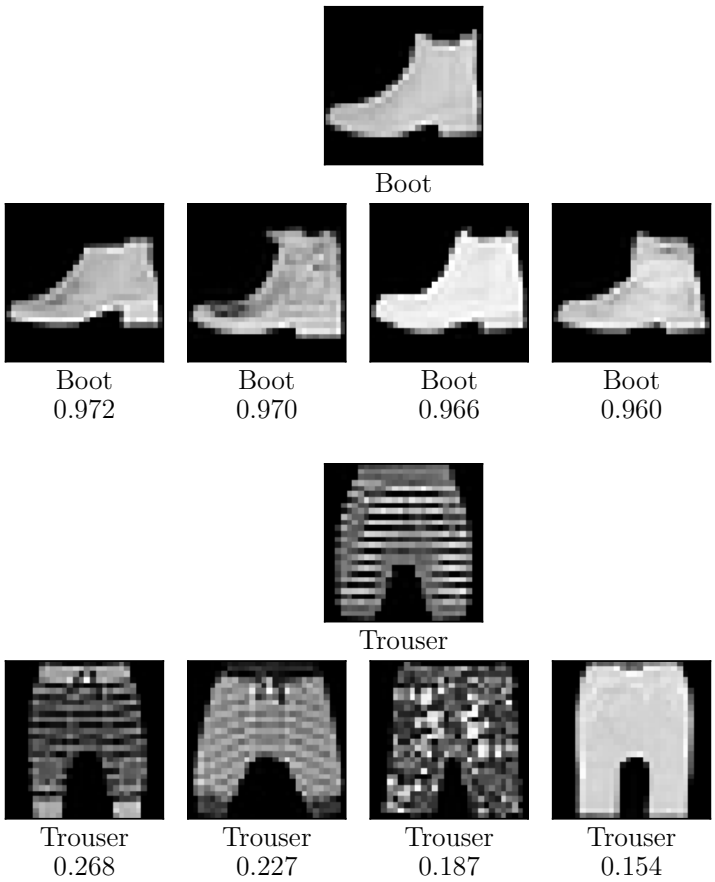

Figure 6.3: Two examples of predictions made by the DWAC model on the Fashion MNIST dataset. An approximate explanation for the model's prediction on the test image (single image) is the four most-highly weighted training images (row of four), along with their weights and labels. The top example shows a prediction that is well supported, with many nearby training instances of the same class, whereas the bottom shows an example where even the nearest training instances are relatively distant.

with its true label, and the four closest images from the proper training set (as measured in the embedded space) are shown below it, along with their weights. In this case, all share the same label, and all contribute approximately equally to the prediction.

As a contrasting example, Figure 6.3 (bottom) shows an example which has poor support among the nearest training points. Although the prediction is correct, and the closest training images appear visually similar, this sort of wide-legged trouser is quite rare in the dataset (most trousers included have narrow legs). As such, the model has

Weight	Sentence	Label
Test	*Drupal 6 dynamic menu item*	*Drupal*
0.994	drupal how to limit the display results of the calendar view	Drupal
0.994	Drupal : Different output for first item in a block	Drupal
0.994	changing user role in drupal	Drupal

Weight	Sentence	Label
Test	*save data from editable division*	*Ajax*
0.214	Pass data from workspace to a function	Matlab
0.133	upload data from excel to access using java	Apache
0.130	Finding incorrectly - formatted email addresses in a CSV file	Excel

Table 6.3: Two examples from the Stack Overflow dataset with approximate explanations from a DWAC model: an easy example with many close neighbours (top) and a more difficult example with no close neighbours (bottom). The first line is the test instance in both cases.

clearly not learned as well how to embed such images into the low-dimensional space. The low weights indicate that one should be skeptical of this prediction.

Images are relatively easy to compare at a glance; text, by contrast, may be more difficult. Nevertheless, the explanations given by DWAC for the predictions on text data are in some cases very meaningful. For the Stack Overflow dataset, for example, many instances are almost trivially easy, in that the label is part of the question. Not surprisingly, these examples tend to have many highly-weighted neighbors which provide a convincing explanation. Such an example is shown in Table 6.3 (top). In other cases, the text is more ambiguous. Table 6.3 (bottom) shows an example with very little support, for which a user might rightly be skeptical of the model, based on both the weights and the explanation.

Finally, because one can use any deep model to compute $f(\boldsymbol{x})$, one is free to choose one with preferred characteristics, including interpretability. For the text classification experiments, I chose a base model originally proposed for interpretable classification of medical texts (Mullenbach et al., 2018). To further unpack the explanation given for

	Income	Cover.	Lend.	Fash.	CIFAR	Subj.	Stack	IMDB	Amazon
k=1	0.85	0.76	0.96	0.95	0.96	0.98	0.81	0.95	0.78
k=5	0.91	0.77	0.98	0.97	0.98	0.99	0.84	0.98	0.87
k=10	0.93	0.77	0.99	0.98	0.98	0.99	0.85	0.98	0.90
k=100	0.96	0.83	0.99	0.99	0.99	1.00	0.89	0.99	0.94

Table 6.4: Impact of considering a subset of the training instances as an approximate explanation: the columns show agreement with the full model on the single most-probable label when basing the prediction on only the k closest training instances.

a prediction, one could, for example, inspect the attention weights for a particular pair of sentences to understand the importance of each word in context.

6.5.3 Approximate explanations

Because the weights on training instances decrease exponentially with distance, the closest training instances will contribute the most to the prediction. In some cases, only relatively few training instances will be required to fully determine the model's predicted label (because beyond this the remaining instances will lack sufficient weight to alter which class will be most highly-weighted). In practice, most test instances tend to require a substantial proportion of the nearest training instances in order to cross this threshold. However, even considering a much smaller number of the closest training instances may still result in high agreement with the prediction based on all of the data. Table 6.4 shows the agreement with the full model if one only considers the top k neighbors to each test instance. For most datasets, this agreement is very high, even for a very small number of neighbours.

6.5.4 Confidence and credibility

The above results only considered the single top-scoring label predicted for each instance; here I extend both models to the conformal setting. To verify the theoretical

expectations of conformal methods, I show that my proposed measure of nonconformity correctly works to maintain a desired error rate in making predictions. Figure 6.4 shows the results for the Fashion MNIST dataset, where I vary ε (the desired maximum error rate) from 0 to 0.2. The top subfigure (a) shows the proportion of predictions on the test set which are *correct* (that is, which contain the true label), for the softmax model using negative probability (probs) as a measure of nonconformity, DWAC using the same measure, and DWAC using my proposed measure of nonconformity (weights) given in Equation 6.13. As can be seen, all three demonstrate correct coverage, with all lines close to but not exceeding the expected proportion ($1 \quad \varepsilon$) across the full range (shown as a dashed line on the top subfigure). Note that this is not the same as accuracy, as some predictions may contain multiple labels.

The second and third subfigures show the same lines for the proportion of predicted label sets that are empty (b) or contain multiple labels (c). The bottom figure shows the mean number of labels in all non-empty predicted label sets. In all cases, the dashed line represents an optimal outcome (that is, a proportion of predicted label sets equal to ε are empty, all other predictions are correct, and no predictions contain multiple labels).

