not get any feedback on their performance. Only the last submission of each team was considered official and was used for the final team ranking.

In phase 1, a total of 47 teams made submissions on the development set for the SI subtask, and 46 teams submitted for the TC subtask. In phase 2, the number of teams who made official submissions on the test set for subtasks SI and TC was 35 and 31, respectively: this is a total of 66 submissions for the two subtasks, which were made by 44 different teams.

Note that we left the submission system open for submissions on the development set (phase 1) after the competition was over. The up-to-date leaderboards can be found on the website of the competition.[6]

---

[6]`http://propaganda.qcri.org/semeval2020-task11/leaderboard.php`

4 Participating Systems

In this section, we focus on a general description of the systems participating on both the SI and the TC subtasks. We pay special attention to the most successful approaches. The subindex on the right of each team represents their official rank in the subtasks. Appendix A includes brief descriptions of all systems.

4.1 Span Identification Subtask

Table 3 shows a quick overview of the systems that took part in the SI subtask.[7] All systems in the top-10 positions relied on some kind of Transformer, in combination with an LSTM or a CRF. In most cases, the Transformer-generated representations were complemented by engineered features, such as named entities and the presence of sentiment and subjectivity clues.

Team **Hitachi**(SI:1) achieved the top performance in this subtask (Morio et al., 2020). They used a BIO encoding, which is typical for related segmentation and labeling tasks (e.g., named entity recognition). They relied upon a complex heterogeneous multi-layer neural network, trained end-to-end. The network uses pre-trained language models, which generate a representation for each input token. They further added part-of-speech (PoS) and named entity (NE) embeddings. As a result, there are three representations for each token, which are concatenated and used as an input to bi-LSTMs. At this moment, the network branches, as it is trained with three objectives: (*i*) the main BIO tag prediction objective and two auxiliary ones, namely (*ii*) token-level technique classification, and (*iii*) sentence-level classification. There is one Bi-LSTM for objectives (*i*) and (*ii*), and there is another Bi-LSTM for objective (*iii*). For the former, they used an additional CRF layer, which helps improve the consistency of the output. A number of architectures were trained independently —using BERT, GPT-2, XLNet, XLM, RoBERTa, or XLM-RoBERTa—, and the resulting models were combined in ensembles.

Team **ApplicaAI**(SI:2) (Jurkiewicz et al., 2020) based its success on self-supervision using the RoBERTa model. They used a RoBERTa-CRF architecture trained on the provided data and used it to iteratively produce silver data by predicting on 500k sentences and retraining the model with both gold and silver data. The final classifier was an ensemble of models trained on the original corpus, re-weighting, and a model trained also on silver data. ApplicaAI was not the only team that reported performance boost when using additional data. Team **UPB**(SI:5) (Paraschiv and Cercel, 2020) decided not to stick to the pre-trained models from BERT–base alone and used masked language modeling to domain-adapt it using 9M articles containing fake, suspicious, and hyperpartisan news articles. Team **DoNotDistribute**(SI:22) (Kranzlein et al., 2020) also opted for generating silver data, but with a different strategy. They report a 5% performance boost when adding 3k new silver training instances. To produce them, they used a library to create near-paraphrases of the propaganda snippets by randomly substituting certain PoS words. Team **SkoltechNLP**(SI:25) (Dementieva et al., 2020) performed data augmentation based on distributional semantics. Finally, team **WMD**(SI:33) (Daval-Frerot and Yannick, 2020) applied multiple strategies to augment the data such as back translation, synonym replacement and TF.IDF replacement (replace unimportant words, based on TF.IDF score, by other unimportant words).

Closing the top-three submissions, Team **aschern**(SI:3) (Chernyavskiy et al., 2020) fine-tuned an ensemble of two differently intialized RoBERTa models, each with an attached CRF for sequence labeling and span character boundary post-processing.

There have been several other promising strategies. Team **LTIatCMU**(SI:4) (Khosla et al., 2020) used a multi-granular BERT BiLSTM model with additional syntactic and semantic features at the word, sentence and document level, including PoS, named entities, sentiment, and subjectivity. It was trained jointly for token and sentence propaganda classification, with class balancing. They further fine-tuned BERT on persuasive language using 10,000 articles from propaganda websites, which turned out to be important. Team **PsuedoProp**(SI:14) (Chauhan and Diddee, 2020) built a preliminary sentence-level classifier using an ensemble of XLNet and RoBERTa, before it fine-tuned a BERT-based CRF sequence tagger to identify the exact spans. Team **BPGC**(SI:21) (Patil et al., 2020) went beyond these multigranularity approaches. Information both at the article and at the sentence level was considered when classifying each word as propaganda or not, by computing and concatenating vectorial representations for the three inputs.

---

[7]Tables 3 and 4 only include the systems for which a description paper was submitted.

	Transformers									Learning Models								Representations											Misc		
Rank. Team	BERT	RoBERTa	XLNet	XLM	XLM RoBERTa	ALBERT	GPT-2	SpanBERT	LaserTagger	LSTM	CNN	SVM	Naïve Bayes	Boosting	Log regressor	Random forest	CRF	Embeddings	ELMo	NEs	Words/n-grams	Chars/n-grams	PoS	Trees	Sentiment	Subjectivity	Rhetorics	LIWC	Ensemble	Data augmentation	Post-processing
1. Hitachi	☑	☑	☑	☑	☑		☑			☑							☑	☑		☑		☑							☑		
2. ApplicaAI		☑															☑													☑	
3. aschern		☑															☑	☑			☑								☑	☑	
4. LTIatCMU	☑									☑								☑		☑	☑		☑	☑	☑	☑		☑			
5. UPB	☑									✔							☑	✔													☑
7. NoPropaganda	☑								☑	✔							✔	☑	☑			☑									
8. CyberWallE	☑									☑		✔											☑		☑		☑				
9. Transformers	☑	☑								☑	☑							☑											☑		
11. YNUtaoxin	☑	✔	✔																												
13. newsSweeper	☑	✔						✔	✔														✔								
14. PsuedoProp	☑	☑	☑				✔																						☑		
16. YNUHPCC	✔									☑								☑													
17. NLFIIT	✔									☑						✔		✔	☑										✔		
20. TTUI	☑	☑																				☑							☑		☑
21. BPGC	✔	☑	✔					✔										☑													
22. DoNotDistribute	☑									✔										☑	☑		☑								☑
23. UTMN													☑					☑							☑						
25. syrapropa								☑										☑			☑										
26. SkoltechNLP	☑									✔	✔					✔			☑		☑										
27. NTUAAILS										☑								☑													
28. UAIC1860												✔	✔			✔	☑	☑			☑	☑									
31. 3218IR											☑							☑													
33. WMD	☑											✔		✔		✔		☑											☑	☑	☑
– UNTLing																	✔	✔					✔		✔	✔	✔				

1. (Morio et al., 2020)
2. (Jurkiewicz et al., 2020)
3. (Chernyavskiy et al., 2020)
4. (Khosla et al., 2020)
5. (Paraschiv and Cercel, 2020)
7. (Dimov et al., 2020)
8. (Blaschke et al., 2020)
9. (Verma et al., 2020)
11. (Tao and Zhou, 2020)
13. (Singh et al., 2020)
14. (Chauhan and Diddee, 2020)
16. (Dao et al., 2020)
17. (Martinkovic et al., 2020)
20. (Kim and Bethard, 2020)
21. (Patil et al., 2020)
22. (Kranzlein et al., 2020)
23. (Mikhalkova et al., 2020)
25. (Li and Xiao, 2020)
26. (Dementieva et al., 2020)
27. (Arsenos and Siolas, 2020)
28. (Ermurachi and Gifu, 2020)
31. (Dewantara et al., 2020)
33. (Daval-Frerot and Yannick, 2020)
– (Krishnamurthy et al., 2020)

Table 3: Overview of the approaches to the span identification subtask. ☑=part of the official submission; ✔=considered in internal experiments. The references to the description papers appear at the bottom.

A large number of the participating teams built systems that rely heavily on engineered features. For instance, Team **CyberWallE** (SI:8) (Blaschke et al., 2020) used features modeling sentiment, rhetorical structure, and POS tags, while team **UTMN** (SI:23) injected the sentiment intensity from VADER and it was among the only teams not relying on deep learning architectures to produce a computationally affordable model.

4.2 Technique Classification Subtask

The same trends as for the snippet identification subtask can be observed in the approaches used for the technique classification subtask: practically, all top-performing approaches used representations produced by some kind of Transformer.

Team **ApplicaAI** (TC:1) achieved the top performance for this subtask (Jurkiewicz et al., 2020). As in their approach to subtask SI, ApplicaAI produced additional silver data for training. This time, they ran their high-performing SI model to spot new propaganda snippets in free text and applied their preliminary TC model to produce extra silver-labeled instances. Their final classifier consisted of an ensemble of models trained on the original corpus, re-weighting, and a model trained also on silver data. In all cases, the input to the classifiers consisted of propaganda snippets and their context.

Rank. Team	Transformers										Learning Models											Representations											Misc		
	BERT	R BERT	RoBERTa	XLNet	XLM	XLM RoBERTa	ALBERT	GPT-2	SpanBERT	DistilBERT	LSTM	RNN	CNN	SVM	Naïve Bayes	Boosting	Log regressor	Random forest	Regression tree	CRF	XGBoost	Embeddings	ELMo	NEs	Words/n-grams	Chars/n-grams	PoS	Sentiment	Rhetorics	Lexicons	String matching	Topics	Ensemble	Data augmentation	Post-processing
1. ApplicaAI			☑																	☑													☑	☑	
2. aschern			☑																			☑											☑		☑
3. Hitachi	☑		☑	☑	☑	☑	☑						☑									☑		☑			☑						☑		
4. Solomon			☑											✓								☑									☑		☑		
5. newsSweeper	✓		☑					✓	✓																		✓								
6. NoPropaganda	☑	☑																																	
7. Inno	✓		☑	✓				✓			✓						✓					✓													
8. CyberWallE	☑																					☑			☑										☑
10. Duth	☑																					☑													
11. DiSaster	☑													✓			☑					☑											☑		
13. SocCogCom	☑							✓										✓										☑	✓						
14. TTUI	☑	☑																				☑											☑		
15. JUST	☑																					☑													
16. NLFIIT	✓										☑									✓		✓	☑										✓		
17. UMSIForeseer	☑																					☑											☑		
18. BPGC	☑												✓	✓			☑					✓		☑	☑							☑	☑		
19. UPB	☑																																		☑
20. syrapropa	☑																☑							☑	☑										☑
21. WMD	☑												✓		✓	✓	✓					☑											☑	☑	
22. YNUHPCC	☑										☑									✓															
24. DoNotDistribute	☑										☑													☑	☑		☑								
25. NTUAAILS											☑											☑													
26. UAIC1860														✓	✓		✓	☑				☑		☑	☑										
27. UNTLing																				☑		☑		☑	☑						☑				

1. (Jurkiewicz et al., 2020)   11. (Kaas et al., 2020)   20. (Li and Xiao, 2020)
2. (Chernyavskiy et al., 2020)   13. (Krishnamurthy et al., 2020)   21. (Daval-Frerot and Yannick, 2020)
3. (Morio et al., 2020)   14. (Kim and Bethard, 2020)   22. (Dao et al., 2020)
4. (Raj et al., 2020)   15. (Altiti et al., 2020)   24. (Kranzlein et al., 2020)
5. (Singh et al., 2020)   16. (Martinkovic et al., 2020)   25. (Arsenos and Siolas, 2020)
6. (Dimov et al., 2020)   17. (Jiang et al., 2020)   26. (Ermurachi and Gifu, 2020)
7. (Grigorev and Ivanov, 2020)   18. (Patil et al., 2020)   27. (Petee and Palmer, 2020)
8. (Blaschke et al., 2020)   19. (Paraschiv and Cercel, 2020)   29. (Verma et al., 2020)
10. (Bairaktaris et al., 2020)

Table 4: Overview of the approaches to the technique classification subtask. ☑=part of the official submission; ✔=considered in internal experiments. The references to the description papers appear at the bottom.

