Dataset of Propaganda Techniques of the State-Sponsored Information Operation of the People's Republic of China

Rong-Ching Chang
g08350003@thu.edu.tw
Tunghai University
Taiwan

Chun-Ming Lai
cmlai@thu.edu.tw
Tunghai University
Taiwan

Kai-Lai Chang
g08350009@thu.edu.tw
Tunghai University
Taiwan

Chu-Hsing Lin
chlin@thu.edu.tw
Tunghai University
Taiwan

ABSTRACT

The digital media, identified as computational propaganda provides a pathway for propaganda to expand its reach without limit. State-backed propaganda aims to shape the audiences' cognition toward entities in favor of a certain political party or authority. Furthermore, it has become part of modern information warfare used in order to gain an advantage over opponents.

Most of the current studies focus on using machine learning, quantitative, and qualitative methods to distinguish if a certain piece of information on social media is propaganda. Mainly conducted on English content, but very little research addresses Chinese Mandarin content. From propaganda detection, we want to go one step further to provide more fine-grained information on propaganda techniques that are applied.

In this research, we aim to bridge the information gap by providing a multi-labeled propaganda techniques dataset in Mandarin based on a state-backed information operation dataset provided by Twitter. We labeled 9,950 tweets in total with 21 propaganda techniques. In addition to presenting the dataset, we apply a multi-label text classification using fine-tuned BERT. We have observed consistency in the promoted message and used techniques by state-backed propaganda operations toward certain entities or topics. Viewing country and political party as an entity, we could view state-backed propaganda detection on different topics as stance detection tasks. Potentially our research could help future research in detecting state-backed propaganda online, especially in a cross-lingual context and cross-platform manner.

CCS CONCEPTS

• **Social and professional topics → Political speech**.

KEYWORDS

propaganda, information operation, social media

MIS2 workshop at KDD 2021, Aug 15, 2021, Virtual

ACM Reference Format:
Rong-Ching Chang, Chun-Ming Lai, Kai-Lai Chang, and Chu-Hsing Lin. 2021. Dataset of Propaganda Techniques of the State-Sponsored Information Operation of the People's Republic of China. In *KDD '21: The Second International MIS2 Workshop: Misinformation and Misbehavior Mining on the Web, Aug 15, 2021, Virtual.* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Propaganda has the purpose of framing and influencing opinions. With the rise of the internet and social media, propaganda has adopted a powerful tool for its unlimited reach, as well as multiple forms of content that can further drive engagement online and offline without disclosing the writers' identity. Computational propaganda is defined as propaganda being created or distributed using computational or technical means [5]. Exploiting social media is considered as one of the low-cost and high-impact techniques in information warfare, driving and manipulating human behavior with various psychological manipulations [1]. How information is conveyed is by using propaganda techniques. Propaganda techniques are not only used for political content, but also for marketing, and religious content for persuasion purpose. Propaganda techniques, commonly used in disinformation and misinformation, are the way that propaganda is conveyed [10], such detection requires for more fine-grained analysis and detection, not only distinguishing if it is propaganda, but characterizing where it might come from. The propaganda activity launched by foreign adversaries could be particularly concerning to a country as the usual goal may include steering discord, spreading fear, influencing beliefs and behaviors, diminishing trust, and threatening the stability of a country [1]. Various state-backed official and unofficial departments, organizations, and agencies were established to address information warfare include the Internet Research Agency of Russia [12], 50 Cent Party [15] [14] of Chinese Communist Party (CCP) and the Public Opinion Brigades of the Communist Party of Vietnam [6].

Most of the recent work has been focused on propaganda detection, in another word, identifying if the information is propaganda or not. This has been done using various methods such as qualitative analysis, quantitative analysis [4], and machine learning [20] [7]. The main features for this detection task could be divided into two parts, content-driven, and network-driven. Some of the current propaganda text corpora open data sets on document levels include Rashkin et al. [18] which labeled texts into trusted, satire, hoax, and propaganda on news. Barrón-Cedeno et al. [3] further increased the corpus [18] with more propaganda articles and metadata of the articles. The currently available fine-grained propaganda technique dataset is the one presented by Da San Martino et al. [10]. From news articles, they labeled 18 propaganda techniques on a word-by-word sentence level, so that the position of where the propaganda technique was applied from start to end was being documented. All of the mentioned data sets are in English. Baisa et al. [2] released a propaganda technique dataset for Czech based on newspapers. Another open data source is Twitter, a popular social media platform, the dataset discloses state-linked information operations that took place on their platform. However, the Twitter dataset is not labeled with propaganda techniques but the Twitter account metadata and media information only. The previously labeled propaganda technique in news article texts could be quite different linguistically compared to texts on social media. Tweets, messages posted on Twitter, tend to be more casual with slang and emojis. They are also shorter as the platform has a text length limit. Da San Martino et al. [9] conducted a survey of computational propaganda, mentioned that there is limited propaganda detection research based on text features due to the lack of annotated data sets. Yet we think text content is an important feature for performing cross-platform detection, in user-identity linking, and in information origin tracing. Since the network feature may differ from platform to platform, text content is more consistent in that regard. To our knowledge, there is no existing propaganda technique dataset for Mandarin Chinese.

