STAT9700 Final Project: Conformal Prediction

Joseph Rudoler

December 11, 2023

Abstract

Conformal prediction is a model-agnostic framework for constructing valid prediction intervals and hypothesis tests without making any assumptions about the underlying data distribution. As modern machine learning systems continue to scale in both size and complexity, the ability to make valid predictions without making assumptions about the training data or model is becoming increasingly important. In particular, the success and ubiquity of deep and over-parameterized neural networks has led to a growing interest in techniques that can provide statistically rigorous uncertainty quantification for models that are effectively a "black box". In this report, we will provide an overview of the conformal inference framework and some important results, discuss some extensions and applications of conformal inference, and finally discuss some potential avenues for future research.

Introduction

Conformal inference (also known as conformal prediction) was first introduced by Vovk et al. (1999)

Foundations

This section will begin with a discussion of some conformal prediction basics, loosely following Angelopoulos and Bates, 2022. This will include an abbreviated proof of the validity of conformal prediction sets for exchangeable data, and an overview of some important applications and extensions of conformal prediction. Next, we will discuss in more detail how conformal prediction can be married to kernel density estimation to obtain prediction sets with asymptotic efficiency guarantees (Lei et al., 2013). We will also discuss how conformal prediction can be used in a cross-validation setting to obtain confidence intervals with rigorous coverage guarantees (Barber et al., 2020). Finally, we will see what happens when we try to apply conformal prediction to non-exchangeable data (Barber et al., 2023).

Conformal prediction basics

The central goal of conformal prediction is to construct prediction sets which have coverage guarantees at a specified confidence level $1-\alpha$. That is, given the task of mapping some inputs $X \in \mathcal{X} \mapsto \mathcal{Y}$, we ideally want to construct a set $C(X) \subset \mathcal{Y}$ such that $\mathbb{P}(Y \in C(X)) \geq 1-\alpha$ for all $X \in \mathcal{X}$. As it turns out, when the data are i.i.d. or exchangeable, this is pretty easy. The definition of conformal prediction sets and the formal statement of the coverage guarantee, for exchangeable data, is as follows:

Theorem 1 (Conformal prediction sets for exchangeable data)**.** *Suppose that* (X_i, Y_i) *for* $i = 1, \ldots, n$ *and* (X_{test}, Y_{test}) *are exchangeable random variables.*
Define

$$\hat{q} = \inf\left\{q : \frac{|i : s(X_i, Y_i) \leq q|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right\}$$

where $\lceil \cdot \rceil$ *is the ceiling function.*
Define a prediction set

$$C(X) = \{y : s(X, y) \leq \hat{q}\}$$

Then

$$\mathbb{P}(Y_{test} \in C(X_{test})) \geq 1 - \alpha$$

Informally, we define a *conformal score* function $s : (\mathcal{X}, \mathcal{Y} \mapsto \mathbb{R})$ which serves as a heuristic measure of how unlikely a given observation is (usually based on a trained model). Then, for a test observation X_{test} we predict the set of all Y which are below approximately the $(1-\alpha)$ quantile of these scores on a calibration dataset. This means that we're choosing a prediction set which contains all Y which are "not too unlikely" given the calibration data. Note that the coverage guarantee holds for any s, but the choice of s will affect the size (and consequently, the usefulness) of the prediction set $C(X)$ - intuitively, the less informative s is, the larger $C(X)$ will need to be to achieve the desired coverage.
The proof of this coverage guarantee (with the simplifying assumption that $s(X_i, Y_i)$ are unique, to avoid handling ties) is as follows:

Proof. Begin with a set of exchangeable pairs $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{\text{test}}, Y_{\text{test}})$. We want to obtain a prediction set for X_{test} which contains the true value Y_{test} with probability at least $1 - \alpha$. Assume without loss of generality that we have a model (e.g. a trained neural network) which can be used to compute a score $s(X, Y)$ for any (X, Y) pair.
We can sort the scores $s(X_i, Y_i)$ for $i = 1, \ldots, n$ in ascending order, and denote the ith smallest score by $s_{(i)}$. We set the threshold

$$\hat{q} = \inf\left\{q : \frac{|i : s(X_i, Y_i) \leq q|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right\}$$

