When Do Simple Text Representations Suffice? A Study of Structural Bag-of-Words and Transformers

Atishay Jain and Chia-Han Chen
University of California San Diego
Department of Computer Science and Engineering
atj003@ucsd.edu, chc016@ucsd.edu

Emily Chen and Samantha Lin
University of California San Diego
Halıcıoğlu Data Science Institute
emc001@ucsd.edu, yul186@ucsd.edu

Abstract

Although transformer-based models dominate many NLP benchmarks, classical lexical representations such as Bag-of-Words (BoW) and TF-IDF remain widely used due to their efficiency, interpretability, and robustness. In this work, we revisit classical Bag-of-Words representations and study how incorporating lightweight structural and stylistic information affects their performance, without introducing the complexity of end-to-end neural models. We propose Structural Bag-of-Words (S-BoW) representations that augment TF-IDF with positional lexical features, function-word distributions, and document-level statistics, and evaluate them across multiple text classification datasets under varying data regimes and distribution shifts. Contrary to our initial expectations, S-BoW does not consistently outperform strong TF-IDF or frozen transformer-based baselines. However, our analysis highlights when and why structural cues fail to provide additional gains, clarifies the practical limits of enriching classical representations, and offers guidance for future work on bridging the gap between simple lexical models and modern neural encoders. Our code and demo are available at https://github.com/ChiaHan1/CSE-261-Project.

1 Introduction

1.1 Problem statement

Recent advances in representation learning have led to the widespread adoption of dense embedding and transformer-based models for text classification. These models often demonstrate strong empirical performance, but their advantages are not uniform across datasets and settings. In particular, factors such as limited training data, domain-specific vocabulary, and distribution shift can reduce the effectiveness of complex representations while increasing computational and deployment costs. This project investigates how far classical Bag-of-Words representations can be extended to capture useful structural and stylistic information while preserving their simplicity and interpretability. We plan to further examine when extended lexical representations remain competitive with frozen transformer-based representations, particularly in low-resource and distribution-shifted settings.

1.2 Significance

Representation choice has a significant impact on model performance, interpretability, and robustness. In many practical text mining applications, deploying large neural models may be unnecessary or undesirable due to limited data, computational constraints, or the need for transparent decision-making. Understanding when simpler representations suffice, and how they can be improved, can lead to more efficient and reliable systems. By revisiting and extending classical lexical representations, this project contributes to a clearer understanding of representation trade-offs and provides insights that are directly relevant to real-world text mining tasks.

This question is especially important because representation choice is often treated as secondary to model choice even though it strongly shapes performance, computational cost, and interpretability. By comparing classical, structurally enriched, and frozen transformer-based representations under a unified framework, this study helps clarify when model sophistication yields meaningful gains and when simpler alternatives are sufficient.

2 Related work

Classical lexical representations such as Bag-of-Words (BoW) and TF-IDF have long served as foundational methods for text classification and information retrieval due to their simplicity, efficiency, and interpretability (Joachims, 1998; Ramos, 2003). Despite their age, these sparse rep-

resentations remain strong baselines and are still widely deployed in practical text mining systems.

Distributed word representations, including Word2Vec and GloVe, introduced dense vector encodings that capture semantic regularities and improved performance across a range of natural language processing tasks (Mikolov et al., 2013; Pennington et al., 2014). These methods form the basis of many embedding-based document representations and are often assumed to dominate sparse lexical approaches.

Transformer-based models further advanced text representation learning by modeling contextual and long-range dependencies, achieving state-of-the-art results on many benchmarks (Vaswani et al., 2017; Devlin et al., 2019). However, these improvements come with increased computational cost, environmental impact, and reduced interpretability, and they can be sensitive to data distribution shifts (Strubell et al., 2019; Belinkov and Glass, 2019).

Several empirical studies have compared sparse and dense representations for text classification, often finding that linear models with BoW or TF-IDF features remain competitive under certain data regimes (Wang and Manning, 2012; Joulin et al., 2017). More recent work has explicitly revisited classical representations in modern contexts, arguing that carefully designed sparse features can still provide strong performance and favorable efficiency trade-offs (Arora et al., 2017; Galke and Scherp, 2022).

Robustness under distribution shift has become an important evaluation criterion in recent NLP research. Prior studies emphasize the need to measure performance degradation when test data differs from training data and highlight substantial robustness gaps in modern models (Oren et al., 2019; Taori et al., 2020). Our evaluation follows this line of work by comparing the robustness of sparse, structured, and frozen transformer-based representations under controlled shift scenarios.