To address such a gap, we present our dataset[1] that focuses on propaganda techniques in Mandarin based on state-linked information operations dataset from the PRC released by Twitter in July 2019. The dataset consists of multi-label propaganda techniques of the sampled tweets. Additionally, we employed a fine-tuned BERT model for the multi-label classification task.

2 PROPAGANDA TECHNIQUES

Below we explained a list of selected propaganda techniques we have considered based on various studies [10] [2] [21]. Using the same assumption as [10], we labeled our data based on the linguistic and language use that can be judged directly without retrieving extra information. The propaganda techniques we considered are as follows:

(1) Presenting Irrelevant Data
Also called Red Herring. Introducing irrelevant information or issues to an argument.
(2) Misrepresentation of Someone's Position (Straw Man)
Substituting one's opinion with a distorted version rather than the original one.
(3) Whataboutism
Defaming the opponents with hypocrisy.
(4) Oversimplification
Overly generalizing information or the complexity of certain issues to favor a party.
(5) Obfuscation, intentional vagueness, confusion
Purposefully being vague with the intention for the audience to develop false recognition toward the subject.
(6) Appeal to authority
Supporting the opinion or clam unconditionally as long as it comes from the government or an expert.
(7) Black-and-white Fallacy
Presenting only two opposite possibilities, one favoring a certain party and one presented by the opponent.
(8) Stereotyping, name-calling, labeling
Labeling the target with the intention of arousing prejudices or making an association with stereotypes.
(9) Loaded Language
Using emotional words to influence audience opinions.
(10) Exaggeration or Minimisation
Overly amplifying or reducing the importance of something.
(11) Flag-waving
Justifying or presenting as a nation or group or idea. In our case, we also consider Flag-waving when one person is presented as their opinion represents the entire nation or group.
(12) Doubt
Questioning or steering uncertainty or trust toward something, an entity, or a group.
(13) Appeal to fear or prejudice
Spreading a sense of anxiety, fear, or panic toward the audience or entity.
(14) Slogans
A brief sentence that includes labeling, stereotyping, or certain cognitive belief.
(15) Thought-terminating cliché
Using simple and generic sentences to discourage detail in discussions.
(16) Bandwagon
Persuading the audience to align with the bigger crowd who appear to have an advantage or better situation, or implying a certain entity will lose or have a worse situation.
(17) Guilt by association or Reductio ad Hitlerum
Associating an opponent or target with the usually disliked object or group.
(18) Repetition
Repeating the same message or idea several times.

Additional to the usual propaganda techniques, we also introduce the following that has been seen in the dataset:

(1) Neutral Political
This includes the international political news that's being written objectively.
(2) Non-Political
This includes the non-political related content, which could be written with a neutral or angry, or happy tone.
(3) Meme humor
This is the content that used sarcastic humor toward an entity.

3 DATA

Twitter disclosed 936 accounts with identified state-backed information operations from the People's Republic of China (PRC) government departments, agencies, and party-supported organizations.

[1]Dataset will be released on https://github.com/annabechang

The dataset was divided into two batches that consist of 744 accounts and 196 accounts separately on Twitter's information operations disclosure. In our study, we sampled tweets from a batch of 744 accounts. The available data disclosed containing account metadata (account created time, a user profile description, user-reported location, etc), tweet metadata (tweet created time, tweet text, tweet language, tweet hashtags, etc), and shared media content (images and videos). In our study, we only focus on the tweet metadata.

The total number of tweets sent by the 744 accounts is 1, 898, 108. We first filter it by language, and duplicates were dropped. The total number of tweets in Chinese contained in the dataset is 74, 277, we randomly selected 9, 950 tweets out of that number for labeling.