Since the scores are sorted (with no ties), we can see that

$$\hat{q} = \begin{cases} s_{(\lceil (n+1)(1-\alpha) \rceil)} & \text{if } \alpha \geq \frac{1}{n+1} \\ +\infty & \text{otherwise} \end{cases}$$

And accordingly construct a prediction set:

$$C(X) = \{y : s(X, y) \leq \hat{q}\}$$

What is the probability that $Y_{\text{test}} \in C(X_{\text{test}})$? This is where the exchangeability assumption is crucial. Since the data are exchangeable, s_{test} is distributed identically to s_i for any $i = 1, \ldots, n$. It follows that s_{test} is equally likely to be fall into any of the $n+1$ intervals between the n calibration scores. Formally, for any integer k we have:

$$\mathbb{P}(s_{\text{test}} \leq s_{(k)}) = \frac{k}{n+1}$$

In particular, we have:

$$\mathbb{P}(s_{\text{test}} \leq s_{(\lceil (n+1)(1-\alpha) \rceil)}) = \frac{\lceil (n+1)(1-\alpha) \rceil}{n+1} \geq 1 - \alpha$$

Which proves the coverage guarantee. □

Part of the "magic" of conformal prediction is that this procedure can be applied post-training to any model, for basically any prediction task. That means we can treat any model as a "black box" and still obtain valid prediction sets without making any assumptions about how the model was parameterized. While we assume that the data are exchangeable (i.e. identically distributed), they need not be independent or to follow any particular distribution. We will see later how this assumption can be relaxed to allow for non-exchangeable data.
To help solidify our intuition for conformal prediction, we include a simple outline of the conformal procedure as described in (Angelopoulos & Bates, 2022) and concurrently expand the example case of multi-label image classification:

1. Given a pre-trained model, we **identify a heuristic notion of uncertainty** for the model's predictions. In image classification, we might use the softmax probabilities $\hat{f}(X)$ output by the a neural network.
2. **Define a conformal score function $s(\cdot)$** based on this uncertainty measure. For our example, we might define $s(X, Y) = 1 - \hat{f}(X)_Y$. This score function is lower for more likely image labels, and higher for less likely image labels.
3. **Set a threshold $\hat{q}$** at the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the scores $s(X_i, Y_i)$ for $i = 1, \ldots, n$.
4. **Construct a prediction set** $C(X_{\text{test}}) = \{y : s(X_{\text{test}}, y) \leq \hat{q}\}$. In image classification, this is the set of all image labels with a score below the $\hat{q}$ quantile of scores on the calibration dataset. In other words, you add the mostly likely labels to the prediction set until their total probability mass under the distribution of scores on the calibration dataset is at least $1 - \alpha$. Then the prediction set contains the true image label with probability at least $1 - \alpha$.

Again, we emphasize that the key to making conformal prediction sets *useful* lies in the choice of the score function s. While any s provides sets which are valid (in the sense that they have the desired coverage), a good choice of s will result in a small prediction sets which are also *adaptive*. That is, the size of the prediction set will depend on how much uncertainty the model has about the prediction.

Split vs Full conformal prediction In the above discussion, we assumed that we had access to a calibration dataset of exchangeable pairs (X_i, Y_i) for $i = 1, \ldots, n$ which is non-overlapping with the training data used to fit the model. This is known as *split conformal prediction* (or *inductive conformal prediction*), and it has the desirable property of allowing a user to evaluate a pre-trained model without doing any additional training. However, in some cases we may not have enough data to split into training and calibration sets. In this case, we can use the full dataset to construct a valid prediction set. This is known as *full conformal prediction* (or *transductive conformal prediction*). Full conformal prediction is computationally more expensive than split conformal prediction, as it involves fitting a new model for each possible value $y \in \mathcal{Y}$ and recomputing the scores. However, it is still possible to obtain valid prediction sets without splitting the data. In full conformal prediction, we construct an augmented dataset of exchangeable pairs (X_i, Y_i) for $i = 1, \ldots, n$ by adding a new observation (X_{n+1}, y) to the training data. We

then fit a model to this augmented dataset and compute the scores $s^y(X_i, Y_i)$ for $i = 1, \ldots, n+1$. The threshold $\hat{q}$ is now dependent on the value y we added to the training data:

$$\hat{q}^y = \inf\left\{q : \frac{|i : s^y(X_i, Y_i) \leq q|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right\}$$