Team **aschern**(TC:2) (Chernyavskiy et al., 2020) was the second best, and it based its success on a RoBERTa ensemble with several interesting techniques. They treated the task as one of sequence classification, using an average embedding of the surrounding tokens and the length of the span as contextual features. They further incorporated knowledge from the span identification task, using transfer learning: namely, they first pre-trained their model for the SI task, and then continued training for the TC task. Finally, they performed task-specific postprocessing in order to increase the consistency for the repetition technique spans and to avoid insertions of techniques inside other techniques.

Team **Hitachi**(TC:3) (Morio et al., 2020) used two distinct feed-forward neural networks (FFNs). The first one is for sentence representation, whereas the second one is for the representation of the tokens in the propaganda span. The propaganda span representation is obtained by concatenating the representation of the begin-of-sentence token, the span start token, the span end token, and the aggregated representation obtained using attention and max-pooling. As for their winning approach to SI, team Hitachi trained on the TC subtask independently with different large-scale pre-trained state-of-the-art language models (BERT, GPT-2, XLNet, XLM, RoBERTa, or XLM-RoBERTa), and then combined the resulting models in an ensemble.

As the top-performing models to subtask TC show, while the two subtasks can be seen as fairly independent, combining them in a reasonable way pays back. Additionally, the context of a propaganda snippet is important to identify the specific propaganda technique it uses. Indeed, other teams tried to make context play a role in their models with certain success. For instance, team **newsSweeper** (TC:5) (Singh et al., 2020) used RoBERTa to obtain and to concatenate representations for both the propaganda snippet and its context (i.e., sentence). Team **SocCogCom** (TC:13) (Krishnamurthy et al., 2020) reduced the context to a window of three words before and after the propaganda snippet.

As in the SI subtask, a number of teams achieved sizable improvements when using various features. For instance, team **BPGC** (TC:18) (Patil et al., 2020) included TF.IDF vectors of words and character n-grams, topic modeling, and sentence-level polarity, among others, to their ensemble model that used BERT and logistic regression. Team **SocCogCom** (TC:13) (Krishnamurthy et al., 2020) integrated semantic-level emotional salience features from CrystalFeel (Gupta and Yang, 2018) and word-level psycholinguistic features from LIWC (Pennebaker et al., 2015). Team **CyberWallE** (TC:8) (Blaschke et al., 2020) added named entities, rhetorical, and question features, while taking special care of repetitions as part of a complex ensemble architecture. According to team **UNTLing** (TC:27) (Petee and Palmer, 2020), considering NEs is particularly useful for propaganda techniques such as *Loaded Language* and *Flag Waving* (e.g., the latter usually includes references to idealized entities) and VAD features were useful for emotion-related propaganda techniques such as *Appeal to fear/prejudice* and *Doubt*. Team **DiSaster** (TC:11) (Kaas et al., 2020) combined BERT with features including frequency of the fragment in the article and in the sentence it appears in, and the inverse uniqueness of words in a span. The goal of these features was to compensate the inability of BERT to deal with distant context, specifically to target the technique *Repetition*. Team **Solomon** (TC:4) also targeted *Repetition* by using dynamic least common sub-sequence, which they used to score the similarity between the fragment and the context. Then, the fragment was considered to be a repetition if the score was greater than a threshold heuristically set with respect to the length of the fragment.

Some other teams decided to perform a normalization of the input texts, thus trying to reduce the representation diversity. This was the case of team **DUTH** (TC:10) (Bairaktaris et al., 2020), which mapped certain words into classes using named entity recognition with focus on person names and gazetteers containing names and variations of names of countries (255 entries), religions (35 entries), political ideologies (23 entries), and slogans (41 entries). The recognized categories were replaced by the category name in the input, before passing the input to BERT.

As the class distribution for subtask TC is heavily skewed, some teams tried balancing techniques. For example, team **Inno** (TC:7) (Grigorev and Ivanov, 2020) experimented with undersampling (i.e. removing some examples from the bigger classes), team **syrapropa** (TC:20) applied a cost adjustment to their BERT-based model, and team **UMSIForeseer** (TC:17) (Jiang et al., 2020) used a mix of oversampling and undersampling, which they combined using a bagging ensemble.

Finally, some teams used an overriding strategy on top of the output of their supervised models. For example, team **CyberWallE** (TC:8) (Blaschke et al., 2020) performed a rule-based label post-processing, and team **syrapropa** (TC:20) applied syntactic rules based on part of speech.

5 Results and Discussion

5.1 Results on the Span Identification Subtask

Table 5 shows the performance of the participating systems both on the testing and on the development partitions on the SI subtask. The baseline for subtask SI is a simple system that randomly generates spans, by first selecting the starting character of a span and then its length. As mentioned in Section 4, practically all approaches relied on Transformers to produce representations and then plugged their output into a sequential model, at the token level. It is worth observing that only three of the top-5 systems on the development set appear also among the top-5 systems on the test set. Indeed, teams **syrapropa** and **PALI** felt down from positions 1 and 2 on the development set to positions 25 and 18 on the test set, which suggests possible overfitting. The performance for the final top-3 systems on the test partition —**Hitachi**, **ApplicaAI**, and **aschern**— reflects robust systems that seem to generalize much better.
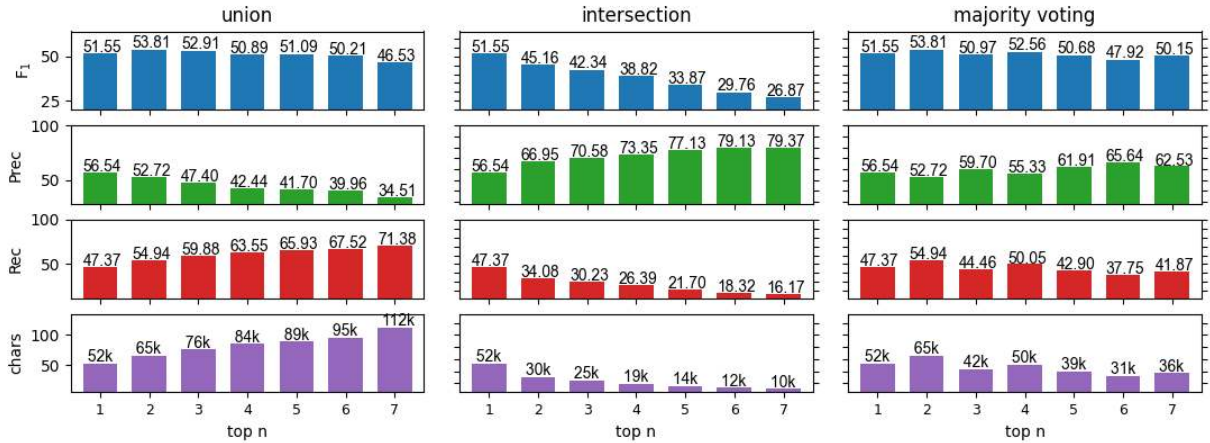
Figure 5: **SI Subtask**: performance (P, R, F_1) when combining the output of the top-7 systems using union, intersection, and majority voting. The bottom plots show the number of characters deemed propagandistic in each combination.

Figure 5 shows the performance evolution when combining the output of the top-performing systems on the test set. All operations are carried out at the character level, i.e., we label characters as propagandistic or not, and then we combine into spans the longest possible sequences of neighboring characters that we labeled as propagandistic. The union and the intersection use the corresponding set operations. In union, a character is considered propagandistic if at least one of the participating systems has recognized it as part of a propaganda span. In intersection, a character is considered propagandistic if all systems have included it as part of a propaganda span. For majority voting, we consider a character propagandistic if more than 50% of the participating systems had included it as part of a propaganda span. We can see in Figure 5 that the precision and the recall trends behave just as we expected: a lower precision (higher recall) is observed when more systems are combined with a union operation, and the opposite is true for the intersection. Despite the loss in terms of precision, computing the union of the top-$[2, 3]$ systems results in a better performance than what each of the top systems could achieve on its own. Such a combination gathers large ensembles of Transformer representations together with self-supervision to produce additional training data and boundary post-processing. If we are interested in a high-precision model, applying the intersection would make more sense, as it reaches a precision of 66.95 when combining the top-2 systems; however, this comes at the cost of a sizable lost of spans, which results in considerable drop in recall. The majority voting combination lies somewhere in between: keeping reasonable levels for both the precision and the recall.

5.2 Results on the Technique Classification Subtask

Table 6 shows the performance of the participating systems on the test set for the TC subtask, and Table 7 reports the results on the development set. The baseline system for subtask TC is a logistic regression classifier using one feature only: the length of the fragment. A similar pattern as for the SI subtask is observed: only two of the top-5 systems on the development set appear among the top-5 systems on the test set, which is a sign of possible overfitting for some of the systems. At the same time, systems that appeared to have a modest performance on the development set could eventually reach a higher position on test. For instance, team **Hitachi**, which was ranked 8th on the development set, ended up in the third position on the test set.

The tables further show the performance for each of the 14 propaganda techniques. In general, the systems show reasonably good performance when predicting *Loaded Language* and *Name Calling or Labeling*. These two classes are the most frequent ones, by a margin, and are also among the shortest ones on average (cf. Figure 3). On the other hand, techniques 13 (*Straw man, red herring*) and 14 (*Bandwagon, reduction ad hitlerum, whataboutism*) are among the hardest to identify. They are also among the least frequent ones.
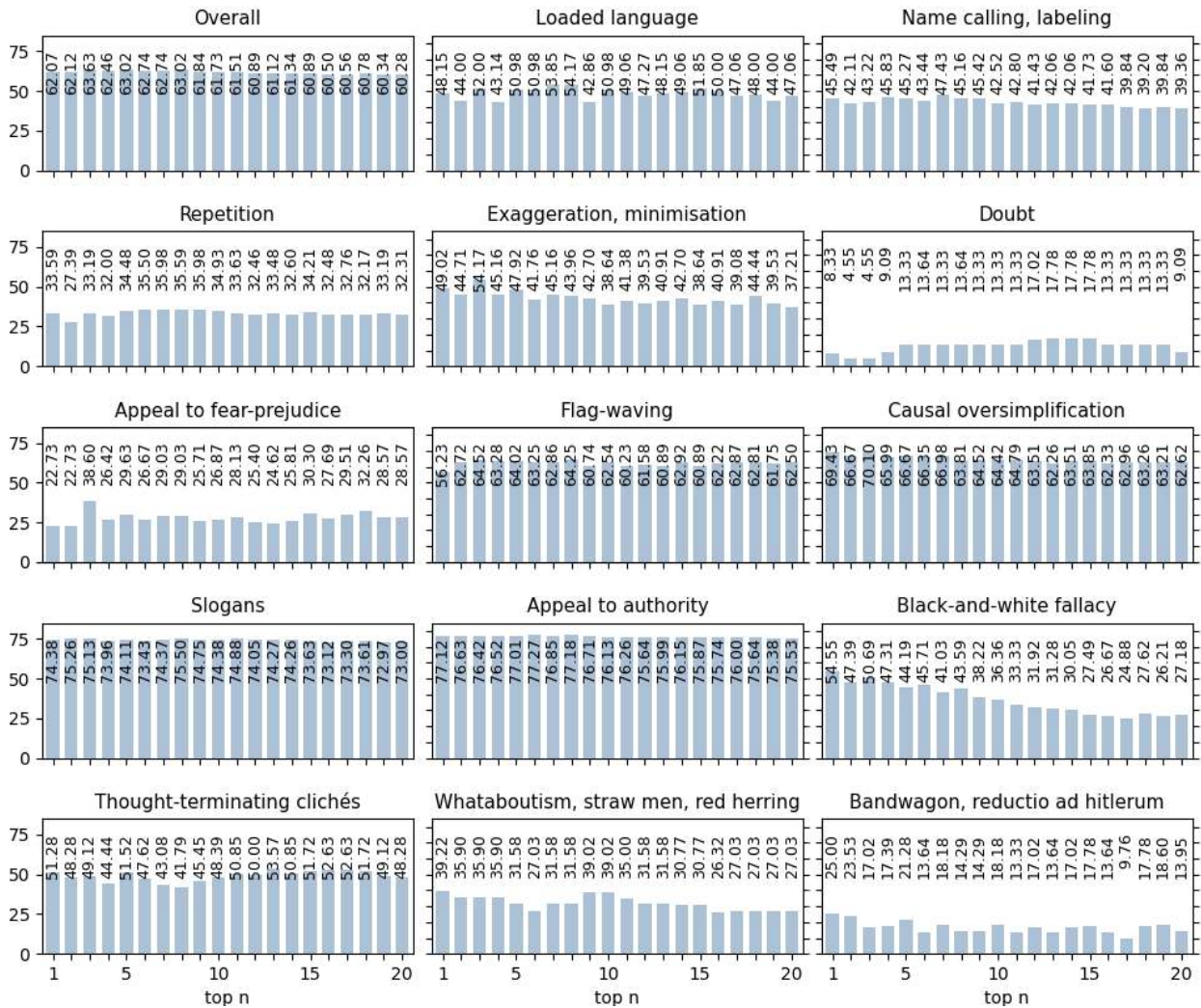
Figure 6: F_1 performance for the technique classification subtask when combining the top-20 systems with majority voting. The plots show the overall F_1 performance (top left) as well as the F_1 performance for each of the 14 propaganda techniques.