As can be seen, the softmax and DWAC models give indistinguishable results when using the same measure of nonconformity. My proposed measure of nonconformity, by contrast, appears to be slightly less efficient, producing slightly more predicted label sets with multiple labels, but also slightly more empty label sets, which represent identifiable errors contributing to the proportion of incorrect predictions.

The advantage of my proposed measure, however, comes in robustness to out-of-domain data. If one trains a model on the Fashion MNIST dataset, and then ask it to make predictions on the original MNIST digits dataset (which has the same size and data format, but consists of hand-written digits rather than items of clothing), one

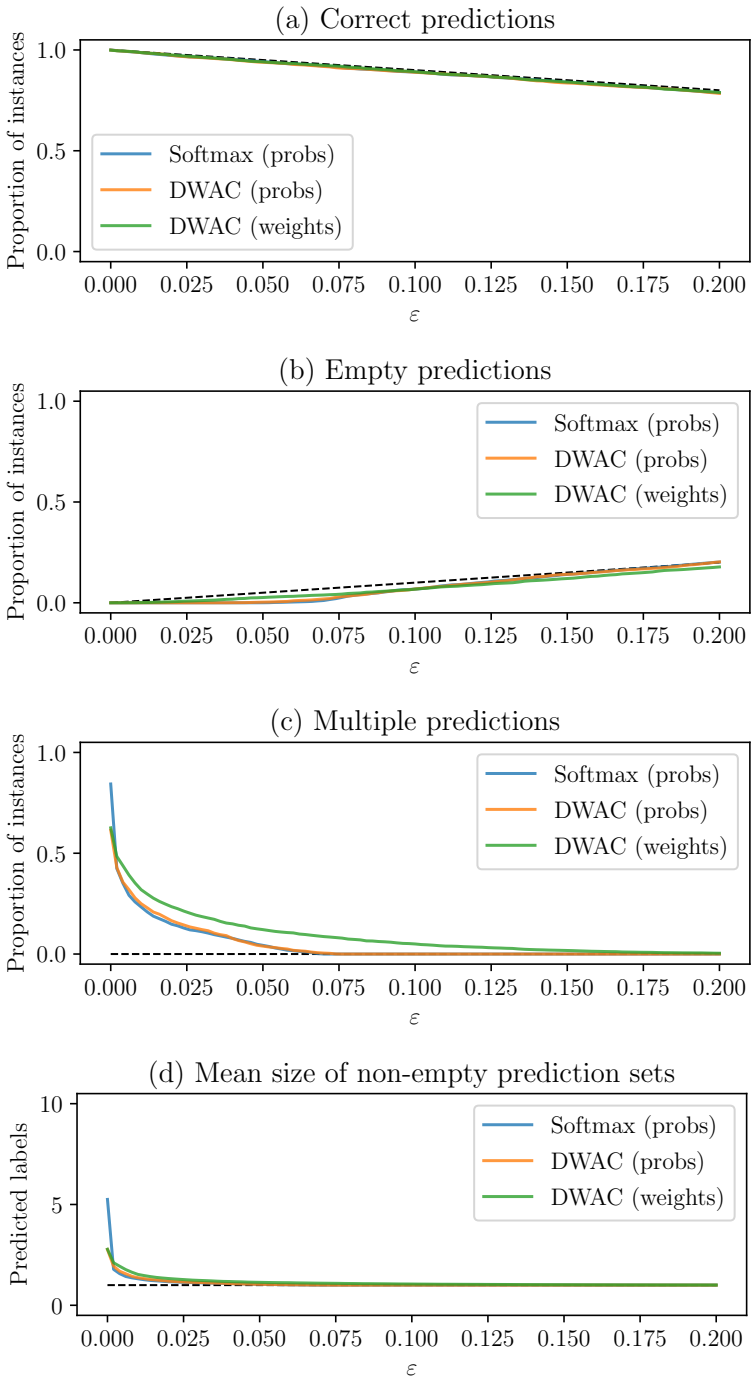

Figure 6.4: Coverage of various models on the Fashion MNIST test data, as I vary the desired maximum error rate (ε). From top to bottom, the subfigures show (a) the proportion of predicted label sets that are correct (contain the true label); (b) that are empty (make no prediction); (c) that contain multiple labels; and (d) the mean number of labels in non-empty prediction sets. The softmax and DWAC models give nearly identical results when using negative probability as a measure of nonconformity (probs). My proposed measure (weights) has an indistinguishable error rate, but is slightly less efficient. The dashed line in each figure represents an optimal response.

would hope that a good model would predict relatively low probability for all such instances. In fact, as has been previously observed (Nguyen et al., 2015a), deep models tend to predict relatively high probabilities, even for out-of-domain data, and this is also true of DWAC models.

Fortunately, the *credibility* score from a conformal predictor provides a meaningful estimate of how much one should trust the corresponding prediction.[10] Both the softmax model (using negative probability as a measure of nonconformity), and the DWAC model (using my proposed measure of nonconformity) give low credibility to the vast majority of out-of-domain examples, as shown in Figure 6.5. The credibility scores from DWAC however, are noticeably shifted closer to zero, indicating that the sum of the weights of the corresponding class is a better measure when one is concerned about the possibility of out-of-domain data. (For in-domain data, the credibility values will be approximately uniformly distributed).

I find similar results when training on CIFAR-10 images, and predicting on the Tiny Images dataset (see Figure 6.6), and an even more extreme difference in the case of the Covertype dataset, where I treat one out of seven classes as out-of-domain data, and train a model on only the six remaining classes (see Figure 6.7) . For that setup, the mean credibility score from the softmax model on the out-of-domain data is 0.9, whereas for the DWAC model with my measure of nonconformity it is 0.2. This indicates that to the softmax model, the held-out images “look like” even more extreme versions of one of the other classes, whereas the DWAC model correctly recognizes that the held-out images are relatively unlike the training instances (relative to the calibration data).

[10]Specifically, as mentioned above, it is equal to one minus the model probability that none of the labels should be given to this instance.

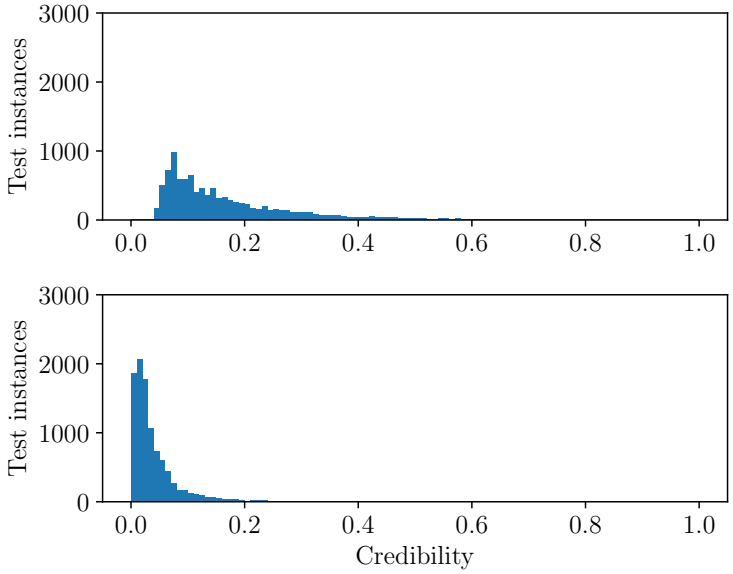

Figure 6.5: Empirical distribution of credibility scores from the softmax (top) and DWAC (bottom) models when trained on Fashion MNIST and tested on MNIST digits (which have the same input format but different content), with the latter using my proposed measure of nonconformity.