And the prediction set becomes:

$$C(X) = \{y : s^y(X_{n+1}, y) \leq \hat{q}^y\}$$

Intuitively, this is testing if the the new observation (X_{n+1}, y) is exchangeable with all of the other observations, under the null hypothesis that y is the true label for X_{n+1}. The next section discusses a connection between full conformal prediction and kernel density estimation which allows us to obtain prediction sets with asymptotic efficiency guarantees.

Kernel Density Estimation

Generating conformalized prediction sets can be cast as a problem of estimating density level sets for an unknown probability distribution (Lei et al., 2013). Consider a sequence of i.i.d. data drawn from some unknown distribution with density $p(y)$. If we want to find a set C for a new observation such that $\mathbb{P}(Y \in C) \geq 1 - \alpha$, this is equivalent to finding an estimated density level set

$$L(t_\alpha) = C_\alpha = \{y : \hat{p}(y) \geq t_\alpha\} \quad \text{where} \quad t_\alpha = \inf\{t : \mathbb{P}(Y \in L(t)) \geq 1 - \alpha\}$$

We will deal with the more general setting in this section (following the original paper), but it is worth noting that in typical machine learning applications, we will actually be interested in finding a set $C(X)$ such that the conditional probability $\mathbb{P}(Y \in C(X)|X) \geq 1 - \alpha$ and the density we want to estimate is $p_{Y|X}(x, y)$.
Note that if $\hat{p}(y)$ were estimated perfectly, we would have no need to conformalize the density level set – we would simply greedily add the most likely observations to the set until we reached the desired coverage. However, in practice we will need to estimate $\hat{p}(y)$ from a finite sample of data and inevitably introduce some error which might cause us to undercover or overcover.
We now consider the standard kernel density estimator (KDE) with kernel fuction K and bandwidth h,

$$\hat{p}(u) = \frac{1}{n}\sum_{i=1}^{n} K_h(u - Y_i)$$

For our score function, we will use an augmented version of the KDE which is estimated on the augmented dataset $Y_1, \ldots, Y_{n+1}$ where $Y_{n+1} = y$:

$$\hat{p}^y(u) = \frac{1}{n+1}\sum_{i=1}^{n+1} K_h(u - Y_i)$$

The paper defines a conformal score a bit differently than we have so far, such that higher conformal scores correspond to more likely labels. This is helpful in order to more naturally frame the prediction set as a density level set. So we have $s^y(y) = \hat{p}^y(y)$, and we set the threshold $\hat{q}^y$ at the $\frac{\lfloor (n+1)\alpha \rfloor}{n}$ quantile of the scores $s^y(X_i, Y_i)$ for $i = 1, \ldots, n$ and take the valid prediction set to be:

$$\hat{C}_\alpha = L(\hat{t}_\alpha) = \{y : s^y(y) \geq \hat{q}^y\}$$

So, our prediction set is all y such that the estimated density level set $L(\hat{t}_\alpha)$ contains y with probability at least $1-\alpha$. (Note that this is equivalent to making the score function $s^y(y) = -\hat{p}^y(y)$, defining $\hat{q}^y$ at the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the scores, taking a prediction set $\hat{C}_\alpha = \{y : s^y(y) \leq \hat{q}^y\}$.)
Since this kernel density estimator is expensive to compute for every possible value of y, Lei et al. propose a faster approximation that sandwiches the kernel density estimator between two simpler estimators (L^- and L^+) and preserves finite-sample validity. That is, they they define the following lemma:

Lemma 1. *Let $\hat{C}_\alpha$ be the conformal prediction set based on the KDE with kernel function K, and assume $\sup_u |K(u)| = K(0)$. Let $Y_{(1)} \dots Y_{(n)}$ be the data in increasing order by their estimated kernel density.*

$$L^- = L\left(\hat{p}(Y_{(\lfloor (n+1)\alpha \rfloor)})\right) \quad \textit{and} \quad L^+ = L\left(\hat{p}(Y_{(\lfloor (n+1)\alpha \rfloor)})\right) - \frac{1}{nh} \sup_{u,u'} |K(u) - K(u')|$$