3 Experimental Design

We conduct a controlled empirical study comparing three families of text representations for document classification: (i) classical sparse lexical models, (ii) structurally enriched Bag-of-Words variants, and (iii) frozen transformer-based encoders. Our goal is to quantify when simple lexical representations suffice relative to more complex alternatives, and to characterize the regimes in which structural cues or transformer features provide consistent benefits.

Dataset	Training set size	Test set size
20 Newsgroups	11314	7532
AG News	120000	7600
Amazon Polarity	100000	100000

Table 1: Training set and test set size of each dataset

3.1 Datasets

We evaluate on three widely used English classification benchmarks that differ in domain, label space, and document style:

- **20 Newsgroups**: a topic classification dataset with 20 semi-specialized newsgroup categories, which emphasizes multi-domain vocabulary and topical variation.
- **AG News**: a four-way news topic classification benchmark covering broadly accessible news headlines and descriptions.
- **Amazon Polarity**: a large-scale sentiment classification dataset with user-generated product reviews and substantial lexical variation.

Together, these datasets span topic-focused and sentiment-focused tasks, as well as both moderate- and large-scale training regimes. They also vary in label granularity, lexical diversity, and dataset scale, which allow a broader evaluation of representation behavior.

3.2 Representations and Classifier

For each dataset, we construct multiple representations under a shared preprocessing pipeline (lowercasing, basic tokenization, and minimal normalization) and train identical linear classifiers for comparability.

Classical lexical representations. We consider several strong sparse baselines that are standard in text classification and information retrieval (Joachims, 1998; Ramos, 2003):

- **Binary BoW**: indicator features marking the presence of each token in a document.
- **Count BoW**: raw term frequency vectors.
- **TF–IDF**: unigram (and optionally bigram) term frequency vectors reweighted by inverse document frequency.

Structural Bag-of-Words (S–BoW). We design S–BoW variants that augment TF–IDF with simple structural and stylistic cues motivated by prior work on text structure and style:

- **Position-aware TF–IDF**, where we compute separate TF–IDF vectors over document segments (e.g., beginning vs. end) and concatenate them to capture positional emphasis.
- **Function-word distributions**, which summarize stylistic patterns via normalized frequencies of common function words.
- **Document-level statistics**, including length and type–token ratio to capture global lexical richness and verbosity.

Frozen transformer-based representations. Finally, we extract document embeddings from a pretrained transformer encoder used strictly as a fixed feature extractor. The representations are extracted once from a pretrained encoder and used as fixed document features. We obtain sequence representations by pooling the final-layer token embeddings, and feed these fixed vectors into the same linear classifier used for BoW models. This setting mirrors common practice in using large pre-trained models as generic feature providers for downstream tasks.

Classifier and training. Across all representations, we train multinomial logistic regression classifiers with cross-entropy loss. Hyperparameters (e.g., regularization strength) are selected via validation on each dataset using the same search space for all representations. This design isolates the effect of the input representation while holding the classifier family and optimization procedure constant, so differences in performance can be attributed primarily to representation quality rather than downstream optimization choices.

Implementation details For all lexical models, the vocabularies were constructed from the training split only. For frozen transformer, we used `bert-base-uncased` and obtained a fixed document vector by mean pooling over the final hidden layer (Devlin et al., 2018). Logistic regression regularization strength was selected from a shared validation grid, and all experiments were repeated over three random seeds.

3.3 Evaluation Protocol

We focus on three complementary aspects of performance.

Standard in-domain accuracy. We report accuracy and macro-F1 on held-out test splits using the original training data for each dataset.

Data regime analysis. To probe sample efficiency, we construct stratified subsamples of the training data at multiple fractions (e.g., 1%, 5%, 10%, 25%, 50%, and 100%) and train separate models on each subset. This yields learning curves for all representations, allowing us to compare their behavior in low- and high-resource regimes.

Robustness to distribution shift. We evaluate robustness under shifts in topic and vocabulary by constructing out-of-domain test settings where the label space is preserved but the topic mix or document sources differ from those seen during training. We measure degradation in accuracy and macro-F1 relative to the in-domain setting.

Ablation of structural components. Within S–BoW, we perform ablation studies that incrementally remove structural features (position-aware TF–IDF, function-word distributions, document-level statistics) to estimate their individual contributions.

All experiments are run with multiple random seeds; we report mean performance and, when appropriate, standard deviations.

4 Results and Analysis

In this section, we summarize the main empirical findings across datasets, data regimes, and robustness settings. Overall, we find that plain TF–IDF remains a very strong baseline, and that our structural extensions do not consistently improve over it.

