Pythoneers at WANLP 2022 Shared Task: Monolingual AraBERT for Arabic Propaganda Detection and Span Extraction

Joseph Attieh
Huawei Technologies Oy., Finland
joseph.attieh@huawei.com

Fadi Hassan
Huawei Technologies Oy., Finland
fadi.hassan@huawei.com

Abstract

In this paper, we present two deep learning approaches that are based on AraBERT, submitted to the Propaganda Detection shared task of the Seventh Workshop for Arabic Natural Language Processing (WANLP 2022). Propaganda detection consists of two main sub-tasks, mainly propaganda identification and span extraction. We present one system per sub-task. The first system is a Multi-Task Learning model that consists of a shared AraBERT encoder with task-specific binary classification layers. This model is trained to jointly learn one binary classification task per propaganda method. The second system is an AraBERT model with a Conditional Random Fields (CRF) layer. We achieved rank 3 on the first sub-task and rank 1 on the second sub-task.

1 Introduction

Social media platforms have been one of the main mediums of communication and source of information for most internet users. These platforms, as useful as they might be, can also be used to deceive and manipulate individuals. This is mostly done through propaganda techniques. Propaganda can be defined as the expression of opinion that is crafted to deliberately manipulate people's beliefs, attitudes, or actions, achieving a set of specified goals (Smith, 2021). This is done by presenting certain arguments to divert the attention of the victims from everything but their own propaganda. Since fallacies and propaganda devices overlap, researchers have defined propaganda techniques in terms of argumentative fallacies (Miller, 1939; Weston, 2018).

Several initiatives were made to detect propaganda on social media. For instance, Da San Martino et al. (2019b) provided a fine-grained propaganda analysis and a corpus of news articles annotated with 18 propaganda techniques. This corpus was employed at SemEval-2020 for propaganda identification (Martino et al., 2020), then at NLP4IF-2020 for span detection respectively (Da San Martino et al., 2019a).

In this paper, we present our solution to the Propaganda 2022 shared task (Alam et al., 2022). The Propaganda 2022 shared task is one of the first shared tasks of its kind and is held with the 7th Arabic Natural Language Processing Workshop (WANLP 2022) co-located with the EMNLP 2022 Conference in Abu Dhabi (Dec 7, 2022). The goal of the task is to build models for identifying propaganda techniques in Arabic tweets. It provides two sub-tasks; the goal of the first sub-task is to detect the propaganda technique used in the tweet (if any), while the goal of the second sub-task is to identify the span of the text covered by each technique.

As mentioned by Da San Martino et al. (2019a), the best-performing systems in the propaganda shared tasks used Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) to generate contextual representations of the text. Therefore, we propose to fine-tune an Arabic variant of BERT called AraBERT for each sub-task. The system submitted to the first sub-task is a multi-task model that performs binary classification per propaganda technique. The system submitted for the second sub-task is an AraBERT model fine-tuned with a Conditional Random Fields (CRF) layer. Both systems achieved top rankings on the leaderboard; the first system ranked third with a micro-averaged F1-Score of 0.602, while the second system ranked first with a micro-averaged F1-Score of 0.396.

This paper is structured as follows: Section 2 describes the data used for each sub-task, as well as the data preprocessing techniques employed. Section 3 gives an overview of the fine-tuning process of BERT models. Section 4 presents the systems submitted to sub-tasks 1 and 2 respectively. In Section 5, we show the results and discuss them briefly. Finally, we present the related work section in Section 6 and conclude the paper with Section 7.

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 534 - 540
December 8, 2022 ©2022 Association for Computational Linguistics

2 Data

2.1 Overall Description

The following propaganda task covers around 20 propaganda techniques, defined in terms of logical argumentative fallacies[1].

2.2 Dataset Split

Both systems presented in this paper are solely trained and validated on the data provided by the organizer. The training sets (i.e., train) for both sub-tasks consist of around 500 tweets each, while the development sets (i.e., dev and dev_test) consist of around 50 tweets each. The first sub-task provides the tweets labeled with the propaganda techniques present in these tweets. It should be noted that multiple propaganda techniques might be present in the same tweet. Tweets with no propaganda technique are labeled with "no technique". The second sub-task presents the tweets with the propaganda methods employed in each tweet with their span (i.e., start and end indexes of the text fragment containing the propaganda technique provided). It should be noted that both sub-tasks share the same tweets. The label distribution amongst the different sets is provided in the results sections in Table 2 for conciseness (the mismatch in the number of labels between the first sub-task and the second sub-task is because every propaganda technique can have multiple spans in the same text).