Then

$$L^- \subseteq \hat{C}_\alpha \subseteq L^+$$

Importantly, this sandwiching technique also provides lower and upper bounds on the efficiency of the kernel density estimator, which itself is harder to analyze. Lei et al. (2013) proceed to show that this procedure is asymptotically efficient at rate $r_n = (\frac{\log n}{n})^{c_{p_\alpha}}$ where c_{p_α} is a constant depending on the smoothness of the true underlying density. By efficient, we mean that probability of the loss $R(\hat{C}_\alpha, C_\alpha) = \mu(\hat{C}_\alpha \setminus C_\alpha) \leq \mu(\hat{C}_\alpha) - \mu(C_\alpha)$ exceeding r_n converges to zero as $n \to \infty$.

It is worth noting as well that while the asymptotic efficiency of the kernel density estimator is only proven for the full conformal prediction procedure, kernel density estimation can be used to obtain valid prediction sets in the split conformal prediction setting as well.

Conformalized Bayes. Before moving on to the next section, we will mention that framing conformal prediction as a density level set estimation problem also connects nicely to Bayesian inference. In particular, we can choose the conformal score function to be the posterior predictive density of the label given the data, and the prediction set will be the density level set with coverage guarantee $1-\alpha$. "Conformalized Bayes" allows us to obtain valid prediction sets for Bayesian models without making any assumptions about the correctness of the model or the data distribution - in this sense it mitigates concerns about misspecified priors. Hoff (2021) shows that under certain conditions, this procedure is Bayes optimal (in the sense that it minimizes the Bayes risk among all valid prediction sets).

Jackknife Conformal Prediction

As in Barber et al., 2020.

Extensions and related work

In this section, we further discuss two crucial extensions of conformal inference to settings with non-exchangeable data (Gibbs & Candes, 2021; Tibshirani et al., 2020). We will also discuss some related work on uncertainty quantification for deep neural networks that lies outside the conformal inference framework

Learn then test

See (Angelopoulos et al., 2022).
Maybe go more specifically into conformal language modeling (Quach et al., 2023).

Application: Adaptive Conformal Prediction for Text Classification

In this section, we will provide a case study of how conformal prediction can be used to obtain prediction sets with rigorous coverage guarantees for text classification with a large language model (LLM). We will compare the naive conformal prediction procedure with the adaptive conformal prediction procedure, and show that the adaptive procedure can result in prediction sets which are much smaller. Moreover, we will show that the adaptive procedure actually satisfies a stronger guarantee of *conditional coverage* as opposed to the *marginal coverage*.

Methods

Model and data

We classify text with the transformer-based language model BART (Lewis et al., 2019), finetuned on the Multi-Genre Natural Language Inference (MNLI) categorized text corpus [1], as proposed by Yin et al. (2019) for zero-shot text classification. We use the HuggingFace implementation of `bart-large-mnli`[2], and set up a zero-shot text classification pipeline with the `transformers` library (Wolf et al., 2019). We note that this model is no longer state of the art, but it is more lightweight (400 million parameters) than the larger and more performant models which have gained prominence (e.g. Meta's Llama2 models range from 7-70 billion parameters), and therefore more suitable for fast inference on a laptop. Since we are interested in demonstrating the utility of conformal prediction, we do not need to use the most powerful model available. We will see that this model results in prediction sets which are fairly large, and it is likely that a more powerful model would result give us better heuristic uncertainty estimates and therefore smaller prediction sets.

Our prediction task is to classify text from a new data source, a Kaggle dataset of text documents belonging to 5 new categories[3]. These categories – Politics, Sport, Technology, Entertainment, and Business – are mostly nonoverlapping with the categories in the MNLI corpus. The one exception being Politics, which is similar to the Government genre in MNLI – we will see that this category is, unsurprisingly, the easiest to classify. Of course, BART's original training data surely contained text related to all of these genres, but it was not specifically trained to classify them.