Once again, we studied the performance when combining more approaches. Figure 6 shows the performance evolution when combining different numbers of top-performing systems on the test set. As this is a multi-class problem, we combine the systems only on the basis of majority voting. In case of a tie, we prefer the more frequent propaganda technique on the training set. When looking at the overall picture, the performance evolution when adding more systems is fairly flat, reaching the top performance when combining the top-3 systems: an F_1 score of 63.63, which represents more than 1.5 points of absolute improvement over the top-1 system. When zooming into each of the fourteen propaganda techniques, we observe that in general the performance peak is indeed reached when considering three systems, e.g., for *Appeal to fear–prejudice*, *Exaggeration, minimisation*, or *Causal oversimplification*. Still for *Doubt*, which is the hardest class to recognize, as many as 13 systems are necessary in order to reach a (still discrete) peak performance of 17.78. Finally, note that there are other classes, such as *Black-and-white fallacy* or *Whatabaotism, straw men, red herring*, for which system combinations do not help.

1389

Team	Test				Development			
	Rnk	F_1	P	R	Rnk	F_1	P	R
Hitachi	1	**51.55**	56.54	47.37	4	**50.12**	42.26	61.56
ApplicaAI	2	49.15	59.95	41.65	3	52.19	47.15	58.44
aschern	3	49.10	53.23	45.56	5	49.99	44.53	56.98
LTIatCMU	4	47.66	50.97	44.76	7	49.06	43.38	56.47
UPB	5	46.06	58.61	37.94	8	46.79	42.44	52.13
Fragarach	6	45.96	54.26	39.86	12	44.27	41.68	47.21
NoPropaganda	7	44.68	55.62	37.34	9	46.13	40.65	53.31
CyberWallE	8	43.86	42.16	45.70	17	42.39	33.45	57.86
Transformers	9	43.60	49.86	38.74	14	43.06	40.85	45.52
SWEAT	10	43.22	52.77	36.59	16	42.51	42.97	42.06
YNUtaoxin	11	43.21	55.62	35.33	11	44.35	40.74	48.67
DREAM	12	43.10	54.54	35.63	19	42.15	42.66	41.65
newsSweeper	13	42.21	46.52	38.63	10	44.45	38.76	52.10
PsuedoProp	14	41.20	41.54	40.87	22	39.32	34.27	46.11
Solomon	15	40.68	53.95	32.66	15	42.86	43.24	42.49
YNUHPCC	16	40.63	36.55	45.74	18	42.27	32.08	61.95
NLFIIT	17	40.58	50.91	33.73	21	39.67	35.04	45.72
PALI	18	40.57	53.20	32.79	2	52.35	49.64	55.37
UESTCICSA	19	39.85	56.09	30.90	13	44.17	43.21	45.18
TTUI	20	39.84	**66.88**	28.37	6	49.59	48.76	50.44
BPGC	21	38.74	49.39	31.88	25	36.79	34.72	39.12
DoNotDistribute	22	37.86	42.36	34.23	24	37.73	32.41	45.12
UTMNandOCAS	23	37.49	37.97	37.03	31	34.35	23.65	62.69
Entropy	24	37.23	41.68	33.63	32	32.89	30.82	35.25
syrapropa	25	36.20	49.53	28.52	1	53.40	39.88	80.80
SkoltechNLP	26	34.07	46.52	26.87	26	36.70	34.99	38.59
NTUAAILS	27	33.60	46.05	26.44	33	31.21	27.95	35.35
UAIC1860	28	33.21	24.49	51.57	34	30.27	20.69	56.37
CCNI	29	29.48	38.09	24.05	35	29.61	29.04	30.21
NCCU-SMRG	30	28.47	17.30	**80.37**	42	15.83	09.12	59.92
3218IR	31	23.47	22.63	24.38	40	20.03	15.10	29.76
WMD	32	20.09	47.11	12.77	27	36.34	33.26	40.05
LS	33	18.18	34.14	12.39	29	35.49	22.41	**85.33**
HunAlize	34	3.19	23.24	1.71	38	24.45	37.75	18.08
YOLO	35	0.72	17.20	0.37	46	0.64	9.36	0.33
Baseline	36	0.31	13.04	0.16	43	1.10	11.00	0.58
Murgila	–	–	–	–	20	41.38	32.96	55.56
TakeLab	–	–	–	–	23	39.06	38.85	39.27
atulcst	–	–	–	–	28	36.29	38.15	34.61
AAA	–	–	–	–	30	34.68	30.61	40.00
CUNLP	–	–	–	–	36	27.78	**58.23**	18.24
IIITD	–	–	–	–	37	25.82	18.81	41.15
UoB	–	–	–	–	39	24.02	22.30	26.04
UBirmingham	–	–	–	–	41	16.95	23.07	13.39
SocCogCom	–	–	–	–	44	0.79	9.97	0.41
Inno	–	–	–	–	45	0.64	9.36	0.33
Raghavan	–	–	–	–	47	0.40	0.20	33.45
California	–	–	–	–	48	0.39	5.92	0.20

Table 5: **Subtask 1: Span Identification (SI) performance on test and development.** The highest scores for each measure are highlighted. (*Note: We found a bug in the evaluation script after the end of the competition. The correct ranking, shown in Appendix B, does not differ substantially from above.*)

Table 6: **Technique classification F_1 performance on the test set.** The systems are ordered on the basis of the final ranking. Columns **1** to **14** show the performance for each of the propaganda techniques (cf. Section 2). The best score for each technique appears highlighted. (*Note: We found a bug in the evaluation script after the end of the competition. The correct ranking, shown in Appendix B, does not differ substantially from above.*)

Rnk	Team	Overall	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	ApplicaAI	**62.07**	77.12	74.38	**54.55**	33.59	56.23	**45.49**	**69.43**	22.73	51.28	48.15	49.02	39.22	25.00	8.33
2	aschern	62.01	77.02	**75.65**	53.38	32.65	59.44	41.78	66.35	25.97	**54.24**	35.29	**53.57**	**42.55**	18.87	14.93
3	Hitachi	61.73	75.64	74.20	37.88	34.58	**63.43**	38.94	68.02	**36.62**	45.61	40.00	47.92	29.41	**26.92**	4.88
4	Solomon	58.94	74.66	70.75	42.53	28.44	61.82	39.39	61.84	19.61	50.75	26.67	42.00	38.10	0.00	4.88
5	newsSweeper	58.44	75.32	74.23	20.69	37.10	56.55	42.80	60.53	19.72	50.75	41.67	25.00	21.62	8.00	13.04
6	NoPropaganda	58.27	**77.17**	73.90	42.71	**37.99**	56.27	38.02	59.30	12.12	42.42	23.26	8.70	23.26	0.00	0.00
7	Inno	57.99	73.31	74.30	24.89	35.39	58.65	45.09	59.41	24.32	43.75	43.14	40.40	29.63	19.36	10.71
8	CyberWallE	57.37	74.68	70.92	47.68	28.34	58.65	39.84	54.38	15.39	39.39	14.63	23.68	23.81	0.00	12.25
9	PALI	57.32	74.29	69.09	24.56	28.57	58.97	36.59	61.62	30.59	39.22	27.59	39.62	40.82	20.90	**28.57**
10	DUTH	57.21	73.71	71.41	20.10	28.24	59.16	33.33	58.95	26.23	34.78	44.44	33.33	27.03	17.78	9.30
11	DiSaster	56.65	74.49	68.10	20.44	30.64	59.12	35.25	58.25	14.63	42.55	51.16	26.67	19.05	4.35	20.41
12	djichen	56.54	73.21	68.38	29.75	31.42	60.00	33.65	56.19	22.79	30.77	37.50	43.81	27.91	18.87	20.83
13	SocCogCom	55.81	72.18	67.34	18.88	34.86	60.40	31.62	54.26	6.35	40.91	28.57	26.51	23.53	10.00	9.76
14	TTUI	55.64	73.22	68.49	21.18	32.20	57.40	41.48	61.68	23.08	37.50	28.24	35.29	25.00	20.29	24.56
15	JUST	55.31	71.96	64.73	21.94	29.57	58.26	37.10	62.56	27.27	33.33	**48.89**	28.89	31.82	28.57	24.49
16	NLFIIT	55.25	72.55	69.30	21.55	30.30	55.66	24.89	63.32	0.00	41.67	29.63	32.10	13.64	0.00	9.30
17	UMSIForeseer	55.14	73.02	70.79	21.49	28.57	57.21	31.97	56.14	0.00	39.22	29.41	0.00	14.29	0.00	9.76
18	BPGC	54.81	71.58	67.51	23.74	33.47	53.78	33.65	58.93	24.18	40.00	30.77	40.00	20.69	20.90	12.50
19	UPB	54.30	70.09	68.86	20.00	30.62	52.55	30.00	55.87	16.95	34.62	20.00	19.72	22.86	4.88	0.00
20	syrapropa	54.25	71.47	68.44	30.77	28.10	56.14	29.77	57.02	21.51	29.03	31.58	30.61	28.57	9.09	19.61
21	WMD	52.01	69.33	64.67	13.89	25.46	53.94	29.20	52.08	5.71	6.90	7.14	0.00	7.41	0.00	5.00
22	YNUHPCC	50.50	68.08	62.33	17.72	21.54	51.04	26.40	55.56	3.45	27.59	29.79	38.38	17.78	15.00	13.79
23	UESTCICSA	49.94	68.23	66.88	27.96	25.44	44.99	22.75	53.14	3.74	41.38	12.77	11.27	28.57	3.70	0.00
24	DoNotDistribute	49.72	68.44	60.65	19.44	27.23	46.25	29.75	53.76	14.89	28.07	22.64	24.49	12.25	9.68	4.55
25	NTUAAILS	46.37	65.79	54.55	18.43	29.66	48.75	28.31	46.47	0.00	13.79	36.36	0.00	11.43	4.08	9.76
26	UAIC1860	41.17	62.33	42.97	11.16	21.01	36.41	22.12	38.78	7.60	11.43	17.39	2.90	5.56	4.26	9.76
27	UNTLing	39.11	62.57	36.74	7.78	11.82	32.65	5.29	40.48	2.86	17.65	4.35	0.00	0.00	0.00	0.00
28	HunAlize	37.10	58.59	15.82	2.09	23.81	31.76	11.83	29.95	7.84	4.55	6.45	8.00	0.00	0.00	0.00
29	Transformers	26.54	47.55	24.06	2.86	0.00	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30	Baseline	25.20	46.48	0.00	19.26	14.42	29.14	3.68	6.20	11.56	0.00	0.00	0.00	0.00	0.00	0.00
31	Entropy	20.39	37.74	15.49	5.83	6.39	12.81	6.32	4.95	7.41	0.00	3.92	2.27	0.00	6.78	0.00
32	IJSE8	19.72	38.07	14.70	4.92	8.23	15.47	7.07	8.57	2.27	0.00	0.00	0.00	0.00	0.00	0.00

