Automatic Classification of Neutralization Techniques in the Narrative of Climate Change Scepticism

Shraey Bhatia[1] **Jey Han Lau**[1] **Timothy Baldwin**[1]

[1] School of Computing and Information Systems,
The University of Melbourne

shraeybhatia@gmail.com, jeyhan.lau@gmail.com, tb@ldwin.net

Abstract

Neutralisation techniques, e.g. *denial of responsibility* and *denial of victim*, are used in the narrative of climate change scepticism to justify lack of action or to promote an alternative view. We collect manual annotations of neutralised techniques used in these texts, and explore semi-supervised models to automatically classify them.

1 Introduction

There is strong consensus in the scientific community on human-induced climate change (Cook et al., 2016; Powell, 2017). Despite this, action on climate change has become an increasingly partisan issue with strong opposition voices discrediting scientists, and spreading scepticism and misinformation. One such source is climate change counter movement organizations, which are an amalgam of lobbyists, big corporations, conservative think tanks, and media corporations (Dunlap and Jacques, 2013; Boussalis and Coan, 2016; Farrell, 2016; McKie, 2018), whose aim is to fuel climate change scepticism (CCS). Public perception is influenced by the narrative presented to them (Fløttum, 2014; Fløttum et al., 2016), and CCS texts use *neutralization* techniques to build counter-climate narratives (McKie, 2018).

The cure can't be worse than the disease/problem is a phrase frequently used by climate change sceptics,[1] and also recently by Donald Trump in reference to COVID-19.[2] Though two widely different issues, *neutralization* is used to justify opposing a policy, lack of action, and thus promotion of either total denial of the problem (Diethelm and McKee, 2009) or its severity. Table 1 presents two examples of neutralization in the context of climate change.

Sure, we should reduce greenhouse gases, but if our climate policies hurt our ability to create more wealth and bring power to the world's poor, then we are ridding the patient of the disease, but only by killing him
It's very convenient for alarmist greens to blame the fires of Australia and California on global warming. In reality, global warming is just a natural cycle and the policies they themselves advocate are the culprits.
The IPCC falsely attributes natural warming and urban warming to greenhouse gas (GHG) emission warming. It ignores the compelling evidence of natural climate change before 1950 that correlates well with indicators of solar activity

Table 1: Neutralization examples

In social science, neutralization is defined as justification/vindication for a deviant behaviour (Sykes and Matza, 1957; Maruna and Copes, 2005; Kaptein and Van Helvoort, 2019). Though initially developed in the field of criminology, it has been widely studied in different fields ranging from lack of corporate social responsibility (Cherry and Sneirson, 2010), to fast fashion (Joy et al., 2012), the tobacco industry (Fooks et al., 2013; Oreskes and Conway, 2010), and CCS (McKie, 2018). McKie (2018) argued that to fully understand the neutralization narrative around CCS, there is a need to break it down into specific techniques (e.g. denial of responsibility vs. denial of victim; see Section 3). Our paper proposes a method to automatically classify these neutralization techniques (henceforth "NT"), as a tool to analyse CCS narrative at scale and help build counter-narratives.

Our contributions in this work are as follows: (1) we introduce the NT (multilabel) classification task; (2) we develop and release a dataset with manual annotations of NT used in CCS texts; and (3) we explore semi-supervised models for the classifica-

[1] https://www.wired.com/story/the-analogy-between-covid-19-and-climate-change-is-eerily-precise/

[2] https://www.business-standard.com/article/international/trump-opposes-perpetual-lockdown-says-cure-cannot-be-worse-than-problem-120101300184_1.html

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2167–2175
June 6–11, 2021.

tion task, resulting in strong results on par with human performance. We release the code and data used in our experiments at: `https://github.com/sb1992/cc-neutralization`.