[1] https://huggingface.co/datasets/multi_nli
[2] https://huggingface.co/facebook/bart-large-mnli
[3] https://www.kaggle.com/datasets/sunilthite/text-document-classification-dataset/data

References

Angelopoulos, A. N., & Bates, S. (2022, December 7). *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*. arXiv: 2107.07511 [cs, math, stat]. Retrieved November 2, 2023, from http://arxiv.org/abs/2107.07511

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., & Lei, L. (2022, September 29). *Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control*. arXiv: 2110.01052 [cs, stat]. https://doi.org/10.48550/arXiv.2110.01052

Barber, R. F., Candes, E. J., Ramdas, A., & Tibshirani, R. J. (2020, May 29). *Predictive inference with the jackknife+*. arXiv: 1905.02928 [stat]. https://doi.org/10.48550/arXiv.1905.02928

Barber, R. F., Candes, E. J., Ramdas, A., & Tibshirani, R. J. (2023, March 16). *Conformal prediction beyond exchangeability*. arXiv: 2202.13415 [stat]. https://doi.org/10.48550/arXiv.2202.13415

Gibbs, I., & Candes, E. (2021). Adaptive Conformal Inference Under Distribution Shift. *Advances in Neural Information Processing Systems*, *34*, 1660–1672. Retrieved November 17, 2023, from https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html

Hoff, P. (2021, May 28). *Bayes-optimal prediction with frequentist coverage control*. arXiv: 2105.14045 [math, stat]. https://doi.org/10.48550/arXiv.2105.14045

Lei, J., Robins, J., & Wasserman, L. (2013). Distribution Free Prediction Sets. *Journal of the American Statistical Association*, *108*(501), 278–287. https://doi.org/10.1080/01621459.2012.751873

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019, October 29). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. arXiv: 1910.13461 [cs, stat]. https://doi.org/10.48550/arXiv.1910.13461

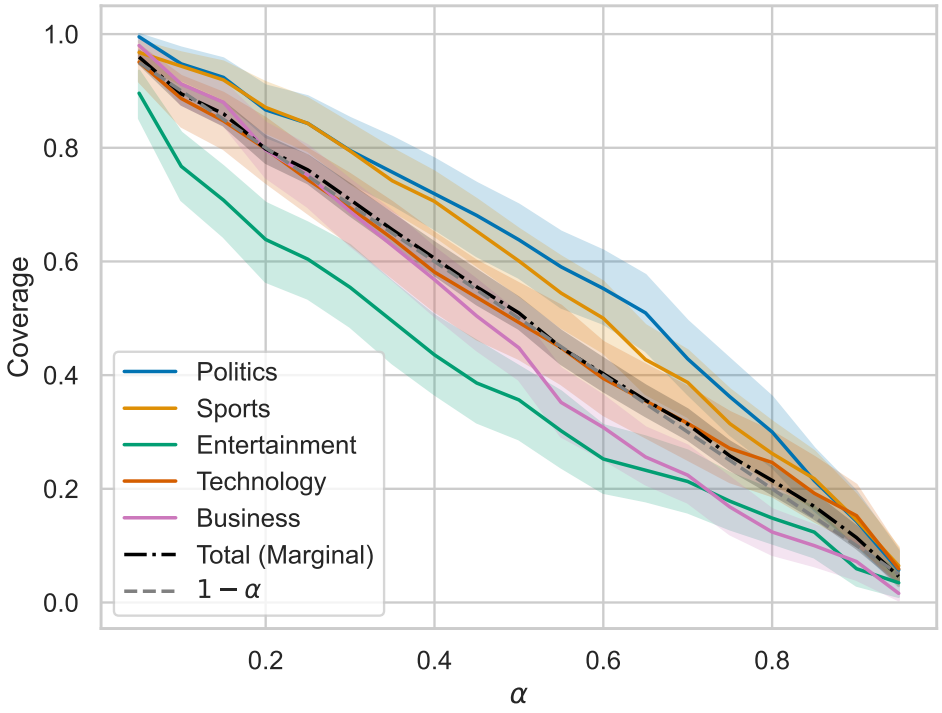
Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., & Barzilay, R. (2023, June 16). *Conformal Language Modeling*. arXiv: 2306.10193 [cs]. https://doi.org/10.48550/arXiv.2306.10193