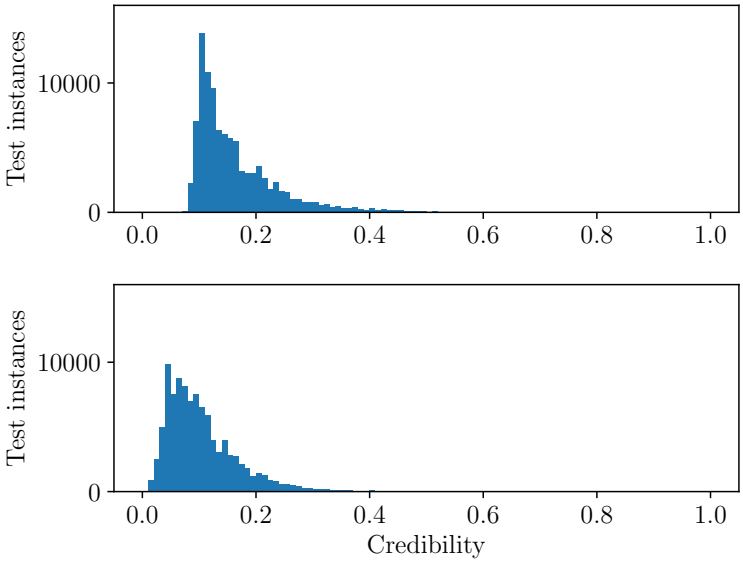

Figure 6.6: Empirical distribution of credibility scores from the softmax (top) and DWAC (bottom) models when trained on CIFAR-10 and tested on Tiny Images, with the latter using our proposed measure of nonconformity.

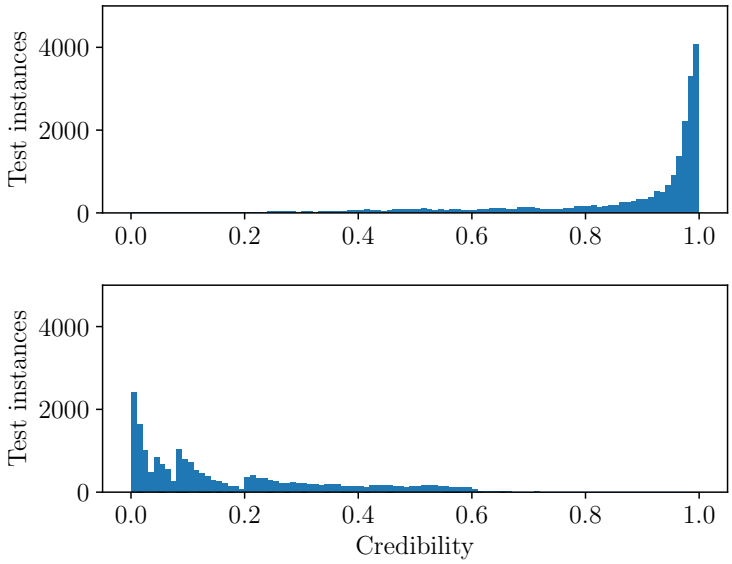

Figure 6.7: Empirical distribution of credibility scores from the softmax (top) and DWAC (bottom) models when trained on 6 of the classes in the Covertype dataset and tested on the 7th class. This represents an extreme example where the credibility scores are skewed towards 1 in the softmax model, given that in-domain data would typically be approximately uniformly distributed.

6.6 Discussion and future work

The idea of interpretability has always been important in statistics and machine learning, but has taken on a renewed urgency with the increased expressive power of deep learning models and the expanded deployment of machine learning systems in society. No single approach to interpretability is likely to solve all problems; here I have focused on adapting existing models so as to make their predictions more transparent, by decomposing them into a sum over training instances, the support for each of which can be inspected. As emphasized by papers that have made use of user studies (Huysmans et al., 2011; Kulesza et al., 2013; Narayanan et al., 2018; Poursabzi-Sangdeh et al., 2018; Yin et al., 2019), careful empirical work is required to evaluate the effectiveness of explanations, and I leave such an evaluation of this approach for future work.

A few recent papers have also sought to formalize the question of when to trust a

classifier. Jiang et al. (2018) present a method for determining how much any existing classifier should be trusted on any given test point, based on the relative distance to the most-likely and second-most-likely clusters, with clusters based on a pruned training set. This appears to be a theoretically well-motivated and empirically effective technique, but is more focused on trust than interpretability. In an unpublished paper, Papernot and McDaniel (2018) propose to train a conventional deep model, but make predictions using a k-nearest neighbours approach, with distance computed using all internal nodes of the network. My approach, by contrast, is to train a model using the same form as will be used for prediction, to make predictions based on a weighted sum over the entire training set, and rely on similarities computed in the low-dimensional space of the final layer. Wallace et al. (2018) apply the method from Papernot and McDaniel (2018) to the problem of text classification and explore the implications for interpretability.

There have also been several papers focused on the problem of predicting whether data is in-domain or out-of-domain (Lee et al., 2018; Liang et al., 2018). Many of these build on Hendrycks and Gimpel (2017), who observed that the predicted probabilities contain some useful signal as to whether data came from in-domain or out-of-domain, and proposed to use this to differentiate between the two by thresholding these probabilities. The authors did not, however, make the connection to conformal methods, which offer a more theoretically sound basis on which to make these decisions, as well as greater flexibility of metrics, beyond just predicted probability.

There are several natural extensions to this work which could be pursued, such as applying a similar architecture to regression or multi-label problems, as well as extending the idea of nonconformity to provide class-conditional guarantees.

6.7 Summary

In this chapter I have demonstrated that even for sophisticated deep learning models, it is possible to create a nearly identical model, with all of the same desirable properties, that nevertheless provides an explanation for any prediction in terms of a weighted sum of training instances. In domains where the training data can be freely inspected, this provides greater transparency by revealing the many components that explicitly contribute to a model's prediction, each of which can in principle be inspected and interrogated. Moreover, this method can build on top of other approaches to interpretability, by choosing an appropriate base model. When an approximate explanation will suffice, then using only a small subset of the training instances provides a natural high-fidelity approximation.

More importantly, representing the prediction in this manner suggests a natural alternative measure of nonconformity, which, as I have shown, provides a more effective measure for detecting out-of-domain examples. Even in cases where training data cannot be shared (due to privacy concerns, for example), this use of conformal methods still allows us to assert a quantitative estimate of the credibility of an individual prediction, one that is far more meaningful than the model's predicted probability.

The experiments in this chapter focused on a variety of data types, rather than a particular social science problem, in part to show the generality of this method. However, the ability to better interrogate why a model is making a particular prediction, or to be able to easily visualize what a model has learned, is of obvious value for model development where transparency and interpretability are important. Moreover, by providing a quantitative estimate of how much a particular prediction should be trusted, we can more effectively know the limits of where our measurements might break down, and when we should be skeptical of a model's predictions. Particularly for

social science applications, where there is real potential for the translation of ideas into policy, the credibility scores from DWACs, or from conformal methods more generally, provide a potentially useful check on the possibility of invalid predictions.