2 Related Work

Sykes and Matza (1957) first introduced the techniques of neutralization, known as the "famous five", as a tool for justification of deviant behaviour. The neutralization techniques inventory has since been expanded to include "metaphor of ledger" (Klockars, 1974), "excuse acceptance" (Minor, 1981) and "no one cares" (Shigihara, 2013). More recently, Kaptein and Van Helvoort (2019) developed a schema which combined them into a hierarchy of categorizations and sub-categorizations. McKie (2018) extended the work of Sykes and Matza (1957) to `CCS`.

Separately, research on fake news and propaganda has primarily operated at the article level, and focused on binary detection (presence vs. absence) (Barrón-Cedeno et al., 2019; Rashkin et al., 2017). Da San Martino et al. (2019) argued for the need for finer granularity in propaganda detection, both in terms of propaganda sub-types and fragment-level detection. In a similar vein, Nakamura et al. (2020) proposed fine-grained classes of fake news to differentiate between misleading, manipulated, or totally false content. More recently in the climate change domain, Luo et al. (2020) released a stance-annotated dataset for global warming, and proposed an opinion framing task to study discourse used in the debate around global warming.

One challenge in building supervised NLP models is the strong dependency on labelled data. To tackle this, one approach is apply transfer learning from pretrained language models (Radford et al., 2019; Peters et al., 2018; Yang et al., 2019; Conneau and Lample, 2019; Devlin et al., 2019). Another approach is semi-supervised learning. Yang et al. (2017) and Gururangan et al. (2019) employed variational autoencoders, and Clark et al. (2018) leveraged cross-view training using a mixture of labelled and unlabelled data. More recently, pretrained models and semi-supervised learning have been combined with great success, e.g. Xie et al. (2020) used `BERT` along with consistency regularization on unlabeled data, Croce et al. (2020) extended the fine-tuning process of `BERT` to a generative adversarial setting, and Chen et al. (2020) used interpolation to mix up the hidden representations of `BERT` to create augmented data for training.

3 Neutralization and Frames

Dunlap and Brulle (2015), Farrell (2016), and Boussalis and Coan (2016) categorised `CCS` arguments into 2 frames: science ("`SCIENCE`") and policy ("`POLICY`"). `SCIENCE` questions the scientific facts, is heavy on denial, or promotes pseudo science, whereas `POLICY` deals with issues of cost and economy (e.g. carbon tax), targets the scientists, or passes the blame for action to other nations. McKie (2018) rebranded the `CCS` arguments on neutralization by adapting Sykes and Matza (1957)'s original `NT` schema, establishing a connection between `NT` and `SCIENCE`/`POLICY` frames. We adopt the definitions and coding schema from McKie (2018), as follows (the first four of which relate to the `SCIENCE` frame, and the last three to the `POLICY` frame, as indicated):

- **Denial of Responsibility (`Deny-Responsibility` ⇝`SCIENCE`)**: climate change is happening, but is a natural cycle and human are not responsible.
- **Denial of Injury1 (`Deny-Injury1` ⇝`SCIENCE`)**: there are no significant harms attributable to climate change, and claims are generally overstated.
- **Denial of Injury2 (`Deny-Injury2` ⇝`SCIENCE`)**: there are benefits in rising rising C02 levels which have a positive effect on the environment.
- **Denial of Victim (`Deny-Victim` ⇝`SCIENCE`)**: there is no evidence of climate change and no climate change victims; total denial of any global warming.
- **Condemnation of the Condemner (`Condemn` ⇝`POLICY`)**: climate change is misrepresented by scientists or manipulated by politicians, the media, environmentalists, etc.
- **Appeal to Higher Loyalties (`Loyalties` ⇝`POLICY`)**: economic progress and development are more important than action on climate change, and hence policies like renewables or carbon taxes are not worth it.
- **Justification by Comparison (`Justify` ⇝`POLICY`)**: our actions are not as important as other countries which pollute more, or there are other more important issues than global

warming.