Uren et al. [19] conducted a detailed quantitative and qualitative analysis on these accounts and suggested that this cluster could be re-purposed spam accounts as they altered the used language through different periods of time. These findings are aligned with ours. Figure 2 shows the top 15 tweet language usage out of 50 total used languages. The top 5 languages used in this cluster of accounts are Indonesian (in), English (en), Portuguese (pt), Chinese (zh), and Tagalog (tl).

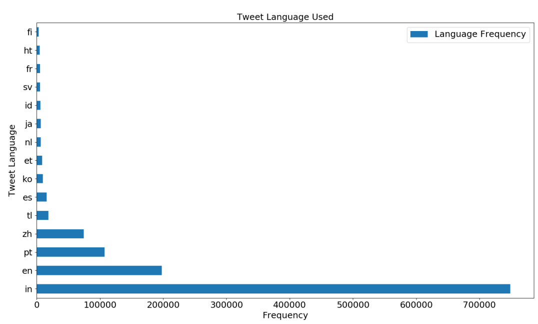

Figure 1: Language usage of more than 10,000 times each year

In Figure 1 we plot the language used more than 10,000 times each year and we can see that Chinese was only used by this cluster of accounts after 2017. The primary used language appears to have been clear cut in different years, which indicate that this cluster could be spam accounts that were created and used by entities with different backgrounds and purposes at different time period.

Manual annotation was done by two annotators separately on different portions of the dataset: the first 4,950 tweets and the rest 5,000 tweets respectively. Both annotators are native mandarin speakers in graduate degree programs. This was designed intentionally to insure the alliance of opinions in the dataset [13]. This design will increase the annotator consistency, reduce noise and have better model performance. In the annotation process, we iterate the process of reviewing, data labeling, and documenting political entities being named in the sentence. After the first 1,000 labeling, we observed a consistent propaganda techniques alignment on the tweets toward certain topics and entities. Based on such observation, we built pre-defined labels according to the political entities mentioned. Human annotators iterate the action of updating the labels from the rule-based labels when necessary, adding

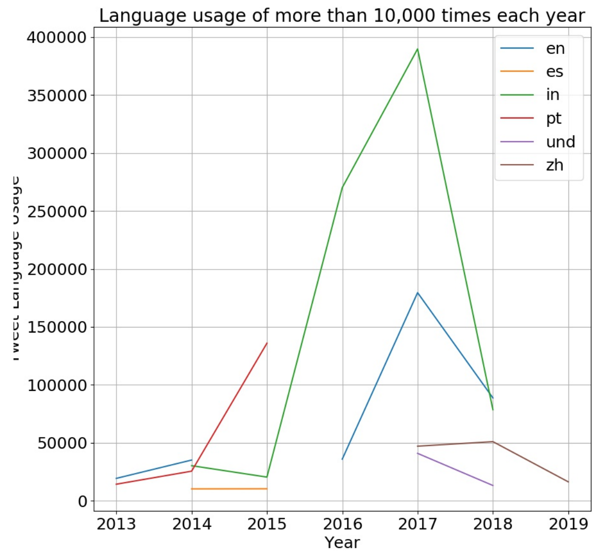

Figure 2: Total language usage in the data set

more entity keywords for rule update, and updating the further unseen data labels based on the updated rules. The category of keywords for the pre-defined labeling rules are shown in Table 1, they can be divided into four categories: exiled or anti-government Chinese, Hong Kong protest, Taiwan independence, International Geo-Political related topics. The usage of keywords is not to be exact but assisting the human annotators.

Table 1: Aggregation of keyword category for the pre-defined labeling rules

Keyword Category	Count
Exiled or anti-government Chinese	5,406
Hong Kong protest	209
International Geo-Political	1,718
Taiwan independence	2

The propaganda techniques are only labeled on the political-related content, there could be non-political content using propaganda techniques but this is not labeled as it was not our focus. Such content will be labeled as a non-political class.

In total, we have 21 different propaganda techniques, we showed a label statistic in Table 2. This is an imbalanced dataset, as the most frequently used label is the non-political content that was used for 6, 117 times. Loaded Language was used the most at 2, 609 times, followed by Whataboutism 2, 509 and Name-Calling 2, 313 are the most used propaganda techniques on political-related content. A few techniques occurred rarely, especially the Thought-terminating cliché was not used. We suspect that this is due to the nature of spam accounts. That is, building a relationship with other accounts was not their primary goal. Thought-terminating clichés might be used more in circumstances where building a relationship with other accounts is one of the target goals. An example of how the dataset was formatted can be seen in Table 3, the tweets were translated for the purpose of display.

