An Improved Method of Automated Nonparametric Content Analysis for Social Science*

Connor T. Jerzak† Gary King‡ Anton Strezhnev§

January 26, 2018

Abstract

Computer scientists and statisticians are often interested in classifying textual documents into chosen categories. Social scientists and others are often less interested in any one document and instead try to estimate the proportion falling in each category. The two existing types of techniques for estimating these category proportions are parametric "classify and count" methods and "direct" nonparametric estimation of category proportions without an individual classification step. Unfortunately, classify and count methods can sometimes be highly model dependent or generate more bias in the proportions as the percent correctly classified increases. Direct estimation avoids these problems, but can suffer when the meaning and usage of language is too similar across categories or too different between training and test sets. We develop an improved direct estimation approach without these problems by introducing continuously valued text features optimized for this problem, along with a form of matching adapted from the causal inference literature. We evaluate our approach in analyses of a diverse collection of 72 data sets, showing that it achieves substantially improved performance compared to existing approaches. As a companion to this paper, we offer easy-to-use software that implements all ideas discussed herein.

*Our thanks to Neal Beck, Aykut Firat, and Ying Lu for data and helpful comments.

†PhD Candidate and Carl J. Friedrich Fellow, 1737 Cambridge Street, Harvard University, Cambridge MA 02138, ConnorJerzak.com, cjerzak@g.harvard.edu.

‡Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; GaryKing.org, King@Harvard.edu, (617) 500-7570.

§PhD Candidate, 1737 Cambridge Street, Harvard University, Cambridge MA 02138, antonstrezhnev.com, astrezhnev@fas.harvard.edu.

1 Introduction

One of the defining characteristics of social science is a focus on population-level generalizations. We discover an interesting puzzle about one election, but try to develop theories that apply to many more. We are interested in the politics of one country, but attempt to understand it as an example of how all countries (or all democracies, or all developing countries, etc.) operate. We survey 1,500 Americans about their political attitudes, but we seek to understand how all Americans, or all people, form attitudes. Social scientists usually leave to journalists whether any one congressional speech supports a particular policy and focus instead on the percent of all speeches that support the policy.

Different emphases across disciplinary areas generate divergent methodological approaches. For example, computer scientists and statisticians often focus on the classification of an individual object (war, person, document, country, social media post, etc.) into a set of mutually exclusive and exhaustive categories. Social scientists use classification methods too, but their interest is more often on aggregate generalizations about populations of objects, such as the percent in each category, rather than any one individual classification, a task that is sometimes called "quantification".[1]

Applying a simple "Classify and Count" approach yields accurate category percentages under a perfect classifier. Perfect classifiers are unrealistic in real applications (Hand, 2006), but they are unnecessary for aggregate accuracy if individual-level errors cancel. Moreover, choosing a classifier by maximizing the percent correctly classified can sometimes drastically increase the bias of aggregate quantities. For example, the decision rule "war never occurs" accurately classifies country-year dyads into war/no war categories with over 99% accuracy, but is obviously misleading for social science research purposes.

Similarly, the proportion of email you receive that lands in your spam folder is a biased estimate of the percent of spam you receive overall because spam filters are tuned to

[1] Estimating category percentages, as opposed to individual classifications, is also of interest in epidemiology, where it is called "prevalence estimation." Interest in the technical area is also growing in computer science, machine learning, computational linguistics, and data mining, where it is variously called "quantification," "class prior estimation," "counting," "class probability re-estimation," and "learning of class balance." See Buck and Gart (1966), Esuli and Sebastiani (2015), Forman (2007), Kar et al. (2016), P. Levy and Kass (1970), Milli et al. (2013), and Tasche (2016) and the unification in Firat (2016).

the fact that people are more annoyed when they miss an important email than when some spam appears in their inbox. This is easy to fix by tuning your spam filter to avoid the bias, or correcting after the fact, but in most applications we do not know the classifier's bias. To be more specific, a method that classifies 60% of documents correctly into one of 8 categories might be judged successful and useful for classification. For example, a success rate of 60% might be considered quite good when using Google or Bing. However, because the individual category percentages still might be off by as much as 40 percentage points, the same classifier may be useless for some social science purposes.