4.1 Overall Performance

Across all three datasets, TF–IDF with a linear classifier achieves competitive or best-in-class performance among the lexical representations. S–BoW variants occasionally match or slightly outperform plain TF–IDF on specific datasets or metrics, but these gains are small and not consistent across tasks. Frozen transformer embeddings typically perform on par with or better than the best lexical model, especially on larger datasets, but the margin over TF–IDF is often modest given their substantially higher computational cost.

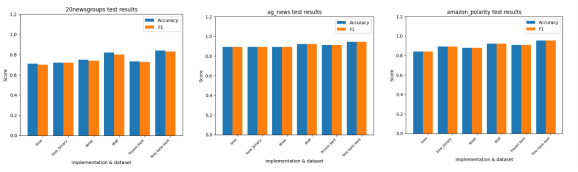

Figure 1: Test accuracy and macro-F1 for all models (BoW variants, TF–IDF, and transformer-based representations) across the three datasets. TF–IDF and fine-tuned transformer models achieve the strongest overall performance.

Model	Accuracy	F1 score
BoW	0.71	0.70
BoW Binary	0.72	0.72
TF-IDF	0.82	0.80
S-BoW	0.75	0.74
Frozen BERT	0.73	0.72
Fine-tuned BERT	0.84	0.83

Table 2: Performance on 20 Newsgroups dataset for all models

The strong TF-IDF performance suggests that these classification tasks remain heavily driven by lexical identity and high-level token salience. In such settings, weighted sparse features already capture much of the discriminative information needed by a linear classifier. This also helps explain why the structural S–BoW variants do not consistently improve performance since the added cues may be intuitively meaningful, but they do not always introduce information that is sufficiently distinct from what TF–IDF already encodes.

To contextualize the gap between feature-based baselines and full end-to-end neural modeling, we also ran a small supplementary experiment with light BERT fine-tuning. We trained a lightly fine-tuned BERT classifier where BERT is fine-tuned for three epochs. As expected, this model achieved strong overall performance and outperformed sparse and frozen-representation baselines. However, fine-tuning introduces substantially greater model capacity and computational cost, and because our primary goal is to compare simple lexical representations against lightweight structural extensions and frozen transformer features under a controlled setup, we treat this result only as a contextual upper-bound rather than a central object of study.

Model	Accuracy	F1 score
BoW	0.89	0.89
BoW Binary	0.89	0.89
TF-IDF	0.92	0.92
S-BoW	0.89	0.89
Frozen BERT	0.91	0.91
Fine-tuned BERT	0.94	0.94

Table 3: Performance on AG News dataset for all models

Model	Accuracy	F1 score
BoW	0.84	0.84
BoW Binary	0.89	0.89
TF-IDF	0.92	0.92
S-BoW	0.88	0.88
Frozen BERT	0.91	0.91
Fine-tuned BERT	0.95	0.94

Table 4: Performance on Amazon Polarity dataset for all models

4.2 Data Regime Analysis

Learning curves reveal that all representations benefit from additional training data, with performance generally saturating as the training fraction approaches 100%. In low-resource regimes (e.g., 1% and 5%), S–BoW often starts from a lower accuracy and macro-F1 than plain TF–IDF, suggesting that the added structural features may introduce sparsity or noise that is harder to estimate from few examples. As the amount of data increases, S–BoW tends to catch up and can sometimes reach similar end performance to other BoW variants, but it does not reliably surpass TF–IDF. On the large and relatively simple Amazon Polarity dataset, some S–BoW configurations come closer to the best lexical and transformer baselines, but the improvements are not robust across seeds or feature choices.

4.3 Robustness to Distribution Shift

Under topic and vocabulary shifts, both S–BoW and TF–IDF exhibit similar patterns of degradation: accuracy and macro-F1 drop substantially on out-of-domain data relative to in-domain test sets. The structural cues we introduce do not meaningfully mitigate this degradation; in some shifted settings they slightly worsen performance, likely because the positional or stylistic patterns they encode are themselves domain-specific. Frozen transformer representations are also affected by distribution shift, but in several cases they show somewhat

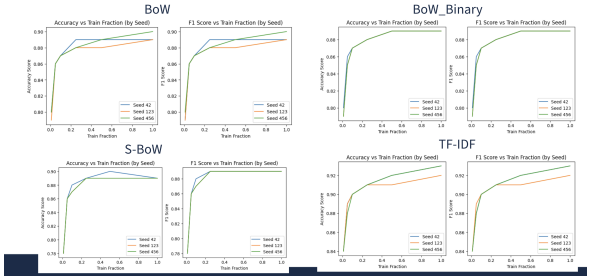

Figure 2: Learning curves on AG News across different training fractions, comparing BoW variants, TF–IDF, and transformer-based representations.