Rnk	Team	Overall	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	ApplicaAI	**70.46**	80.42	74.11	70.67	**60.93**	63.57	**48.49**	**82.56**	47.06	62.65	**42.86**	20.00	43.48	**41.67**	80.00
2	aschern	68.11	80.00	73.85	67.39	59.36	**66.19**	45.24	74.25	30.44	63.89	28.57	32.43	33.33	38.46	40.00
8	Hitachi	65.19	77.99	71.16	46.98	56.58	62.96	36.96	81.40	50.00	67.61	12.50	22.22	19.05	40.91	75.00
16	Solomon	59.55	75.92	71.98	33.48	49.65	51.28	39.08	74.68	34.29	45.90	0.00	20.83	32.26	20.83	0.00
5	FakeSolomon	67.17	79.29	74.38	66.04	50.75	61.22	42.72	74.36	31.25	70.00	9.52	18.75	43.48	6.25	**88.89**
10	newsSweeper	62.75	76.23	71.35	45.14	55.32	56.94	44.90	74.29	47.83	60.53	16.67	6.90	16.67	6.45	57.14
11	NoPropaganda	60.68	77.06	72.49	47.69	38.71	50.36	33.33	75.74	48.00	53.17	0.00	0.00	8.33	0.00	0.00
13	Inno	60.11	76.57	70.62	34.67	43.75	52.86	40.82	77.97	34.15	59.70	11.77	35.71	18.18	15.39	61.54
7	CyberWallE	66.42	76.62	**81.00**	**73.29**	52.70	53.85	30.61	73.68	21.05	51.35	18.18	21.43	17.39	0.00	22.22
6	PALI	66.89	78.01	76.34	61.59	48.61	58.46	40.45	79.53	42.11	58.33	41.67	35.56	**45.71**	27.03	75.00
23	DUTH	57.86	76.12	70.32	23.08	46.15	48.72	34.41	64.43	42.86	55.39	9.52	11.11	15.39	11.11	50.00
9	DiSaster	62.84	76.73	77.33	47.11	48.28	56.21	35.29	75.15	23.08	35.09	30.00	0.00	9.09	5.56	28.57
26	djichen	57.57	76.04	70.53	24.89	44.60	48.28	36.78	76.36	31.82	51.43	13.33	20.00	10.81	4.55	0.00
28	SocCogCom	57.01	70.65	64.38	31.78	45.67	53.99	32.91	77.11	28.57	30.19	21.43	21.43	12.90	6.45	57.14
17	TTUI	58.98	75.08	71.00	32.07	41.96	56.92	34.69	77.53	33.33	50.00	16.22	19.05	20.69	29.09	66.67
24	JUST	57.67	74.97	72.36	23.64	49.64	48.75	31.07	73.75	23.81	29.03	0.00	0.00	0.00	5.88	0.00
20	NLFIIT	58.51	74.46	70.48	33.33	41.48	53.13	35.14	75.00	34.29	60.53	13.33	27.78	24.24	24.14	60.00
19	UMSIForeseer	58.70	74.44	70.95	31.76	51.80	47.95	31.58	75.86	**52.63**	43.33	10.00	0.00	0.00	10.00	0.00
18	BPGC	58.80	75.45	70.05	29.75	49.64	52.86	32.43	75.82	27.03	52.31	19.05	18.75	23.08	9.52	0.00
21	UPB	58.33	74.02	71.64	24.16	50.00	46.15	29.89	70.73	34.29	50.00	7.41	8.70	26.09	6.06	57.14
3	syrapropa	67.83	77.24	77.00	50.00	51.39	56.74	40.82	78.89	41.86	68.49	20.00	13.79	24.24	24.24	60.00
33	WMD	52.31	71.57	57.70	40.00	35.12	39.77	26.83	63.10	6.90	15.69	19.05	0.00	0.00	0.00	40.00
29	YNUHPCC	56.16	70.57	69.45	29.75	36.36	51.39	26.09	73.05	27.27	40.58	6.45	18.18	26.09	21.05	54.55
25	UESTCICSA	57.57	74.15	65.04	48.37	39.66	46.48	38.71	63.23	26.67	58.82	0.00	0.00	36.36	9.52	0.00
30	DoNotDistribute	54.00	71.20	67.33	29.38	31.93	45.95	25.58	72.61	25.00	28.99	0.00	0.00	22.86	9.76	0.00
32	NTUAAILS	53.25	69.76	59.90	27.19	43.64	51.61	21.33	73.03	17.14	27.12	6.67	7.14	0.00	8.00	44.44
34	UAIC1860	43.84	57.88	40.37	14.12	23.66	42.03	7.02	60.00	0.00	4.55	10.53	0.00	0.00	0.00	33.33
37	UNTLing	40.92	59.45	33.33	11.83	12.28	35.42	11.94	51.70	26.67	25.00	8.33	0.00	5.88	5.88	0.00
36	HunAlize	41.02	56.13	40.75	5.32	22.22	41.18	9.23	53.50	33.33	0.00	0.00	0.00	0.00	0.00	0.00
41	Transformers	30.10	48.50	28.62	12.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
43	Baseline	26.53	40.58	0.00	38.50	11.68	19.20	9.38	8.28	7.23	0.00	0.00	0.00	0.00	4.88	0.00
27	Entropy	57.10	75.22	66.50	28.19	45.16	53.60	31.11	74.85	29.27	46.88	13.33	18.18	20.00	10.00	0.00
4	FLZ	67.17	**81.09**	71.12	70.78	23.38	65.46	20.00	64.08	46.15	64.00	44.44	**37.04**	28.57	6.67	75.00
12	Fragarach	60.49	77.12	70.50	38.21	50.69	51.03	40.91	74.12	34.04	45.71	14.29	12.90	16.67	6.06	57.14
14	NerdyBirdies	59.83	76.23	73.55	33.05	48.65	52.06	36.04	70.30	47.62	48.57	15.39	8.70	20.00	13.33	0.00
15	CUNLP	59.55	74.38	73.30	47.62	43.66	50.69	34.29	71.90	35.90	55.07	16.22	15.79	19.51	8.89	25.00
22	Tianyi	58.23	75.39	71.22	26.43	44.72	50.36	34.57	75.61	28.57	49.32	13.33	11.77	0.00	57.14	57.14
31	Murgila	53.53	71.21	66.14	24.55	42.55	37.29	22.47	69.51	30.00	40.58	17.65	13.33	31.58	22.64	42.86
35	hseteam	41.40	63.65	42.82	22.40	32.26	26.29	11.11	57.52	7.69	26.67	0.00	7.27	0.00	4.76	0.00
38	HenryAtDuderstadt	34.15	52.97	4.10	0.00	0.00	34.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
39	XJPaccelerationMASTER	32.46	48.76	0.00	0.00	0.00	33.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
40	SWEAT	30.57	46.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.00	0.00	0.00
42	NCCU-SMRG	29.26	37.92	34.46	25.53	14.69	36.92	15.09	45.07	9.76	13.33	12.50	16.67	6.67	12.50	23.08
44	LS	26.23	39.20	0.00	39.31	12.95	19.67	9.09	8.28	8.79	0.00	0.00	0.00	6.06	0.00	0.00
45	SkoltechNLP	26.23	39.20	0.00	39.31	12.95	19.67	9.09	8.28	8.79	0.00	0.00	0.00	6.25	0.00	0.00
46	TakeLab	25.31	42.92	24.70	1.37	0.00	6.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
47	UTMNandOCAS	23.42	40.53	19.74	18.67	4.44	0.00	10.53	0.00	0.00	0.00	0.00	23.08	8.70	0.00	0.00

Table 7: **Technique classification F$_1$ performance on the development set.** The systems are ordered based on the final ranking on the test set (cf. Table 6), whereas the ranking is the one on the development set. Columns **1** to **14** show the performance on each class (cf. Section 2). The best score for each class is bold.

6 Related Work

Propaganda is particularly visible in the context of "fake news" on social media, which have attracted a lot of research recently (Shu et al., 2017). Thorne and Vlachos (2018) surveyed fact-checking approaches to fake news and related problems, and Li et al. (2016) focused on truth discovery in general. Two recent articles in *Science* offered a general discussion on the science of "fake news" (Lazer et al., 2018) and the process of proliferation of true and false news online (Vosoughi et al., 2018).

We are particularly interested here in how different forms of propaganda are manifested in text. So far, the computational identification of propaganda has been tackled mostly at the article level. Rashkin et al. (2017) created a corpus, where news articles are labeled as belonging to one of the following four categories: *propaganda*, *trusted*, *hoax*, or *satire*. The articles came from eight sources, two of which were propagandistic. The labels were obtained using distant supervision, assuming that all articles from a given news source share the label of that source, which introduces noise (Horne et al., 2018). Barrón-Cedeño et al. (2019b) experimented with a binary version of the problem: *propaganda* vs. *no propaganda*. See (Da San Martino et al., 2020a) for a recent survey on computational propaganda detection.

In general, propaganda techniques serve as a means to persuade people, often in argumentative settings. While they may increase the rhetorical effectiveness of arguments, they naturally harm other aspects of argumentation quality (Wachsmuth et al., 2017). In particular, many of the span propaganda techniques considered in this shared task relate to the notion of *fallacies*, i.e. arguments whose reasoning is flawed in some way, often hidden and often on purpose (Tindale, 2007). Some recent work in computational argumentation has dealt with such fallacies. Among these, Habernal et al. (2018) presented and analyzed a corpus of web forum discussions with *Ad hominem* fallacies, and Habernal et al. (2017) introduced *Argotario*, a game that educates people to recognize fallacies. Argotario also had a corpus as a by-product, with 1.3k arguments annotated for five fallacies, including *Ad hominem*, *Red herring*, and *Irrelevant authority*, which are related to some of our propaganda techniques (cf. Section 2). Unlike these corpora, the news articles in our corpus are annotated with fourteen propaganda techniques. Moreover, instead of labeling entire arguments, our annotation aims at identifying the minimal text spans related to a technique.

