Dictionary-Assisted Supervised Contrastive Learning

Patrick Y. Wu[1], **Richard Bonneau**[1,2,4,5], **Joshua A. Tucker**[1,2,3], and **Jonathan Nagler**[1,2,3]
[1] Center for Social Media and Politics, New York University
[2] Center for Data Science, New York University
[3] Department of Politics, New York University
[4] Department of Biology, New York University
[5] Courant Institute of Mathematical Sciences, New York University
{pyw230, bonneau, joshua.tucker, jonathan.nagler}@nyu.edu

Abstract

Text analysis in the social sciences often involves using specialized dictionaries to reason with abstract concepts, such as perceptions about the economy or abuse on social media. These dictionaries allow researchers to impart domain knowledge and note subtle usages of words relating to a concept(s) of interest. We introduce the dictionary-assisted supervised contrastive learning (DASCL) objective, allowing researchers to leverage specialized dictionaries when fine-tuning pretrained language models. The text is first keyword simplified: a common, fixed token replaces any word in the corpus that appears in the dictionary(ies) relevant to the concept of interest. During fine-tuning, a supervised contrastive objective draws closer the embeddings of the original and keyword-simplified texts of the same class while pushing further apart the embeddings of different classes. The keyword-simplified texts of the same class are more textually similar than their original text counterparts, which additionally draws the embeddings of the same class closer together. Combining DASCL and cross-entropy improves classification performance metrics in few-shot learning settings and social science applications compared to using cross-entropy alone and alternative contrastive and data augmentation methods.[1]

1 Introduction

We propose a supervised contrastive learning approach that allows researchers to incorporate dictionaries of words related to a concept of interest when fine-tuning pretrained language models. It is conceptually simple, requires low computational resources, and is usable with most pretrained language models.

Dictionaries contain words that hint at the sentiment, stance, or perception of a document (see, e.g., Fei et al., 2012). Social science experts often craft these dictionaries, making them useful when the underlying concept of interest is abstract (see, e.g., Brady et al., 2017; Young and Soroka, 2012). Dictionaries are also useful when specific words that are pivotal to determining the classification of a document may not exist in the training data. This is a particularly salient issue with small corpora, which is often the case in the social sciences.

However, recent supervised machine learning approaches do not use these dictionaries. We propose a contrastive learning approach, dictionary-assisted supervised contrastive learning (DASCL), that allows researchers to leverage these expert-crafted dictionaries when fine-tuning pretrained language models. We replace all the words in the corpus that belong to a specific lexicon with a fixed, common token. When using an appropriate dictionary, keyword simplification increases the textual similarity of documents in the same class. We then use a supervised contrastive objective to draw together text embeddings of the same class and push further apart the text embeddings of different classes (Khosla et al., 2020; Gunel et al., 2021). Figure 1 visualizes the intuition of our proposed method.

The contributions of this project are as follows.

- We propose keyword simplification, detailed in Section 3.1, to make documents of the same class more textually similar.
- We outline a supervised contrastive loss function, described in Section 3.2, that learns patterns within and across the original and keyword-simplified texts.
- We find classification performance improvements in few-shot learning settings and social science applications compared to two strong baselines: (1) RoBERTa (Liu et al., 2019) / BERT (Devlin et al., 2019) fine-tuned with cross-entropy loss, and (2) the supervised contrastive learning approach detailed in Gunel et al. (2021), the most closely related approach to DASCL. To be clear, although BERT and

[1] Our code is available at https://github.com/SMAPPNYU/DASCL.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10217 - 10235
December 7-11, 2022

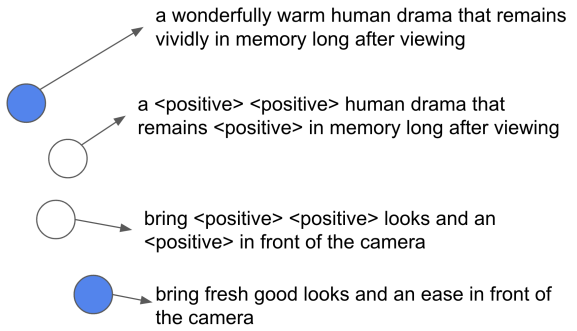


Figure 1: The blue dots are embeddings of the original reviews and the white dots are the embeddings of the keyword-simplified reviews from the SST-2 dataset (Wang et al., 2018). Both reviews are positive, although they do not overlap in any positive words used. The reviews are more textually similar after keyword simplification. Using $\text{BERT}_{\text{BASE-UNCASED}}$ out-of-the-box, the cosine similarity between the original reviews is .654 and the cosine similarity between the keyword-simplified reviews is .842. Although there are some issues with using cosine similarity with BERT embeddings (see, e.g., Ethayarajh, 2019; Zhou et al., 2022), we use it as a rough heuristic here.

RoBERTa are not state-of-the-art pretrained language models, DASCL can augment the loss functions of state-of-the-art pretrained language models.

2 Related Work

Use of Pretrained Language Models in the Social Sciences. Transformers-based pretrained language models have become the de facto approach when classifying text data (see, e.g., Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020), and are seeing wider adoption in the social sciences. Terechshenko et al. (2021) show that RoBERTa and XLNet (Yang et al., 2019) outperform bag-of-words approaches for political science text classification tasks. Ballard et al. (2022) use BERTweet (Nguyen et al., 2020) to classify tweets expressing polarizing rhetoric. Lai et al. (2022) use BERT to classify the political ideologies of YouTube videos using text video metadata. DASCL can be used with most pretrained language models, so it can potentially improve results across a range of social science research.

Usage of Dictionaries. Dictionaries play an important role in understanding the meaning behind text in the social sciences. Brady et al. (2017) use a moral and emotional dictionary to predict whether tweets using these types of terms increase their diffusion within and between ideological groups. Simchon et al. (2022) create a dictionary of politically polarized language and analyze how trolls use this language on social media. Hopkins et al. (2017) use dictionaries of positive and negative economic terms to understand perceptions of the economy in newspaper articles. Although dictionary-based classification has fallen out of favor, dictionaries still contain valuable information about usages of specific or subtle language.

Text Data Augmentation. Text data augmentation techniques include backtranslation (Sennrich et al., 2016) and rule-based data augmentations such as random synonym replacements, random insertions, random swaps, and random deletions (Wei and Zou, 2019; Karimi et al., 2021). Shorten et al. (2021) survey text data augmentation techniques. Longpre et al. (2020) find that task-agnostic data augmentations typically do not improve the classification performance of pretrained language models. We choose dictionaries for keyword simplification based on the concept of interest underlying the classification task and use the keyword-simplified text with a contrastive loss function.

Contrastive Learning. Most works on contrastive learning have focused on self-supervised contrastive learning. In computer vision, images and their augmentations are treated as positives and other images as negatives. Recent contrastive learning approaches match or outperform their supervised pretrained image model counterparts, often using a small fraction of available annotated data (see, e.g., Chen et al., 2020a; He et al., 2020; Chen et al., 2020b; Grill et al., 2020). Self-supervised contrastive learning has also been used in natural language processing, matching or outperforming pretrained language models on benchmark tasks (see, e.g., Fang et al., 2020; Klein and Nabi, 2020).

Our approach is most closely related to works on supervised contrastive learning. Wen et al. (2016) propose a loss function called center loss that minimizes the intraclass distances of the convolutional neural network features. Khosla et al. (2020) develop a supervised loss function that generalizes NT-Xent (Chen et al., 2020a) to an arbitrary number of positives. Our work is closest to that of Gunel et al. (2021), who also use a version of NT-Xent extended to an arbitrary number of positives with pretrained language models. Their supervised contrastive loss function is detailed in Section A.1.

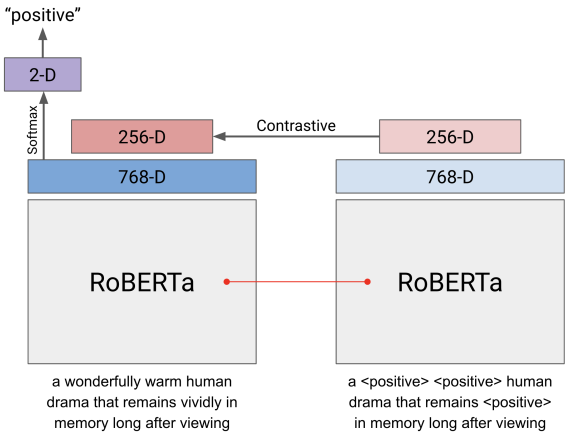


Figure 2: Overview of the proposed method. Although RoBERTa is shown, any pretrained language model will work with this approach. The two RoBERTa networks share the same weights. The dimension of the projection layer is arbitrary.

