Exploring Data Augmentation for Classification of Climate Change Denial: Preliminary Study

Jakub Piskorski[1], Nikolaos Nikolaidis[2], Nicolas Stefanovitch[3], Bonka Kotseva[4], Irene Vianini[5], Sopho Kharazi[5] and Jens P. Linge[3]

[1] *Polish Academy of Sciences, Warsaw, Poland*
[2] *Trasys International, Brussels, Belgium*
[3] *European Commission Joint Research Centre, Ispra, Italy*
[4] *CRI, Luxembourg, Luxembourg*
[5] *Piksel SRL, Ispra, Italy*

Abstract

In order to address the growing need of monitoring climate-change denial narratives in online sources, NLP-based methods have the potential to automate this process. Here, we report on preliminary experiments of exploiting Data Augmentation techniques for improving climate change denial classification. We focus on a selection of both known techniques, and augmentation transformations not reported elsewhere that replace certain type of named entities with high probability of preserving labels. We also introduce a new benchmark dataset consisting of text snippets extracted from online news labeled with fine-grained climate change denial types.

Keywords

text classification, climate change denial, machine learning, data augmentation

1. Introduction

To better understand climate change (CC) denial, it is crucial to collect, analyse and classify narratives that oppose the scientific consensus of anthropogenic global warming. The sheer volume of misinformation on climate change, makes automation key in order to help tackle with this infodemic. AI-based solutions can help to label already known narratives and identify novel narratives in content from news or social media. They also enable trend analysis to point out emerging topics over time. This is of particular interest for journalists, fact-checking organisations and government authorities as it allows addressing specific areas e.g. by publishing rebuttals or designing public awareness campaigns.

In this paper we report on a preliminary study of exploiting Data Augmentation (DA) for improving CC denial classification and elaborate on the creation of a new benchmark dataset consisting of text snippets extracted from online news labeled with CC denial type. In particular, we explore a selection of known techniques and others that have not been reported elsewhere

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'22 Workshop, Stavanger (Norway), 10-April-2022

jpiskorski@gmail.com (J. Piskorski); nikolaidis.nikolaos@ext.ec.europa.eu (N. Nikolaidis); nicolas.stefanovitch@ec.europa.eu (N. Stefanovitch); bonka.kotseva@ext.ec.europa.eu (B. Kotseva); irene.vianini@ext.ec.europa.eu (I. Vianini); sopho.kharazi@ext.ec.europa.eu (S. Kharazi); jens.linge@ec.europa.eu (J. P. Linge)

(specific name-entity type replacements) with a focus on transformations with high probability of preserving labels. The main drive behind this research is two-fold: First, it emerges from the need to rapidly develop a production-level component for CC denial text classification for Europe Media Monitor (EMM)[1], a large-scale media monitoring platform used by EU institutions, and, secondly, from the scarcity of annotated data for the task at hand. The experiments reported in this paper build mainly on top of the only publicly available text corpus of CC contrarian claims, which is labeled using a fine-grained taxonomy presented in [1]. We also present a preliminary evaluation of a some models on a new EMM-derived news snippet corpus reusing the same taxonomy. The findings contained in this paper are not of general nature, but rather specific to the exploited data and the domain, paving the way for future in-depth explorations.

The paper starts with an overview of related work in Section 2. Next, the DA techniques exploited in our study is described in Section 3, whereas Section 4 introduces a news-derived corpus of text snippets related to CC denial. The DA techniques performance evaluation is presented in Section 5. Section 6 provides detailed analysis of the behaviour of two specific named-entity replacement-based DA techniques. Finally, we present our conclusions in Section 7.

2. Related Work

Only recently the CC debate has received more attention in the NLP community in the context of developing solutions for making sense of the vast amount of textual data produced on this topic [2]. A corpus of manually-tagged blog posts on CC in terms of scepticism and acceptance of CC is presented in [3]. In 2016 a SemEval task on stance detection of tweets, where ”CC is a real concern” was organized [4]. In [5] an annotated news corpus for stance toward ”climate change is a real concern” and related experiments are presented, whereas [6] introduced a dataset for sentence-based climate change topic detection. Finally, [7] reported on a collection of tweets used to study the public discourse around CC.

To the best of our knowledge only two textual corpora with CC denial and disinformation labels exists, namely, the corpus of ca. 30K text paragraphs containing contrarian claims about climate change extracted from conservative think-tank websites and contrarian blogs (4C corpus) [1], and a collection of ca. 500 news articles with known CC misinformation scraped from web pages of CC counter movement organisations [8]. Given that the latter corpus is not publicly accessible at the moment, we exploit the former 4C corpus, and the associated taxonomy of climate contrarianism in our study.