Examples of these 7 neutralization techniques are given in Table 2. `CCS` texts often use multiple `NT` together in their narrative (hence motivating a multilabel classification task), as seen in the second example in Table 1 where `Condemn` (`POLICY`) is used to blame the *alarmist greens* and `Deny-Responsibility` (`SCIENCE`) is used to highlight that global warming is a *natural cycle*. Similarly, in third example as well we see `Condemn` (`POLICY`) is used to accuse the IPCC (Intergovernmental Panel on Climate Change,[3] in conjunction with `Deny-Responsibility` (`SCIENCE`) to point out climate change being a natural and linked to solar activity.

4 Dataset

We construct our neutralisaton techniques dataset from 3 sources: (1) paragraphs extracted from `CCS` documents (Bhatia et al., 2020); (2) `CCS` sentences/paragraphs from McKie (2018);[4] and (3) anti-global warming opinions (sentences) from Luo et al. (2020).[5] This results in a mixture of sentences and paragraphs, resulting in diversity in the dataset (with longer snippets expected to have more multi-labelling). We henceforth call these text snippets "sentences" for brevity.

Our dataset has a total of 8000 sentences, of which 785 were annotated (and the remainder used as unlabelled data). We formulate the task as a multi-label classification problem where an annotator selects `NONE`, or one or more `NT` labels.

To make the task easier for annotators, we split it into 2 `NT` annotation subtasks based on the two frames: (1) the `SCIENCE` frame (`Deny-Responsibility`, `Deny-Injury1`, `Deny-Injury2`, `Deny-Victim`, or `NONE`); and (2) the `POLICY` frame (`Condemn`, `Loyalties`, `Justify`, or `NONE`). We combine annotations by taking a majority vote within each frame, and label a sentence as `NONE` only if it is the majority-class for both sub-tasks (i.e. none of the `NT` labels are majority-assigned for either frame). We collect human judgements using Amazon Mechanical Turk with 9 sentences forming a single HIT, one of which acts as a quality control in the form of a labelled data instance from McKie (2018). Each HIT was annotated by a minimum of 5 and maximum of 10 annotators. For further details of the annotation process, see Section 8.

We present statistics of the labelled data in Table 3. Interestingly, we see 3 large classes of `NT`—`Deny-Victim`, `Condemn`, and `Loyalties`—implying that most `CCS` narratives completely deny climate change, condemn the scientists, and prioritise the economy.

5 Automatic Classification

We experiment with SVM as a baseline and then explore several BERT-based supervised and semi-supervised models for classification (Devlin et al., 2019). As it is a multilabel classification problem, we add a number of one-vs-rest classification layers (one for each class) on top of BERT, and update all parameters during fine-tuning.

`SVM`: Standard linear-kernel `SVM` used in one vs. rest mode, and adapted to a multilabel setting.

`BERT`: Standard supervised BERT fine-tuned using the labelled data.

`MTEXT`: A semi-supervised BERT-based model based on Chen et al. (2020) extended to a multi-label setting. `MTEXT` combines the hidden representation of 2 training instances (drawn from both labelled and unlabelled instances) via interpolation to create a large number of augmented data samples. The supervised objective ($\mathcal{L}_s$) uses standard cross-entropy loss whereas the unsupervised objective uses consistency loss ($\mathcal{L}_{cl}$) in the form of KL-divergence. $\mathcal{L}_{cl}$ is computed both on labelled and unlabelled data, where the labels for the unlabelled data are inferred in a self-training manner. To encourage sharp probabilities for unsupervised instances, an entropy minimization loss $\mathcal{L}_{em}$ is added, yielding the overall objective $\mathcal{L}_{nt}$ = $w_1\mathcal{L}_s + w_2\mathcal{L}_{cl} + w_3\mathcal{L}_{em}$, where w_x are tunable hyper-parameters.

$\texttt{MTEXT}_{\texttt{multi}}$: As we see in Section 3, `NT` is associated with `SCIENCE` and `POLICY` frames. We experiment with adding these frames (including the `NONE` class, 3 in total) as an auxiliary objective, creating another supervised loss ($\mathcal{L}_{frame}$).[6] The final objective is $\mathcal{L}_{nt} + \alpha\mathcal{L}_{frame}$, where α is a tunable hyper-parameter.