3 Method

The approach consists of keyword simplification and the contrastive objective function. Figure 2 shows an overview of the proposed framework.

3.1 Keyword Simplification

The first step of the DASCL framework is keyword simplification. We select a set of M dictionaries $\mathcal{D}$. For each dictionary $d_i \in \mathcal{D}$, $i \in \{1, ..., M\}$, we assign a token t_i. Then, we iterate through the corpus and replace any word w_j in dictionary d_i with the token t_i. We repeat these steps for each dictionary. For example, if we have a dictionary of positive words, then applying keyword simplification to

```
a wonderfully warm human drama that remains
    vividly in memory long after viewing
```

would yield

```
a <positive> <positive> human drama that remains
    <positive> in memory long after viewing
```

There are many off-the-shelf dictionaries that can be used during keyword simplification. Table 4 in Section A.2 contains a sample of dictionaries reflecting various potential concepts of interest.

3.2 Dictionary-Assisted Supervised Contrastive Learning (DASCL) Objective

The dictionary-assisted supervised contrastive learning loss function resembles the loss functions from Khosla et al. (2020) and Gunel et al. (2021). Consistent with Khosla et al. (2020), we project the final hidden layer of the pretrained language model to an embedding of a lower dimension before using the contrastive loss function.

Let $\Psi(x_i)$, $i \in \{1, ...N\}$, be the L_2-normalized projection of the output of the pretrained language encoder for the original text and $\Psi(x_{i+N})$ be the corresponding L_2-normalized projection of the output for the keyword-simplified text. $\tau > 0$ is the temperature parameter that controls the separation of the classes, and $\lambda \in [0, 1]$ is the parameter that balances the cross-entropy and the DASCL loss functions. We choose λ and directly optimize τ during training. In our experiments, we use the classifier token as the output of the pretrained language encoder. Equation 1 is the DASCL loss, Equation 2 is the multiclass cross-entropy loss, and Equation 3 is the overall loss that is optimized when fine-tuning the pretrained language model. The original text and the keyword-simplified text are used with the DASCL loss (Eq. 1); only the original text is used with the cross-entropy loss. The keyword-simplified text is not used during inference.

$$\mathcal{L}_{\text{DASCL}} = -\frac{1}{2N}\sum_{i=1}^{2N}\frac{1}{2N_{y_i}-1}\times \sum_{j=1}^{2N}\mathbb{1}_{i\neq j, y_i=y_j}\log\left[\frac{\exp(\Psi(x_i)\cdot\Psi(x_j)/\tau)}{\sum_{k=1}^{2N}\mathbb{1}_{i\neq k}\exp(\Psi(x_i)\cdot\Psi(x_k)/\tau)}\right] \tag{1}$$

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=0}^{C} y_{i,c}\cdot\log\hat{y}_{i,c} \tag{2}$$

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{CE} + \lambda\mathcal{L}_{DASCL} \tag{3}$$

4 Experiments

4.1 Few-Shot Learning with SST-2

SST-2, a GLUE benchmark dataset (Wang et al., 2018), consists of sentences from movie reviews and binary labels of sentiment (positive or negative). Similar to Gunel et al. (2021), we experiment with SST-2 with three training set sizes: N=20, 100, and 1,000. Accuracy is this benchmark's primary metric of interest; we also report average precision. We use RoBERTa$_{\text{BASE}}$ as the pretrained language model. For keyword simplification, we use the opinion lexicon (Hu and Liu, 2004), which contains dictionaries of positive and negative words. Section A.3.3 further describes these dictionaries.

We compare DASCL to two other baselines: RoBERTa$_{\text{BASE}}$ using the cross-entropy (CE) loss function and the combination of the cross-entropy and supervised contrastive learning (SCL) loss functions used in Gunel et al. (2021). We also experiment with augmenting the corpus with the

Loss	N	Accuracy	Avg. Precision
CE	20	.675 ± .066	.791 ± .056
CE w/ DA	20	.650 ± .051	.748 ± .050
CE+SCL	20	.709 ± .077	.826 ± .068
CE+DASCL	20	**.777 ± .024**	**.871 ± .014**
CE+DASCL w/ DA	20	.697 ± .075	.796 ± .064
CE	100	.822 ± .019	.897 ± .023
CE w/ DA	100	.831 ± .032	.904 ± .031
CE+SCL	100	.833 ± .042	.883 ± .043
CE+DASCL	100	**.858 ± .017**	**.935 ± .012**
CE+DASCL w/ DA	100	.828 ± .020	.908 ± .012
CE	1000	.903 ± .006	**.962 ± .007**
CE w/ DA	1000	.899 ± .005	.956 ± .006
CE+SCL	1000	.905 ± .005	.960 ± .011
CE+DASCL	1000	**.906 ± .006**	.959 ± .009
CE+DASCL w/ DA	1000	.904 ± .004	.960 ± .011

Table 1: Accuracy and average precision over the SST-2 test set in few-shot learning settings. Results are averages over 10 random seeds with standard deviations reported. DA refers to **d**ata **a**ugmentation, where the keyword-simplified text augments the training corpus.

Loss	N	Accuracy	Avg. Precision
L2 Logit	100	.614	.479
CE	100	.673 ± .027	.593 ± .048
CE w/ DA	100	.663 ± .030	.576 ± .058
CE+SCL	100	.614 ± .000	.394 ± .043
CE+DASCL	100	.705 ± .013	**.645 ± .016**
CE+DASCL w/ DA	100	**.711 ± .013**	.644 ± .027
L2 Logit	1000	.624	.482
CE	1000	.716 ± .012	.662 ± .030
CE w/ DA	1000	.710 ± .011	.656 ± .024
CE+SCL	1000	.722 ± .009	.670 ± .022
CE+DASCL	1000	.732 ± .011	.671 ± .025
CE+DASCL w/ DA	1000	**.733 ± .008**	**.681 ± .021**
L2 Logit	Full	.681	.624
CE	Full	.753 ± .012	.713 ± .015
CE w/ DA	Full	.752 ± .011	.708 ± .017
CE+SCL	Full	.756 ± .011	.723 ± .009
CE+DASCL	Full	.759 ± .006	**.741 ± .010**
CE+DASCL w/ DA	Full	**.760 ± .008**	.739 ± .014

Table 2: Accuracy and average precision over the economic media test set (Barberá et al., 2021) when using 100, 1000, and all labeled examples from the training set for fine-tuning. Except for logistic regression, results are averages over 10 random seeds with standard deviations reported.

keyword-simplified text (referred to as "data augmentation," or "DA," in results tables). In other words, when data augmentation is used, both the original text and the keyword-simplified text are used with the cross-entropy loss.

We use the original validation set from the GLUE benchmark as the test set, and we sample our own validation set from the training set of equal size to this test set. Further details about the data and hyperparameter configurations can be found in Section A.3. Table 1 shows the results across the three training set configurations.

DASCL improves results the most when there are only a few observations in the training set. When N=20, using DASCL yields a 10.2 point improvement in accuracy over using the cross-entropy loss function (p<.001) and a 6.8 point improvement in accuracy over using the SCL loss function (p=.023). Figure 3 in Section A.3.8 visualizes the learned embeddings using each of these loss functions using t-SNE plots. When the training set's size increases, the benefits of using DASCL decrease. DASCL only has a slightly higher accuracy when using 1,000 labeled observations, and the difference between DASCL and cross-entropy alone is insignificant (p=.354).