Data Augmentation (DA) is a family of techniques aiming at creation of additional training data in order to alleviate problems related to insufficient and imbalanced data and low data variability, with the overall goal of model performance improvement. Recently, DA has gained attention in the NLP domain, and a wide range of DA techniques has been elaborated and explored, including, i.a., simple word substitution, deletion, and insertion [9], sub-structure substitution [10], back-translation [11, 12], contextual augmentation [13], data noising [14, 15], injection of noise into the embedding space [16], interpolating the vector representations of text and labels [17], etc. A survey on DA techniques for text classification is presented in [18], whereas [19] provides a more general overview of DA in the broader area of NLP .

[1] https://emm.newsbrief.eu/

3. Data Augmentation

For the sake of carrying out DA experiments we have selected a range of known and 2 variants of some known techniques, in particular, focusing on transformations with high probability of preserving labels by the automatically created instances. The list of DA techniques encompasses:

COPY: simply creates copies of the existing instances in the training dataset.

DATE: randomly changes all dates, e.g., month and day-of-the-week names.

DEL-ADJ-ADV: deletes up to a maximum of 1/3 of all adjectives and adverbs in the text, provided that they are preceded by nouns and verbs resp. Here, the assumption is that such transformation preserves the label assigned to the text.

PUNCT: inserts various punctuation marks randomly selected from (‘.’, ‘;’, ‘?’, ‘:’, ‘!’, ‘,’) into randomly selected positions in the text, where the number of insertions is a randomly selected number between 1 and 1/3 of the length of the text (in words). This simple DA technique introduced recently in [15] proved to outperform many other simple DA techniques.

GEO: randomly replaces all occurrences of toponyms referring to a populated place with another randomly chosen toponym from GEONAMES-based[2] gazetteer of about 200K populated cities.

PER-ORG: randomly replaces occurrences of mentions of person and organisation names matched using the JRC NAME VARIANT database [20] (containing large fraction of entities whose mentions appear in the news) with some other names therefrom (not spelling variants of the replaced names). The current version of JRC NAME VARIANT contains circa 3 million names.

SYN: randomly replaces verbs and adjectives with their synonyms. It picks the top-10 tokens (verbs/adjectives) whose deletion maximizes the cosine distance from the resulting sentence's embedding to that of the original sentence and replaces them with semantically close words. For the first part, we exploit USE embedding [21] and for the second, we approximate the semantic proximity of words with wikipedia pre-trained FASTTEXT embeddings [22][3].

SYN-REV: same process as above, but differs in picking the top-10 tokens whose deletion minimizes the cosine distance of the sentence's embedding.

BACK-TRANSL: consists of translating the input text to some other language and then translating back the translation into English [11, 12]. Here, we translated to French, German and Polish and then back to English using an in-house NMT-based solution [24].

Some examples of the application of the DA techniques enumerated above are provided in Table 8 in Annex A. While most of these techniques were reported elsewhere, GEO and PER-ORG, i.e., replacement of specific types of named entities, to the best of our knowledge, were not explicitly explored. Based on empirical observations, the application of these transformations result in label preservation with high probability, although the transformed texts might appear ‘unrealistic’ due to random name replacement. Furthermore, since the replacement is based on a lexicon look-up, the transformation might result in replacing entities of other type by mistake, but, again, based on empirical observations, this does not have high impact on the label.

Additionally, we explored ways of combining the DA techniques enumerated above, incl.: (a)

[2] https://www.geonames.org/

[3] We exploit the GENSIM interface [23]

ALL: combination the results of all the above DA techniques created separately, (b) **ALL-KB:**, a variant of ALL, but combining only DA techniques based on knowledge-based resources, i.e., PUNCT, DEL-ADJ-ADV, DATE, GEO and PER-ORG, (c) **ALL-KB-STACKED:** resulting from running the techniques used in ALL-KB in a pipeline (in the order as above) that modifies progressively the same input text, and (d) **BEST-3** combination of the 3 DA techniques, whose results were merged (not stacked), and which yield best gain in performance (see Section 5).

4. EMM-derived CC denial text snippet corpus

In order to establish a benchmark corpus for the news domain and to test the classification performance, we relied on EMM. Articles taken from a limited set of news sources that disinformation experts had identified as frequently spreading misinformation. In order to limit the dataset to articles on CC, we queried for articles containing keywords related to the topic such as: '*climate change*', '*global warming*', '*greenhouse gas[es]*', '*greenhouse effect[s]*' and limited the publication date to the whole of 2021. Out of these, a random subset of 2500 articles was sampled. For each article, we generated a snippet made of the title and of up to the first 500 characters. The corpus was manually annotated by five disinformation experts, using the Codebook defined in [1]. 1118 snippets were annotated, 42.7% of which are tagged with a class indicating a CC denial narrative, while the second half has been tagged as *No claim*, i.e, not containing any CC denial claim captured by the Codebook. In some snippets, while inflammatory language superficially similar to CC denial was used, the texts actually embrace polemical stance on CC inaction. When stance was ambiguous, the snippet was discarded, whereas the remaining snippets containing activists stance were assigned the label *No claim*.