The tasks of estimating category percentages (quantification) or classifying individual documents (classification) both begin by analyzing a small subset of documents with (usually hand-coded) category labels. Classification methods normally require these labeled and unlabeled document sets to be drawn from the same population, so the class probabilities can be calibrated. Commonly, however, the labeled set is created in one time period and a sequence of unlabeled sets are collected during subsequent time periods, each with potentially different distributions. For example, scholars may hand-label a set of social media posts about a presidential candidate into the 10 reasons people like or do not like this person. Then, for each day after the hand coding, a researcher may try to estimate the percent of social media posts in each of these categories. The methods of quantification we discuss here are designed to accommodate these situations even though these are the circumstances where the assumptions behind classification methods are violated.

We build on the only nonparametric method developed for estimating multi-category proportions that does not resort to individual classification as a first step. This methodology was developed in King and Lu (2008), with applications in public health, and in Hopkins and King (2010), with applications to text analysis in political science; and it was extended in King, Lu, and Shibuya (2010) and King, Pan, and Roberts (2013, Appendix B). A U.S. Patent has been issued for the technology (King, Hopkins, and Lu, 2012) and licensed by a university to a firm originally formed to implement an industrial strength version (Crimson Hexagon, Inc.). Over 1,600 scholarly articles in several scholarly fields have cited these works (according to Google scholar). The method as come

to be known by the name "readme," which is the widely-used open source software that implements it (Hopkins, King, Knowles, and Melendez, 2013).

We begin by developing the intuition behind readme's nonparametric methodology, and highlight situations where it can perform poorly. We then outline an approach for improving performance via two techniques, both of which involve better representing the meaning of the text. First, we develop an algorithm that chooses a feature space to discriminate between categories with as many non-redundant or independent features as possible. Unlike principal components analysis, independents component analysis, random projections, t-distributed stochastic neighborhood embeddings, or others designed for exploration, visualization, or classification, our approach is the first to generate a feature space optimized for quantification in a way that takes category information into account. And second, we overcome semantic change and changes in language use over time by adapting matching techniques developed in the causal inference literature.

We summarize the readme estimator and its key assumptions in Section 2. Through analytical and simulation evidence, we uncover the factors that affect the bias and variance of readme in Section 3. Section 4 then introduces our new methodology. In Section 5, we compare our approach to readme in out-of-sample empirical evaluations in 7,200 data sets, derived from subsets of 72 corpora (and repeated with 16 different evaluation protocols). We discuss what can go wrong and how to avoid it in Section 6. We do not claim that our approach will perform better than readme or other methods in all data sets; the well-known "ping pong theorem" shows this is impossible (i.e., every method includes enough tweakable options that any contest among these methods is usually won by the researcher who goes last; Hoadley 2001). Our more specific claim is that our approach will normally outperform other approaches in real data, under the real-world conditions we describe below. The results are encouraging. Section 7 concludes; mathematical proofs and evaluation protocols appear in the appendix.

2 Readme: Estimation without Classification

We now describe readme in a manner that conveys its logic, while also laying the groundwork for our subsequent improvements. The technique begins with a *text-to-numbers* step, that maps the entire set of textual documents into a numerical feature space. In the second *estimation* step, a statistical method is applied to the numerical summaries, to estimate the category proportions of interest. Our running example is textual documents, where humans hand code labels for documents by reading, but the methodology also applies to any other set of objects (such as people, deaths, attitudes, buildings, books, or many other things) for which the goal is estimating category proportions.