Following Gururangan et al. (2020), we also experiment with adaptive pretraining for BERT,

[3] https://www.ipcc.ch/about/

[4] Extracted from the appendix of their thesis.

[5] Opinions which disagree with the statement: *climate change/global warming is a serious concern.*

[6] $\mathcal{L}_{frame}$ is implemented as multilabel loss, as a sentence can have both `SCIENCE` and `POLICY` frames.

Argument or Example	NT	Frame
There's no indication this is anything other than natural variability, with humans not playing a part	Deny-Responsibility	SCIENCE
There is a very real probability that global warming has been overestimated by computer models, and won't be too bad	Deny-Injury1	SCIENCE
CO2 is plant food and good for the planet, as it is essential for plants in photosynthesis	Deny-Injury2	SCIENCE
Despite forecasts of warming, the world has actually been cooling, so global warming is a hoax	Deny-Victim	SCIENCE
An avalanche of global warming alarmism is about to hit, thanks to environmentalists, the media, and a few scientists	Condemn	POLICY
So-called "new renewable energy technologies" are extremely expensive and rely on huge subsidies, pushing up energy costs	Loyalties	POLICY
New Zealand's actions should be less ambitious than Australia's because Australia is a wealthier country	Justify	POLICY

Table 2: Examples of counter climate arguments and their frames.

NT	%	Sentence Length
Deny-Responsibility	11.47	44.43
Deny-Injury1	8.78	44.71
Deny-Injury2	9.67	41.56
Deny-Victim	23.18	42.31
Condemn	35.67	49.89
Loyalties	21.23	48.87
Justify	4.01	50.09
NONE	7.52	36.31

Table 3: Distribution across classes.

Model	P	R	F
BERT	0.57	0.62	0.59
BERT*	0.60	0.64	0.62
MTEXT	0.62	0.71	0.66
MTEXT*	0.63	0.71	0.67
$\text{MTEXT}_{\text{multi}}$	**0.64**	**0.73**	**0.68**
$\text{MTEXT}^*_{\text{multi}}$	0.62	0.71	0.67
SVM	0.78	0.39	0.49
Human	0.69	0.72	0.70

Table 4: NT multi-label classification performance. "P", "R", and "F" denote micro-precision, micro-recall and micro-F1 respectively.

i.e. before we fine-tune BERT to our task, we pretrain the off-the-shelf BERT using the masked language model objective on CCS documents (Bhatia et al., 2020). Models with adaptive pretraining are marked with '*', e.g. $\text{MTEXT}^*_{\text{multi}}$.

At test time, we add two extra post-processing rules for the NONE class: (1) it is automatically selected if all other classes are predicted to be absent; and (2) it is never selected if any other classes are predicted to be present.

6 Experiments

We split the labelled data into train/dev/test with 450/135/200 sentences. The semi-supervised models (MTEXT variations) also have access to the unlabelled 7215 sentences. We use the uncased BERT-base as the pretrained model for all experiments. We detail the full training details and hyperparameters in supplementary material.

We present micro-precision, micro-recall and micro-F1 results for the test-set in Table 4. To provide an upper bound, we also present estimated human performance, which is computed by randomly isolating a worker's annotations, and calculating agreement with the rest for the test instances (repeated 100 times to reduce variance, and micro-averaged).

We first look at the (fully) supervised results, and see that the baseline BERT performs the worst, but adaptive pretraining (BERT*) boosts results.