Chapter 7

Conclusion

7.1 Summary of contributions

Machine learning and natural language processing have repeatedly demonstrated their value in interdisciplinary research, and the social sciences are no exception. However, productive interdisciplinary work requires consideration of the concerns and priorities of the social sciences, including an emphasis on validity, reliability, reproducibility, interpretability, and cost. This thesis has brought together a line of work on methods for both supervised and unsupervised settings which try to take these desiderata seriously. While no single method will ever be the right solution for all problems, these chapters include examples of models designed for specific purposes, such as estimating label proportions, as well as more generically useful approaches, such as modeling documents with metadata. In all cases, the goal has been to work towards collaborative research, by incorporating thinking from the social sciences about research methods, while aiming to make machine learning and natural language processing useful to scholars from a broader range of disciplines.

Specific contributions of this thesis include:

Chapter 3 presented a graphical model designed for automatically inferring archetypal character representations and story types in an unsupervised manner. These inferred values were shown to be useful in predicting annotations of the framing and tone of news articles about U.S. immigration, including when using a novel type of feature evaluation based on Bayesian optimization.

Chapter 4 introduced a model which generalizes multiple topic model variants in a unified framework, using neural variational inference. In addition to showing that this allows for more flexible exploration, by allowing metadata associated with documents to be incorporated in multiple ways, this chapter also argued for using neural variational inference as the foundation of a more accessible form of modeling, with the potential to expand the range of tools available to scholars in other disciplines.

Chapter 5 focused on the use of text classification as tool for measurement when the goal is to estimate label proportions, as is often the case in text-as-data research. It characterized two different sets of assumptions that might be made when researchers need to confront domain shift (*intrinsic* vs *extrinsic* labels), and argued that most past work targeted at accounting for this shift is based on an assumption which is inappropriate for most annotation scenarios. For the *extrinsic* setting, it was shown that the goal should be a well-calibrated model, and that we can improve upon a naive approach by using a different criterion for model selection.

Finally, Chapter 6 demonstrated how a simple change to any deep learning classifier can create a model which is equally accurate, but with more transparently interpretable predictions, by making predictions explicitly in terms of training instances. Furthermore, this approach suggested a novel measure of

non-conformity, one which was shown to be more useful in detecting out-of-domain data. These *deep weighted averaging classifiers* (DWACs) provide a useful approach when researchers are concerned with the possibility of encountering out-of-domain instances, or want to be able to provide an *explanation* for the decisions made by a classifier.

7.2 Recurring themes

Several themes which recur throughout this work, as summarized below:

Text-as-data: As emphasized in Chapter 2, human-generated natural language holds enormous potential for learning about society. Indeed, a great deal of social science is based on text in some form, including interviews and open-ended survey responses. Historically, making use of such data required interpretation by humans to extract important insights and patterns, which is neither reproducible nor scalable. By drawing on methods from machine learning and natural language processing, we can more effectively leverage the text that is available from these sources, as well as larger-scale cultural products, such as news articles, wikis, social media, and so on. Most chapters in this thesis (especially 3, 4, and 5) proposed ways of learning from text, with the idea of *framing* providing a running example.

Exploration and measurement: Social science is typically theorized in terms of concepts (formally, constructs), but these concepts must be operationalized in terms of properties that can be measured, along with appropriate instruments. Particularly when working with complicated data such as open ended text, the measurement problem becomes especially difficult. While machine learning and natural language

processing provide many possible ways of quantifying text, there is still enormous scope for designing custom models (as in Chapter 3), as well as for expanding the space of tools that are easily useable by scholars in other disciplines (as in Chapter 4). The ideas of exploration and measurement appear throughout, but Chapter 5 is specifically concerned with the problem of using text classification as an instrument, while Chapter 6 explores how we can make such instruments more transparent and trustworthy.

Learning from social science: Just as social sciences can benefit from tools and insights derived from machine learning and natural language processing, computer science can also benefit from ideas in the social sciences. Chapter 2 provided an overview of the main concerns surrounding research methods in social science, and discussed the importance of taking these concerns seriously, especially as algorithmic systems are coming to play an increasingly prominent role in society. Chapter 3 emphasized the importance of collaborating with domain experts for theoretical grounding, validation, and interpretation of results. Other chapters addressed various properties valued by social scientists, such as reliability, calibration (Chapter 5), transparency, and interpretability (Chapter 6). For additional recent work related to this thesis addressing the issues of cost and reproducibility, please refer to Gururangan et al. (2019) and Dodge et al. (2019).

7.3 Implications for computational social science

The majority of this thesis has focused on methods, without always making explicit the connection to particular, substantive research questions in social science. This is in part to emphasize generality, but also because many social science questions merit the kind of in-depth investigation that would not easily fit into this sort of thesis (and often

warrant a thesis or a book-length treatment of their own). Nevertheless, I hope that the ideas contained within will be useful for computational social science research and the text-as-data community in particular.

Chapter 3 provided an example of an approach to exploration and measurement using probabilistic graphical models, demonstrating how a model can be designed to instantiate ideas from other fields (e.g., the idea of archetypal character types in news) and produce meaningful representations. In particular, the personas model provided a means of both discovering the sorts of character representations and story types that were present in a corpus, as well as providing a set of measurements of the corpus (i.e., as features), which were found to be predictive of framing. This sort of approach has wide-spread applicability to a variety of different types of data and questions, though as emphasized, it suffers from the need for a certain level of expertise in the area of graphical models in order to design an appropriate model and inference algorithm.

As a step towards mitigating that limitation, neural variational inference was used in Chapter 4 as a means of unifying a set of related models into a common framework, thereby providing the foundation for painless model customization by users, even by those with less expertise in statistics and machine learning. The resulting model, SCHOLAR, provides an alternative to other widely-used topic model variants, such as the structural topic model (Roberts et al., 2014), one which provides greater scalability, and the ability to incorporate richer covariates or additional prior knowledge in the form of pre-trained word vectors. As above, this sort of approach to modeling is useful primarily as a way of exploring a large collection of documents (possibly for the purpose of hypothesis generation), but also provides a way of making measurements of text (Wallach, 2016). In particular, follow on work (Gururangan et al., 2019) has demonstrated that an extension of this model is highly effective as a way to learn useful document representations for semi-supervised text classification in the limited-resources setting. Further

work is required in order to make neural variational inference a broadly accessible way of exploring text corpora, including questions of interface design and how to handle more complex models. Nevertheless, there is an opportunity here to use it as the basis of a user-friendly system for text analysis in the social sciences.

The idea of using text classification as a tool for measurement was addressed more directly in Chapter 5. This remains an important paradigm for any setting in which we would like to augment manual annotation (i.e., human coding) of documents. As shown, however, there are two important take-aways. First, if the goal is *only* to estimate the label proportions in a fixed corpus of documents, then for most applications, the *simple random sampling* approach of sampling and manually labeling a modest number of documents is likely to be far better, in terms of validity, reliability, and cost, than any approach based on dictionary methods or supervised learning. However, for cases in which we care about individual document labels, or only have access to part of the corpus during annotation, some form of supervised learning will often be necessary. As emphasized, if we must confront the problem of domain shift (as in the case of generalizing to future data), then it is important to consider the data generating process (i.e., intrinsic vs. extrinsic labels), and to recognize that validation is the only guarantee against potential error.