4.2 *New York Times* Articles about the Economy

Barberá et al. (2021) classify the tone of *New York Times* articles about the American economy as positive or negative. 3,119 of the 8,462 labeled articles (3,852 unique articles) in the training set are labeled positive; 162 of the 420 articles in the test set are labeled positive. Accuracy is the primary metric of interest; we also report average precision. In addition to using the full training set, we also experiment with training sets of sizes 100 and 1,000. We use the positive and negative dictionaries from Lexicoder (Young and Soroka, 2012) and dictionaries of positive and negative economic terms (Hopkins et al., 2017). Barberá et al. (2021) use logistic regression with L_2 regularization. We use RoBERTa$_{\text{BASE}}$ as the pretrained language model. Section A.4 contains more details about the data, hyperparameters, and other evaluation metrics. Table 2 shows the results across the three training set configurations.

When N=100, DASCL outperforms cross-entropy only, cross-entropy with data augmentation, and SCL on accuracy (p<.005 for all) and average precision (p<.01 for all). When N=1000, DASCL outperforms cross-entropy only, cross-entropy with data augmentation, and SCL on accuracy (p<.05 for all) and average precision (but not statistically significantly). DASCL performs statistically equivalent to DASCL with data augmentation across all metrics when N=100 and 1000.

When using the full training set, RoBERTa$_{\text{BASE}}$ is a general improvement over logistic regression. Although the DASCL losses have slightly higher accuracy than the other RoBERTa-based models, the differences are not statistically significant. Us-

ing DASCL yields a 2.8 point improvement in average precision over cross-entropy ($p<.001$) and a 1.8 improvement in average precision over SCL ($p<.001$). Figure 4 in Section A.4.9 visualizes the learned embeddings using each of these loss functions using t-SNE plots.

4.3 Abusive Speech on Social Media

The OffensEval dataset (Zampieri et al., 2020) contains 14,100 tweets annotated for offensive language. A tweet is considered offensive if "it contains any form of non-acceptable language (profanity) or a targeted offense." Caselli et al. (2020) used this same dataset and more narrowly identified tweets containing "hurtful language that a speaker uses to insult or offend another individual or group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions." We focus on this dataset, AbusEval, because of its greater conceptual difficulty. 2,749 of the 13,240 tweets in the training set are labeled abusive, and 178 of the 860 tweets in the test set are labeled abusive. Caselli et al. (2021) pretrain a $\text{BERT}_{\text{BASE-UNCASED}}$ model, HateBERT, using the Reddit Abusive Language English dataset. Macro F1 and F1 over the positive class are the primary metrics of interest; we also report average precision. In addition to using the full training set, we also experiment with training sets of sizes 100 and 1,000. Section A.5 contains more details about the data and hyperparameters.

We combine DASCL with $\text{BERT}_{\text{BASE-UNCASED}}$ and HateBERT. Alorainy et al. (2019) detect cyberhate speech using threats-based othering language, focusing on the use of "us" and "them" pronouns. Following their strategy, we look at the conjunction of sentiment using Lexicoder and two dictionaries of "us" and "them" pronouns, which may suggest abusive speech. Table 3 compares the performance of $\text{BERT}_{\text{BASE-UNCASED}}$ and HateBERT with cross-entropy against $\text{BERT}_{\text{BASE-UNCASED}}$ and HateBERT with cross-entropy and DASCL.

When $N=100$, BERT with DASCL outperforms BERT on macro F1 ($p=.008$), F1 over the positive class ($p=.011$), and average precision ($p=.003$); when $N=1000$, BERT with DASCL outperforms BERT on macro F1 ($p=.021$), F1 over the positive class ($p=.028$), and average precision ($p=.007$). HateBERT with DASCL performs statistically on par with HateBERT across all metrics for $N=100$ and $N=1000$. BERT with DASCL performs statistically equivalent to HateBERT when $N=100$ and $N=1000$ on all metrics, except on F1 over the positive class when $N=1000$ ($p=.030$).

Model	N	Macro F1	F1, Pos	Avg. Precision
BERT	100	.293±.083	.334±.025	.233±.040
HateBERT	100	**.513±.120**	.346±.049	**.303±.059**
BERT w/ DASCL	100	.427±.112	**.362±.017**	.284±.022
HateBERT w/ DASCL	100	.449±.110	.345±.033	.281±.045
BERT	1000	.710±.018	.523±.034	.608±.024
HateBERT	1000	.711±.028	.512±.053	.626±.017
BERT w/ DASCL	1000	**.729±.014**	**.559±.032**	**.637±.016**
HateBERT w/ DASCL	1000	.704±.010	.501±.020	.626±.016
BERT	Full	.767±.008	.636±.012	.703±.006
HateBERT	Full	.768±.005	.630±.009	.695±.005
BERT w/ DASCL	Full	**.779±.010**	**.653±.014**	**.716±.005**
HateBERT w/ DASCL	Full	.766±.007	.623±.011	.695±.006

Table 3: Macro F1, F1, and average precision over the AbusEval test set (Caselli et al., 2020) when using 100, 1000, and all labeled examples from the training set for fine-tuning. Results are averages over 10 random seeds with standard deviations reported.

When using the full training set, BERT with DASCL improves upon the macro F1, F1 over the positive class, and average precision compared with both BERT (macro F1: $p=.010$; F1: $p=.010$; AP: $p<.001$) and HateBERT (macro F1: $p=.007$; F1: $p<.001$; AP: $p<.001$). Figure 5 in Section A.5.8 visualizes the learned embeddings using BERT and BERT with DASCL using t-SNE plots.

5 Conclusion

We propose a supervised contrastive learning approach that allows researchers to leverage specialized dictionaries when fine-tuning pretrained language models. We show that using DASCL with cross-entropy improves classification performance on SST-2 in few-shot learning settings, on classifying perceptions about the economy expressed in *New York Times* articles, and on identifying tweets containing abusive language when compared to using cross-entropy alone or alternative contrastive and data augmentation methods. In the future, we aim to extend our approach to other supervised contrastive learning frameworks, such as using this method to upweight difficult texts (Suresh and Ong, 2021). We also plan to expand this approach to semi-supervised and self-supervised settings to better understand core concepts expressed in text.

Limitations

We aim to address limitations to the supervised contrastive learning approach described in this paper in future works. We first note that there are no experiments in this paper involving multiclass or multilabel classification; all experiments involve only binary classification. Multiclass or multilabel classification may present further challenges when categories are more nuanced. We expect improvements in classification performance when applied to multiclass or multilabel classification settings, but we have not confirmed this.

Second, we have not experimented with the dimensionality of the projection layer or the batch sizes. At the moment, the projection layer is arbitrarily set to 256 dimensions, and we use batch sizes from previous works. Future work aims to study how changing the dimensionality of this projection layer and the batch size affects classification outcomes.

Third, we have used the DASCL objective with RoBERTa and BERT, but have not used it with the latest state-of-the-art pretrained language models. We focused on these particular pretrained language models because they are commonly used in the social sciences and because of computational constraints.

Fourth, we have not examined how the quality or size of the dictionary may affect classification outcomes. A poorly constructed dictionary may lead to less improvement in classification performance metrics or may even hurt performance. Dictionaries with too many words or too few words may also not lead to improvements in classification performance metrics. Future work aims to study how the quality and size of dictionaries affect the DASCL approach.

Fifth, we have not explored how this method can be used to potentially reduce bias in text classification. For example, we can replace gendered pronouns with a token (such as "<pronoun>"), potentially reducing gender bias in analytical contexts such as occupation.

Lastly, we have not explored how keyword simplification may be useful in a self-supervised or semi-supervised contrastive learning setting. This may be particularly helpful for social scientists who are often interested in exploring core concepts or perspectives in text rather than classifying text into specific classes.

Ethics Statement

Our paper describes a supervised contrastive learning approach that allows researchers to leverage specialized dictionaries when fine-tuning pretrained language models. While we did not identify any systematic biases in the particular set of dictionaries we used, any dictionary may encode certain biases and/or exclude certain groups. This can be particularly problematic when working with issues such as detecting hate speech and abusive language. For example, in the context of abusive language, if words that attack a particular group are (purposely or unintentionally) excluded from the dictionaries, those words would not be replaced. This may under-detect abusive text that attacks this specific group.

This paper does not create any new datasets. The OffensEval/AbusEval dataset contains sensitive and harmful language. Although we did not annotate or re-annotate any tweets, we are cognizant that particular types of abusive language against certain groups or identities may not have been properly annotated as abusive, or certain types of abusive language may have been excluded from the corpus entirely.