Tibshirani, R. J., Barber, R. F., Candes, E. J., & Ramdas, A. (2020, July 6). *Conformal Prediction Under Covariate Shift*. arXiv: 1904.06019 [stat]. https://doi.org/10.48550/arXiv.1904.06019

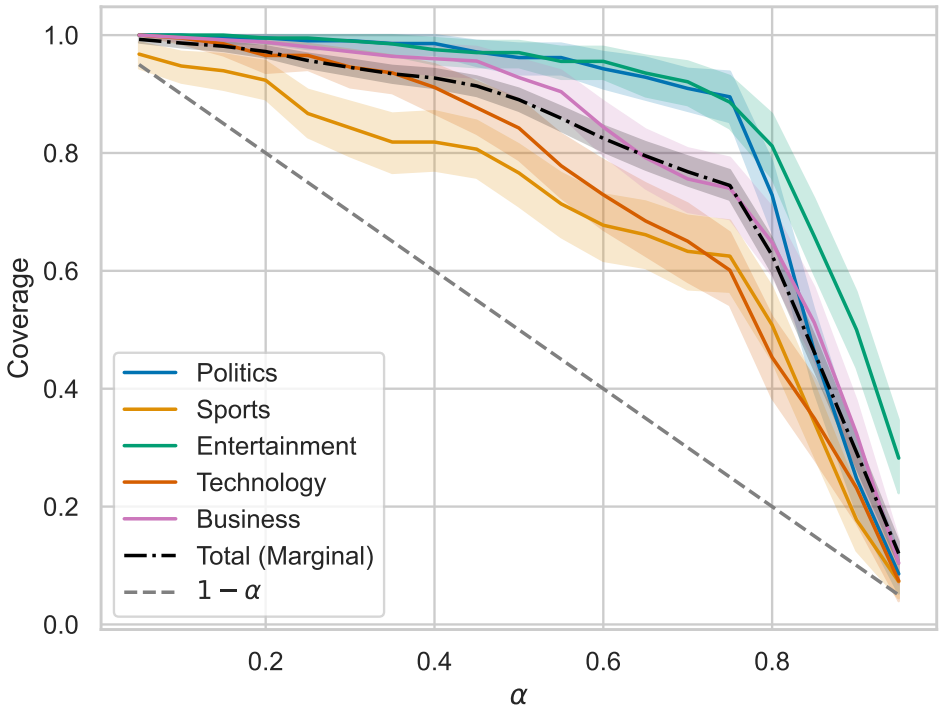
Vovk, V., Gammerman, A., & Saunders, C. (1999). Machine-Learning Applications of Algorithmic Randomness. *Proceedings of the Sixteenth International Conference on Machine Learning*, 444–453.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2019, October 9). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv.org. Retrieved December 12, 2023, from https://arxiv.org/abs/1910.03771v5

Yin, W., Hay, J., & Roth, D. (2019, August 31). *Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach*. arXiv: 1909.00161 [cs]. https://doi.org/10.48550/arXiv.1909.00161