In the present SemEval task, we departed from the eighteen propaganda techniques and the corpus described in (Da San Martino et al., 2019b; Yu et al., 2019).[8] We used the news articles included in that corpus in a pilot task that ran in January 2019, the *Hack the News Datathon*,[9] as well as in a previous shared task, held as part of the *2019 Workshop on NLP4IF: Censorship, Disinformation, and Propaganda*.[10] Both the datathon and the shared task tackled the identification of propaganda techniques as one overall task (along with a binary sentence-level propaganda classification task), i.e. without splitting it into subtasks as we did here. As detailed in the overview paper of Da San Martino et al. (2019a), the best-performing systems in the shared task used BERT-based contextual representations. Other systems used contextual representations based on RoBERTa, Grover, and ELMo, or context-independent representations based on lexical, sentiment, readability, and TF-IDF features. As in the task at hand, ensembles were also popular. Still, the most successful submissions achieved an F_1-score of 24.88 only (and only 10.43 in the datathon). This is why, here we decided to split the task into subtasks in order to allow researchers to focus on one subtask at a time. Moreover, we merged some of the original 18 propaganda techniques to reduce data sparseness issues.

Other related shared tasks include the FEVER 2018 and 2019 tasks on *Fact Extraction and VERification* (Thorne et al., 2018), the SemEval 2017 and 2019 tasks on determining the veracity of rumors (Derczynski et al., 2017; Gorrell et al., 2019) and the SemEval 2019 task on *Fact-Checking in Community Question Answering Forums* (Mihaylova et al., 2019). Also, the CLEF 2018–2020 *CheckThat!* labs' shared tasks (Nakov et al., 2018; Elsayed et al., 2019a; Elsayed et al., 2019b; Barrón-Cedeño et al., 2020a; Barrón-Cedeño et al., 2020b), which featured tasks on automatic identification (Atanasova et al., 2018; Atanasova et al., 2019) and verification (Barrón-Cedeño et al., 2018; Hasanain et al., 2019; Hasanain et al., 2020; Shaar et al., 2020) of claims in political debates and in social media.

---

[8]You can also try the `Prta` system (Da San Martino et al., 2020b) online at: `http://www.tanbih.org/prta`

[9]`http://www.datasciencesociety.net/hack-news-datathon/`

[10]`http://www.netcopia.net/nlp4if/2019/`

7 Conclusion and Future Work

We have described SemEval-2020 Task 11 on Detection of Propaganda Techniques in News Articles. The task attracted the interest of a number of researchers: 250 teams signed up to participate, and 44 made submissions on the test dataset. We received 35 and 31 submissions for subtask SI and subtask TC, respectively. Overall, subtask SI (segment identification) was easier and all systems managed to improve over the baseline. However, subtask TC (technique classification) proved to be much more challenging, and some teams could not improve over our baseline.

In future work, we plan to extend the dataset to cover more examples as well as more propaganda techniques. We further plan to develop similar datasets for other languages.

Acknowledgments

References

Ola Altiti, Malak Abdullah, and Rasha Obiedat. 2020. JUST at SemEval-2020 Task 11: Detecting propaganda techniques using BERT pretrained model. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Anastasios Arsenos and Georgios Siolas. 2020. NTUAAILS at SemEval-2020 Task 11: Propaganda detection and classification with biLSTMs and ELMo. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 1: Check-worthiness. In *CLEF 2018 Working Notes*, Avignon, France.

Pepa Atanasova, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019. Overview of the CLEF-2019 CheckThat! Lab on automatic identification and verification of claims. Task 1: Check-worthiness. In *CLEF 2019 Working Notes*, Lugano, Switzerland.

Anastasios Bairaktaris, Symeon Symeonidis, and Avi Arampatzis. 2020. DUTH at SemEval-2020 Task 11: BERT with entity mapping for propaganda classification. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 2: Factuality. In *CLEF 2018 Working Notes*, Avignon, France.

Alberto Barrón-Cedeño and Paolo Rosso. 2009. On automatic plagiarism detection based on n-grams comparison. In *Proceedings of the 31th European Conference on IR Research*, ECIR '09, pages 696–700, Toulouse, France.

Alberto Barrón-Cedeño, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019a. Proppy: A system to unmask propaganda in online news. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI'19, pages 9847–9848, Honolulu, HI, USA.

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

---

[11]http://propaganda.qcri.org
[12]http://tanbih.qcri.org

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, Shaden Shaar, and Zien Sheikh Ali. 2020a. Overview of CheckThat! 2020 — automatic identification and verification of claims in social media. In *Proceedings of the 11th International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, CLEF '2020, Thessaloniki, Greece.

Alberto Barrón-Cedeño, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020b. CheckThat! at CLEF 2020: Enabling the automatic identification and verification of claims in social media. In *Proceedings of the 42nd European Conference on Information Retrieval*, ECIR '20, pages 499–507, Lisbon, Portugal.

Verena Blaschke, Maxim Korniyenko, and Sam Tureski. 2020. CyberWallE at SemEval-2020 Task 11: An analysis of feature engineering for ensemble models for propaganda detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Bolsover and Howard. 2017. Computational propaganda and political big data: Toward a more critical research agenda. *Big Data*, 5(4):273–276.

Akemi Takeoka Chatfield, Christopher G. Reddick, and Uuf Brajawidagda. 2015. Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided Twitter networks. In *Proceedings of the 16th Annual International Conference on Digital Government Research*, pages 239–249, Phoenix, AZ, USA.

Aniruddha Chauhan and Harshita Diddee. 2020. PsuedoProp at SemEval-2020 Task 11: Propaganda span detection using BERT-CRF and ensemble sentence level classifier. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Wei-Te Chen and Will Styler. 2013. Anafora: A web-based general purpose annotation tool. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Demonstration Session)*, NAACL-HLT '13, pages 14–19, Atlanta, GA, USA.

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 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019a. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF '19, pages 162–170, Hong Kong, China.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019b. 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 '19, pages 5636–5646, Hong Kong, China.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020a. A survey on computational propaganda detection. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence*, IJCAI-PRICAI '20, pages 4826–4832, Yokohama, Japan.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020b. Prta: A system to support the analysis of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL '20, pages 287–293.

Lavinia Dan. 2015. Techniques for the translation of advertising slogans. In *Proceedings of the International Conference Literature, Discourse and Multicultural Dialogue*, pages 13–23, Tirgu-Mures, Mures.

Jiaxu Dao, Jin Wang, and Xuejie Zhang. 2020. YNU-HPCC at SemEval-2020 Task 11: LSTM network for detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Guillaume Daval-Frerot and Weis Yannick. 2020. WMD at SemEval-2020 Tasks 7 and 11: Assessing humor and propaganda using unsupervised data augmentation. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

1395

Daryna Dementieva, Igor Markov, and Alexander Panchenko. 2020. SkoltechNLP at SemEval-2020 Task 11: Exploring unsupervised text augmentation for propaganda detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, pages 60–67, Vancouver, Canada.

Dimas Sony Dewantara, Indra Budi, and Muhammad Okky Ibrohim. 2020. 3218IR at SemEval-2020 Task 11: Conv1D and word embedding in propaganda span identification at news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Ilya Dimov, Vladislav Korzun, and Ivan Smurov. 2020. NoPropaganda at SemEval-2020 Task 11: A borrowed approach to sequence tagging and text classification. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019a. CheckThat! at CLEF 2019: Automatic identification and verification of claims. In *Proceedings of the 41st European Conference on Information Retrieval*, ECIR '19, pages 309–315, Cologne, Germany.

Tamer Elsayed, Preslav Nakov, Alberto Barrón-Cedeño, Maram Hasanain, Reem Suwaileh, Giovanni Da San Martino, and Pepa Atanasova. 2019b. Overview of the CLEF-2019 CheckThat!: Automatic identification and verification of claims. In *Proceedings of the 10th International Conference of the CLEF Association*, CLEF '19, pages 301–321, Lugano, Switzerland.

Vlad Ermurachi and Daniela Gifu. 2020. UAIC1860 at SemEval-2020 Task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, IJCNLP '17, pages 774–782, Taipei, Taiwan.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2015. Quantifying controversy in social media. *ACM Transactions on Social Computing*, 1(1):1–27, jan.

Monika Glowacki, Vidya Narayanan, Sam Maynard, Gustavo Hirsch, Bence Kollanyi, Lisa-Maria Neudert, Phil Howard, Thomas Lederer, and Vlad Barash. 2018. News and political information consumption in Mexico: Mapping the 2018 Mexican presidential election on Twitter and Facebook. Technical Report COMPROP DATA MEMO 2018.2, Oxford University, Oxford, UK, June.

Jean Goodwin. 2011. Accounting for the force of the appeal to authority. In *Proceedings of the 9th International Conference of the Ontario Society for the Study of Argumentation*, pages 1–9, Windsor, Canada.

Genevieve Gorrell, Ahmet Aker, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, SemEval '19, pages 845–854, Minneapolis, MN, USA.

Dmitry Grigorev and Vladimir Ivanov. 2020. Inno at SemEval-2020 Task 11: Leveraging pure transfomer for multi-class propaganda detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

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*, SemEval '18, pages 256–263, New Orleans, LA, USA.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP '17, pages 7–12, Copenhagen, Denmark.

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*, NAACL-HLT '18, pages 386–396, New Orleans, LA, USA.

Maram Hasanain, Reem Suwaileh, Tamer Elsayed, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Overview of the CLEF-2019 CheckThat! Lab on automatic identification and verification of claims. Task 2: Evidence and factuality. In *CLEF 2019 Working Notes*, Lugano, Switzerland.

Maram Hasanain, Fatima Haouari, Reem Suwaileh, Zien Sheikh Ali, Bayan Hamdan, Tamer Elsayed, Alberto Barrón-Cedeño, Giovanni Da San Martino, and Preslav Nakov. 2020. Overview of CheckThat! 2020 Arabic: Automatic identification and verification of claims in social media. In *Working Notes of CLEF 2020*, CLEF '2020, Thessaloniki, Greece.

Renee Hobbs and Sandra Mcgee. 2008. Teaching about Propaganda: An Examination of the Historical Roots of Media Literacy. *Journal of Media Literacy Education*, 6(62):56–67.

Benjamin D. Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Proceedings of the Twelfth International Conference on Web and Social Media*, ICWSM '18, pages 518–527, Stanford, CA, USA.

John Hunter. 2015. Brainwashing in a large group awareness training? The classical conditioning hypothesis of brainwashing. Master's thesis, University of Kwazulu-Natal, South Africa.

Ins. 1938. How to detect propaganda. In *Propaganda Analysis. Volume I*, chapter 2, pages 210–218. Publications of the Institute for Propaganda Analysis, New York, NY, USA.

Yunzhe Jiang, Cristina Gârbacea, and Qiaozhu Mei. 2020. UMSIForeseer at SemEval-2020 Task 11: Propaganda detection by fine-tuning BERT with resampling and ensemble learning. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Garth S. Jowett and Victoria O'Donnell. 2012a. *Propaganda and Persuasion*. SAGE, Los Angeles, CA, USA, 5th edition.

Garth S. Jowett and Victoria O'Donnell. 2012b. What is propaganda, and how does it differ from persuasion? In *Propaganda and Persuasion*, chapter 1, pages 1–48. Sage Publishing.

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 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Anders Friis Kaas, Viktor Torp Thomsen, and Barbara Plank. 2020. Team DiSaster at SemEval-2020 Task 11: Combining BERT and hand-crafted features for identifying propaganda techniques in news. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Sopan Khosla, Rishabh Joshi, Ritam Dutt, Alan W. Black, and Yulia Tsvetkov. 2020. LTIatCMU at SemEval-2020 Task 11: Incorporating multi-level features for multi-granular propaganda span identification. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Moonsung Kim and Steven Bethard. 2020. TTUI at SemEval-2020 Task 11: Propaganda detection with transfer learning and ensembles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Michael Kranzlein, Shabnam Behzad, and Nazli Goharian. 2020. Team DoNotDistribute at SemEval-2020 Task 11: Features, finetuning, and data augmentation in neural models for propaganda detection in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Gangeshwar Krishnamurthy, Raj Kumar Gupta, and Yinping Yangi. 2020. SocCogCom at SemEval-2020 Task 11: Detecting propaganda techniques in news articles using semantic-level emotional salience features. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

David M.J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Jinfen Li and Lu Xiao. 2020. HybridPro at SemEval-2020 Task 11: A hybrid model for propagandistic technique classification in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A survey on truth discovery. *SIGKDD Explor. Newsl.*, 17(2):1–16, February.