Acknowledgements

We gratefully acknowledge that the Center for Social Media and Politics at New York University is supported by funding from the John S. and James L. Knight Foundation, the Charles Koch Foundation, Craig Newmark Philanthropies, the William and Flora Hewlett Foundation, the Siegel Family Endowment, and the Bill and Melinda Gates Foundation. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise. We thank the members of the Center for Social Media and Politics for their helpful comments when workshopping this paper. We would also like to thank the anonymous reviewers for their valuable feedback in improving this paper.

References

Quinn Albaugh, Julie Sevenans, Stuart Soroka, and Peter John Loewen. 2013. The automated coding of policy agendas: A dictionary-based approach. Retrieved from https://www.almendron.com/tribuna/wp-content/uploads/2017/05/CAP2013v2.pdf.

Wafa Alorainy, Pete Burnap, Han Liu, and Matthew L.

Williams. 2019. "The enemy among us": Detecting cyber hate speech with threats-based othering language embeddings. *ACM Trans. Web*, 13(3).

American Defamation League. 2022. Hate on display hate symbols database. Retrieved from https://www.adl.org/resources/hate-symbols/search. Accessed: 2022-08-28.

Muhammad Zubair Asghar, Shakeel Ahmad, Maria Qasim, Syeda Rabail Zahra, and Fazal Masud Kundi. 2016. SentiHealth: creating health-related sentiment lexicon using hybrid approach. *SpringerPlus*, 5(1).

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Andrew O. Ballard, Ryan DeTamble, Spencer Dorsey, Michael Heseltine, and Marcus Johnson. 2022. Dynamics of polarizing rhetoric in congressional tweets. *Legislative Studies Quarterly*.

Pablo Barberá, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1):19–42.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-It 2018)*.

Margaret M. Bradley and Peter J. Lang. 1999. Affective Norms for English Words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida. Retrieved from https://pdodds.w3.uvm.edu/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf.

William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Jeffrey T. Chang, Hinrich Schütze, and Russ B. Altman. 2002. Creating an Online Dictionary of Abbreviations from MEDLINE. *Journal of the American Medical Informatics Association*, 9(6):612–620.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning.

Yoonjung Choi and Janyce Wiebe. 2014. +/-EffectWordNet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191, Doha, Qatar. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 233–240, New York, NY, USA. Association for Computing Machinery.

Matías Dell' Amerlina Ríos and Agustín Gravano. 2013. Spanish DAL: A Spanish dictionary of affect in language. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 21–28, Atlanta, Georgia. Association for Computational Linguistics.

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

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. CERT: Contrastive self-supervised learning for language understanding.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 4647–4657, New York, NY, USA. Association for Computing Machinery.

Geli Fei, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2012. A dictionary-based approach to identifying aspects implied by adjectives for opinion mining. In *Proceedings of COLING 2012: Posters*, pages 309–318, Mumbai, India. The COLING 2012 Organizing Committee.

Sarah Fioroni, Ariel Hasell, Stuart Soroka, and Brian Weeks. 2022. Constructing a dictionary for the automated identification of discrete emotions in news content.

Konstantina Georgelou, Antje Hildebrandt, and Katerina Paramana. 2017. The PSi manifesto lexicon: An online discursive platform. *Global Performance Studies*, 1(1).

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Hatebase. 2022. Hatebase is a collaborative, regionalized repository of multilingual hate speech. Retrieved from `https://hatebase.org/`. Accessed: 2022-08-28.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

Daniel J. Hopkins, Eunji Kim, and Soojong Kim. 2017. Does newspaper coverage influence or reflect public perceptions of the economy? *Research & Politics*, 4(4):1–7.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

C. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. AEDA: An easier data augmentation technique for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2748–2754, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Svetlana Kiritchenko, Saif Mohammad, and Mohammad Salameh. 2016. SemEval-2016 task 7: Determining sentiment intensity of English and Arabic phrases. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 42–51, San Diego, California. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016a. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif M. Mohammad. 2016b. Sentiment composition of words with opposing polarities. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1102–1108, San Diego, California. Association for Computational Linguistics.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland. Association for Computational Linguistics.

Tassilo Klein and Moin Nabi. 2020. Contrastive self-supervised learning for commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7517–7523, Online. Association for Computational Linguistics.

Angela Lai, Megan A. Brown, James Bisbee, Richard Bonneau, Joshua A. Tucker, and Jonathan Nagler. 2022. Estimating the ideology of political youtube videos.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Shayne Longpre, Yu Wang, and Chris DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4401–4411, Online. Association for Computational Linguistics.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Xuan Lu and Qiaozhu Mei. 2022. Covid-core twitter dataset. Retrieved from `https://midas.umich.edu/twitter-decahose-data/`. Accessed: 2022-08-28.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

C. Martindale. 1975. *The Romantic Progression: The Psychology of Literary History*. A Halsted Press Book. Hemisphere Publishing Corporation.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Saif Mohammad. 2011. Even the abstract have color: Consensus in word-colour associations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–373, Portland, Oregon, USA. Association for Computational Linguistics.

Saif Mohammad. 2018a. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad. 2018b. Word affect intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Antonio Moreno-Ortiz, Javier Fernandez-Cruz, and Chantal Pérez Chantal Hernández. 2020. Design and evaluation of SentiEcon: a fine-grained economic/financial sentiment lexicon from a corpus of business news. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5065–5072, Marseille, France. European Language Resources Association.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Nuno Oliveira, Paulo Cortez, and Nelson Areal. 2016. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, 85:62–73.

PeaceTech Lab. 2022. Our hate speech lexicons. Retrieved from `https://www.peacetechlab.org/hate-speech-lexicons`. Accessed: 2022-08-28.

Kathryn Pearson and Logan Dancey. 2011. Speaking for the underrepresented in the house of representatives: Voicing women's interests in a partisan era. *Politics & Gender*, 7(4):493–519.

Kevin Quealy. 2021. The complete list of trump's twitter insults (2015-2021). Retrieved from `https://www.nytimes.com/interactive/2021/01/19/upshot/trump-complete-insult-list.html`. Accessed: 2022-08-28.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463, Denver, Colorado. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kashish Ara Shakil, Kahkashan Tabassum, Fawziah S. Alqahtani, and Mudasir Ahmad Wani. 2021. Analyzing user digital emotions from a holy versus non-pilgrimage city in saudi arabia on twitter platform. *Applied Sciences*, 11(15).

Connor Shorten, Taghi M. Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of Big Data*, 8.

Alexandra A. Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2021. Trumping hate on twitter? online hate speech in the 2016 u.s. election campaign and its aftermath. *Quarterly Journal of Political Science*, 16(1):71–104.

Almog Simchon, William J Brady, and Jay J Van Bavel. 2022. Troll and divide: the language of online polarization. *PNAS Nexus*, 1(1). Pgac019.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 26–34, Antwerp, Belgium. Association for Computational Linguistics.

Jacopo Staiano and Marco Guerini. 2014. Depeche mood: a lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–433, Baltimore, Maryland. Association for Computational Linguistics.

Varsha Suresh and Desmond Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhanna Terechshenko, Fridolin Linder, Vishakh Padmakumar, Michael Liu, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2021. A comparison of methods in political science text classification: Transfer learning language models for politics.

Paul Thompson, John McNaught, Simonetta Montemagni, Nicoletta Calzolari, Riccardo del Gratta, Vivian Lee, Simone Marchi, Monica Monachini, Piotr Pezik, Valeria Quochi, CJ Rupp, Yutaka Sasaki, Giulia Venturi, Dietrich Rebholz-Schuhmann, and Sophia Ananiadou. 2011. The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics*, 12(1).

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Isabelle van der Vegt, Maximilian Mozes, Bennett Kleinberg, and Paul Gill. 2021. The grievance dictionary: Understanding threatening language use. *Behavior Research Methods*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Mudasir Ahmad Wani, Nancy Agarwal, Suraiya Jabin, and Syed Zeeshan Hussain. 2018. User emotion analysis in conflicting versus non-conflicting regions using online social networks. *Telematics and Informatics*, 35(8):2326–2336.

Michael Weaver. 2019. "Judge Lynch" in the court of public opinion: Publicity and the de-legitimation of lynching. *American Political Science Review*, 113(2):293–310.