2.3 Dataset Preprocessing

2.3.1 Sub-task 1

The first sub-task is a multi-label classification task. We first standardize the text by removing non-Arabic words, emojis, and URLs from the tweets. Then, we proceed by tokenizing the tweets using the AraBERT tokenizer.

2.4 Sub-task 2

The second sub-task is a sequence tagging task. Therefore, we encode the input text based on the spans that represent the propaganda techniques. We experimented with different encoding schemes, displayed in Table 1. Preliminary experiments conducted with these encoding schemes showed that the *BIO data format* results in better performance for the task [2]. Therefore, we employ this format for the data.

[1]The propaganda techniques are defined in the following link: https://propaganda.qcri.org/annotations/definitions.html

[2]Results are not reported for conciseness.

Table 1: Encoding formats (**LL = Loaded Language** and **NC = Name calling/Labeling**)

Data Format	**Notations**		**Encoding** صدمة في تركيا بعد هذا القرار الروسي
BIO	B	first token in a span	B-LL O O O
	I	token in a span	O B-NC I-NC
	O	token outside of a span	
BIOUL	B	first token in a span	U-LL O O O
	I	non-first and non-last token in a span	O B-NC L-NC
	O	token outside of a span	
	U	unit-length span (span same size as token)	
	L	last token in a multi-token span	
IO	I	token in a span	I-LL O O O
	O	token outside a span	O I-NC I-NC

3 Fine-tuning BERT

As mentioned previously, the first sub-task is a multi-label text classification task, while the second sub-task is a sequence tagging task. We choose to fine-tune a pre-trained Bidirectional Encoder Representation from Transformer (BERT) model (Devlin et al., 2019) for each of these sub-tasks. This is usually done by adding an appropriate output layer to the BERT encoder and training the parameters of the network to predict correctly for the corresponding sub-task. It is a direct application of Transfer Learning, as the knowledge from the pre-trained model is transferred to the downstream task.

Therefore, finding an appropriate pre-trained model to fine-tune highly affects the performance of the model on the sub-task. Since we are dealing with Arabic tweets, we choose to build our systems using the Arabic pre-trained language model called AraBERT (Antoun et al., 2020). The specific model employed in both sub-tasks is the *bert-large-arabertv02-twitter*. It is based on *AraBERTv0.2-large*, first pre-trained on publicly available large-scale raw Arabic text, and then pre-trained again on 60M Multi-Dialect Tweets.

For the first sub-task, we propose to employ Multi-Task Learning to fine-tune AraBERT on the multi-label text classification task. As for the second sub-task, we propose to employ a CRF layer to fine-tune BERT for the sequence tagging task. All models have been trained on NVIDIA Tesla Volta V100.

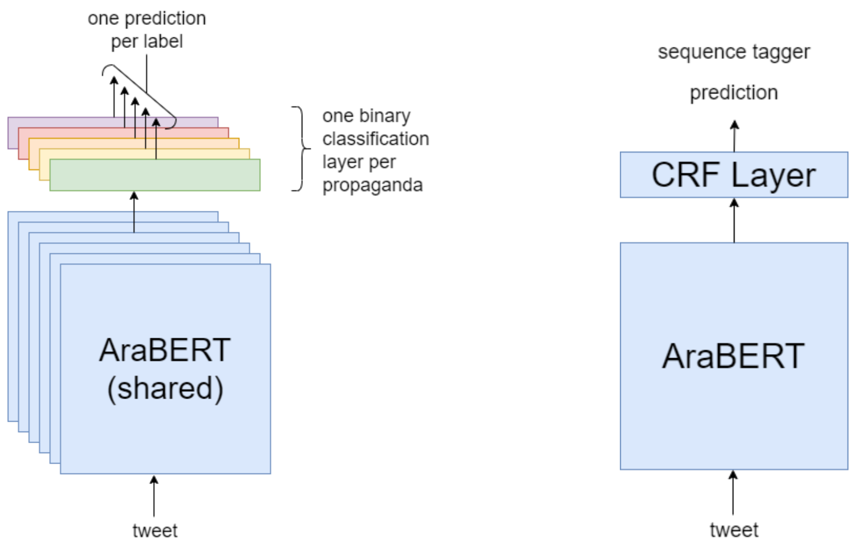

Figure 1: Diagrams for both systems 1 and 2 submitted for sub-task 1 and 2 respectively.