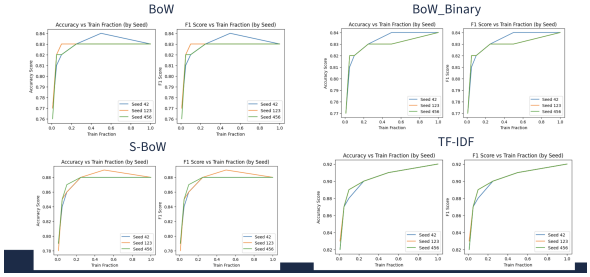

Figure 3: Learning curves on Amazon Polarity across different training fractions, highlighting how S–BoW compares to other BoW variants and TF–IDF in a large binary classification task.

smaller relative drops, consistent with their broader pre-training.

4.4 Ablation of Structural Components

Ablation experiments indicate that our current positional weighting scheme can even hurt performance relative to plain TF–IDF in some settings: removing position-aware TF–IDF sometimes yields small but consistent gains. Function-word distributions and simple document-level statistics have more mixed effects, providing slight improvements on certain datasets but negligible or negative changes elsewhere. Taken together, these results suggest that naïve structural augmentations to BoW do not reliably capture task-relevant information beyond what is already encoded in TF–IDF.

5 Discussion and Conclusions

5.1 Discussion

Our study highlights the enduring strength of simple lexical representations for document classification. Despite incorporating a range of lightweight structural and stylistic cues, S–BoW does not consistently outperform carefully tuned TF–IDF or frozen transformer-based features across datasets,

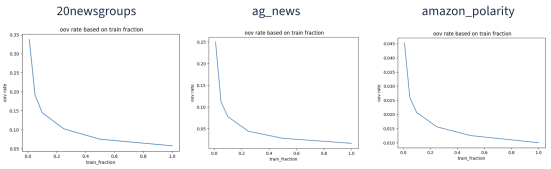

Figure 4: Learning curves under an out-of-vocabulary-focused setting, illustrating how different representations respond to increased lexical novelty.

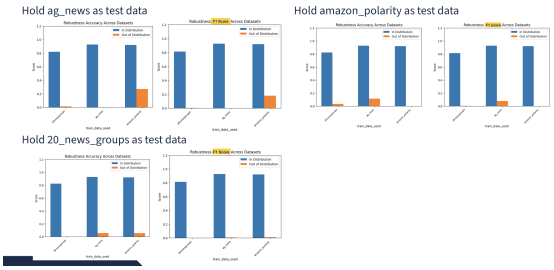

Figure 5: Robustness of TF–IDF under distribution shift, measured by accuracy and macro-F1 drops between in-domain and out-of-domain test settings.

data regimes, and robustness settings. This negative result is nevertheless informative: it clarifies that many intuitive structural features are either redundant with existing lexical statistics or insufficiently aligned with the underlying task signal. However, a limitation of this study is that the structural cues we evaluate are intentionally simple and mostly surface-level. As a result, the negative results should be interpreted as evidence against these lightweight augmentations, not against all possible forms of structural representation.

5.2 Future Work

Our findings point to several concrete directions for future investigation.

Deeper structural representations. The surface-level cues explored in this work (positional bins, function-word frequencies, document statistics) are provably expressible within or redundant with TF–IDF weighting. A natural next step is to evaluate structural signals that sparse lexical models *cannot* capture, such as dependency-parse features (e.g., subject–verb–object triples) (de Marneffe et al., 2006), discourse-level structure derived from Rhetorical Structure Theory (RST) parses (Ji and Eisenstein, 2014), or learned soft positional encodings over sparse features (Vaswani et al., 2017). These representations encode relational and hierarchical information that is absent from any bag-of-

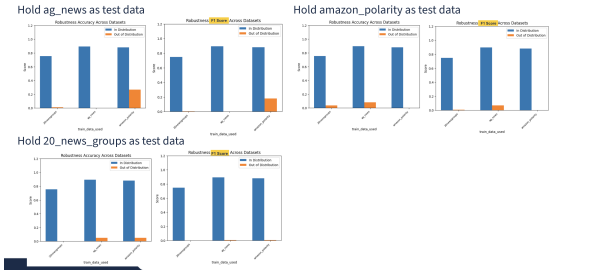

Figure 6: Robustness of S–BoW under distribution shift, compared to TF–IDF and other BoW variants.