Jason Wei and Kai Zou. 2019. 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)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *Computer Vision – ECCV 2016*, pages 499–515, Cham. Springer International Publishing.

Cynthia Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological Reports*, 105(2):509–521. PMID: 19928612.

Cynthia M. Whissell. 1989. Chapter 5 - the dictionary of affect in language. In Robert Plutchik and Henry Kellerman, editors, *The Measurement of Emotions*, pages 113–131. Academic Press.

Tobias Widmann and Maximilian Wich. 2022. Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in german political text. *Political Analysis*, page 1–16.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1065–1072, Sydney, Australia. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–423, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Gunel et al. (2021)'s Supervised Contrastive Learning Objective

Equation 4 is the supervised contrastive learning objective from Gunel et al. (2021). The dictionary-assisted supervised contrastive learning objective in Equation 1 is similar to Equation 4 except Equation 1 is extended to include the keyword-simplified text.

$$\mathcal{L}_{SCL} = \sum_{i=1}^{N} -\frac{1}{N_{y_i} - 1} \times \sum_{j=1}^{N} \mathbb{1}_{i \neq j} \mathbb{1}_{y_i = y_j} \log \left[\frac{\exp(\Phi(x_i) \cdot \Phi(x_j)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{i \neq k} \exp(\Phi(x_i) \cdot \Phi(x_k)/\tau)} \right] \tag{4}$$

A.2 Examples of Dictionaries and Lexicons

Table 4 contains a sample of dictionaries across various use cases and academic fields that can be potentially used with DASCL. There is no particular order to the dictionaries, with similar dictionaries clustered together. We did not include any non-open source dictionaries.

A.3 Additional Information for the Few-Shot Learning Experiments with SST-2

A.3.1 Data Description: Few-Shot Training Sets, Validation Set, and Test Set

The SST-2 dataset was downloaded using Hugging Face's Datasets library (Lhoest et al., 2021). The test set from SST-2 does not contain any labels, so we use the validation set from SST-2 as our test set. We create our own validation set by randomly sampling a dataset equivalent in size to the original validation set. Our validation set and few-shot learning sets were sampled with no consideration to the label distributions of the original training or validation sets.

When $N = 20$, there are 11 positive examples and 9 negative examples. When $N = 100$, there are 60 positive examples and 40 negative examples. When $N = 1000$, there are 558 positive examples and 442 negative examples. Our validation set has 486 positive examples and 386 negative examples. Lastly, our test set has 444 positive examples and 428 negative examples.

A.3.2 Text Preprocessing Steps

The only text preprocessing step taken is that non-ASCII characters are removed from the dataset. The text is tokenized using a byte-level BPE tokenizer (Liu et al., 2019).

A.3.3 Dictionaries Used During Keyword Simplification

We used the opinion lexicon from Hu and Liu (2004). This lexicon consists of two dictionaries: one with all positive unigrams and one with all negative unigrams. There are 2,006 positive words

Dictionary Name	Type of Words or Phrases	Source
Opinion Lexicon	Positive/negative sentiment	Hu and Liu (2004)
Lexicoder Sentiment Dictionary	Positive/negative sentiment	Young and Soroka (2012)
SentiWordNet	Sentiment	Baccianella et al. (2010)
ANEW	Emotions	Bradley and Lang (1999)
EmoLex	Emotions	Mohammad and Turney (2013)
DepecheMood	Emotions	Staiano and Guerini (2014)
Moodbook	Emotions	Wani et al. (2018)
Moral & Emotion Dictionaries	Moralization and Emotions	Brady et al. (2017)
Emotion Intensity Lexicon	Emotion intensity	Mohammad (2018b)
VAD Lexicon	Valence, arousal, and dominance of words	Mohammad (2018a)
SCL-NMA	Negators, modals, degree adverbs	Kiritchenko and Mohammad (2016a)
SCL-OPP	Sentiment of mixed polarity phrases	Kiritchenko and Mohammad (2016b)
Dictionary of Affect in Language	Affect of English words	Whissell (1989, 2009)
Spanish DAL	Affect of Spanish words	Dell' Amerlina Ríos and Gravano (2013)
Subjectivity Lexicon	Subjectivity clues	Wilson et al. (2005)
Subjectivity Sense Annotations	Subjectivity disambiguation	Wiebe and Mihalcea (2006)
Arguing Lexicon	Patterns representing arguing	Somasundaran et al. (2007)
+/-Effect Lexicon	Positive/negative effects on entities	Choi and Wiebe (2014)
AEELex	Arabic and English emotion	Shakil et al. (2021)
Discrete Emotions Dictionary	Emotions in news content	Fioroni et al. (2022)
Yelp/Amazon Reviews Dictionaries	Sentiments in reviews	Kiritchenko et al. (2014)
VADER	Sentiments on social media	Hutto and Gilbert (2014)
English Twitter Sentiment Lexicon	Sentiments on social media	Rosenthal et al. (2015)
Arabic Twitter Sentiment Lexicon	Sentiments on social media	Kiritchenko et al. (2016)
Hashtag Emotion Lexicon	Emotions associated with Twitter hashtags	Mohammad and Kiritchenko (2015)
Political Polarization Dictionary	Political polarizing language	Simchon et al. (2022)
ed8	Affective language in German political text	Widmann and Wich (2022)
Policy Agendas Dictionary	Keywords by policy area	Albaugh et al. (2013)
"Women" Terms	Terms related to the category "women"	Pearson and Dancey (2011)
Trump's Twitter Insults	Insults used by President Trump	Quealy (2021)
Hate Speech Lexicon	Hate speech	Davidson et al. (2017)
PeaceTech Lab's Hate Speech Lexicons	Hate speech	PeaceTech Lab (2022)
Hatebase	Hate speech	Hatebase (2022)
Hurtlex	Hate speech	Bassignana et al. (2018)
Hate on Display Hate Symbols	Hate speech and symbols	American Defamation League (2022)
Hate speech on Twitter	Hate speech on Twitter	Siegel et al. (2021)
Reddit hate lexicon	Hate speech on Reddit	Chandrasekharan et al. (2017)
Pro/Anti-Lynching	Pro/anti-lynching terms	Weaver (2019)
Grievance Dictionary	Terms related to grievance	van der Vegt et al. (2021)
Economic Sentiment Terms	Economic sentiment in newspapers	Hopkins et al. (2017)
Loughran-McDonald Dictionary	Financial sentiment	Loughran and McDonald (2011)
SentiEcon	Financial and economic sentiment	Moreno-Ortiz et al. (2020)
Financial Phrase Bank	Phrases expressing financial sentiment	Malo et al. (2014)
Stock Market Sentiment Lexicon	Stock market sentiment	Oliveira et al. (2016)
BioLexicon	Linguistic information of biomedical terms	Thompson et al. (2011)
SentiHealth	Health-related sentiment	Asghar et al. (2016)
MEDLINE Abbreviations	Abbreviations from medical abstracts	Chang et al. (2002)
COVID-CORE Keywords	COVID-19 keywords on Twitter	Lu and Mei (2022)
PSi Lexicon	Performance studies keywords	Georgelou et al. (2017)
Concreteness Ratings	Concreteness of words and phrases	Brysbaert et al. (2013)
Word-Colour Association Lexicon	Word-color associations	Mohammad (2011)
Regressive Imagery Dictionary	Primordial vs. conceptual thinking	Martindale (1975)
Empath	Various lexical categories	Fast et al. (2016)
WordNet	Various lexical categories	Miller (1995)

Table 4: A sample of dictionaries that can potentially be used with DASCL. There is no particular order to the dictionaries, with similar dictionaries clustered together. We did not include any non-open source dictionaries.

and 4,783 negative words. We replaced the positive words with the token "<positive>". We replaced the negative words with the token "<negative>".

A.3.4 Number of Parameters and Runtime

This experiment uses the RoBERTa$_{\text{BASE}}$ pretrained language model, which contains 125 million parameters (Liu et al., 2019). When using DASCL, we also had an additional temperature parameter, τ, that was directly optimized. With the hyperparameters described in Section A.3.5 and using an NVIDIA V100 GPU, it took approximately 2.1 seconds to train over 40 batches using cross-entropy (CE) alone, 2.2 seconds to train over 40 batches using CE+SCL, and 3.3 seconds to train over 40 batches using CE+DASCL.