4 Systems

4.1 System 1 - Multi-Task Learning

For the first sub-task, we propose to use multi-task learning to perform multi-label text classification. We propose to encode more knowledge in AraBERT by training the model to predict different types of propaganda techniques, one technique at a time. In other words, AraBERT is fine-tuned to perform **n** binary classification, where n corresponds to the number of propaganda techniques. BERT will learn weights that will allow it to represent the text appropriately for the task, while at the same time fine-tuning the different binary classification layers to distinguish between the different techniques.

The Multi-Task model consists of a single shared AraBERT encoder. The pre-trained AraBERT model is fine-tuned using n task-specific classification heads (i.e., binary classification layers). Each classification head consists of a Dropout layer of probability 0.1 followed by a linear layer that maps the pooled embeddings of the AraBERT encoder to the number of predicted classes (2 classes at a time, since predicting each propaganda technique is a binary classification task). We use the cross-entropy loss to compute the loss on the outcome of every classifier head. Since the losses assess different measures, we chose to fine-tune one loss at a time per batch.

As mentioned earlier, the dataset used is a relatively small dataset, which makes the task more difficult to achieve. We train the model using the Adam optimizer (Kingma and Ba, 2015), with a learning rate of 10^5. After a couple of experiments, we set the batch size to 8 for the first 2 epochs, then to 1 for 2 epochs. This training scenario ensured that the model learns from the dataset without over-fitting (since the gradients would be computed differently throughout the different epochs).

As seen in Table 2, the dataset used suffers from class imbalance. Therefore, we propose to randomly sample (with replacement) 2000 sentences per propaganda label value from the training set (i.e., for the Smears classification head, we sample 2000 samples with a negative label and 2000 samples with a positive label). In other terms, the training set used for this model consists of 2000 tweets for every label. This will guarantee that all classes participate in the training process equally.

4.2 System 2 - CRF Layer

For the second subtask, we propose to fine-tune BERT using a Conditional Random Fields (Lafferty et al., 2001) layer. In general, CRFs are a generalization of Bayesian Networks and are used in applications in which the contextual information of the neighbors affects the current prediction (e.g., sequence labeling task). First, we encode the input text using the AraBERT model, and then we pass the output to the CRF layer to predict the label of the spans using the BIO data format. The model is trained to perform a multi-class classification, as the model will predict whether every token in the text is either the first token in the span (B-<type>), inside the span (I-<type>) or outside the span (O),

Table 2: Label distribution and F1 scores for both sub-tasks 1 and 2

Propaganda Techniques	**Sub-task 1**						**Sub-task 2**					
	TRAIN	**DEV**	**DEV TEST**	**TEST**	**DEV TEST F1 Micro**	**TEST F1 Micro**	**TRAIN**	**DEV**	**DEV TEST**	**TEST**	**DEV TEST F1**	**TEST F1**
Loaded Language	289	28	31	223	75.0	69.34	446	46	42	326	36.42	43.25
Name calling/Labeling	186	35	27	142	73.07	66.25	244	44	33	163	31.15	45.21
Smears	84	12	16	50	80.76	82.34	85	12	15	50	51.16	38.09
Appeal to fear/prejudice	47	7	3	25	88.46	90.71	48	7	4	25	18.18	42.23
Exaggeration/Minimisation	41	10	12	23	76.92	90.71	44	10	16	26	0	0
Slogans	28	1	1	7	98.07	97.73	44	1	1	6	0	5.40
Doubt	27	1	2	19	94.23	95.04	29	1	2	19	0	45.16
Glittering generalities (Virtue)	25	7	2	1	96.15	98.45	25	7	2	1	40	26.67
Appeal to authority	21	7	2	1	96.16	99.07	21	7	1	1	56.93	0
Obfuscation, Intentional vagueness, Confusion	9	3	1	6	98.07	97.83	9	3	1	6	0	0
Repetition	7	2	1	3	98.07	98.45	9	2	1	3	0	0
Thought-terminating cliché	6	1	1	0	100	100	6	1	1	0	0	100
Flag-waving	5	2	2	10	96.15	96.59	5	2	2	9	0	0
Causal Oversimplification	4	1	1	4	98.07	98.76	4	1	1	4	0	0
Whataboutism	3	1	1	0	98.07	100	3	1	1	0	0	100
Black-and-white Fallacy/Dictatorship	2	1	2	7	96.15	97.83	2	1	2	7	0	0
Presenting Irrelevant Data (Red Herring)	1	0	0	0	100	99.33	1	0	0	0	100	100
Misrepresentation of Someone's Position (Straw Man)	0	0	0	1	100	99.69	0	0	0	1	100	100
Reducto ad hitlerum	0	0	0	0	100	100	0	0	0	0	100	100
Bandwagon	0	0	0	0	100	100	0	0	0	0	100	100
No techniques	95	7	8	44	84.61	79.87	0	0	0	0	100	100
OVERALL	880	126	113	566	**59.07**	**60.2**	1025	146	125	647	**27.95**	**39.55**