Finally, Chapter 6 made use of a variety of types of data, but is nevertheless highly relevant. First, in developing text classifiers as tools for measurement, it is important to be able to understand and communicate how they are working. Deep learning has brought such large gains in accuracy in numerous areas of natural language processing, that it would be ill-advised to ignore it. However, given the problems associated with deep learning as discussed in Chapter 6, it may be unappealing for many social scientists. DWACs provide, at a minimum, a more direct way of inspecting the basis for each prediction, which will help to determine if the model is doing something reasonable

during model development.

In addition, social science research is noteworthy in that many findings have the potential to be translated into policy. Any tool which is developed for the purpose of measurement (e.g., of text) has the potential to be deployed as a way of making predictions. Given the problems of domain shift discussed above, and the potential harms involved, especially when making predictions about people (e.g., based on their social media posts), it would be highly valuable to know when a model might be making a prediction that should not be trusted. The credibility scores from DWACs, or from conformal methods more generally, provide just such a mechanism; especially using DWACs, this allows us to identify data which may be unlike the training data, and for which we should be skeptical of the model's prediction.

Of course, as emphasized in the introduction, this kind of observational text-as-data research is only one part the universe of computational social science. Going beyond purely observational data, some of the ideas presented here could also be applied to the results of open-ended survey responses, for example, and could be incorporated into causal analysis in experimental work (see, for example, Fong and Grimmer, 2016).

7.4 Directions for future work

Each of the main chapters of this thesis addressed a particular problem and made a specific contribution. Nevertheless, there is great scope for potential future work.

The idea of *personas* served as the foundation for the model presented in Chapter 3, and was found to be useful for predicting annotations about the framing of news articles. Indeed, one important aspect of framing is the set of entities that are present in text, and the ways in which they are depicted. Further work along these lines seems

like a particularly promising direction for future research into framing. Although effective and interpretable, the model presented here only made use of a very limited set of textual evidence for inferring entity representations. A model which made use of greater context would be a natural extension, as would one which made use of additional metadata about the shared identity of entities depicted across documents or by different sources.

Developing a more comprehensive persona model might be difficult using traditional inference techniques, but neural variational inference, as used in Chapter 4 provides a potentially useful foundation upon which to base a model. Indeed, demonstrating that neural variational inference could easily be extended to the type of hierarchical model of text used for studying personas would be highly useful, as it would suggest greater generalizability. While there is excellent work happening in the domain of hierarchical VAE models (e.g., Sønderby et al., 2016), as well as probabilistic programming and automated inference (e.g., Tran et al., 2016), these methods have not yet seen much use in the social sciences when dealing with text.

In addition, further consideration of the trade-offs inherent in this sort of approach to inference would be valuable. Experiments in Chapter 4 suggest that inference using VAEs is inferior (in terms of model fit) to traditional collapsed Gibbs sampling with hyperparameter updating, at least for standard models like LDA, and understanding this difference would be useful. Similarly, the randomness in VAE-based inference creates potential problems in terms of reproducibility, though this could perhaps be mitigated by some sort of initialization, such as the spectral initializer used in the structural topic model (Roberts et al., 2014).

Going beyond entities, there is great scope for deeper analysis of framing, especially given the recent gains obtained on a wide range of tasks using contextual embeddings (Peters et al., 2018; Devlin et al., 2019). Promising directions include metaphor, ar-

gumentation, and narrative. Metaphor is a frequently-cited component of framing, but one which has been difficult to study computationally. Similarly, arguments are perhaps the quintessential aspect of framing (i.e., "to promote a particular problem definition, causal interpretation ...", Entman, 1993). Many codes in social science codebooks take the form of more or less specific ideas or claims (e.g., "immigration is good for the economy"), which may be expressed in many different ways. Given continued advances such as contextual embeddings, recognizing such specific ideas in text seems quite promising, given sufficient data. Finally, linking together these metaphors and claims into larger narratives which change over time is perhaps the ultimate objective, and could potentially benefit from work in computational narrative analysis.

The identification of ideas in text would be useful beyond the analysis of framing, and is a promising direction of research. A limitation, however, is the lack of evaluation data, as identifying every mention of an idea in a large corpus remains challenging. Moreover, the fact that the similarity of ideas is often implicit means that a more interactive approach, one integrating active learning and proper interface design, would allow a domain expert to navigate a corpus, gradually refining the concept by example. Additional methods of visualizing the relations between ideas, such as Tan et al. (2017), would also be useful.

Both calibration and interpretability are still underdeveloped areas which need further work. It is clear that classifier accuracy can benefit massively from pretraining (even in the low-resource setting; Gururangan et al., 2019), but we don't yet have a full appreciation of what this sort of transfer learning implies for calibration, or the potential for introducing unwanted biases into measurement. Many ideas have been proposed for interpretable models, but there is a great need for more empirical studies with real users to understand how people make sense of such information (Poursabzi-Sangdeh et al., 2018; Yin et al., 2019). Similar evaluation would be useful for thinking about

the credibility scores returned by conformal methods and the value of transparent predictions more broadly.

Finally, there is a great deal of work to be done in studying the second-order effects of these technologies. As machine learning and natural language processing are increasingly being deployed in social settings, such as recommender systems, voice assistants, and social media monitoring, we should expect that these interventions will modify the environment in which they operate. Fortunately, these fields of study also provide the means to study these kinds of downstream consequences at scale. Collaborative work in computational social science is in the best position of all fields to make sense of changes that are taking place in society due to the expansion of algorithmic interventions, and this presents a huge opportunity for further development of datasets, methods, analyses, interventions, and insights.

Bibliography

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J. Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014)*. 51

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional spaces. In *Proceedings of International Conference on Database Theory*. 103

David Aldous. 1985. Exchangeability and related topics. In *École d'Été St Flour 1983*, pages 1–198. Springer-Verlag. 36

Charles E. Antoniak. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6). 34

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596. 92

Josh Attenberg and Foster Provost. 2011. Inactive learning?: Difficulties employing active learning in practice. *SIGKDD Explorations Newsletter*, 12(2):36–41. 96

Ramnath Balasubramanyan, William W. Cohen, Doug Pierce, and David P. Redlawsk. 2012. Modeling polarizing topics: When do different political communities respond differently to the same news? In *Proceedings of ICWSM*. 56

Jason Baldridge and Miles Osborne. 2004. Active learning and the total cost of annota-

tion. In *Proceedings of EMNLP*. 96

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of ACL*. 31, 33, 35, 38

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of NAACL*. 51

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of ACL*. 38, 51

Pablo Barbera, Amber E. Boydstun, Suzanna Linn, Jonathan Nagler, and Ryan McMahonk. 2019. Automated text classification of news articles: A practical guide. In submission. 24

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpreting blackbox models via model extraction. *CoRR*, abs/1705.08504. 99

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of NAACL*. 28, 30

Frank R. Baumgartner, Suzanna L. De Boef, and Amber E. Boydstun. 2008. *The decline of the death penalty and the discovery of innocence*. Cambridge University Press. 28