A.3.5 Hyperparameter Selection and Configuration Details

We take our hyperparameter configuration directly from Gunel et al. (2021). For each configuration, we set the learning rate to 1×10^{-5} and used a batch size of 16. When using the SCL objective, in line with Gunel et al. (2021), we set $\lambda = 0.9$ and $\tau = 0.3$. When using the DASCL objective, we also set $\lambda = 0.9$ and initialized $\tau = 0.3$. We trained for 100 epochs for all few-shot learning settings.

A.3.6 Model Evaluation Details

The model from the epoch with the highest accuracy over our own validation set was chosen as the final model for each random seed. We report accuracy, which is the main metric of interest with this benchmark, and average precision. Average precision is used to summarily quantify the precision-recall tradeoff, and is viewed as the area under the precision-recall curve (Davis and Goadrich, 2006). Average precision is defined as

$$AP = \sum_n (R_n - R_{n-1}) P_n$$

where P_n and R_n are the precision and recall at the nth threshold.

A.3.7 Results over the Validation Set

Table 5 reports the accuracy and the average precision over the validation set for the SST-2 few-shot setting experiments. The validation set was used for model selection, so the reported results over the validation set are from the model with the highest accuracy achieved on the validation set across the 100 epochs.

Loss	N	Accuracy	Avg. Precision
CE	20	.680 ± .043	.768 ± .062
CE w/ DA	20	.668 ± .025	.729 ± .034
CE+SCL	20	.707 ± .049	.797 ± .060
CE+DASCL	20	.743 ± .016	.839 ± .023
CE+DASCL w/ DA	20	.700 ± .048	.765 ± .061
CE	100	.832 ± .015	.905 ± .021
CE w/ DA	100	.849 ± .023	.915 ± .027
CE+SCL	100	.848 ± .020	.905 ± .029
CE+DASCL	100	.872 ± .010	.942 ± .010
CE+DASCL w/ DA	100	.842 ± .020	.927 ± .010
CE	1000	.900 ± .005	.958 ± .010
CE w/ DA	1000	.905 ± .006	.959 ± .005
CE+SCL	1000	.904 ± .038	.958 ± .016
CE+DASCL	1000	.907 ± .004	.961 ± .011
CE+DASCL w/ DA	1000	.908 ± .005	.965 ± .008

Table 5: Accuracy and average precision over the SST-2 validation set in few-shot learning settings. Results are averages over 10 random seeds with standard deviations reported.

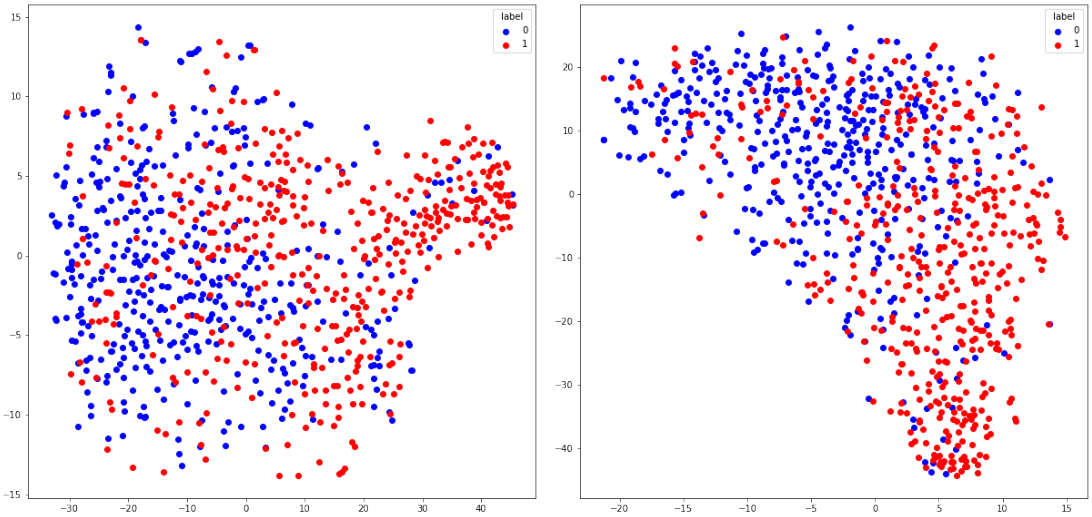
A.3.8 t-SNE Plots of the Learned Classifier Token Embeddings for the Test Set, N = 20

We use t-SNE (van der Maaten and Hinton, 2008) plots to visualize the learned classifier token embeddings, "<s>", over the SST-2 test set when using the cross-entropy objective alone, using the cross-entropy objective with the supervised contrastive learning (SCL) objective (Gunel et al., 2021), and using the cross-entropy objective with the dictionary-assisted supervised contrastive learning (DASCL) objective. These plots are in Figure 3. We see that DASCL draws embeddings of the same class closer and pushes embeddings of different classes farther apart compared to using cross-entropy alone or using cross-entropy with SCL.

A.4 Additional Information for the *New York Times* Articles about the Economy Experiments

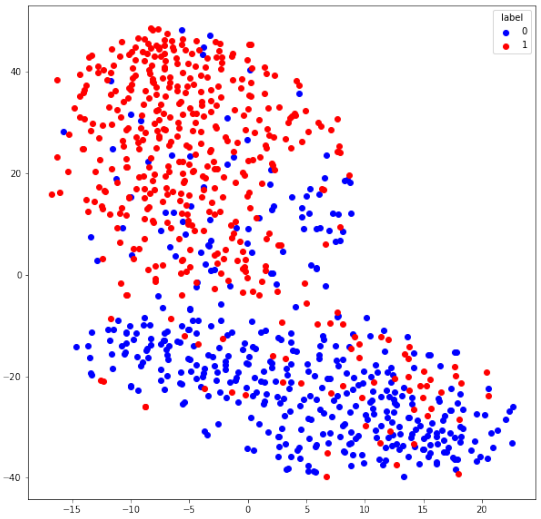
A.4.1 Data Description: Few-Shot Training Set, Validation Set, and Test Set

The data for the *New York Times* articles was downloaded from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/MXKRDE. The test set was created using the replication files included at the link. In the original code, there was an error with overlapping training and test sets. We removed the duplicated observations from the training set. Because a single article could be annotated multiple times by different annotators, our validation set was created



(a) CE

(b) CE+SCL



(c) CE+DASCL

Figure 3: t-SNE plots of the classifier token embeddings on the SST-2 test set fine-tuned using a training set size of 20. The loss configuration is noted below each plot. Blue are negative examples and red are positive examples.

using 15% of the unique number of articles in the training data. 452 of the 1,317 labeled articles in the validation set are labeled positive. For our few-shot training sets, when $N = 100$, there were 41 positive examples. When $N = 1000$, there were 363 positive examples. Our validation set and few-shot learning sets were sampled with no consideration to the label distributions of the original training or validation sets.

A.4.2 Text Preprocessing Steps

The only preprocessing was removing HTML tags that occasionally appeared in the text. The text is tokenized using a byte-level BPE tokenizer.

A.4.3 Dictionaries Used During Keyword Simplification

We used two sets of dictionaries during keyword simplification. We first used Lexicoder, downloaded from `http://www.snsoroka.com/data-lexicoder/`. It is a dictionary specifically designed to study sentiment in news coverage (Young and Soroka, 2012). The dictionary is split into four separate sub-dictionaries: positive words, negative words, "negative" positive words (e.g., "not great"), and "negative" negative words (e.g., "not bad"). There are 1,709 positive words, 2,858 negative words, 1,721 negative positive words, and 2,860 negative negative words. We replaced positive words and negative negative words with the token "<positive>". We replaced negative words and negative positive words with the token "<negative>".

The second dictionary we used was the 21 economic terms from Hopkins et al. (2017). The 6 positive economic terms (in stemmed form) are "bull*", "grow*", "growth*", "inflat*", "invest*", and "profit*". The 15 negative economic terms (in stemmed form) are "bad*", "bear*", "debt*", "drop*", "fall*", "fear*", "jobless*", "layoff*", "loss*", "plung*", "problem*", "recess*", "slow*", "slump*", and "unemploy*". We replaced the positive economic words with the token "<positive_econ>". We replaced the negative economic words with the token "<negative_econ>".

