Dictionary-Assisted Supervised Contrastive Learning

Patrick Y. Wu and **Joshua A. Tucker** and **Jonathan Nagler**
New York University
Center for Social Media and Politics

Abstract

We propose dictionary-assisted supervised contrastive learning (DASCL), a framework that allows the researcher to incorporate substantive knowledge when finetuning pretrained language models. The researcher first selects a dictionary of terms relevant to the concept of interest during the preprocessing step. Then, any words in the corpus that appears in the dictionary are replaced with a common, fixed string. This preprocessing step highlights similarities between texts in the corpus. Second, we propose an objective function that contrasts the original texts with the keyword-simplified texts in the same class during the finetuning stage of a pretrained language model. This contrastive loss function draws closer together the text embeddings in the same class and pushes further apart the text embeddings in different classes. Combining this contrastive loss function and the cross-entropy loss function leads to improvements in performance metrics for both benchmark natural language processing datasets and social science applications.

1 Introduction

We propose a contrastive learning approach that improves the performance of the automated classification of text. It is conceptually simple, requires low computational resources, and is usable with most pretrained language models. The approach is particularly useful when the concept of interest underlying the labeling process is abstract or complex.

Dictionaries often contain words that hint at or reveal the sentiment, stance, etc. of a document (see, e.g., Fei et al., 2012). Domain 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 certain 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 an issue in fields such as the social sciences.

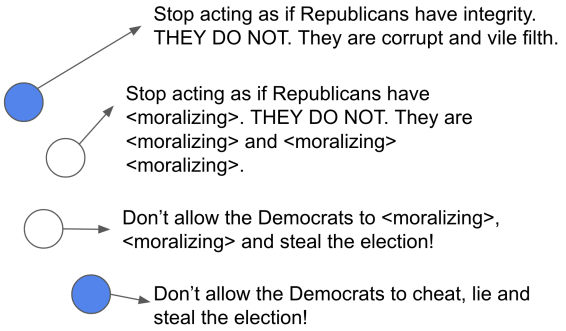

Figure 1: The blue dots are embeddings of the original tweets. Both tweets are examples of moralizing tweets, although they do not overlap in the moralizing words used. By keyword simplifying both tweets, they appear more similar, meaning that their embeddings are closer. A contrastive learning objective is used to draw these keyword-simplified text closer together, which effectively draws closer together embeddings of the original tweets.

The latest machine learning approaches, especially pretrained language models, outperform dictionary methods (Barberá et al., 2021). We propose a contrastive learning approach that incorporates the substantive knowledge inherent in dictionaries with pretrained language models. Intuitively, we replace all the words in the corpus that belong to a specific lexicon with a fixed, common string (e.g., replacing all words from a negative lexicon with the string "<negative>"). Keyword simplification, when using an appropriate dictionary, increases the similarity in the documents of the same class. We then use a supervised contrastive objective to draw together text embeddings in the same class and push further apart the text embeddings of different classes (Khosla et al., 2020; Gunel et al., 2020). See Figure 1 for a visualization of the intuition behind our proposed method.

The contributions of this project are as follows.

- We propose keyword simplification, which is described in greater detail in Section 3.1, to make documents of the same classes more similar.
- We outline a supervised contrastive loss function, described in Section 3.2, that learns patterns within and across the original texts and keyword-simplified texts.
- We find classification performance improvements in few-shot learning settings and social science data applications compared to two strong baselines: RoBERTa-Base (Liu et al., 2019) finetuned with cross-entropy and the supervised contrastive learning approach detailed in Gunel et al. (2020), the most closely related approach to DASCL.

2 Related Work

Use of Pretrained Language Models. 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). Pretrained language models have seen increasing usages in other domains such as the social sciences. Terechshenko et al. (2021) show that RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019) outperform the bag-of-words approaches typically used for political science text classification tasks. Ballard et al. (2022) use BERTweet (Nguyen et al., 2020) to classify tweets expressing polarizing rhetoric. Wu and Mebane (2022) use BERT to classify tweets about the electoral process during the 2016 U.S. general election. The method we propose can be used with any pretrained language model, so it can potentially improve results across a wide range of substantive research projects.

Usage of Dictionaries. In the social sciences, dictionaries play an important role in understanding meaning behind text. Brady et al. (2017) use a moral dictionary and an 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) examine if newspaper coverage lead or follow the public's perceptions of the economy; to understand perceptions of the economy in newspaper articles, they use two dictionaries of positive and negative terms about the economy. Although the latest deep learning techniques almost always outperform dictionary-based approaches, they still contain valuable information about concepts of interest.