Antonio Bella, Maria Jose Ramirez-Quintana, Jose Hernandez-Orallo, and Cesar Ferri. 2010. Quantification via probability estimators. In *IEEE International Conference on Data Mining*. 77, 82

Aurélien Bellet, Amaury Habrard, and Marc Sebban. 2013. A survey on metric learning for feature vectors and structured data. *CoRR*, abs/1306.6709. 100

Shai Ben-David, Shai Shalev-Shwartz, and Ruth Urner. 2012. Domain adaptation – can quantity compensate for quality? *Annals of Mathematics and Artificial Intelligence*, 70:185–202. 80

Robert D. Benford and David A. Snow. 2000. Framing processes and social movements: An overview and assessment. *Annual Review of Sociology*, 26:611–639. 27, 28

Adrian Benton and Mark Dredze. 2018. Deep Dirichlet multinomial regression. In *Proceedings of NAACL*. 74

James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. 2015. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science and Discovery*, 8(1). 46, 47

Adam J. Berinsky and Donald R. Kinder. 2006. Making sense of issues through media frames: Understanding the Kosovo crisis. *Journal of Politics*, 68(3):640–656. 28

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. 2009. Importance weighted active learning. In *Proceedings of ICML*. 96

Anol Bhattacherjee. 2012. *Social Science Research: Principles, Methods, and Practices*. Global Text Project. 20, 21

Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2009. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10. 85

Jock A. Blackard and Denis J. Dean. 1999. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24. 113

David M. Blei. 2014. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232. 16, 19

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *JACM*, 57(2). 57

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of ICML*. 52, 57

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022. 19, 33, 47, 52

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of NeurIPS*. 26

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of COLT*. 99

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. Applications of topic models. *Foundations and Trends in Information Retrieval*, 11(2-3):143–296. 19, 52

Amber E. Boydstun, Dallas Card, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2014. Tracking the development of media frames within and across policy issues. *Proceedings of APSA*. 29

Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Wadsworth. 98

Jochen Bröcker. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519. 83

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computional Linguistics*, 18(4):467–479. 47

Yuri Burda, Roger B. Grosse, and Ruslan R. Salakhutdinov. 2016. Importance weighted autoencoders. *CoRR*, abs/1509.00519. 62

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. 26

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic

model and its supervised extension. In *Proceedings of AAAI*. 74

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of ACL*. 28

Dallas Card, Justin H. Gross, Amber E. Boydstun, , and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of EMNLP*. 30

Dallas Card and Noah A. Smith. 2018. The importance of calibration for estimating proportions from annotations. In *Proceedings of NAACL*. 76

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural models for documents with metadata. In *Proceedings of ACL*. 52, 63

Dallas Card, Michael Zhang, and Noah A. Smith. 2019. Deep weighted averaging classifiers. In *Proceedings of ACM FAT**. 97

Ali Taylan Cemgil. 2009. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, pages 4:1–4:17. 55

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of ACL*. 51

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of NeurIPS*. 55, 66

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of AAAI*. 67

Eunsol Choi, Chenhao Tan, Lillian Lee, Cristian Danescu-Niculescu-Mizil, and Jennifer Spindel. 2012. Hedge detection as a lens on framing in the GMO debates: A position paper. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 70–79. 30

Dennis Chong and James N. Druckman. 2007. Framing theory. *Annual Review of Political Science*, 10(1):103–126. 27, 28

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of CVPR*. 100

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of KDD*. 27

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297. 99

Thomas M. Cover and Peter E. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27. 99

Paul D'Angelo and Jim A. Kuypers. 2010. *Doing News Framing Analysis*. Routledge. 27

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Proceedings of NAACL*. 47

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Proceedings of ACL*. 65

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*. 77

Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. 2007. Information-theoretic metric learning. In *Proceedings of ICML*. 100

Morris H. DeGroot and Stephen E. Fienberg. 1983. The comparison and evaluation of forecasters. *The Statistician: Journal of the Institute of Statisticians*, 32:12–22. 83

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Daniel Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of NAACL*. 28

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*. 19, 136

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. In *Proceedings of EMNLP*. 68

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of EMNLP*. 49, 132

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *CoRR*. 22, 98

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *TACL*, 5:471–486. 22

Ellen A. Drost. 2011. Validity and reliability in social science research. *Education Research and Perspectives*, 38(1). 20

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of ICML*. 52, 53, 56, 60

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58. 13, 27, 137

Robert M. Entman. 2007. Framing bias: Media in the distribution of power. *Journal of Communication*, 57(1):163–173. 30

Michael D. Escobar and Mike West. 1994. Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, 90:577–588. 34

Andrea Esuli and Fabrizio Sebastiani. 2015. Optimizing text quantifiers for multivariate loss functions. *ACM Trans. Knowl. Discov. Data*, 9(4). 77, 95

Ronen Feldman. 2013. Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82–89. 30

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Daniel Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. In *Proceedings of EMNLP*. 28

Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of Twitter research ethics. *Social Media + Society*, 4(1). 26

Lucie Flekova and Iryna Gurevych. 2015. Personality profiling of fictional characters using sense-level links between lexical resources. In *Proceedings of EMNLP*. 51

Christian Fong and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *Proceedings of ACL*. 81, 135

George Forman. 2005. Counting positives accurately despite inaccurate classification. In *Proceedings of ECML*. 77, 82, 86

George Forman. 2008. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206. 77

William A. Gameson and Andre Modigliani. 1989. Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology*, 95(1):1–37. 28

Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, 78(1):35–71. 31

Todd Gitlin. 1980. *The Whole World is Watching*. Berkeley: University of California Press. 13, 27

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report. 89

Jacob Goldberger, Geoffrey E. Hinton, Sam T. Roweis, and Ruslan R. Salakhutdinov.

2004. Neighbourhood components analysis. In *Proceedings of NeurIPS*. 100

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of ICLR*. 98

Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine*, 8(341). 22

David Graff and C Cieri. 2003. English gigaword corpus. *Linguistic Data Consortium*. 66

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of ACL*. 30

Glenn Greenwald. 2014. *No Place to Hide: Edward Snowden, The NSA, and the U.S. surveillance state*. Picador. 31

Justin Grimmer, Solomon Messing, and Sean J. Westwood. 2012. How words and money cultivate a personal vote: The effect of legislator credit claiming on constituent credit allocation. *American Political Science Review*, 106(4). 77

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297. 12, 15, 17, 18, 24, 25, 76

Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2018. A survey of methods for explaining black box models. *CoRR*, abs/1802.01933. 98

Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. In *Proceedings of AAAI*. 22

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of ICML*. 97

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A. Smith. 2019. Variational pretraining for semi-supervised text classification. In *Proceedings of ACL*. 19, 23, 68,

74, 132, 133, 137

Eric Hardisty, Jordan L. Boyd-Graber, and Philip Resnik. 2010. Modeling perspective using adaptor grammars. In *Proceedings of EMNLP*. 30

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of NeurIPS*. 20, 27

Mareike Hartmann, Tallulah Jansen, Isabelle Augenstein, and Anders Søgaard. 2019. Issue framing in online discussion fora. In *Proceedings of NAACL*. 28