The statistics of the current version of the corpus[4] are provided in Annex A in Table 7.

5. Classification Experiments

We have experimented with two ML paradigms, namely: (a) linear SVM using the algorithm described in [25] and Liblinear library[5], with 3-6 character n-grams as binary features, using vector normalization and $c = 1.0$ resulting from parameter optimization, and (b) RoBERTA$_{large}$ architecture [26] using *batch size=32*, *learning rate 1e-5* and *class weighting*.

Prior to carrying out ML experiments we cleaned the original 4C corpus [1] due to some problems, i.a., (a) some entries were included in both training and test data, and often having different labels, and (b) some entries were corrupt, i.e., missing texts, non parseable content. We used this modified version of the 4C corpus, containing ca. 30 entries less. The 4C dataset is highly imbalanced, i.e., more than 60% of the instances are labeled as *No claim*, whereas 14 classes constitute ca. 1-2% of the entire dataset each (see Table 6 in Annex A for statistics.).

The results of the evaluation of SVM and RoBERTa$_{large}$ on the 4C corpus without any DA are presented in Table 1, where we explored SVM both with and without class weighting. The performance of the baseline RoBERTa$_{large}$ is similar to the one of its counterpart reported in [1].

As regards DA techniques, we have augmented all instances of all **CC-denial** classes, whereas the *No claim* class was not augmented. Each to-be-augmented instance was augmented $l \in \{1, 2, 4\}$ times and the experiments have been repeated 3 times. The gain/loss obtained for SVM-

[4] Please note that this corpus is ongoing active development and will be continuously extended.

[5] https://www.csie.ntu.edu.tw/~cjlin/liblinear

Table 1
The results obtained with baseline models: SVM and RoBERTa-based transformers on 4C corpus.

SVM		weighted SVM		RoBERTa$_{large}$	
Accuracy	macro F_1	Accuracy	macro F_1	Accuracy	macro F_1
78.2	59.6	75.0	64.9	86.7	77.5

and RoBERTa-based models for all DA techniques is reported in Table 2 and 3 resp., with the best results per measure and number of augmentations marked in bold. In all experiments all original training data was used as well. BEST-3 refers to a combination of 3 DA techniques, each run separately, which yield best gain in performance and were: (a) PUNCT, BACK-TRANSL, GEO for SVM, and (b) PUNCT, GEO, PER-ORG for weighted SVM and RoBERTa.

Table 2
The gain in accuracy and macro F_1 obtained by using different DA techniques with SVM-based model. The figures in brackets refer to the SVM version without class weighting.

	1 augmentation		2 augmentations		4 augmentations	
DA Method	Accuracy gain	macro F_1 gain	Accuracy gain	macro F_1 gain	Accuracy gain	macro F_1 gain
COPY	+1.5 (**+0.9**)	+0.6 (+2.1)	+3.0 (+0.6)	+0.8 (+1.8)	+3.9 (+0.3)	**+0.9** (+1.7)
DATE	+0.2 (+0.3)	-0.4 (+0.9)	+0.3 (+0.1)	-0.2 (+0.5)	+0.5 (0.0)	-0.3 (+0,4)
DEL-ADJ-ADV	+1.2 (+0.5)	-0.1 (+1.1)	+2.1 (+0.2)	+0.3 (+1.2)	+3.1 (+0,4)	+0.3 (+1.1)
PUNCT	+1.9 (+0.6)	+0.4 (+1.9)	+3.4 (+0.2)	+1.1 (+1.0)	+4.3 (+0.5)	+1.2 (+1.6)
GEO	+1.1 (+0.6)	+0.1 (+2.0)	+2.0 (+0.6)	+0.5 (+2.1)	+3.0 (**+0.9**)	+0.6 (+2.7)
PER-ORG	+0.6 (+0,7)	+0.6 (+1.8)	+1.4 (+0.4)	+0.7 (+1.5)	+2.1 (+0.3)	+0.6 (+1.1)
SYN	+1.7 (+0.4)	-0.3 (+0.7)	+3.0 (+0.3)	-0.2 (+0.4)	+3.9 (+0.4)	-0.6 (+0.1)
SYN-REV	+0.7 (-0.8)	-1.5 (-1.7)	+2.4 (-0.1)	-0.9 (-0.2)	+3.9 (0.0)	-0.5 (-0.3)
BACK-TRANSL	+0.6 (+0.6)	-0.8 (+2.0)	+1.7 (+0.6)	-0.3 (+2.1)	+2.2 (+0.7)	-0.1 (+2.4)
ALL-KB	**+3.3** (**+0.9**)	**+1.6** (+2.3)	+4.0 (+1.0)	+1.0 (+2.8)	**+4.6** (+0.6)	0.1 (+2.1)
ALL-KB-STACKED	+1.7 (+0.8)	+0.4 (+2.5)	+2.9 (+0.8)	+0.9 (+2.5)	+3.8 (+0.8)	+0.7 (+2.6)
ALL	+3.2 (-0.2)	-0.1 (+2.7)	**+4.3** (+0.8)	+0.5 (+2.7)	**+4.6** (+0.7)	+0.2 (+1.9)
BEST-3	+2.4 (**+0.9**)	+1.4 (**+3.8**)	+3.7 (**+1.1**)	**+1.4** (**+3.8**)	**+4.6** (+0,7)	**+0.9** (**+3.1**)