Notation Consider two sets of textual documents, L, *labeled* with a category number and N^{L} documents, and U, *unlabeled* with N^{U} documents, where $N = N^{\text{L}} + N^{\text{U}}$. When there is no ambiguity, we use i as a generic index for a document in either set and N as a generic description of either set size. Each document falls into category c in a set of mutually exclusive and exhaustive categories ($c \in \{1, \ldots, C\}$), but the category label is only observed in the labeled set. We write $D_i = c$ to denote that document i falls into category c. Denote $N_c^{\text{L}} = \sum_{i=1}^{N^{\text{L}}} \mathcal{I}(D_i = c)$ as the number of documents in category c in the labeled set, N_c^{U} as the (unobserved) number in c in the unlabeled set, and N_c generically for either set in category c.

The proportion of unlabeled documents in each category c is $\pi_c^{\text{U}} = \text{mean}_{i \in U}[\mathcal{I}(D_i = c)]$ (where for set A with cardinality $\#A$, the mean over i of function $g(i)$ is $\text{mean}_{i \in A}[g(i)] = \frac{1}{\#A} \sum_{i=1}^{\#A} g(i)$). The vector of proportions $\pi^{\text{U}} \equiv \{\pi_1^{\text{U}}, \ldots, \pi_C^{\text{U}}\}$, which represents our quantities of interest, forms a simplex, i.e. $\pi_c^{\text{U}} \in [0, 1]$ for all c and $\sum_{c=1}^{C} \pi_c^{\text{U}} = 1$. We also define the analogous (but observed) category proportions for the labeled set π^{L}.

Text to Numbers In this first step, we map the entire labeled and unlabeled corpora, with the document as the unit of analysis, into a constructed space of textual features. Many ways of performing this mapping can be created, and we propose a new one below optimized for quantification. For readme, Hopkins and King (2010) began with a set of

k unigrams, each a binary indicator for the presence (coded 1) or absence (0) of a chosen word or word stem in a document. The number of possible strings of these zeros and ones, called a *word stem profile*, is $W = 2^k$.

The readme approach then computes a W-length *feature vector* S^{L} by sorting the labeled documents into the W mutually exclusive and exhaustive word stem profiles, and computing the proportion of documents that fall in each. To make the definition of $S^{\text{L}} = \{S^{\text{L}}_w\}$ more precise and easier to generalize later, begin with the $N^{\text{L}} \times W$ *document-feature matrix* $F = \{F_{iw}\}$ with rows for documents, and columns for features which in this case are unique word stem profiles. Each element of this matrix, F_{iw}, is a binary indicator for whether document i is characterized by word stem profile w. Then elements of S^{L} are column means of F: $S^{\text{L}}_w = \text{mean}_{i \in L}(F_{iw})$. Then the same procedure is applied, with the same word stem profiles, to the unlabeled set, which we denote S^{U}.

We also define a W-length *conditional feature vector* as $X^{\text{L}}_c = \{X^{\text{L}}_{wc}\}$, which results from the application of the same procedure within category c in the labeled set, and $X^{\text{U}}_c = \{X^{\text{U}}_{wc}\}$ within category c in the unlabeled set (X^{U}_c is unobserved because c is unknown in the unlabeled set). These can be computed from F^c, a document-feature matrix representing only documents in category c. We also collect these vectors for all categories into two $W \times C$ matricies, $X^{\text{L}} = \{X^{\text{L}}_1, \ldots, X^{\text{L}}_C\}$ and $X^{\text{U}} = \{X^{\text{U}}_1, \ldots, X^{\text{U}}_C\}$, respectively.