Text Data Augmentation. There is a wide range of text data augmentation techniques, including backtranslation (Sennrich et al., 2016) and rule-based data augmentations (Wei and Zou, 2019). Rule-based augmentations include random synonym replacements, random insertions, random swaps, and random deletions. Karimi et al. (2021) find that even inserting punctuation randomly yields classification performance improvements. Shorten et al. (2021) surveys text data augmentation techniques for deep learning. Longpre et al. (2020) find that these task-agnostic data augmentations typically do not improve the classification performance of pretrained language models. Our approach, on the other hand, is task-specific: we choose dictionaries to make the keyword-simplifications based on the concept of interest underlying the classification task.

Contrastive Learning. Most of the works on contrastive learning have focused on self-supervised contrastive learning, especially in computer vision. This setting treats images and their augmentations as positives and other images and their augmentations as negatives. Recent contrastive learning approaches match or outperform their supervised pretrained image model counterparts, often using a small fraction of available labeled 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 (Fang et al., 2020; Klein and Nabi, 2020).

Our work is most closely related to several 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. (2020), which also uses a generalized version of NT-Xent that is extended to an arbitrary number of positives. They treat all pairs

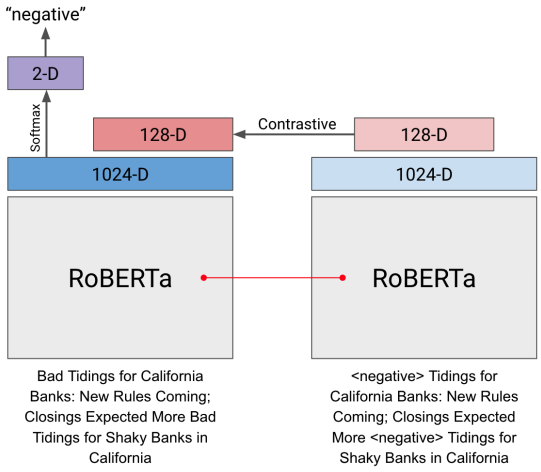

Figure 2: Overview of the proposed framework. Although RoBERTa$_{\text{LARGE}}$ is shown in this diagram, any pretrained language model will work with this approach. The dimensions of the projection layer is also arbitrary.

of observations in the same class as positives and all pairs of observations in different classes as negatives. Their supervised contrastive loss function is defined as follows.

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

3 Method

The approach consists of two steps: keyword simplification and using a contrastive objective function to draw closer together the text embeddings of the same class and pushes further apart the text embeddings of different classes. Figure 2 shows an overview of the proposed framework.

3.1 Keyword Simplification

The first step of the dictionary-assisted supervised contrastive learning (DASCL) framework is keyword simplification. Keyword simplification involves taking a dictionary of keywords and replacing all occurrences of keywords with a common, fixed word. More specifically, we select a set of M dictionaries $\mathcal{D}$. For each dictionary $d_i \in \mathcal{D}$, $i \in \{1, ..., M\}$, we assign a string t_i. Then, iterate through the corpus and replace any word w_j that is in dictionary d_i with the string t_i. We repeat these steps for each dictionary. For example, if we have a dictionary of negative words, then applying keyword simplification to

```
Bad tidings for California Banks: New
    Rules Coming; Closings Expected
```

would yield

```
<negative> tidings for California Banks:
 New Rules Coming; <negative> Expected
```

3.2 Dictionary-Assisted Supervised Contrastive Learning (DASCL) Objective

The dictionary-assisted supervised contrastive loss function resembles the loss functions from Khosla et al. (2020) and Gunel et al. (2020). 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 projection of the l_2-normalized output of the pretrained language encoder for the original text and let $\Psi(x_{i+N})$ be the corresponding projection of the l_2-normalized output for the keyword-simplified text. τ is the temperature hyperparameter that controls separation of the classes, and λ is the hyperparameter balances the cross-entropy and the DASCL loss functions. c denotes the class, with C total number of classes. Equation 2 is the dictionary-assisted supervised contrastive learning loss, Equation 3 is the multiclass cross-entropy loss, and Equation 4 is the overall loss that is optimized when finetuning the pretrained language model.

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

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=0}^{C} y_{i,c} \cdot \log \hat{y}_{i,c} \quad (3)$$

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{CE} + \lambda \mathcal{L}_{DASCL} \quad (4)$$

During training, we use the original text and the keyword-simplified text. The keyword-simplified text is not needed during inference.

4 Experiments

We use DASCL with three experimental setups. The experiments focus on text classification tasks that aim to automatically classify either stance or sentiment.