Matej Martinkovic, Samuel Pecar, and Marian Simko. 2020. NLFIIT at SemEval-2020 Task 11: Neural network architectures for detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479, sep.

Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, SemEval '19, pages 860–869, Minneapolis, MN, USA.

Elena Mikhalkova, Nadezhda Ganzherli, Anna Glazkova, and Yulia Bidulia. 2020. UTMN at SemEval-2020 task 11: A kitchen solution to automatic propaganda detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Clyde R. Miller. 1939. The Techniques of Propaganda. From "How to Detect and Analyze Propaganda," an address given at Town Hall. The Center for learning.

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 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Robert Muller. 2018. Indictment of Internet Research Agency. pages 1–37.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. In *Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, CLEF '18, pages 372–387, Avignon, France.

Andrei Paraschiv and Dumitru-Clementin Cercel. 2020. UPB at SemEval-2020 Task 11: Propaganda detection with domain-specific trained BERT. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Rajaswa Patil, Somesh Singh, and Swati Agarwal. 2020. BPGC at SemEval-2020 Task 11: Propaganda detection in news articles with multi-granularity knowledge sharing and linguistic features based ensemble learning. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015.

Maias Petee and Alexis Palmer. 2020. UNTLing at SemEval-2020 Task 11: Detection of propaganda techniques in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 997–1005, Beijing, China.

Mayank Raj, Ajay Jaiswal, Rohit R.R, Ankita Gupta, Sudeep Sahoo, Vertika Srivastava, and Kim Yeon Hyang. 2020. Solomon at SemEval-2020 Task 11: Novel ensemble architechture for fine-tuned propaganda detection in news articles. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

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*, EMNLP 17, pages 2931–2937, Copenhagen, Denmark.

Monika L. Richter. 2017. The Kremlin's platform for 'useful idiots' in the West: An overview of RT's editorial strategy and evidence of impact. *Kremlin Watch*, page 53.

Shaden Shaar, Alex Nikolov, Nikolay Babulkov, Firoj Alam, Alberto Barrón-Cedeño, Tamer Elsayed, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Giovanni Da San Martino, and Preslav Nakov. 2020. Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media. In *Working Notes of CLEF 2020*, CLEF '2020, Thessaloniki, Greece.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36, September.

Paramansh Singh, Siraj Sandhu, Subham Kumar, and Ashutosh Modi. 2020. newsSweeper at SemEval-2020 Task 11: Context-aware rich feature representations for propaganda classification. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Xin Tao and Xiaobing Zhou. 2020. YNUtaoxin at SemEval-2020 Task 11: Identification fragments of propaganda technique by neural sequence labeling models with different tagging schemes and pre-trained language model. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Cristina Tardáguila, Fabrício Benevenuto, and Pablo Ortellado. 2018. Fake news is poisoning brazilian politics. WhatsApp can stop it. https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html.

Gabriel H Teninbaum. 2009. Reductio ad hitlerum: Trumping the judicial Nazi card. *Michigan State Law Review*, 2009.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 3346–3359, Santa Fe, NM, USA.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 809–819, New Orleans, LA, USA.

Christopher W. Tindale. 2007. *Fallacies and Argument Appraisal. Critical Reasoning and Argumentation*. Cambridge University Press.

Robyn Torok. 2015. Symbiotic radicalisation strategies: Propaganda tools and neuro linguistic programming. In *Proceedings of the Australian Security and Intelligence Conference*, pages 58–65, Perth, Australia.

Ekansh Verma, Vinodh Motupalli, and Souradip Chakraborty. 2020. Transformers at SemEval-2020 Task 11: Propaganda fragment detection using diversified bert architectures based ensemble learning. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, SemEval '20, Barcelona, Spain.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '17, pages 176–187, Valencia, Spain.

Douglas Walton. 2013. The straw man fallacy. In *Methods of Argumentation*, pages 249–286. Cambridge University Press.

Anthony Weston. 2000. *A Rulebook for Arguments*. Hackett Student Handbooks.

Seunghak Yu, Giovanni Da San Martino, and Preslav Nakov. 2019. Experiments in detecting persuasion techniques in the news. In *Proceedings of the NeurIPS 2019 Joint Workshop on AI for Social Good*, NeurIPS '19, Vancouver, Canada, December.

A   Summary of all Submitted Systems

This appendix includes a brief summary of all systems for both subtasks. We present the teams in alphabetical order. The subindex on the right of each team represents its official test rank in the subtasks. The teams appearing in Tables 5, 6, or 7 but not here did not submit a paper describing their approach.

Team 3218IR (Dewantara et al., 2020)(SI:31)   used a one-dimensional convolutional neural network (CNN) with word embeddings, whose number of layers and filters as well as kernel and pooling sizes were all tuned empirically.

Team ApplicaAI (Jurkiewicz et al., 2020)(SI: 2, TC: 1)   applied self-supervision using the RoBERTa model. For the SI subtask, they used a RoBERTa-CRF architecture. The model trained using this architecture was then iteratively used to produce silver data by predicting on 500k sentences and retraining the model with both gold and silver data. As for subtask TC, ApplicaAI opted for feeding their models with propagandas snippets in context. Full sentences are shaped as the input with the specific propaganda in them. Once again, silver data was used, taking advantage of the spans detected by their SI model and labeling with their preliminary TC model. The final classifier was an ensemble of models trained on the original corpus, re-weighting, and a model trained also on silver data.

Team aschern (Chernyavskiy et al., 2020)(SI: 3, TC: 2)   tackled both subtasks. For SI, they fine-tuned an ensemble of two differently intialized RoBERTa models, each with an attached CRF for sequence labeling and simple span character boundary post-processing. A RoBERTa ensemble was also used for TC, treating the task as sequence classification but using an average embedding of the surrounding tokens and the length of a span as contextual features. They further used transfer learning, to pass knowledge about the SI subtask to help the TC subtask. Finally, specific postprocessing was done to increase the consistency of the repetition technique spans and to avoid insertions of techniques in other techniques.

Team BPGC (Patil et al., 2020)(SI: 21, TC: 18)   used a multi-granularity approach to address subtask SI. Information about the article and the sentence was considered when classifying each word as propaganda or not, by means of computing and concatenating vectorial representations for the three inputs. For subtask TC, they used an ensemble of BERT and logistic regression, complemented with engineered features which, as stated by the authors, were particularly useful for the smaller classes. Such features include TF.IDF vectors of words and character n-grams, topic modeling, and sentence-level polarity, among others. Different learning models were explored for both tasks, including LSTM and CNN, together with diverse Transformers to build ensembles of classifiers.

Team CyberWallE (Blaschke et al., 2020)(SI: 8, TC: 8)   used BERT embeddings for subtask SI, as well as manual features modeling sentiment, rhetorical structure, and POS tags, which were eventually fed into a bi-LSTM to produce IO labels, followed by some post-processing to merge neighboring spans. For subtask TC, they extracted the pre-softmax layer of BERT and further added extra features (rhetorical, named entities, question), while taking special care of repetitions as part of a complex ensemble architecture, followed by label post-processing.

Team DiSaster (Kaas et al., 2020)(TC:11)   used a combination of BERT and hand-crafted features, including frequency of the fragment in the article and in the sentence it appears in and the inverse uniqueness of words in a span. The goal of the features is to compensate the inability of BERT to deal with distant context, specifically to target the technique *Repetition*.

Team DoNotDistribute (Kranzlein et al., 2020)(SI: 22, TC: 24)   opted for a combination of BERT-based models and engineered features (e.g., PoS, NEs, frequency within propaganda snippets in the training set). A reported performance increase of 5% was obtained by producing 3k new silver training instances. A library was used to create near-paraphrases of the propaganda snippets by randomly substituting certain PoS words.

Team DUTH (Bairaktaris et al., 2020)(TC:10) pre-processed the input including URL normalization, number and punctuation removal, as well as lowercasing. They further mapped certain words into classes using named entity recognition with focus on person names and gazetteers containing names and variations of names of countries (255 entries), religions (35 entries), political ideologies (23 entries), and slogans (41 entries). The recognized categories were replaced by the category name in the input, before passing the input to BERT.

Team Hitachi (Morio et al., 2020)(SI: 1, TC: 3) used BIO encoding for subtask SI, which is typical for related segmentation and labeling tasks such as named entity recognition. They used a complex heterogeneous multi-layer neural network, trained end-to-end. The network used a pre-trained language model, which generates a representation for each input token. To this were added part-of-speech (PoS) and named entity (NE) embeddings. As a result, there were three representations for each token, which were concatenated and used as an input to bi-LSTMs. At this moment, the network branches as it is trained with three objectives: (*i*) the main BIO tag prediction objective, and two auxiliary objectives, namely (*ii*) token-level technique classification, and (*iii*) sentence-level classification. There is one Bi-LSTM for objectives (*i*) and (*ii*), and there is another Bi-LSTM for objective (*iii*). For the former, there is an additional CRF layer, which helps improve the consistency of the output. *For subtask TC*, there are two distinct FFNs, feeding input representation, which are obtained in the same manner as for subtask SI. One of the two FFNs is for sentence representation, and the other one is for the representation of tokens in the propaganda span. The propaganda span representation is obtained by concatenating representation of the begin-of-sentence token, span start token, span end token, and aggregated representation by attention and max-pooling. *For both subtasks*, these architectures were trained independently with different BERT, GPT-2, XLNet, XLM, RoBERTa, or XLM-RoBERTa Transformers; and the resulting models were combined in ensembles.

Team Inno (Grigorev and Ivanov, 2020)(TC:7) used RoBERTa with cost-sensitive learning for subtask TC. They experimented with undersampling, i.e. removing examples from the bigger classes, as well as with modeling the context. They also tried various pre-trained Transformers, but obtained worse results.

Team JUST (Altiti et al., 2020)(TC:15) based its approach to the task on the BERT uncased pre-trained language model, which used 12 transformer layers that were trained for 15 epochs.

Team LTIatCMU (Khosla et al., 2020)(SI:4) used a multi-granular BERT BiLSTM for subtask SI. It used additional syntactic, semantic and pragmatic affect features at the word, sentence and document level. It was jointly trained on token and sentence propaganda classification, with class balancing. In addition, BERT was fine-tuned to persuasive language on about 10,000 articles from propaganda websites, which turned out to be important in their experiments.

Team newsSweeper (Singh et al., 2020)(SI: 13, TC: 5) used BERT with BIOE encoding for subtask SI. For the TC subtask, their official run used RoBERTa to obtain representations for the span and for the sentence, which they concatenated. The team further experimented (*i*) with other Transformers (BERT, RoBERTa, SpanBERT, and GPT-2), (*ii*) with other sequence labeling schemes (P/NP, BIO, BIOES), (*iii*) with concatenating different hidden layers of BERT to obtain a token representation, and (*iv*) with POS tags, as well as (*v*) with different neural architectures.

Team NLFIIT (Martinkovic et al., 2020)(SI: 17, TC: 16) used various combinations of neural architecture and embeddings and found out that ELMo combined with BiLSTM (and self attention for subtask TC) yielded the best performance.

Team NoPropaganda (Dimov et al., 2020)(SI: 7, TC: 6) used the LasetTagger model with the BERT-base encoder for subtask SI. R-BERT was used for subtask TC.