As regards weighted SVM, one can observe that overall highest gain in macro F_1 was obtained with the ALL-KB setting (+1.6) with a 1-per-instance augmentation, and BEST-3 obtained highest gain for 2 and 4 augmentations (1.4 and 0.9 resp.). PUNCT appears to be the best stand-alone DA technique with some gains above 1.0. Applying simple copying (COPY) beats many other DA techniques (macro F_1 improved by up to +0.9), although it is outperformed by the ones mentioned earlier. The two new DA techniques, i.e., GEO and PER-ORG yield positive gain in all set-ups, while the usage of DATE, SYN, SYN-REV and BACK-TRANSL in a stand-alone mode does not appear to be beneficial, i.e., close to zero gain or deterioration. The DA gains for unweighted SVM are higher, but since the best setting (BEST-3) for unweighted SVM case is worse than the weighted SVM baseline, we do not analyze it any further.

As regards the RoBERTa$_{large}$-based models, one can observe that DA consistently deteriorates the accuracy on average, whereas for the most of the basic DA techniques there is little or

Table 3
The gain in accuracy and macro F_1 for DA techniques with RoBERTa-based model.

DA Method	1 augmentation Accuracy gain	 macro F_1 gain	2 augmentations Accuracy gain	 macro F_1 gain	4 augmentations Accuracy gain	 macro F_1 gain
COPY	-3.2	-0.8	-1.4	-0.1	-1.0	+0.4
DATE	-6.0	-3.0	-6.1	-2.8	-4.8	-2.2
DEL-ADJ-ADV	-4.7	-1.9	-2.4	-0.5	-1.3	-0.3
PUNCT	-2.1	-0.1	-0.9	+0.6	-0.7	+0.4
GEO	-4.4	-2.2	-2.0	-0.1	-1.6	+0.1
PER-ORG	-5.5	-3.1	-4.3	-2.0	-2.8	-0.8
SYN	-2.4	-0.5	-1.6	+0.1	-0.8	-0.1
SYN-REV	-3.1	-1.4	-1.5	-0.3	**-0.7**	+0.5
BACK-TRANSL	**0.0**	0.0	-1.2	-2.5	-1.7	**+0.9**
ALL-KB-STACKED	-2.7	-0.5	-1.5	**+0.6**	-1.0	+0.1
ALL-KB	-1.9	-0.1	-1.3	+0.1	-0.8	+0.2
ALL	-0.4	**+0.7**	**-0.4**	-0.4	-0.9	-0.3
BEST-3	-1.1	+0.6	-0.8	+0.5	-1.2	-0.5

no again at all in terms of macro F_1 with BACK-TRANSL exhibiting the highest gain (+0.9), followed by PUNCT (+0.6). The composite DA techniques perform on average better, with highest gain of 0.7 for ALL, which is higher than when applying simple COPY (0.4). Such results are consistent with recent literature exploring data augmentation techniques with RoBERTa in the related field of propaganda techniques classification [27].

RoBERTa's deterioration could be possibly explained by potential overfitting to the full sentence structure due to too similar sentences, given neural networks tendency to overfit [18]. However, we also observe that this phenomenon diminishes with more augmentations. While DATE should have the least impact on the label, it showed the most important and consistent drop in performances. A better understanding of this behaviour requires further investigation.