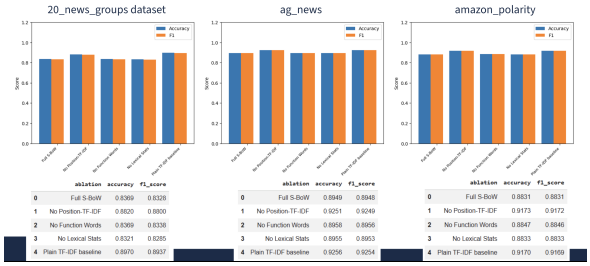

Figure 7: Ablation study of S–BoW components, showing the impact of removing positional, function-word, and document-level features on performance.

words variant and may provide the non-redundant signal needed to improve over TF–IDF baselines.

Task-aware feature selection. Our ablation results show that structural features can *hurt* performance when concatenated indiscriminately. Prior work has demonstrated the importance of principled feature selection for text classification (Yang and Pedersen, 1997). Future work could apply task-driven feature selection, for example using mutual information or chi-squared filtering to retain only class-discriminative structural features before concatenation, or using sparse group lasso regularization (Yuan and Lin, 2006) to allow the classifier to zero out entire feature groups that are uninformative for a given task.

Controlled distribution shift analysis. Our cross-dataset robustness experiments reveal that structural cues are themselves domain-specific, but the shifts we evaluate are coarse-grained. Recent benchmarks have highlighted the need for fine-grained characterization of distribution shifts in machine learning (Koh et al., 2021), and studies on pretrained models have shown that robustness varies substantially across shift types (Hendrycks et al., 2020). More controlled experiments that isolate specific shift types, such as vocabulary shift (same topics, different register), topic shift (same register, different topics), or temporal shift (same source, different time periods), would clarify which representation components are most sensitive to which kinds of distributional change and whether normalization strategies can mitigate the fragility of structural features.

Hybrid sparse–dense representations. Our results identify a persistent gap between strong TF–IDF baselines and fine-tuned transformers, with frozen transformer features falling between the two. A promising direction is to explore hybrid representations that concatenate sparse lexical features with low-dimensional projections of frozen transformer embeddings, following the spirit of learned sparse–dense models in information retrieval (Formal et al., 2021), or to use knowledge distillation (Hinton et al., 2015) from a fine-tuned transformer into a sparse linear model, thereby transferring contextual knowledge into a lightweight, interpretable, and deployable system.

Broader evaluation settings. This study is limited to English-language document classification. Extending the evaluation to morphologically rich languages, where BoW representations face additional challenges from agglutination and inflection (Tsvetkov et al., 2015), and to tasks beyond classification, such as information retrieval, clustering, or extractive summarization, would test whether the patterns we observe generalize to settings where structural sensitivity may play a larger role. Cross-lingual evaluation frameworks such as XTREME (Hu et al., 2020) could provide standardized benchmarks for such comparisons.

5.3 Conclusion

Our findings underscore the importance of treating TF–IDF as a genuinely strong and competitive baseline, rather than a strawman, when evaluating new representation learning methods. The failure of lightweight structural augmentations to consistently improve over TF–IDF clarifies that bridging the gap between sparse lexical models and modern neural encoders will require structural representations that are deeper, task-adapted, and robust to domain variation. By releasing our code and experimental pipeline, we hope to facilitate further investigations into when and how simple text representations suffice in the presence of powerful neural encoders.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. *ICLR*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. In *Transactions of the ACL*.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of SIGIR*.

Lukas Galke and Ansgar Scherp. 2022. Bag-of-words vs. graph vs. sequence in text classification. In *Proceedings of ACL*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, and 1 others. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of ACL*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, and 1 others. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of ICML*.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of ACL*.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of EACL*.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, and 1 others. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of ICML*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS*.

Yonatan Oren, Shiori Sagawa, and Tatsunori Hashimoto. 2019. Distributionally robust language modeling. In *Proceedings of EMNLP*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*.

Juan Ramos. 2003. Using tf-idf to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *Proceedings of ACL*.

Rohan Taori, Ludwig Schmidt, and 1 others. 2020. Measuring robustness to natural distribution shifts in image classification. In *Proceedings of ICML*.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, and Chris Dyer. 2015. Lexicon-free cross-lingual transfer of morpho-syntactic features. In *Proceedings of NAACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, and 1 others. 2017. Attention is all you need. In *Proceedings of NeurIPS*.

Sida Wang and Christopher Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of ACL*.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of ICML*.

Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.