Moving on to semi-supervised models (MTEXT, MTEXT*, $\text{MTEXT}_{\text{multi}}$ and $\text{MTEXT}^*_{\text{multi}}$), we see consistent gains, highlighting the benefits of using unlabelled data. $\text{MTEXT}_{\text{multi}}$ with its multi-task objective gives a small but appreciable gain over MTEXT, producing performance that is on par with human performance. Interestingly, adaptive

Model	**Deny-Responsibility**	Deny-Injury1	Deny-Injury2	**Deny-Victim**	**Condemn**	**Loyalties**	Justify	NONE
BERT	0.51	0.13	0.52	0.62	0.72	0.72	0.00	0.20
BERT*	0.57	0.13	0.49	0.64	0.73	0.80	0.00	0.20
MTEXT	0.68	0.40	0.73	0.65	0.73	0.76	0.30	0.30
$\mathtt{MTEXT}_{\mathtt{multi}}$	0.62	0.50	0.73	0.70	0.73	0.80	0.35	0.30
SVM	0.30	0.08	0.56	0.38	0.68	0.67	0.00	0.00
Human	0.64	0.62	0.88	0.65	0.77	0.81	0.61	0.56

Table 5: F1 breakdown across classes. The 4 largest classes (Deny-Responsibility, Deny-Victim, Condemn and Loyalties) are bolded.

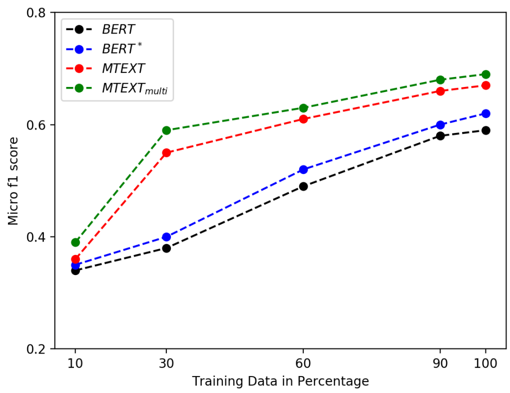

Figure 1: micro-F1 performance over increasing amounts of training data.

pretraining (MTEXT* and $\mathtt{MTEXT}^*_{\mathtt{multi}}$) does not seem to help here, we suspect because both techniques are based on the same idea, i.e. to improve performance by leveraging additional unlabelled data. SVM- a simple baseline has the lowest performance

To better understand how "data efficient" these models are, we present micro-F1 over varying amounts of labelled training data in Figure 1. We see that MTEXT and $\mathtt{MTEXT}_{\mathtt{multi}}$ outperform BERT and BERT* substantially with only 30% training data (135 instances) and maintain their the strong performance as data quantity increases.

Finally, we present a breakdown of F1 scores for each class in Table 5. Adaptive learning mostly improves the two large classes (Deny-Responsibility and Loyalties) for BERT vs. BERT*. When we incorporate semi-supervised learning (MTEXT and $\mathtt{MTEXT}_{\mathtt{multi}}$), we see large improvements for all the small classes (Deny-Injury1, Deny-Injury2, and Justify), suggesting that semi-supervised learning benefits the smaller classes most. Similar to Table 4 to get an estimated upper bound we also present human F1 scores for each class. Looking at those scores, we observe that the gap with human performance is higher for the smaller classes (Deny-Injury1, Deny-Injury2, and Justify) even for our best model, highlighting the limitations of semi-supervised learning.

6.1 Technical Details

For the supervised BERT models, we use the following fine-tuning hyper-parameters: batch size=10, epoch =3, learning rate=0.0005, number of epochs =3 and use BERT-base-uncased as the base model. For semi-supervised MTEXT based models, we use following hyper-parameters: labelled batch size=2, unlabelled batch size=5, sharpening temperature=0.6, the beta distribution parameter = 0.2,[7] learning rate=0.00005, w_1 =1, w_2 = 1, w_3 =0.8 in $w_1\mathcal{L}_s + w_2\mathcal{L}_{cl} + w_3\mathcal{L}_{em}$, α for auxiliary objective $\alpha\mathcal{L}_{frame}$ = 0.8, and perform data augmentation for unlabelled data using German and Russian as pivot languages, similar to Chen et al. (2020). For SVM, we use unigrams and bigrams as features with tf-idf weighting and the regularization parameter $C = 10$. More training details are provided in supplementary material.