Interestingly, we have observed that PUNCT, SYN, and SYN-REV were the three basic DA techniques with highest variance (up to ca. 1.0 difference in the gain for macro F_1 across different experiments) and the same could be observed for the composite methods that do include these basic DA methods. In particular, given that the simple PUNCT method performs overall best across the different settings one could explore in future potential improvements that could be gained through some tuning, e.g., limiting the positions in which punctuation signs are inserted and/or studying whose punctuation sign insertion results in higher gains in performance.

We have applied the baseline and some DA-boosted models on the EMM-derived corpus described in Section 4, whose performance is summarized in Table 4. The deterioration in performance vis-a-vis 4C corpus evaluation could be mainly due to the different nature of the EMM corpus (text structure and writing style). Noteworthy, the evaluation on the EMM dataset revealed that the models trained using DA consistently outperform the baseline models. As regards the RoBERTa-based data-augmented models the gain ranges from -0.7 to +2.8 and -0.7 to +4.6 in accuracy and macro F_1 scores, respectively, with the vast majority being positive. The boost is the result of higher recall in the DA trained models. For the sake of completeness, the confusion matrix for the RoBERTa$_{large}$ model boosted with BACK-TRANSL augmentation (reported in Table 4) is provided in Figure 1 Annex A

Table 4
The performance of the baseline and some DA-based models on EMM-derived corpus.

Baseline models	Accuracy	macro F_1	DA-based models	Accuracy	macro F_1
SVM	63.9	36.8			
weighted SVM	62.7	46.4	weighted SVM + ALL-KB	65.1	48.4
RoBERTa$_{large}$	73.7	59.4	RoBERTa$_{large}$ + BACK-TRANSL	75.2	64.0

6. Data Augmentation Impact on Reducing the Bias

In order to better understand the behavior of the DA techniques relying on proper name replacement, namely GEO and PER-ORG, we performed additional experiments with alternate versions, and analysing the distribution of names entities. This is motivated by the finding that texts containing disinformation are often very specific about the entities involved. These alternate techniques are characterized by a different sampling strategy of the entities to be inserted. In contrast to the GEO and PER-ORG experiments, the replaced named entities are not taken from a larger pool of entities, but instead, are taken from the pool of the entities that are detected in the texts. We respectively define the additional experiments **GEO-SP** and **PER-ORG-SP** which correspond to the GEO and PER-ORG experiments using this modified sampling on the CC-denial classes only; **GEO-SP-ALL** and **PER-ORG-SP-ALL**, where this randomization procedure is applied to the CC-denial classes as well as to the *No claim* class; and finally **GEO-SP-STRICT** and **PER-ORG-STRICT**, where the instances of all classes are perturbed and only perturbed data is used. These experiments were only performed with weighted SVM, using only one augmentation. We report the results in Table 5. We also compare the augmented dataset and the original dataset using the Jensen-Shannon (JS) divergence on two distributions: (a) of the replaced entities, and (b) the labels associated with these entities.

In the GEO and PER-ORG experiments, the entities in the instances of the CC-denial classes were replaced with entities drawn from a much larger pool, practically removing these original entities from the augmented data. The clearly lower performance of the *-STRICT experiment, notably in terms of macro F_1 seems to indicate that some classes rely heavily on the presence of certain entities in order to be correctly predicted. This experiment is the only one not containing the original data at all, and the distribution of replaced entities diverges the most from the original dataset. Most of the errors are due to CC-denial texts being predicted as *No claim*, with the classes 4_* having the most issues, this is coherent as these classes are the most linked to policies, and therefore to the corresponding actors.

The *-SP experiments, where only CC-denial classes get augmented, show a small increase in performances. The increase in performance is notable in the *-SP-ALL experiments, where the *No claim* class also gets augmented. The distribution of entities diverges more than in the case of *-SP, but the distribution of labels associated with these entities diverges less. The combination of both the original dataset and the fully transformed one seems to yield the

Table 5
The gain in accuracy and macro F_1 obtained by using sampling from the same pool for named-entities based DA techniques with weighted SVM-based model, and the Jensen-Shannon divergence for distributions of replaced entities and labels thereof.

DA Method	Accuracy	macro F_1	JS divergence	
			Rep. Ent.	Label of Rep. Ent.
(none)	75.0	64.9	-	-
GEO	76.6 (+1.6)	66.9 (+2.0)	0.002	0.0
PER-ORG	75.7 (+0.7)	65.8 (+1.8)	0.0	0.0
GEO-SP	75.8 (+0.8)	64.9 (+0.0)	0.018	0.012
GEO-SP-ALL	77.8 (+2.8)	65.4 (+0.5)	0.069	0.0
GEO-SP-STRICT	75.3 (+0.3)	61.1 (-3.8)	0.253	0.0
PER-ORG-SP	75.9 (+0.9)	65.5 (+0.6)	0.022	0.012
PER-ORG-SP-ALL	77.9 (+2.9)	66.1 (+1.2)	0.067	0.0
PER-ORG-SP-STRICT	73.5 (-2.5)	50.8 (-14.1)	0.262	0.0

best compromise between generalization and fitting to particular entities in the test dataset. Exploring this interplay is an interesting direction for future works. Randomly swapping named entities could change an actual disinformation claim into factual information or vice versa. It is out of the scope of the classifier to deal with fact checking, however, it is important to reckon the competing interest between a classifier that generalises well to unseen claims on new entities and better fitting to the known narratives.