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of IJCNLP*. 31

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of EMNLP*. 51

Junxian He, Zhiting Hu, Taylor Berg-Kirkpatrick, Ying Huang, and Eric P. Xing. 2017. Efficient correlated topic modeling with topic embedding. In *Proceedings of KDD*. 74

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of ICLR*. 98, 126

Philipp Hennig, David Stern, Ralf Herbrich, and Thore Graepel. 2012. Kernel topic models. In *Proceedings of AISTATS*. 59

Edward S. Herman and Noam Chomsky. 1988. *Manufacturing consent. The political economy of the mass media*. Vintage. 28, 31

Geoffrey E Hinton and Ruslan R Salakhutdinov. 2009. Replicated softmax: An undirected topic model. In *Proceedings of NeurIPS*. 74

Daniel Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):220–247. 12, 15, 76, 77, 82, 83, 93

Daniel J. Hopkins and Jonathan Mummolo. 2017. Assessing the breadth of framing

effects. *Quarterly Journal of Political Science*, 12(1):33–57. 28

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of NAACL*. 95

Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. 2006. Correcting sample selection bias by unlabeled data. In *Proceedings of NeurIPS*. 85

Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decis. Support Syst.*, 51(1). 125

Ozan İrsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *Proceedings of EMNLP*. 51

Mohit Iyyer, Peter Enns, Jordan L. Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of ACL*. 30

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of NAACL*. 51

Heinrich Jiang, Been Kim, and Maya R. Gupta. 2018. To trust or not to trust a classifier. *CoRR*, abs/1805.11783. 126

Michael I. Jordan and Tom M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science*, 349:255–260. 11

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*. 47

Daniel Kahneman. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux. 13

Katherine A. Keith and Brendan O'Connor. 2018. Uncertainty-aware generative models

for inferring document class prevalence. In *Proceedings of EMNLP*. 95

Been Kim, Cynthia Rudin, and Julie A Shah. 2014. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Proceedings of NeurIPS*. 100

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*. 114

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of ICLR*. 53, 57, 61, 62, 63

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017a. Human decisions and machine predictions. Technical Report 23180, National Bureau of Economic Research. 26

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review*, 105(5):491–495. 15

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017b. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of ITCS*. 20

Alexej Klushyn, Nutan Chen, Richard Kurle, Botond Cseke, and Patrick van der Smagt. 2019. Learning hierarchical priors in VAEs. *ArXiv*, abs/1905.04982. 75

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *Proceedings of ICML*. 100

Ron Kohavi. 1996. Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid. In *Proceedings of KDD*. 113

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Technical report, University of Pennsylvania. 24

Klaus Krippendorff. 2012. *Content analysis: an introduction to its methodology*. SAGE. 18, 24, 76

Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto. 113

Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, and David Blei. 2015. Automatic variational inference in Stan. In *Proceedings of NeurIPS*. 75

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. 2016. Automatic differentiation variational inference. ArXiv:1603.00788. 75

Todd Kulesza, Simone Stumpf, Margaret M. Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? ways explanations impact end users' mental models. *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. 125

Brian Kulis. 2013. Metric learning: A survey. *Foundations and Trends R in Machine Learning*, 5(4):287–364. 100

John D. Lafferty and David M. Blei. 2006. Correlated topic models. In *Proceedings of NeurIPS*. 58

Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of KDD*. 98

George Lakoff, Howard Dean, and Don Hazen. 2008. *Don't Think of an Elephant!: Know Your Values and Frame the Debate*. Chelsea Green Publishing. 27

Hugo Larochelle and Stanislas Lauly. 2012. A neural autoregressive topic model. In *Proceedings of NeurIPS*. 74

Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically driven neural language model. In *Proceedings of ACL*. 74

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Lszl Barabsi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational

social science. *Science*, 323(5915):721–723. 11, 16

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. 2018. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *Proceedings of ICLR*. 126

Jing Lei, James M. Robins, and Larry A. Wasserman. 2014. Distribution free prediction sets. *J Am Stat Assoc*, 108(510):278–287. 104

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of EMNLP*. 99

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of KDD*. 31

Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of ICLR*. 126

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of CoNNL*. 30

Zachary C. Lipton. 2016. The mythos of model interpretability. In *ICML Workshop on Human Interpretability in Machine Learning*. 22

Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *Proceedings KDD*. 98

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of NeurIPS*. 99

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of ACL*. 66, 113

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR*,

9:2579–2605. 116

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL*. 38, 47

Winter Mason, Jennifer Wortman Vaughan, and Hanna Wallach. 2014. Computational social science and social computing. *Machine Learning*, 95(3). 16

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of SIGIR*. 88, 113

Jon D. McAuliffe and David M. Blei. 2008. Supervised topic models. In *Proceedings of NeurIPS*. 52, 53, 55, 63

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of ICML*. 53, 55, 57, 61, 62

David Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of UAI*. 56, 74

Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship*. Addison-Wesley publishing company, Inc. 80

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of AAAI*. 100

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of NAACL*. 114, 119

Èlizbar A. Nadaraya. 1964. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142. 103

Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale

Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *CoRR*, abs/1802.00682. 125

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of ACL*. 66

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. 2015a. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of CVPR*, pages 427–436. 98, 123

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015b. Improving topic models with latent feature word representations. In *Proceedings of ACL*. 65

Dong Nguyen, A. Seza Dogruöz, Carolyn Penstein Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics*, 42:537–593. 20

Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of EMNLP*. 84, 112

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Proceedings of NeurIPS*. 31, 57

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. 2015c. Tea party in the house: A hierarchical ideal point topic model and its application to Republican legislators in the 112th congress. In *Proceedings of ACL*. 28, 31, 52, 57

Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, , and Jure Leskovec. 2015. QUOTUS: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of WWW*. 31

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of ICML*. 84

Helen Nissenbaum. 2009. *Privacy in Context: Technology, Policy, and the Integrity of*

Social Life. Stanford University Press. 26

Brendan O'Connor. 2014. *Statistical Text Analysis for Social Science*. Ph.D. thesis, Carnegie Mellon University. 18

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of ICWSM*. 51

Brendan O'Connor, David Bamman, and Noah A. Smith. 2011. Computational text analysis for social science: Model assumptions and complexity. In *NeurIPS Workshop on Computational Social Science and the Wisdom of Crowds*. 12, 15, 16, 18, 52, 76

Brendan O'Connor, Brandon M. Stewart, and Noah A. Smith. 2013. Learning to extract international relations from political context. In *Proceedings of ACL*. 51

Paul Ohm. 2010. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57. 26

John William Paisley, David M. Blei, and Michael I. Jordan. 2014. Bayesian nonnegative matrix factorization with stochastic variational inference. In *Handbook of Mixed Membership Models and Their Applications*, pages 205–224. 55

Sinno J. Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359. 78

Zhongdang Pan and Gerald M. Kosicki. 1993. Framing analysis: An approach to news discourse. *Political communication*, 10(1):55–75. 30

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*. 113

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135. 30

Nicolas Papernot and Patrick McDaniel. 2018. Deep k-nearest neighbors: Towards

confident, interpretable and robust deep learning. *CoRR*. 126

Martin Pelikan. 2005. Bayesian optimization algorithm. In *Hierarchical Bayesian optimization algorithm*, pages 31–48. Springer. 46

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. 2014. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053. 82

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke S. Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL*. 19, 136

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. 85

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2018. Manipulating and measuring model interpretability. *CoRR*, abs/1802.07810. 125, 137

Daniel Preotiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle H. Ungar. 2017. Beyond binary labels: Political ideology prediction of twitter users. In *Proceedings of ACL*. 30

Kevin M. Quinn Burt L. Monroe Michael Colaresi Michael H. Crespin Dragomir R. Radev. 2009. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1). 22

Alec Radford, Rafal Józefowicz, and Ilya Sutskever. 2018. Learning to generate reviews and discovering sentiment. *CoRR*, abs/1704.01444. 19

Piyush Rai, Avishek Saha, III Hal Daumé, and Suresh Venkatasubramanian. 2010. Domain adaptation meets active learning. In *Proceedings of NAACL*. 96

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009.

Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of EMNLP*. 74

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of ACL*. 51

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNNL*. 51

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna K. Jerebko, Charles Florin, Gerardo Hermosillo, Luca Bogoni, and Linda Moy. 2009. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In *Proceedings of ICML*. 95

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2018. Do CIFAR-10 classifiers generalize to CIFAR-10? *CoRR*, abs/1806.00451. 98

Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of ICML*. 53, 57, 61

Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of NAACL*. 99

Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 2018a. Anchors: High-precision model-agnostic explanations. In *Proceedings of AAAI*. 99

Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 2018b. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings ACL*. 99

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from Twitter. In *Proceedings of KDD*. 51

Molly Roberts, Brandon Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Gadarian, Bethany Albertson, and David Rand. 2014. Structural topic models for open ended survey responses. *American Journal of Political Science*, 58:1064–1082.

53, 133, 136

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of UAI*. 52

Francisco J. R. Ruiz, Michalis K. Titsias, and David M. Blei. 2016. The generalized reparameterization gradient. In *Proceedings of NeurIPS*. 75

Matthew J. Salganik. 2017. *Bit by Bit: Social Research in the Digital Age*, open review edition edition. Princeton University Press, Princeton, NJ. 12, 16, 23, 26, 27

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of ACL*. 26, 77

Craig Saunders, Alexander Gammerman, and Vladimir Vovk. 1999. Transduction with confidence and credibility. In *Proceedings of IJCAI*. 104, 106

Anne Schneider and Helen Ingram. 1993. Social construction of target populations: Implications for politics and policy. *The American Political Science Review*, 87(2):334–347. 31

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of ACL*. 38, 47

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of FAT**. 27

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? *CoRR*, abs/1611.07450. 99

Burr Settles. 2012. *Active Learning*. Morgan & Claypool. 96

Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421. 104, 106

David A. Smith, Ryan Cordell, and Elizabeth Maddock Dillon. 2013. Infectious texts: modeling text reuse in nineteenth-century newspapers. In *IEEE International Conference on Big Data*. 31

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical Bayesian optimization of machine learning algorithms. In *Proceedings of NeurIPS*. 46

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*. 95

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*. 51

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. 51

Casper K. Sønderby, Tapani Raiko, Lars Maaløe, Søren K. Sønderby, and Ole Winther. 2016. Ladder variational autoencoders. In *Proceedings of NeurIPS*. 75, 136

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of ACL*. 51

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *Proceedings of ICLR*. 53, 55, 57, 58, 60, 62, 65, 68

Amos J. Storkey. 2009. When training and test sets are different: Characterising learning transfer. In Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Candela Sugiyama Schwaighofer Lawrence Lawrence, editors, *Dataset Shift in Machine Learning*, chapter 1, pages 3–28. MIT Press. 80

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy

considerations for deep learning in nlp. In *Proceedings of ACL*. 23

Masashi Sugiyama, Makoto Yamada, Paul von Bünau, Taiji Suzuki, Takafumi Kanamori, and Motoaki Kawanabe. 2011. Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Networks*, 24(2). 85

Kai S. Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of ACL*. 51

Chenhao Tan, Dallas Card, , and Noah A. Smith. 2017. Friendships, rivalries, and trysts: Characterizing relations between ideas in text. In *Proceedings of ACL*. 137

Rob Tibshirani. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288. 98

Dustin Tran, Alp Kucukelbir, Adji B. Dieng, Maja Rudolph, Dawen Liang, and David M. Blei. 2016. Edward: A library for probabilistic modeling, inference, and criticism. ArXiv:1610.09787. 75, 136

Oren Tsur, Dan Calacci, and David Lazer. 2015. Frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of ACL*. 28, 31

Berk Ustun and Cynthia Rudin. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391. 26, 98

Baldwin Van Gorp. 2010. Strategies to take subjectivity out of framing analysis. In Paul D'Angelo and Jim A. Kuypers, editors, *Doing News Framing Analysis: Empirical and Theoretical Perspectives*, chapter 4, pages 84–109. Routledge. 31

Vladimir Vapnik. 1998. *Statistical learning theory*. Wiley. 99

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Proceedings of NeurIPS*.

100

Vladimir Vovk. 2012. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*. 106

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Springer. 104, 106

Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of NAACL*. 30

Eric Wallace, Shi Feng, and Jordan Boyd-Graber. 2018. Interpreting neural networks with nearest neighbors. In *Proceedings of the Workshop on Analyzing and Interpreting Neural Networks for NLP*. 126

Hanna Wallach. 2016. Interpretability and measurement. EMNLP Workshop on Natural Language Processing and Computational Social Science. 15, 19, 22, 52, 55, 133

Hanna Wallach. 2018. Computational social science ≠ computer science + social data. *Communications of the ACM*, 61(3). 12, 15, 16

Hanna Wallach, David M. Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Proceedings of NeurIPS*. 55

Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. In *Proceedings of AISTATS*. 26, 98

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of ACL*. 45

Geoffrey S. Watson. 1964. Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, 26(4):359–372. 103

Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. 2006. Distance metric learning for large margin nearest neighbor classification. In *Proceedings of NeurIPS*. 100

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer

learning. *Journal of Big Data*, 3(1). 78

Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. In *CoRR*. 113

Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. 2002. Distance metric learning, with application to clustering with side-information. In *Proceedings of NeurIPS*. 100

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *Proceedings of NAACL*. 113

Yan Yan, Rómer Rosales, Glenn Fung, Subramanian Ramanathan, and Jennifer G. Dy. 2013. Learning from multiple annotators with varying expertise. *Machine Learning*, 95:291–327. 95

Ming Yin, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of CHI*. 125, 137

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R Routledge, and Noah A Smith. 2011. Predicting a scientific community's response to an article. In *Proceedings of EMNLP*. 64

Dani Yogatama, Lingpeng Kong, and Noah A. Smith. 2015. Bayesian optimization of text representations. In *Proceedings EMNLP*. 46

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of KDD*. 95

Jing Zhang, Xindong Wu, and Victor S. Sheng. 2016. Learning from crowdsourced labeled data: a survey. *Artif. Intell. Rev.*, 46:543–576. 95

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017.

Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the EMNLP*. 26, 83

Jun Zhu, Amr Ahmed, and Eric P. Xing. 2009. MedLDA: Maximum margin supervised topic models for regression and classification. In *Proceedings of ICML*. 74