where <type> represents the type of propaganda technique.

In the training process, we employ the negative log-likelihood loss, which is more suitable for this type of task than cross-entropy loss. We train the model using the Adam optimizer, with a batch size of 32 for 13 epochs.

5 Results and Discussion

Table 2 reports the size of the training set, development sets (dev and dev_test), and the testing set. Furthermore, it presents the Micro-averaged F1 Score on the dev_test and test sets for both tasks. We did not report the Macro-Averaged F1 Score as it is not the official metric of the task.

We conduct the analysis on the original training set. As mentioned previously, the training set is quite small (around 500 samples for training, covering 880 total labels). We notice that 51% of the tweets contain one propaganda technique, while 29% contain two propaganda techniques, and 20% of the tweets have more than three propaganda techniques. This makes the task quite challenging, as there might be instances with more than one propaganda technique present at the same time, while others with no propaganda technique at all. Therefore, treating the task as multiple binary classification techniques is suitable as we are able to independently predict the presence of different techniques, while at the same time learning their co-occurrence information through sharing the same base model.

For sub-task 1, the model's performance on the test set was on par with its performance on the dev_test set (similar F1-Scores achieved per label, and overall). For sub-task 2, the model generalized very well and scored a much higher F1-Score on the test set compared to the dev_test set.

We analyze these results with respect to the distribution of the samples among the different labels. We notice that 85% of the labels in the training set are covered by 9 propaganda techniques. Furthermore, the rest of the techniques have less than 10 samples in the training set. These samples might not be good representatives of their propaganda techniques that the multi-task model can generalize from. Perhaps training the multi-task model to achieve a higher performance on the 9 most common techniques would have resulted in a more accurate performance of the system. There is also a need to increase the number of instances of the propaganda techniques that rarely occur in the training set. This can be done using a data augmentation method guided using domain knowledge. On a last note, both systems were tested on the Straw Man propaganda technique that did not occur in any set.

6 Related Work

In this section, we present some of the previous work conducted for propaganda detection, also covering the Conference and Labs of the Evaluation Forum (CLEF) *CheckThat!* lab that employs fact-checking (where the propaganda sentences can be viewed as fake claims). Researchers provided multiple datasets to tackle the propaganda detection task. For instance, Rashkin et al. (2017) collected news articles from reliable and unreliable sources, and labeled them using distant supervision to four classes: propaganda, trusted, hoax, or satire. Habernal and Gurevych (2017) presented a corpus of 1.3k arguments annotated with five fallacies. Furthermore, Da San Martino et al. (2019c) presented a corpus of news articles annotated with 18 propaganda techniques. The annotations identify the minimal fragments related to the propaganda technique (i.e., the span), instead of flagging the whole sentence.