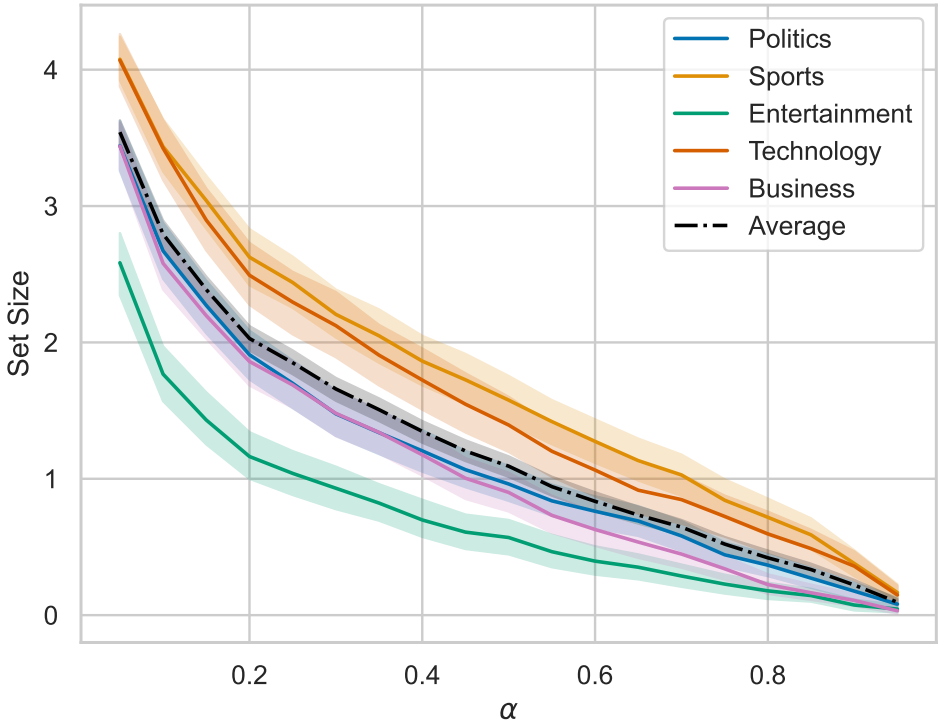


(a) Nonadaptive conformal prediction

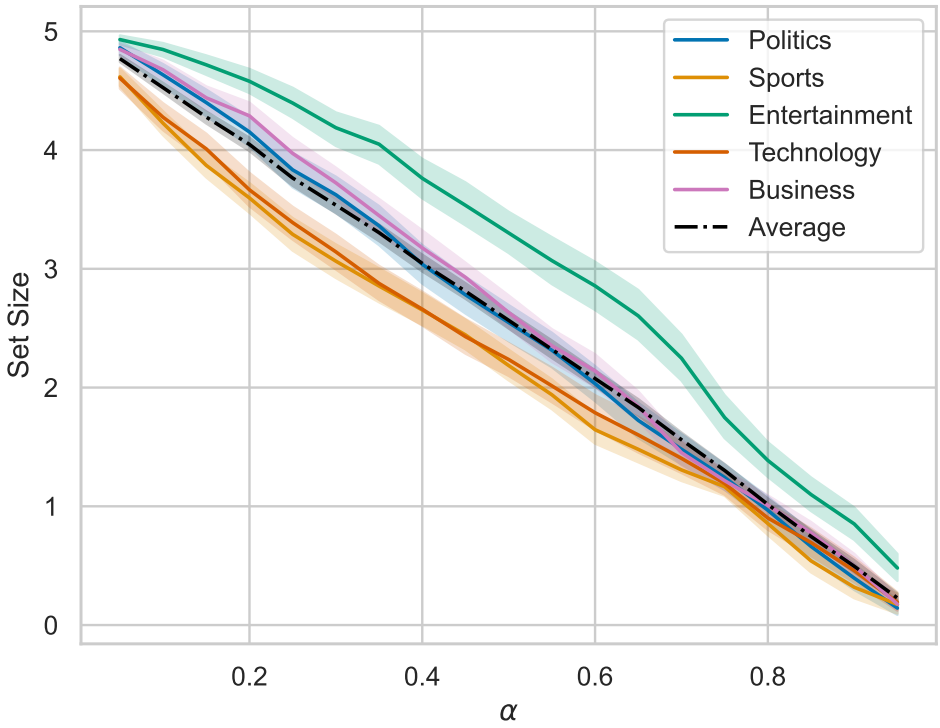


(b) Adaptive conformal prediction

Figure 1: Comparison of coverage for nonadaptive and adaptive conformal prediction for text classification. While both procedures have marginal coverage guarantees of $1 - \alpha$, the adaptive procedure has conditional coverage guarantees of $1 - \alpha$ for all groups. The nonadaptive procedure undercovers for some groups and overcovers for others.



(a) Nonadaptive conformal prediction



(b) Adaptive conformal prediction

Figure 2: Comparison of set size for nonadaptive and adaptive conformal prediction for text classification. The adaptive procedure guarantees coverage for the hardest groups by making larger prediction sets for those groups.