A.4.4 Number of Parameters and Runtime

This experiment uses the RoBERTa$_{\text{BASE}}$ pretrained language model, which contains 125 million parameters (Liu et al., 2019). When using DASCL, we also had an additional temperature parameter, τ, that was directly optimized. With the hyperparameters described in Section A.4.5 and using an NVIDIA V100 GPU, it took approximately 5.7 seconds to train over 40 batches using cross-entropy (CE) alone, 5.7 seconds to train over 40 batches using CE+SCL, and 10.7 seconds to train over 40 batches using CE+DASCL.

A.4.5 Hyperparameter Selection and Configuration Details

We selected hyperparameters using the validation set. We searched over the learning rate and the temperature initialization; we used $\lambda = 0.9$ for all loss configurations involving contrastive learning. We used a batch size of 8 because of resource constraints. We fine-tuned RoBERTa$_{\text{BASE}}$ for 5 epochs.

For the learning rate, we searched over $\{5 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}\}$; for the temperature, τ, initialization, we searched over $\{0.07, 0.3\}$. We fine-tuned the model and selected the model from the epoch with the highest accuracy. We repeated this with three random seeds, and selected the hyperparameter configuration with the highest average accuracy. We used accuracy as the criterion because Barberá et al. (2021) used accuracy as the primary metric of interest. The final learning rate across all loss configurations was 5×10^{-6}. The final τ initialization for both SCL and DASCL loss configurations was 0.07. We used these same hyperparameters when we limited the training set to 100 and 1,000 labeled examples.

A.4.6 Model Evaluation Details

During fine-tuning, the model from the epoch with the highest accuracy over the validation set was chosen as the final model for each random seed. We report accuracy, which is the main metric of interest with this dataset, and average precision. For a definition of average precision, see Section A.3.6.

The results in Table 2 for logistic regression using the full training set differ slightly from their paper because of an error in overlapping training and test sets in the original splits.

A.4.7 Additional Classification Metrics: Precision and Recall

Table 6 contains additional classification metrics—precision and recall—for the test set when using 100, 1,000, and all labeled examples from the training set for fine-tuning.

Loss	N	Precision	Recall
L2 Logit	100	.000	.000
CE	100	.646 ± .079	.359 ± .126
CE+DA	100	.656 ± .096	.312 ± .170
SCL	100	.000 ± .000	.000 ± .000
DASCL	100	.653 ± .044	.519 ± .055
DASCL+DA	100	.690 ± .051	.475 ± .093
L2 Logit	1000	.542	.160
CE	1000	.670 ± .028	.526 ± .074
CE+DA	1000	.658 ± .029	.527 ± .089
SCL	1000	.674 ± .017	.543 ± .046
DASCL	1000	.684 ± .032	.575 ± .069
DASCL+DA	1000	.682 ± .036	.592 ± .062
L2 Logit	Full	.689	.315
CE	Full	.690 ± .036	.663 ± .055
CE+DA	Full	.696 ± .033	.643 ± .056
SCL	Full	.700 ± .036	.652 ± .053
DASCL	Full	.709 ± .021	.640 ± .041
DASCL+DA	Full	.718 ± .031	.628 ± .060

Table 6: Precision and recall over the economic media test set (Barberá et al., 2021) when using 100, 1000, and all labeled examples from the training set for fine-tuning. Except for the logistic regression model, results are averages over 10 random seeds with standard deviations reported.

A.4.8 Results over the Validation Set

Table 7 reports the accuracy, precision, recall, and average precision over the validation set for the economic media data. The validation set was used for model selection, so the reported results over the validation set are from the model with the highest accuracy achieved on the validation set across the 5 epochs.

A.4.9 t-SNE Plots of the Learned Classifier Token Embeddings for the Test Set

We use t-SNE plots to visualize the learned classifier token embeddings, "<s>", over the *New York Times* articles about the economy test set when using the cross-entropy objective alone, using the cross-entropy objective with the supervised contrastive learning (SCL) objective (Gunel et al., 2021), and using the cross-entropy objective with the dictionary-assisted supervised contrastive learning (DASCL) objective. These plots are in Figure 4. We see that DASCL pushes embeddings of different classes farther apart compared to using cross-entropy alone or using cross-entropy with SCL.

A.5 Additional Information for the AbusEval Experiments

A.5.1 Data Description: Few-Shot Training Set, Validation Set, and Test Set

The data was downloaded from https://github.com/tommasoc80/AbuseEval. Because there was no validation set, we created our own validation set by sampling 15% of the training set. 399 of the 1,986 tweets in the validation set are labeled abusive. For our few-shot training sets, when $N = 100$, 17 tweets are labeled abusive. When $N = 1000$, 210 tweets are labeled abusive. Our validation set and few-shot learning sets were sampled with no consideration to the label distributions of the original training or validation sets.

A.5.2 Text Preprocessing Steps

We preprocessed the text of the tweets in the following manner: we removed all HTML tags, removed all URLs (even the anonymized URLs), removed the anonymized @ tags, removed the retweet ("RT") tags, and removed all "&" tags. The text is tokenized using the WordPiece tokenizer (Devlin et al., 2019).

A.5.3 Dictionaries Used During Keyword Simplification

We used two sets of dictionaries during keyword simplification. For the first dictionary, we used Lexicoder. For a description of the Lexicoder dictionary, see Section A.4.3. We used the same token-replacements as described in Section A.4.3.

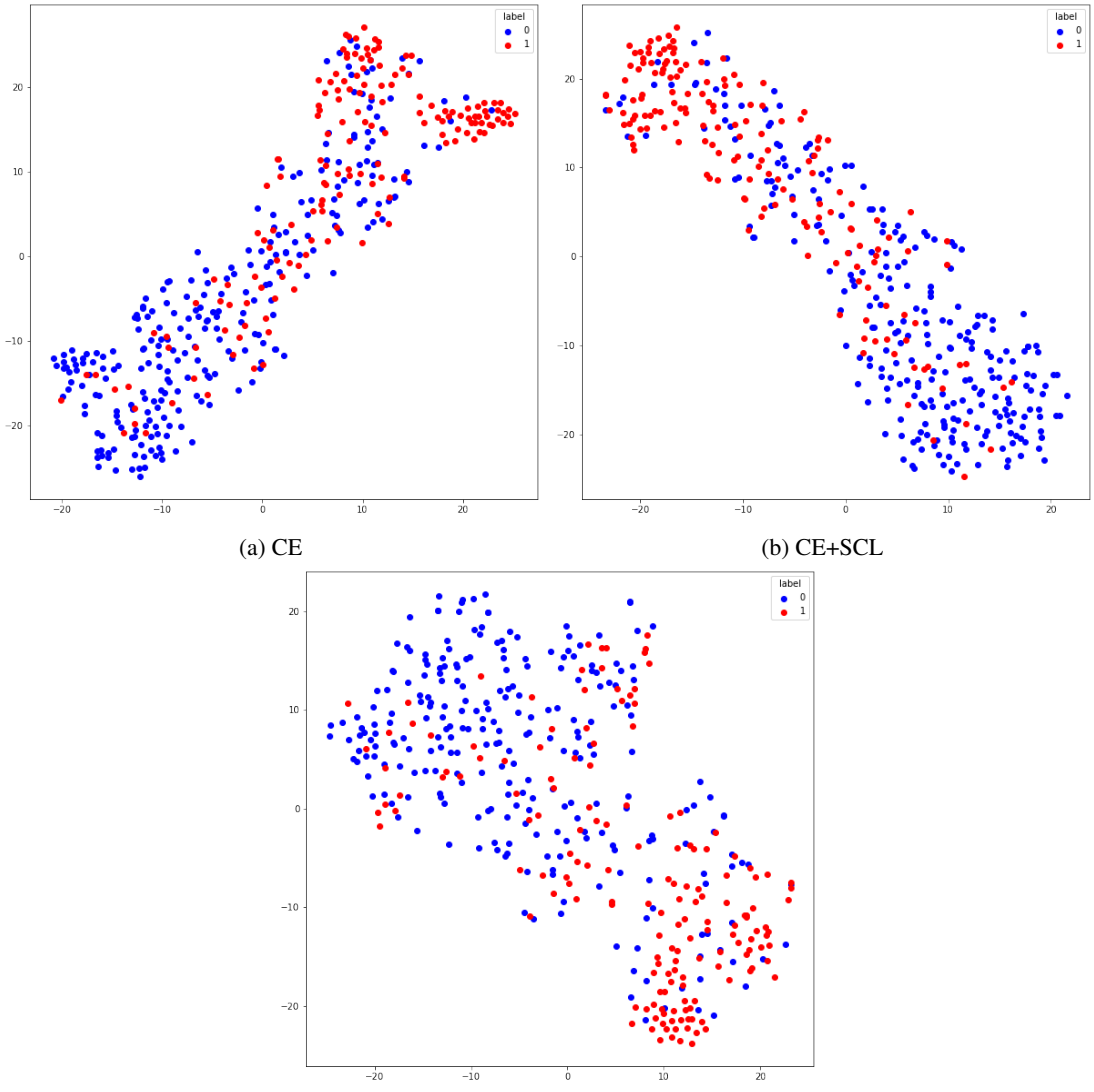
The second dictionary used was a dictionary of "us" and "them" pronouns. These pronouns are intended to capture directed or indirected abuse. The "us" pronouns are "we're", "we'll", "we'd", "we've", "we", "me", "us", "our", "ours", and "let's". The "them" pronouns are "you're", "you've", "you'll", "you'd", "yours", "your", "you", "theirs", "their", "they're", "they", "them", "people", "men", "women", "man", "woman", "mob", "y'all", and "rest." This dictionary is loosely based on suggested words found in Alorainy et al. (2019).