The *-SP experiments exhibit a performance on par or lower than the their equivalents without their characteristic sampling. For GEO-SP there is a clear performance gap with respect to GEO in terms of macro F_1. The reason why the divergence of GEO appears lower than GEO-SP is because it does not take into account newly introduced entities in GEO. Overall, both GEO, which introduces new entities, and GEO-SP, which changes the distribution of labels associated with existing entities, tend to improve the macro F_1 and accuracy.

7. Conclusions

We reported on preliminary experiments of using DA techniques for improving climate change denial classification. The evaluation on the 4C corpus yielded a boost with data augmentation up to 1.6 and 0.9 gain in macro F_1 for SVM- and RoBERTa-based classifiers resp. For the vast majority of the DA techniques respective SVM-based models resulted in gain, whereas for most of the RoBERTa-based models a loss was observed. Analysing the new EMM-derived test dataset introduced in this paper with ca. 1K snippets, DA techniques lead to up to 4.6 point gains in macro F_1 vis-a-vis baseline model. The overall performance is nevertheless worse than on the 4C corpus, which was expected due to the different nature of the sources considered.

We provided a more in-depth analysis of the behaviour of two DA techniques not reported earlier, which randomly replace toponyms and person/organisation names, and which were among the ones that resulted in higher gains in macro F_1 for SVM-based models.

We believe the reported findings will boost the NLP research in the climate change domain. We also make the cleaned version of the 4C and the new EMM-derived corpus publicly accessible[6].

[6] https://github.com/jpiskorski/CC-denial-resources

References

[1] T. G. Coan, C. Boussalis, J. Cook, M. O. Nanko, Computer-assisted classification of contrarian claims about climate change, Scientific Reports 11 (2021).

[2] M. Stede, R. Patz, The climate change debate and natural language processing, in: Proceedings of the 1st Workshop on NLP for Positive Impact, Association for Computational Linguistics, Online, 2021, pp. 8–18.

[3] N. Diakopoulos, A. X. Zhang, D. Elgesem, A. Salway, Identifying and analyzing moral evaluation frames in climate change blog discourse., in: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, 2014, pp. 583––586.

[4] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, C. Cherry, SemEval-2016 task 6: Detecting stance in tweets, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, San Diego, California, 2016, pp. 31–41.

[5] Y. Luo, , D. Card, D. Jurafsky, Detecting stance in media on global warming, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020.

[6] F. S. Varini, J. L. Boyd-Graber, M. Ciaramita, M. Leippold, Climatext: A dataset for climate change topic detection, 2020.

[7] A. Al-Rawi, D. O'Keefe, O. Kane, A.-J. Bizimana, Twitter's fake news discourses around climate change and global warming, Frontiers in Communication 6 (2021).

[8] S. Bhatia, J. H. Lau, T. Baldwin, You are right. I am ALARMED - but by climate change counter movement, CoRR (2020). arXiv:2004.14907.

[9] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6382–6388.

[10] H. Shi, K. Livescu, K. Gimpel, Substructure substitution: Structured data augmentation for NLP, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 3494–3508.

[11] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 86–96.

[12] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, Q. V. Le, Qanet: Combining local convolution with global self-attention for reading comprehension., CoRR abs/1804.09541 (2018).

[13] S. Kobayashi, Contextual augmentation: Data augmentation by words with paradigmatic relations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 452–457.

[14] Z. Xie, S. I. Wang, J. Li, D. Lévy, A. Nie, D. Jurafsky, A. Y. Ng, Data noising as smoothing in neural network language models, in: 5th International Conference on Learning Repre-

sentations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017.

[15] A. Karimi, L. Rossi, A. Prati, AEDA: An easier data augmentation technique for text classification, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2748–2754.

[16] A. Karimi, L. Rossi, A. Prati, Adversarial training for aspect-based sentiment analysis with bert, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 8797–8803.

[17] H. Zhang, M. Cissé, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, CoRR abs/1710.09412 (2017).

[18] M. Bayer, M. Kaufhold, C. Reuter, A survey on data augmentation for text classification, CoRR abs/2107.03158 (2021). URL: https://arxiv.org/abs/2107.03158. arXiv:2107.03158.