On another hand, CLEF provided the *CheckThat*! lab that supported the automatic identification and verification of claims in its multiple editions that are held every year (Atanasova et al., 2018; Barrón-Cedeño et al., 2018; Atanasova et al., 2019; Hasanain et al., 2019, 2020; Shaar et al., 2020; Nakov et al., 2021; Shaar et al., 2021b,a; Nakov et al., 2021, 2022a,b). The Lab provided multiple tasks around Fact-checking, with the following tasks: claim detection, claim matching, evidence retrieval, and claim verification. We briefly describe each task. The claim detection task estimates the check-worthiness of the claim by predicting which claims should be prioritized for fact-checking. The claim matching task determines whether a new claim is similar to a claim that has already been fact-checked; if a similar claim is found, there is no need to fact check the new claim again. The evidence retrieval task finds information that can verify a claim, by asking the participants to rank the set of evidence based on their usefulness for fact-checking a certain claim. Finally, the claim verification task is a Verdict Prediction task in which the claim is either deemed factually true, half-true or false based on the retrieved evidence.

7 Conclusion

In this paper, we introduced two AraBERT-based systems to tackle propaganda identification and span detection. We conclude that identifying propaganda techniques in Arabic tweets is a challenging task. The most challenging aspect of this task lies in the small dataset used (504 samples covering 880 labels) as well as the multi-propaganda aspect of the tweets. Even though the proposed systems did not employ any data augmentation technique, they achieved ranks 3 and 1 on sub-tasks 3 and 1. In future work, we propose to focus the training on the binary classification heads that handle propaganda issues that are more commonly faced by users on social media (such as Loaded Language and Name calling/Labeling). Focusing our attention on these classification heads would help build models that will protect the users from the most present propaganda attacks on the web.

References

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Preslav Nakov, and Giovanni Da San Martino. 2022. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, Abu Dhabi, UAE. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Pepa Atanasova, Lluís Màrquez i Villodre, Alberto Barrón-Cedeño, T. 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. *ArXiv*, abs/1808.05542.

Pepa Atanasova, Preslav Nakov, Georgi Karadzhov, Mitra Mohtarami, and Giovanni Da San Martino. 2019. Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. task 1: Check-worthiness. In *CLEF*.

Alberto Barrón-Cedeño, T. Elsayed, Reem Suwaileh, Lluís Màrquez i Villodre, 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*.

Giovanni Da San Martino, Alberto Barron-Cedeno, 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*, pages 162–170.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019b. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179.

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! 2020i arabic: Automatic identification and verification of claims in social media. In *CLEF*.

Maram Hasanain, Reem Suwaileh, T. Elsayed, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. task 2: Evidence and factuality. In *CLEF*.

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

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. *arXiv preprint arXiv:2007.08024*.

Clyde Raymond Miller. 1939. *How to detect and analyze propaganda*. Town Hall, Incorporated.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Hamdy Mubarak, Alex Nikolov, and Yavuz Selim Kartal. 2022a. Overview of the clef-2022 checkthat! lab task 1 on identifying relevant claims in tweets. In *CLEF*.

Preslav Nakov, Giovanni Da San Martino, Firoj Alam, Shaden Shaar, Hamdy Mubarak, and Nikolay Babulkov. 2022b. Overview of the clef-2022 checkthat! lab task 2 on detecting previously fact-checked claims. In *CLEF*.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mucahid Kutlu, and Yavuz Selim Kartal. 2021. Overview of the clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *CLEF*.

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

Shaden Shaar, Fatima Haouari, Watheq Mansour, Maram Hasanain, Nikolay Babulkov, Firoj Alam, Giovanni Da San Martino, Tamer Elsayed, and Preslav Nakov. 2021a. Overview of the clef-2021 checkthat! lab task 2 on detecting previously fact-checked claims in tweets and political debates. In *CLEF*.

Shaden Shaar, Maram Hasanain, Bayan Hamdan, Zien Sheikh Ali, Fatima Haouari, Alex Nikolov, Mucahid Kutlu, Yavuz Selim Kartal, Firoj Alam, Giovanni Da San Martino, Alberto Barrón-Cedeño, Rubén Míguez, Javier Beltrán, Tamer Elsayed, and Preslav Nakov. 2021b. Overview of the clef-2021 checkthat! lab task 1 on check-worthiness estimation in tweets and political debates. In *CLEF*.

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 *CLEF*.

Bruce Lannes Smith. 2021. Propaganda.

Anthony Weston. 2018. *A rulebook for arguments*. Hackett Publishing.