Table 2: Data set label statistics

Symbo	Propaganda Techniques	Frequency
1	Presenting Irrelevant Data	13
2	Straw Man	2
3	Whataboutism	2,509
4	Oversimplification	37
5	Obfuscation	12
6	Appeal to authority	50
7	Black-and-white	265
8	Name Calling	2,313
9	Loaded Language	2,609
10	Exaggeration or Minimisation	114
11	Flag-waving	81
12	Doubt	147
13	Appeal to fear or prejudice	141
14	Slogans	37
15	Thought-terminating cliché	0
16	Bandwagon	64
17	Reductio ad Hitlerum	83
18	Repetition	60
19	Neutral Political	915
20	Non-Political	6,117
21	Meme humor	5

Table 3: Data set sample display

Tweetid	Translated Tweet	Propaganda Techniques
990189929 836699648	The truth and hypocrisy under the false democratic face of Guo Wengui, the clown jumping beam, is now undoubtedly exposed!	3,8,9
114879827 6281364480	We must severely punish the rioters and return Hong Kong to peaceful	8,9,13,14

4 MULTI-LABEL PROPAGANDA TECHNIQUE CLASSIFICATION

In this section, we describe our methodology in designing, fine-tuning, and provide the result of our BERT-based multi-label classification result.

Bidirectional Encoder Representations from Transformers *BERT* [11] a language representation model has delivered state-of-the-art results in several NLP tasks. In our case, the research problem in our case is a multi-label task where given one sentence, there are one to multiple labels that could apply.

We used the bert-base-chinese pre-trained model provided by Huggingface [22] for both tokenization and pre-training the model. The bert-base-chinese pre-trained model is trained based on both simplified Chinese and traditional Chinese [8], which fits our use case. In our model design, we used a BERT model followed by a dropout and linear layer for regularization and classification purposes. We have 21 different labels defined in our propaganda technique labels with 1 of them without occurrence. We set the number of dimensions for the linear layer to 20. The output of the linear layer is what we used to determine the accuracy of the models.

The max input length was set to 100 with a training batch size of 2 and a validation batch size of 2 using the data loader from Pytorch [17]. We chose to use BCEWithLogitsLoss, which combines a Sigmoid layer and BCELoss, from Pytorch [17] as our loss function. Adam [16] was used as an optimizer. We ran it for 2 epochs with a learning rate equal to $1e-05$. We trained on a Linux machine with GeForce RTX 2070 GPU, and 16 Intel(R) Core(TM) i9-9900K CPU.

5 EVALUATION

The training and testing size was set to 80% and 20% respectively. The results are shown in the Table 4. We only trained it for 2 epochs yet we saw the loss decreased drastically from 0.71102 to 0.05953. In the experiment, we trained for more than 2 epochs; however, the accuracy did not improve. Thus 2 epochs appear to be optimal in our experiment. The evaluation metrics used were accuracy, micro-averaged F1-score, and macro-averaged F1-score. Micro-averaged F1-score aggregate all the samples to compute the average metric of the True Positives our of the Predicted Positives. Macro-averaged F1-score aggregated each class and compute the metrics based on each class. In our case, our accuracy is 0.80352 with micro-averaged F1-score of 0.85431 and macro-averaged F1-score of 0.20803. This indicates that our model performed well in predicting overall samples, however, the performance on each label varied a lot. This is expected as our dataset is skewed, some labels have many data while a few labels have very little data labeled in the dataset.

Table 4: Classification results

Measurement Name	Performance
Loss : Epoch 0	0.71102
Loss : Epoch 1	0.05953
Accuracy	0.80352
F1 Score (Micro)	0.85431
F1 Score (Macro)	0.20803

Two main activity directions of the dataset were to target opponents of the CPC, such as exiled Chinese, human rights lawyers, relevant personnel and to vilify the protesters against the national security law in Hong Kong. This finding was aligned with what was found in [19] [5], where the spam accounts flooded content in Mandarin with the purpose of dominating search results on Twitter when it comes to certain topics. By doing so the propaganda operators wanted the search results to be skewed toward a perspective that favored the CCP and eschewed the certain community.