[19] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A survey of data augmentation approaches for NLP, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 968–988.

[20] M. Ehrmann, G. Jacquet, R. Steinberger, Jrc-names: Multilingual entity name variants and titles as linked data, Semantic Web 8 (2017) 283–295.

[21] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal sentence encoder for English, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 169–174. URL: https://aclanthology.org/D18-2029. doi:10.18653/v1/D18-2029.

[22] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[23] R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, pp. 45–50. http://is.muni.cz/publication/884893/en.

[24] C. Oravecz, K. Bontcheva, D. Kolovratník, B. Bhaskar, M. Jellinghaus, A. Eisele, etranslation's submissions to the WMT 2021 news translation task, in: L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno-Yepes, P. Koehn, T. Kocmi, A. Martins, M. Morishita, C. Monz (Eds.), Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021, Association for Computational Linguistics, 2021, pp. 172–179.

[25] K. Crammer, Y. Singer, On the learnability and design of output codes for multiclass problems, in: Proceedings of the Thirteenth Annual Conference on Computational Learning Theory, COLT '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000, p. 35–46.

[26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[27] V. Gupta, R. Sharma, Nlpiitr at semeval-2021 task 6: Roberta model with data augmentation for persuasion techniques detection, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 1061–1067.

A. Supplementary Information

The statistics for the 4C (Contrarian Claims about Climate Change) and the news-derived text snippet corpus are presented in Table 6 and 7 resp. Please note that both datasets cover only a fraction of types (18 out of 27) of the CC contrarian claim taxonomy [1].

Table 6
The training and test dataset statistics of the 4C (Contrarian Claims about Climate Change) corpus. The ‘code’ column contains the original codes from the 4C taxonomy.

		Training data		Test data	
Code	class Name	Number	%	Number	%
0_0	Other (*No claim*)	18110	69.56%	1754	60.40%
1_1	Ice isn’t melting	370	1.42%	51	1.76%
1_2	Heading into Ice Age	163	0.63%	21	0.72%
1_3	Weather is cold	254	0.98%	30	1.03%
1_4	Hiatus in Warming	537	2.06%	69	2.38%
1_6	Sea level rise is exaggerated	210	0.81%	26	0.90%
1_7	Extremes aren’t increasing	474	1.82%	65	2.24%
2_1	It’s natural cycles	875	3.36%	124	4.27%
2_3	No evidence of Greenhouse effect	377	1.45%	48	1.65%
3_1	Sensitivity is low	230	0.88%	26	0.90%
3_2	No species impact	375	1.44%	49	1.69%
3_3	Not a pollutant	358	1.38%	46	1.58%
4_1	Policies are harmful	364	1.40%	64	2.20%
4_2	Policies are ineffective	211	0.81%	34	1.17%
4_4	Clean energy won’t work	272	1.04%	39	1.34%
4_5	We need energy	202	0.78%	36	1.24%
5_1	Science is unreliable	1525	5.86%	225	7.75%
5_2	Movement is unreliable	1127	4.33%	197	6.78%

Table 7
The statistics of the EMM-derived corpus of text snippets on CC denial. The 'code' column contains the original codes from the 4C taxonomy.

Code	class Name	Number	%
0_0	Other (*No claim*)	641	57.33%
1_1	Ice isn't melting	14	1.25%
1_2	Heading into Ice Age	14	1.25%
1_3	Weather is cold	17	1.52%
1_4	Hiatus in Warming	10	0.89%
1_6	Sea level rise is exaggerated	4	0.69%
1_7	Extremes aren't increasing	16	1.43%
2_1	It's natural cycles	27	2.42%
2_3	No evidence of Greenhouse effect	15	1.34%
3_1	Sensitivity is low	7	0.63%
3_2	No species impact	13	1.16%
3_3	Not a pollutant	13	1.16%
4_1	Policies are harmful	55	4.92%
4_2	Policies are ineffective	27	2.42%
4_4	Clean energy won't work	11	0.98%
4_5	We need energy	9	0.81%
5_1	Science is unreliable	35	3.13%
5_2	Movement is unreliable	183	16.37%

Table 8
Examples of the results of applying the various Data Augmentation techniques.