7 Conclusion

We draw on social science literature in introducing the notion of "neutralisation", in the context of climate change sceptics. We collect annotations of neutralisation techniques in text relating to climate change, and experiment with supervised and semi-supervised BERT-based models.

8 Ethical Considerations

8.1 Mechanical Turk

To pass quality control for a given HIT, the annotator has to select the correct class for the quality control sentence (which is not flagged in any way to the annotator, and presented in random order);

[7]We use a small value here to ensure the generated data in the model is similar to labelled data with small noise regularization

the annotations from a given HIT are not used to determine consensus labelling if their average pass rate across all HITs attempted is ≤ 0.7. We collect additional annotations by releasing the task internally to a small number of local workers.[8] Each HIT was paid at USD$0.61, and took an average of 5 minutes to complete. This amounts to $7.32 per hour, which is slightly above US federal minimum wage ($7.25).

References

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

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2020. You are right. I am ALARMED–but by climate change counter movement. *arXiv preprint arXiv:2004.14907*.

Constantine Boussalis and Travis G Coan. 2016. Text-mining the signals of climate change doubt. *Global Environmental Change*, 36:89–100.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics.

Miriam A Cherry and Judd F Sneirson. 2010. Beyond profit: Rethinking corporate social responsibility and greenwashing after the BP oil disaster. *Tulane Law Review*, 85:983.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

John Cook, Naomi Oreskes, Peter T Doran, William RL Anderegg, Bart Verheggen, Ed W Maibach, J Stuart Carlton, Stephan Lewandowsky, Andrew G Skuce, Sarah A Green, et al. 2016. Consensus on consensus: a synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, 11(4):048002.

[8]The task is restricted to workers with an approval of 97%+, based in the US, Canada, UK, Australia, or New Zealand.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119.

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)*, pages 5640–5650.

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.

Pascal Diethelm and Martin McKee. 2009. Denialism: what is it and how should scientists respond? *The European Journal of Public Health*, 19(1):2–4.

Riley E Dunlap and Robert J Brulle. 2015. *Climate change and society: Sociological perspectives*. Oxford University Press.

Riley E Dunlap and Peter J Jacques. 2013. Climate change denial books and conservative think tanks: Exploring the connection. *American Behavioral Scientist*, 57(6):699–731.

Justin Farrell. 2016. Network structure and influence of the climate change counter-movement. *Nature Climate Change*, 6(4):370.

Kjersti Fløttum. 2014. Linguistic mediation of climate change discourse. *ASp. la revue du GERAS*, (65):7–20.

Kjersti Fløttum, Trine Dahl, and Vegard Rivenes. 2016. Young Norwegians and their views on climate change and the future: findings from a climate concerned and oil-rich nation. *Journal of Youth Studies*, 19(8):1128–1143.

Gary Fooks, Anna Gilmore, Jeff Collin, Chris Holden, and Kelley Lee. 2013. The limits of corporate social responsibility: techniques of neutralization, stakeholder management and political CSR. *Journal of Business Ethics*, 112(2):283–299.

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A Smith. 2019. Variational pretraining for semi-supervised text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5880–5894.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Annamma Joy, John F Sherry Jr, Alladi Venkatesh, Jeff Wang, and Ricky Chan. 2012. Fast fashion, sustainability, and the ethical appeal of luxury brands. *Fashion Theory*, 16(3):273–295.

Muel Kaptein and Martien Van Helvoort. 2019. A model of neutralization techniques. *Deviant Behavior*, 40(10):1260–1285.

Carl B J Klockars. 1974. *The Professional Fence*. The Free Press.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. DeSMOG: Detecting stance in media on global warming. In *Findings of EMNLP 2020*.

Shadd Maruna and Heith Copes. 2005. What have we learned from five decades of neutralization research? *Crime and justice*, 32:221–320.

Ruth McKie. 2018. *Rebranding the Climate Change Counter Movement through a Criminological and Political Economic Lens*. Ph.D. thesis, Northumbria University.