6 DISCUSSION

In this paper, we presented the first propaganda technique dataset of state-backed information operation accounts from PRC for Mandarin based on a dataset released by Twitter. We applied 21 propaganda techniques and we annotated a total of 9, 950 sentences under a multi-label setting. Machine learning models driven by propaganda research can be particularly benefited by our data set. As we labeled political content with propaganda techniques while giving the nonpolitical items a label. Our dataset can be used to train classifiers for political and non-political in Mandarin as well.

Upon the organization structure of PRC, different departments and agencies may lunch online operations targeting the same or different groups of audiences, with different linguistic characteristics. Thus, this data set's linguistic feature or the propaganda techniques may not apply to all.

7 CONCLUSION

We presented a new dataset on propaganda techniques from the state-backed information operation accounts from PRC in Mandarin. We trained a fine-tuned BERT model to perform multi-label classification on our dataset. In the times where information on social media is part of information warfare strategies. Our dataset could be beneficial in propaganda, political research, and beyond.

By considering the country, political party, or authority as an entity, we could initially view state-backed propaganda on different topics as a stance detection of texts from such an entity. And propaganda techniques could be viewed as a writing style feature. This could help future research in clustering and identifying how likely it is that the information is coming from the same entity or agency.

One state could launch several propaganda texts that have a similar stance or opinion in different countries with different languages. Thus we hope to see our dataset inspire or provide useful information on multilingual or cross-platform state-back propaganda research, using the propaganda techniques as the universal features across languages.

REFERENCES

[1] Media Ajir and Bethany Vailliant. 2018. Russian Information Warfare: Implications for Deterrence Theory. *Strategic Studies Quarterly* 12, 3 (2018), 70–89. http://www.jstor.org/stable/26481910

[2] Vít Baisa, Ondřej Herman, and Ales Horak. 2019. Benchmark Dataset for Propaganda Detection in Czech Newspaper Texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. INCOMA Ltd, Varna, 77–83.

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

[4] David M Beskow and Kathleen M Carley. 2020. Characterization and comparison of Russian and Chinese disinformation campaigns. In *Disinformation, misinformation, and fake news in social media*. Springer, Springer, Cham, 63–81.

[5] Gillian Bolsover and Philip Howard. 2017. Computational propaganda and political big data: Moving toward a more critical research agenda.

[6] Samantha Bradshaw and Philip Howard. 2017. *Troops, trolls and troublemakers: A global inventory of organized social media manipulation*. Oxford Internet Institute.

[7] Rong-Ching Chang and Chu-Hsing Lin. 2021. Detecting Propaganda on the Sentence Level during the COVID-19 Pandemic. In *2021 Cryptology and Information Security Conference*. Cryptology and Information Security Conference, Taiwan, 1.

[8] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. *Pre-training with whole word masking for chinese bert.* arXiv preprint arXiv:1906.08101.

[9] 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 e-prints.

[10] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. AACL, Hong Kong, 5640–5650.

[11] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] Renee DiResta, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, and Ben Johnson. 2019. *The tactics & tropes of the Internet Research Agency.* Congress of the United States.

[13] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama, 1–14.

[14] Rongbin Han. 2015. Manufacturing consent in cyberspace: China's "fifty-cent army". *Journal of Current Chinese Affairs* 44, 2 (2015), 105–134.

[15] Gary King, Jennifer Pan, and Margaret E Roberts. 2017. How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American political science review* 111, 3 (2017), 484–501.

[16] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. arXiv:arXiv:1412.6980

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., Vancouver, Canada, 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[18] 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*. ACL, Copenhagen, Denmark, 2931–2937.

[19] Tom Uren, Elise Thomas, and Jacob Wallis. 2019. Tweeting Through the Great Firewall: Preliminary Analysis of PRC-linked Information Operations on the Hong Kong Protests. *Australia Strategic Policy Institute: International Cyber Policy Center, Barton* 1, 1 (2019), 1–37.

[20] Nishan Chathuranga Wickramarathna, Thiruni D Jayasiriwardena, Malith Wijesekara, Pasindu Bawantha Munasinghe, and Gamage Upeksha Ganegoda. 2020. A Framework to Detect Twitter Platform Manipulation and Computational Propaganda. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, IEEE, Colombo, Sri Lanka, 214–219.

[21] Wikipedia contributors. 2021. Propaganda techniques — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Propaganda_techniques&oldid=1020793767 [Online; accessed 30-May-2021].

[22] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6