Team NTUAAILS (Arsenos and Siolas, 2020)(SI: 27, TC: 25) used a residual biLSTM fed with pre-trained ELMo embeddigns for subtask SI. A biLSTM was used for subtask TC as well, but this time fed with GloVe word embeddings

Team PsuedoProp (Chauhan and Diddee, 2020)(SI:14) focused on subtask SI. They pre-classified sentences as propaganda or not using an ensemble of XLNet and RoBERTa, before fine-tuning a BERT-based CRF sequence tagger to identify the exact spans.

Team SkoltechNLP (Dementieva et al., 2020)(SI: 25, TC: 26) fine-tuned BERT for SI, expanding the original training set through data augmentation techniques based on distributional semantics.

Team SocCogCom (Krishnamurthy et al., 2020)(TC:13) approached subtask TC using BERT/ALBERT together with (*i*) semantic-level emotional salience features from CrystalFeel (Gupta and Yang, 2018), and (*ii*) word-level psycholinguistic features from the LIWC lexicon (Pennebaker et al., 2015). They further modeled the context, i.e. three words before and after the target propaganda snippet.

Team Solomon (Raj et al., 2020)(TC:4) addressed subtask TC with a system that combines a transfer learning model based on fine-tuned RoBERTa (integrating fragment and context information), an ensemble of binary classifiers for the smaller classes and a novel system to specifically handle *Repetition*: they used dynamic least common sub-sequence to assess the similarity between the fragment and the context, and then the fragment was considered to be a repetition if the score was greater than a threshold heuristically set with respect to the length of the fragment.

Team syrapropa (Li and Xiao, 2020)(SI: 25, TC: 20) fine-tuned SpanBERT, a variant of BERT for span detection, on the context of spans in terms of the surrounding non-propaganda text for subtask SI. For subtask TC, they used a hybrid model that consists of several submodels, each specializing in some of the relations. These models include (*i*) BERT, (*ii*) BERT with cost adjustment to address class imbalance, and (*iii*) feature-rich logistic regression. The latter uses features such as length, TF.IDF-weighted words, repetitions, superlatives, and lists of fixed phrases targeting specific propaganda techniques. The output from the hybrid model was further post-processed using some syntactic rules based on part of speech.

Team Transformers (Verma et al., 2020)(SI: 9, TC: 29) explored a manifold of models to address the SI subtask. They considered residual biLSTMs fed with ELMo representations as well as different variations of BERT and RoBERTa with CNNs

Team TTUI (Kim and Bethard, 2020)(SI: 20, TC: 14) proposed an ensemble of fine-tuned BERT and RoBERTa models. They observed that feeding as input to the neural network a chunk of multiple, possibly overlapping sentences yielded the best performance. Moreover, for subtask SI, they applied a post-processing to remove gaps in the predictions between adjacent words. For subtask TC, they showed that modeling the context did not help in their experiments.

Team UAIC1860 (Ermurachi and Gifu, 2020)(SI: 28, TC: 26) used traditional text representation techniques: character n-grams, word2vec embeddings, and TF.IDF-weighted word-based features. For both subtasks, these features were used in a Random Forest classifier. Additional experiments with Naïve Bayes, Logistic Regression and SVMs yielded worse results.

Team UMSIForeseer Jiang et al. (2020)(TC:17) focused on subtask TC. They fine-tuned BERT on the labeled training spans, using a mix of oversampling and undersampling that is leveraged using a bagging ensemble learner.

Team UNTLing (Petee and Palmer, 2020)(TC:27) used a logistic regression classifier for subtask TC with a number of features, including bag-of-words, embeddings, NE and VAD lexicon features. Their analysis highlights that NE are useful for *Loaded Language* and *Flag Waving*. The VAD features were useful for emotion-related techniques such as *Appeal to fear/prejudice* and *Doubt*. They performed some experiments on the development set for subtask SI after the deadline. They used CRF with a number of features including PoS, syntactic dependency between the token and the previous/next word, BoW of preceding/following tokens, and the GloVe embedding of the token.

Team UPB (Paraschiv and Cercel, 2020)(SI: 5, TC: 19) used models based on BERT-base. Rather than just using the pre-trained models, they used masked language models to domain-adapt it with 9M-articles with fake, suspicious, and hyperpartisan news articles. They used the same domain-adapted model for both subtasks. They further used CRF for subtask SI, and a softmax for subtask TC.

Team UTMN (Mikhalkova et al., 2020)(SI:23) addressed subtask SI by representing the texts as a concatenation of tokens and context embeddings, together with sentiment intensity from VADER. They avoided deep learning architectures in order to produce a computationally affordable model, namely logistic regression.

Team WMD (Daval-Frerot and Yannick, 2020)(SI: 33, TC: 21) used an ensemble of BERT-based models, LSTMs, SVMs, gradient boosting, and random forest together with character and word-level embeddings. In addition, they used a number of techniques for data augmentation: back-translation, synonym replacement and TF.IDF replacement, i.e., replacing unimportant words, according to their TF.IDF score, with other unimportant words.

Team YNU-HPCC (Dao et al., 2020)(SI: 16, TC: 22) participated in both subtasks using GloVe and BERT embeddings in combination with LSTMs, BiLSTMs, and XGBoost.

Team YNUtaoxin (Tao and Zhou, 2020)(SI:11) used BERT, RoBERTa and XLNet on subtask SI focusing on determining the optimal input sentence length for the networks.

B Errata

After the shared task has ended, we found a bug in the code for our evaluation tools, which affected both subtasks. Overall, its impact was limited, and the ranking computed with the fixed code did not change substantially, in particular for the top-ranked teams. Tables 8 and 9 show the corrected scores on the test sets for subtasks SI and TC, respectively. Any reference to the task results should refer to these numbers.

Team	Rnk	Test F_1	P	R
Hitachi	1	**51.74**	55.76	48.27
ApplicaAI	2	49.88	59.33	43.02
aschern	3	49.59	52.57	46.93
LTIatCMU	4	48.16	50.35	46.15
UPB	5	46.63	57.50	39.22
Fragarach	6	46.43	53.44	41.05
NoPropaganda	7	45.17	55.05	38.29
YNUtaoxin	8	43.80	54.60	36.57
Transformers	9	43.77	49.05	39.52
SWEAT	10	43.69	52.13	37.61
DREAM	11	43.60	53.67	36.71
CyberWallE	12	43.59	40.99	46.54
newsSweeper	13	42.20	45.30	39.49
PsuedoProp	14	41.81	41.24	42.41
Solomon	15	41.26	53.69	33.51
NLFIIT	16	41.10	50.17	34.81
TTUI	17	40.76	**66.37**	29.41
PALI	18	40.73	52.10	33.44
YNUHPCC	19	40.46	36.35	45.63
UESTCICSA	20	40.41	55.58	31.74
BPGC	21	38.89	48.50	32.45
DoNotDistribute	22	37.92	41.47	34.92
UTMNandOCAS	23	37.71	37.12	38.31
Entropy	24	37.31	40.82	34.35
syrapropa	25	36.92	49.22	29.53
SkoltechNLP	26	34.36	45.77	27.51
NTUAAILS	27	34.36	45.62	27.55
UAIC1860	28	32.67	23.86	51.78
CCNI	29	29.68	37.73	24.46
NCCU-SMRG	30	27.69	16.70	**80.94**
3218IR	31	23.28	21.95	24.79
WMD	32	20.51	45.44	13.24
LS	33	18.14	33.20	12.47
HunAlize	34	3.15	22.39	1.69
YOLO	35	0.74	18.32	0.38
Baseline	36	0.32	13.04	0.16

Table 8: **Subtask 1: Span Identification (SI) performance on the test set using the fixed scorer.** The highest score for each measure is highlighted.

Rnk	Team	Overall	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	ApplicaAI	**63.74**	48.15	**47.06**	8.33	50.98	22.73	56.23	36.64	**70.47**	78.27	75.80	**58.79**	**58.97**	39.22	28.13
2	aschern	63.30	35.29	41.78	14.93	**53.57**	25.97	59.44	37.55	66.35	**78.32**	**76.68**	56.94	54.24	**42.55**	22.64
3	Hitachi	63.13	40.00	38.94	4.88	47.92	**36.62**	**63.43**	37.38	69.04	76.69	76.22	42.42	52.63	29.41	26.92
4	NoPropaganda	59.83	23.26	38.02	0.00	8.70	12.12	56.27	**39.43**	60.30	77.99	76.61	48.14	51.52	27.91	0.00
5	Solomon	59.39	26.67	39.39	4.88	42.00	19.61	61.82	32.00	61.84	75.30	70.75	42.53	50.75	38.10	0.00
6	CyberWallE	58.99	14.63	39.84	12.25	26.32	15.39	59.24	33.20	56.22	75.81	71.60	52.98	45.46	23.81	0.00
7	newsSweeper	58.44	41.67	42.80	13.04	25.00	19.72	56.55	37.10	60.53	75.32	74.23	20.69	50.75	21.62	8.00
8	Inno	57.99	43.14	45.09	10.71	40.40	24.32	58.65	35.39	59.41	73.31	74.30	24.89	43.75	29.63	19.36
9	djichen	57.71	37.50	34.58	20.83	43.81	22.79	60.00	32.18	56.19	73.99	70.43	36.36	30.77	27.91	18.87
10	PALI	57.43	27.59	36.59	**28.57**	39.62	30.59	58.97	28.57	61.62	74.29	69.43	25.44	39.22	40.82	20.90
11	DUTH	57.21	44.44	33.33	9.30	33.33	26.23	59.16	28.24	58.95	73.71	71.41	20.10	34.78	27.03	17.78
12	DiSaster	56.65	**51.16**	35.25	20.41	26.67	14.63	59.12	30.64	58.25	74.49	68.10	20.44	42.55	19.05	4.35
13	SocCogCom	55.81	28.57	31.62	9.76	26.51	6.35	60.40	34.86	54.26	72.18	67.34	18.88	40.91	23.53	10.00
14	TTUI	55.64	28.24	41.48	24.56	35.29	23.08	57.40	32.20	61.68	73.22	68.49	21.18	37.50	25.00	20.29
15	JUST	55.36	48.89	37.10	24.49	31.11	27.27	58.26	29.57	62.56	71.96	64.73	21.94	33.33	31.82	**28.57**
16	NLFIIT	55.25	29.63	24.89	9.30	32.10	0.00	55.66	30.30	63.32	72.55	69.30	21.55	41.67	13.64	0.00
17	UMSIForeseer	55.14	29.41	31.97	9.76	0.00	0.00	57.21	28.57	56.14	73.02	70.79	21.49	39.22	14.29	0.00
18	BPGC	54.81	30.77	33.65	12.50	40.00	24.18	53.78	33.47	58.93	71.58	67.51	23.74	40.00	20.69	20.90
19	UPB	54.30	20.00	30.00	0.00	19.72	16.95	52.55	30.62	55.87	70.09	68.86	20.00	34.62	22.86	4.88
20	syrapropa	54.25	31.58	29.77	19.61	30.61	21.51	56.14	28.10	57.02	71.47	68.44	30.77	29.03	28.57	9.09
21	WMD	52.35	7.14	29.20	5.00	0.00	5.71	54.42	26.36	52.08	69.76	64.67	14.82	6.90	7.41	0.00
22	YNUHPCC	50.50	29.79	26.40	13.79	38.38	3.45	51.04	21.54	55.56	68.08	62.33	17.72	27.59	17.78	15.00
23	UESTCICSA	49.94	12.77	22.75	0.00	11.27	3.74	44.99	25.44	53.14	68.23	66.88	27.96	41.38	28.57	3.70
24	DoNotDistribute	49.72	22.64	29.75	4.55	24.49	14.89	46.25	27.23	53.76	68.44	60.65	19.44	28.07	12.25	9.68
25	NTUAAILS	46.37	36.36	28.31	9.76	0.00	0.00	48.75	29.66	46.47	65.79	54.55	18.43	13.79	11.43	4.08
26	UAIC1860	41.17	17.39	22.12	9.76	2.90	7.60	36.41	21.01	38.78	62.33	42.97	11.16	11.43	5.56	4.26
27	UNTLing	39.11	4.35	5.29	0.00	0.00	2.86	32.65	11.82	40.48	62.57	36.74	7.78	17.65	0.00	0.00
28	HunAlize	37.10	6.45	11.83	8.00	8.00	7.84	31.76	23.81	29.95	58.59	15.82	2.09	4.55	0.00	0.00
29	Transformers	26.54	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	47.55	24.06	2.86	0.00	0.00	0.00
30	Baseline	25.20	0.00	3.68	0.00	0.00	11.56	29.14	14.42	6.20	46.48	0.00	19.26	0.00	0.00	0.00
31	IJSE8	20.62	0.00	7.07	0.00	0.00	2.27	17.07	10.70	8.57	39.32	15.39	4.92	0.00	0.00	0.00
32	Entropy	20.50	3.92	7.12	0.00	2.27	7.41	12.81	6.39	4.95	38.02	15.14	5.83	0.00	0.00	6.78

Table 9: **Technique classification F$_1$ performance on the test set using the fixed scorer.** The systems are ordered based on the final ranking. Columns **1** to **14** show the performance for each of the fourteen propaganda techniques (cf. Section 2). The best score for each technique is shown in bold.