4.1 Few-Shot Learning: SST-2

We use the combination of the cross-entropy and DASCL loss functions with SST-2, a GLUE benchmark dataset (Wang et al., 2018). The corpus consists of sentences from movie reviews and binary labels of sentiment (positive or negtive). In a similar fashion to Gunel et al. (2020), we experiment with SST-2 in a few-shot learning setting. We experiment with three training set sizes: $N = 20$, $N = 100$, and $N = 1000$; we also examine performance using the full training dataset. When using DASCL, we use RoBERTa-Base as the pretrained language model. We compare DASCL to two other baselines: RoBERTa-Base using a cross-entropy loss function and the combination of the cross-entropy and supervised contrastive loss (SCL) function used in Gunel et al. (2020); the latter is described in Equation 1.

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. (2020), we set $\lambda = 0.9$ and $\tau = 0.3$. When using the DASCL objective, we also set $\lambda = 0.9$ and $\tau = 0.3$. We trained for 100 epochs when the training set was $N = 20$ and $N = 100$, and we trained for 50 epochs when the training set was $N = 1000$. We report results over the validation split from the original data. We created our own validation dataset by randomly sampling a dataset equivalent in size of the validation split. We used the model from the epoch that had the highest accuracy over the custom validation dataset. Table 1 shows the results across the three training set configurations.

DASCL improves results the most when the training data is small. DASCL represents an 8.5 point improvement in accuracy over using the standard cross-entropy loss function with RoBERTa-Base and a 5.1 point improvement in accuracy over using the supervised contrastive loss function proposed in Gunel et al. (2020). When training data size increases, DASCL's benefits decrease.

4.2 *New York Times* Articles about the Economy

Barberá et al. (2021) classified tone of *New York Times* articles about the American economy as either positive or negative. There are $3,637$ articles in the training set and 420 articles in the training set. Barberá et al. (2021) use a logistic regression with L_2 regularization. We used RoBERTa-Base as the pretrained language model, a learning rate of 1×10^{-5}, and a batch size of 8. When using SCL and DASCL, we set $\lambda = 0.9$ and $\tau = 0.3$. It is important to note that these results are preliminary.

Again, we see that using RoBERTa$_{\text{Base}}$ with cross-entropy and DASCL as the loss outperforms the other baseline models. Further iterations are needed to confirm this result.

References

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

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.

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.

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.

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.

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

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.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal

Model	Loss	Training N	Accuracy
RoBERTa$_{\text{BASE}}$	CE	20	67.5 ± 6.6
RoBERTa$_{\text{BASE}}$	CE and SCL	20	70.9 ± 7.7
RoBERTa$_{\text{BASE}}$	CE and DASCL	20	$\mathbf{76.0 \pm 5.0}$
RoBERTa$_{\text{BASE}}$	CE	100	82.2 ± 1.9
RoBERTa$_{\text{BASE}}$	CE and SCL	100	83.3 ± 4.2
RoBERTa$_{\text{BASE}}$	CE and DASCL	100	$\mathbf{84.8 \pm 1.8}$
RoBERTa$_{\text{BASE}}$	CE	1000	90.3 ± 0.6
RoBERTa$_{\text{BASE}}$	CE and SCL	1000	90.5 ± 0.5
RoBERTa$_{\text{BASE}}$	CE and DASCL	1000	$\mathbf{90.6 \pm 0.5}$

Table 1: Results over the SST-2 validation dataset in few-shot learning settings ($N = 20$, 100, and $1,000$). The average accuracy over 10 random seeds and the standard deviation are reported. SCL refers to the supervised contrastive learning approach detailed in Gunel et al. (2020). DASCL refers to the dictionary-assisted supervised contrastive learning approach detailed in this paper.

Model	Loss	Accuracy	Precision	Recall
Logistic Regression	Log Loss with L_2 Penalty	71.0	71.3	41.4
RoBERTa$_{\text{BASE}}$	CE	75.7	69.0	67.3
RoBERTa$_{\text{BASE}}$	CE and SCL	75.0	67.9	66.7
RoBERTa$_{\text{BASE}}$	CE and DASCL	78.3	71.5	72.8

Table 2: Results over the test dataset for economic media (Barberá et al., 2021). Because of resource constraints, the results are over one random seed. These results are not finalized and are subject to change.

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 Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning.

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

Daniel J. Hopkins, Eunji Kim, and Soojong Kim. 2017. Does newspaper coverage influence or reflect public perceptions of the economy? *Research & Politics*, 4(4):2053168017737900.

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.

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.

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.

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.

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.

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.

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

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

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. Working Paper.

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.

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.

Patrick Y. Wu and Walter R. Mebane, Jr. 2022. MARMOT: A deep learning framework for constructing multimodal representations for vision-and-language tasks. *Computational Communication Research*, 4(1):275–322.

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.