ORIGINAL	*In Istanbul, the snow could easily reach up to 30 cm in June, Mayor Kadir Topba announced.*
DA Technique	Output
DATE	*In Istanbul, the snow could easily reach up to 30 cm in April, Mayor Kadir Topba announced.*
DEL-ADJ-ADV	*In Istanbul, the snow could reach up to 30 cm in June, Mayor Kadir Topba announced.*
PUNCT	*In Istanbul, the snow; could easily reach up to? 30 cm in June, Mayor Kadir Topba: announced.*
GEO	*In Porto Alegre, the snow could easily reach up to 30 cm in June, Mayor Kadir Topba announced.*
PER-ORG	*In Istanbul, the snow could easily reach up to 30 cm in June, Mayor Stephen King announced.*
SYN	*In Istanbul, the snow could easily be up to 30 cm in June, Mayor Kadir Topba said.*
BACK-TRANSL	*In Istanbul, Mayor Kadir Topba announced that the snow could easily be up to 30 cm high in June.*

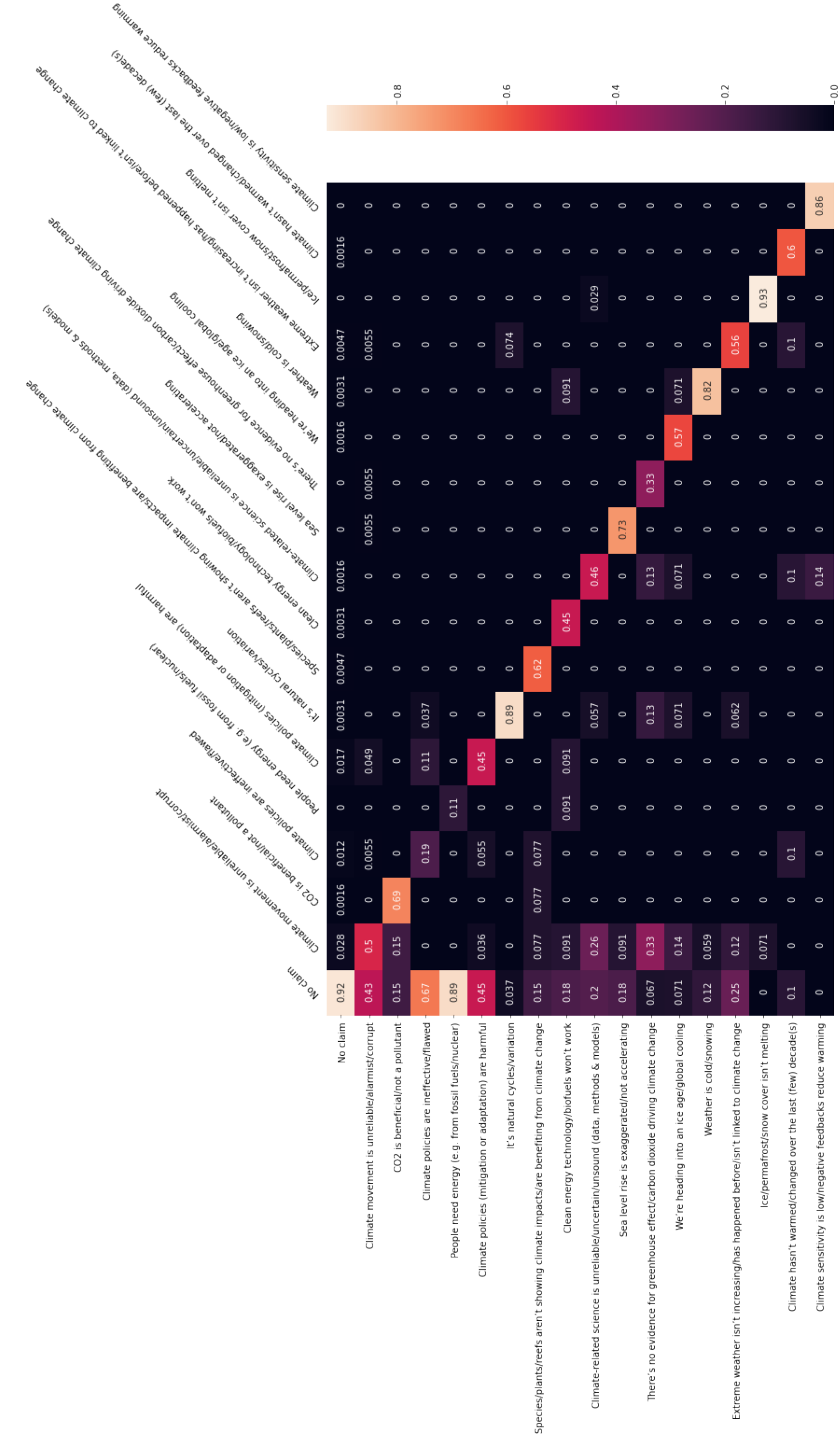

Figure 1: Confusion matrix for the results of RoBERTa$_{large}$ model trained using augmented data with BACK-TRANSL and tested on the EMM-derived corpus.