C   Annotation Instructions

Below, we show a series of snapshots of the actual annotation instructions and propaganda techniques definitions and examples that we showed to the human annotators. These are also available online:

- `http://propaganda.qcri.org/annotations/`

- `https://propaganda.qcri.org/annotations/definitions.html`

Figure 7: Instruction for the annotators.
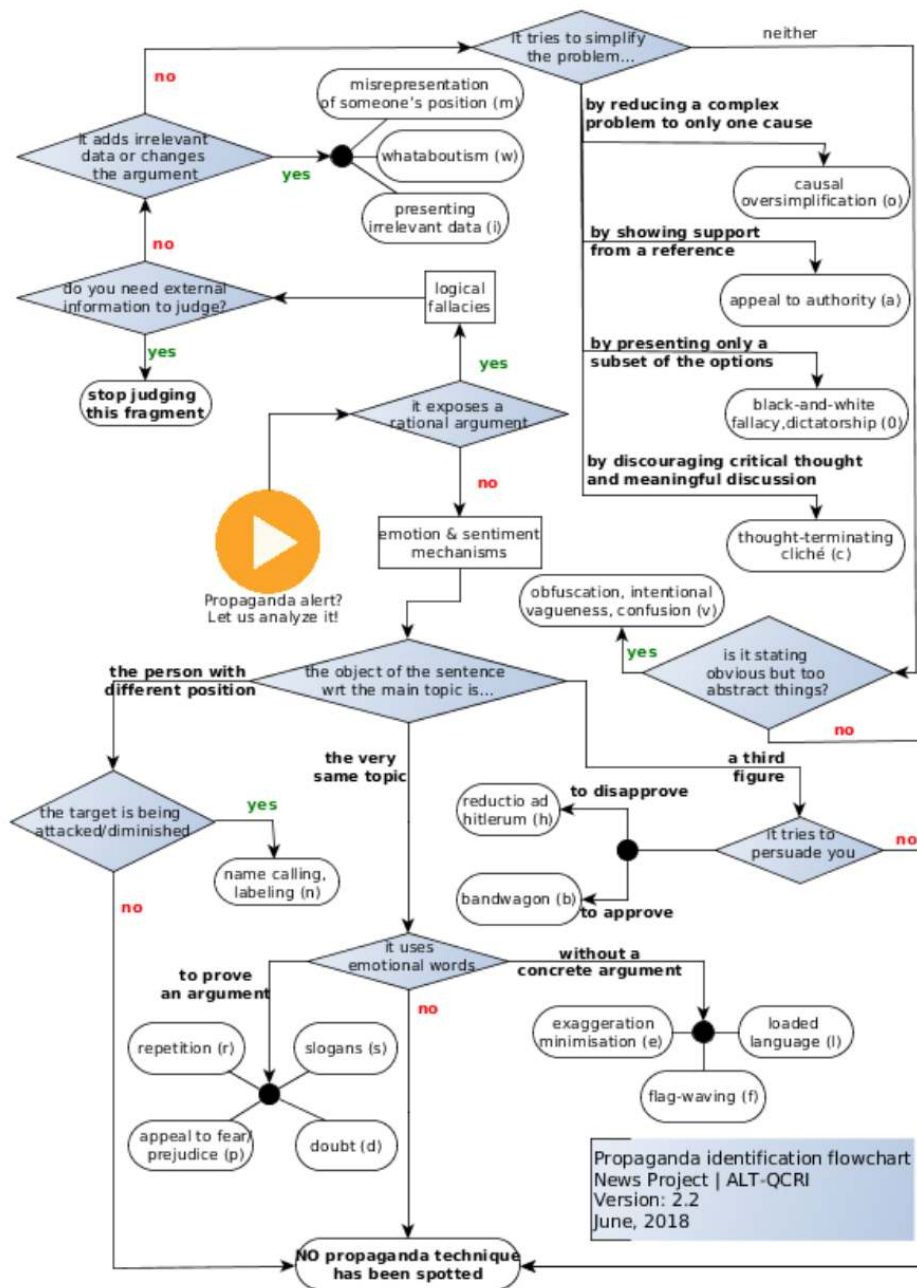
Figure 8: Annotation instructions: hierarchical diagram to guide the choice of technique.

Definitions

1. Presenting Irrelevant Data (Red Herring)

Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made.

Example 1: In politics, defending one's own policies regarding public safety - "I have worked hard to help eliminate criminal activity. What we need is economic growth that can only come from the hands of leadership."

Example 2: "You may claim that the death penalty is an ineffective deterrent against crime -- but what about the victims of crime? How do you think surviving family members feel when they see the man who murdered their son kept in prison at their expense? Is it right that they should pay for their son's murderer to be fed and housed?"

2. Misrepresentation of Someone's Position (Straw Man)

When an opponent's proposition is substituted with a similar one which is then refuted in place of the original proposition.

Example: Zebedee: What is your view on the Christian God?
Mike: I don't believe in any gods, including the Christian one.
Zebedee: So you think that we are here by accident, and all this design in nature is pure chance, and the universe just created itself?
Mike: You got all that from me stating that I just don't believe in any gods?
Explanation: Mike made one claim: that he does not believe in any gods. From that, we can deduce a few things, like he is not a theist, he is not a practicing Christian, Catholic, Jew, or a member of any other religion that requires the belief in a god, but we cannot deduce that he believes we are all here by accident, nature is chance, and the universe created itself.

3. Whataboutism

A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

Example 1: a nation deflects criticism of its recent human rights violations by pointing to the history of slavery in the United States.

Example 2: "Qatar spending profusely on Neymar, not fighting terrorism"

4. Causal Oversimplification

Assuming a single cause or reason when there are actually multiple causes for an issue.
It includes transferring blame to one person or group of people without investigating the complexities of the issue

Example 1: "President Trump has been in office for a month and gas prices have been skyrocketing. The rise in gas prices is because of President Trump."

Example 2: The reason New Orleans was hit so hard with the hurricane was because of all the immoral people who live there.

Explanation: This was an actual argument seen in the months that followed hurricane Katrina. Ignoring the validity of the claims being made, the arguer is blaming a natural disaster on a group of people.

Example 3: if France had not have declared war on Germany then world war two would have never happened.

5. Obfuscation, Intentional vagueness, Confusion

Using words which are deliberately not clear so that the audience may have its own interpretations.
For example when an unclear phrase with multiple definitions is used within the argument and, therefore, it does not support the conclusion.

Example: It is a good idea to listen to victims of theft. Therefore if the victims say to have the thief shot, then you should do that.
Explanation: the definition for "listen to" is equivocated here. In the first case it means listen to their personal account of the experience of being a victim of theft. Empathize with them. In the second case "listen to" means carry out a punishment of their choice.

6. Appeal to authority

Stating that a claim is true simply because a valid authority or expert on the issue said it wa true, without any other supporting evidence offered. We consider the special case in which the reference is not an authority or an expert in this technique, altough it is referred to as Testimonial in literature.

Example: Richard Dawkins, an evolutionary biologist and perhaps the foremost expert in the field, says that evolution is true. Therefore, it's true.
Explanation: Richard Dawkins certainly knows about evolution, and he can confidently tell us that it is true, but that doesn't make it true. What makes it true is the preponderance of evidence for the theory.
Example 2: "According to Serena Williams, our foreign policy is the best on Earth. So we are in the right direction."
Details: since there is a chance that any authority can be wrong, it is reasonable to defer to an authority to support a claim, but the authority should not be the only justification to accept the claim, otherwise the Appeal-to-Authority fallacy is committed.

7. Black-and-white Fallacy, Dictatorship

Presenting two alternative options as the only possibilities, when in fact more possibilities exist. As an the extreme case, tell the audience exactly what actions to take, eliminating any other possible choices (Dictatorship).

Example 1: You must be a Republican or Democrat. You are not a Democrat. Therefore, you must be a Republican
Example 2: I thought you were a good person, but you weren't at church today.
Explanation: The assumption here is that if one doesn't attend chuch, one must be bad. Of course, good people exist who don't go to church, and good church-going people could have had a really good reason not to be in church.
Example 3: There is no alternative to war

8. Name calling or labeling

Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or loves, praises.

Examples: "Republican congressweasels", "Bush the Lesser" (note that lesser does not refer to "the second", but it is pejorative)

9. Loaded Language

Using specific words and phrases with strong emotional implications (either positive or negative) to influence an audience.

Example 1: "[...] a lone lawmaker's childish shouting. ".
Example 2: "how stupid and petty things have become in Washington"

10. Exaggeration or Minimisation

Either representing something in an excessive manner: making things larger, better, worse (e.g., "the best of the best", "quality guaranteed") or making something seem less important or smaller than it really is (e.g., saying that an insult was just a joke).

Example 1: "Democrats bolted as soon as Trump's speech ended in an apparent effort to signal they can't even stomach being in the same room as the president "
Example 2: "We're going to have unbelievable intelligence"
Example 3: I was not fighting with her; we were just playing.

11. Flag-waving

Playing on strong national feeling (or to any group; e.g., race, gender, political preference) to justify or promote an action or idea

Example 1: "patriotism mean no questions" (this is also a [slogan](#))
Example 2: "entering this war will make us have a better future in our country."

12. Doubt

Questioning the credibility of someone or something.

Example: A candidate talks about his opponent and says: Is he ready to be the Mayor?

13. Appeal to fear/prejudice

Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative.
In some cases the support is built based on preconceived judgements.

Example 1: "either we go to war or we will perish" (this is also aBlack and White
  fallacy))
Example 2: "we must stop those refugees as they are terrorists"

14. Slogans

A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals.

Example 1: "The more women at war . . . the sooner we win."
Example 2: "Make America great again!"

15. Thought-terminating cliché

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.

Examples: It is what it is; It's just common sense; You gotta do what you gotta do;
  Nothing is permanent except change; Better late than never; Mind your
  own business; Nobody's perfect; It doesn't matter; You can't change
  human nature.

16. Bandwagon

Attempting to persuade the target audience to join in and take the course of action because "everyone else is taking the same action".

Example 1: Would you vote for Clinton as president? 57% say yes
Example 2: 90% of citizens support our initiative. You should.

17. 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.

Example 1: "Do you know who else was doing that ? Hitler!"
Example 2: "Only one kind of person can think in that way: a communist."

18. Repetition

Repeating the same message over and over again so that the audience will eventually accept it.