W William Minor. 1981. Techniques of neutralization: A reconceptualization and empirical examination. *Journal of Research in Crime and Delinquency*, 18(2):295–318.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6149–6157.

Naomi Oreskes and Erik M Conway. 2010. Defeating the merchants of doubt. *Nature*, 465(7299):686–687.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

James Powell. 2017. Scientists reach 100% consensus on anthropogenic global warming. *Bulletin of Science, Technology & Society*, 37(4):183–184.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

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.

Amanda M Shigihara. 2013. It's only stealing a little a lot: Techniques of neutralization for theft among restaurant workers. *Deviant Behavior*, 34(6):494–512.

Gresham M Sykes and David Matza. 1957. Techniques of neutralization: A theory of delinquency. *American sociological review*, 22(6):664–670.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2020. Unsupervised data augmentation for consistency training. In *Proceedings of NeurIPS 2020*.

Yun Yang, Fengtao Nan, Po Yang, Qiang Meng, Yingfu Xie, Dehai Zhang, and Khan Muhammad. 2019. GAN-based semi-supervised learning approach for clinical decision support in health-iot platform. *IEEE Access*, 7:8048–8057.

Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. 2017. Improved variational autoencoders for text modeling using dilated convolutions. In *International Conference on Machine Learning*, pages 3881–3890.

Supplementary Material

1 Training Details

For the supervised BERT models, we use the following fine-tuning hyper-parameters: batch size=10, epoch =3, learning rate=0.0005, number of epochs =3 and use BERT-base-uncased as the base model. We tune our decision boundary threshold to classify the presence of a label based on development set and are 0.2 for DOR, 0.2 for DOI1, 0.2 for DOI2, 0.3 for DOV, 0.3 for COC, 0.3 for AHL, 0.2 for JBC, and 0.2 for NONE.

For semi-supervised MTEXT based models, we use following hyper-parameters: labelled batch size=2, unlabelled batch size=5, sharpening temperature=0.6, epoch =3, the beta distribution parameter = 0.2,[1], learning rate=0.00005, w_1 =1, w_2 =1, w_3 =0.8 in $w_1\mathcal{L}_s$ + $w_2\mathcal{L}_{cl}$ + $w_3\mathcal{L}_{em}$, α for auxiliary objective $\alpha\mathcal{L}_{frame}$ = 0.8. mixing layers as 7,9,12 and use BERT-base-uncased. We tune our decision boundary threshold to classify the presence of a label based on development set and are 0.75 for DOR, 0.70 for DOI1, 0.70 for DOI2, 0.80 for DOV, 0.85 for COC, 0.80 for AHL, 0.70 for JBC, and 0.60 for NONE. We use 2 augmentations (based on back translation) with Russian and German as the intermediate language.

2 Other Details

- Computing Infrastructure: We use RTX 2080 Ti and GTX 1080. In MTEXT based models we use 2 gpus when trained with RTX 2080 ti and 3 gpus when trained with GTX 1080. BERT based models are trained on a single gpu.
- Average run time. BERT based models are quite quick and a minute per epoch to fine tune whereas MTEXT based models take around 20 minutes for each epoch.
- As all the models are based on BERT-Base-uncased the number of parameters are around 110M [2]
- Validation performance of the various models are given in Table 1

[1] we use a small value here to ensure the generated data in the model is similar to labelled data with small noise regularization

[2] strictly speaking number of parameters in $\mathrm{MTEXT}_{\mathtt{multi}}$ will be slightly more due to auxiliary objective but is insignificant in overall picture

Model	**F**
BERT	0.59
BERT*	0.61
MTEXT	0.62
MTEXT*	0.64
$\texttt{MTEXT}_{\texttt{multi}}$	0.66
$\texttt{MTEXT}^{*}_{\texttt{multi}}$	0.65

Table 1: NT multi-label classification performance on validation data. “F” denote micro-F1 respectively.

- Hyperparameter tuning was done using manual search and the criteria used was micro-F1 on validation set.
- Parameters used for final set of experiments are given in the above section of Training Details