A.5.4 Number of Parameters and Runtime

This experiment uses the $\text{BERT}_{\text{BASE-UNCASED}}$ pretrained language model, which contains 110 million parameters (Liu et al., 2019). When using DASCL, we also had an additional temperature parameter, τ, that was directly optimized. With the hyperparameters described in Section A.5.5 and using an NVIDIA V100 GPU, it took approximately

Loss	Accuracy	Precision	Recall	Avg. Precision
CE	$.723 \pm .004$	$.644 \pm .022$	$.438 \pm .054$	$.605 \pm .007$
CE w/ DA	$.726 \pm .004$	$.662 \pm .035$	$.423 \pm .059$	$.607 \pm .007$
CE+SCL	$.727 \pm .003$	$.659 \pm .037$	$.438 \pm .065$	$.611 \pm .006$
CE+DASCL	$.724 \pm .006$	$.655 \pm .020$	$.416 \pm .037$	$.610 \pm .007$
CE+DASCL w/ DA	$.724 \pm .004$	$.662 \pm .031$	$.406 \pm .051$	$.609 \pm .006$

Table 7: Accuracy, precision, recall, and average precision over the validation set for economic media (Barberá et al., 2021). Results are averages over 10 random seeds with standard deviations reported.



(a) CE

(b) CE+SCL

(c) CE+DASCL

Figure 4: t-SNE plots of the classifier token embeddings on the *New York Times* articles about the economy test set fine-tuned using the full training set. The loss configuration is noted below each plot. Blue are negative examples and red are positive examples.

2.6 seconds to train over 40 batches using cross-entropy (CE) alone and 4.9 seconds to train over 40 batches using CE+DASCL.

A.5.5 Hyperparameter Selection and Configuration Details

We selected hyperparameters using the validation set. We searched over the learning rate and the temperature initialization; again, we used $\lambda = 0.9$ for all loss configurations involving contrastive learning. In line with Caselli et al. (2021), we used a batch size of 32. We fine-tuned $\text{BERT}_{\text{BASE-UNCASED}}$ and HateBERT for 5 epochs.

For the learning rate, we searched over $\{1 \times 10^{-6}, 2 \times 10^{-6}, 3 \times 10^{-6}, 4 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}\}$; for the temperature, τ, initialization, we searched over $\{0.07, 0.3\}$. We fine-tuned the model and selected the model from the epoch with the highest F1 over the positive class. We repeated this with three random seeds, and selected the hyperparameter configuration with the highest average F1 over the positive class. We used the F1 score over the positive class as the criterion because it is one of the metrics of interest in Caselli et al. (2021). The final learning rate across all loss configurations was 2×10^{-6}. The final τ initialization for both SCL and DASCL loss configurations was 0.3. We note that our hyperparameter search yielded a different set of hyperparameters from Caselli et al. (2021). We used these same hyperparameters when we limited the training set to 100 and 1,000 labeled examples.

A.5.6 Model Evaluation Details

During fine-tuning, the model from the epoch with the highest F1 over the validation set was chosen as the final model for each random seed. We report macro F1 and F1, the main metrics of interest with this dataset, and average precision. For a definition of average precision, see Section A.3.6.

A.5.7 Results over the Validation Set

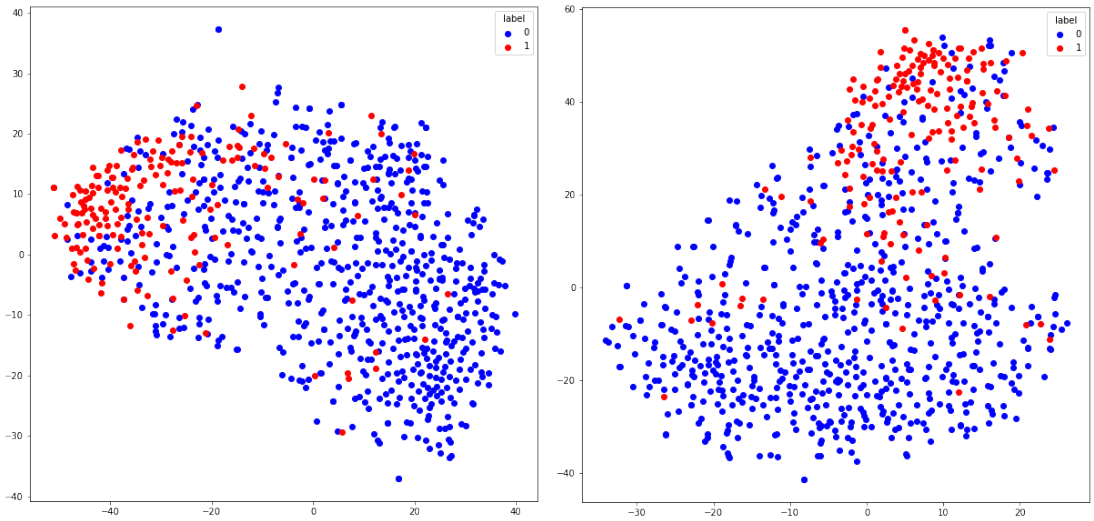
Table 8 reports the macro F1, F1, and average precision over the validation set for AbusEval. The validation set was used for model selection, so the reported results over the validation set are from the model with the highest F1 achieved on the validation set across the 5 epochs.

A.5.8 t-SNE Plots of the Learned Classifier Token Embeddings for the Test Set

We use t-SNE plots to visualize the learned classifier token embeddings, "<s>", over the AbusEval test set when using BERT and when using BERT with DASCL. These plots are in Figure 5. We see that using DASCL with BERT pushes embeddings of different classes farther apart compared to using BERT alone.

Model	Macro F1	F1, Pos	Avg. Precision
BERT	$.756 \pm .006$	$.632 \pm .008$	$.681 \pm .014$
HateBERT	$.754 \pm .009$	$.635 \pm .008$	$.708 \pm .005$
BERT+ DASCL	$.759 \pm .007$	$.639 \pm .007$	$.683 \pm .012$
HateBERT+ DASCL	$.755 \pm .005$	$.635 \pm .005$	$.706 \pm .006$

Table 8: The macro F1, F1, and average precision over the AbusEval validation set (Caselli et al., 2020). Results are averages over 10 random seeds with standard deviations reported.



(a) BERT

(b) BERT with DASCL

Figure 5: t-SNE plots of the classifier token embeddings on the AbusEval test set fine-tuned using the full training set. The model is noted below each plot. Blue are examples of non-abusive tweets and red are examples of abusive tweets.