Estimation Our goal is to estimate the vector of unlabeled set category proportions $\pi^{\text{U}} = \{\pi^{\text{U}}_1, \ldots, \pi^{\text{U}}_C\}$ given S^{L}, S^{U}, and X^{L}. Begin with an accounting identity (i.e., true by definition), $S^{\text{U}}_w = \sum_{c=1}^{C} X^{\text{U}}_{wc} \pi^{\text{U}}_c, \ \forall \ w$, or equivalently in matrix form:

$$S^{\text{U}} = X^{\text{U}} \pi^{\text{U}} \tag{1}$$

We can solve this expression for the quantity of interest as in linear regression, $\pi^{\text{U}} = (X^{\text{U}\prime} X^{\text{U}})^{-1} X^{\text{U}\prime} S^{\text{U}}$. However, π^{U} cannot be directly computed this way since we do not observe the "regressor" X^{U}. So instead readme estimates π^{U} by using X^{L}, which is observed in the labeled set and yields $\widehat{\pi^{\text{U}}} = (X^{\text{L}\prime} X^{\text{L}})^{-1} X^{\text{L}\prime} S^{\text{U}}$ (or a modified version of this expression that explicitly preserves the simplex constraint).

Readme works with the above estimator with any choice of k word stems. Hopkins and

King (2010) then randomly select many subsets of $k \approx 16$ word stems, run the algorithm for each, and average the results. This step is one way to reduce the dimensionality of the text. By averaging across word stem profiles, the variance of the final estimator is also reduced. We return to this step and improve it in Section 4.

Assumptions First, since the unlabeled conditional feature matrix X^{U} is unobserved, Hopkins and King (2010) assume $X^{\mathrm{U}} = X^{\mathrm{L}}$. However, this rigid assumption turns out not to be entirely necessary. Instead, we can get the same statistical result by expressing the labeled conditional feature matrix as an unbiased and consistent estimator of the unlabeled conditional feature matrix:

$$E(X^{\mathrm{L}}) = X^{\mathrm{U}}, \qquad \lim_{N^{\mathrm{L}} \to \infty} X^{\mathrm{L}} = X^{\mathrm{U}}. \tag{2}$$

(This assumption can be violated due to semantic change, which can happen if the labeled set is hand coded at one time, and the unlabeled set is collected at another time or in another place for which the meaning of language differs. We address semantic change in Section 4.) Assumption 2 about the *conditional* distribution of features and categories is considerably weaker than the assumption made by classifiers, which is that (a) the *joint* distribution of features and categories is the same in the labeled and unlabeled sets, (b) the measured features span the space of all predictors of D, and (c) the estimated model nests the true model as a special case (Hand, 2006). Because the correct model linking features to categories is unknown ex ante, this assumption is difficult to satisfy. On the contrary, readme does not need to assume a model for S since the relationship between the unconditional and conditional feature vectors follows directly from the laws of probability applied in Equation 1.

Second, to ensure $\widehat{\pi^{\mathrm{U}}}$ is uniquely defined, we must assume the matrix X^{L} is of full rank, which translates into (a) feature choices that lead to $W > C$ and (b) the lack of perfect collinearity among the columns of X^{L}. Assumption (a) is easy to control by generating a sufficient number of features from the text. Assumption (b) is only violated if the feature distributions in documents across different categories are identical, which is unlikely with a sufficient number of coded documents. (We prove in Section 3 that

high collinearity, which can result if categories are weakly connected to the features or documents are labeled with error, can exacerbate the bias of the readme estimator, which provides an important clue to generate improvements.)

3 Statistical Properties

We now analyze the statistical properties of the readme estimator. Our goal is to understand the situations when readme performs poorly so we can design improvements (in Section 4). We show here, through analytical calculations (Section 3.1) and simulations (Section 3.2), that three situations can degrade readme performance.

First is *semantic change*, which is the difference in the meaning of language between the labeled and unlabeled sets. Authors and speakers frequently morph the semantic content of their prose to be clever, get attention, be expressive, curry political favor, evade detection, persuade, or rally the masses. For these or other purposes, the content, form, style, and meaning of every symbol, object, or action in human language can always be contested. We address two types of semantic change that impact readme: *emergent discourse*, where new words and phrases, or the meanings of existing words and phrases, appear in the unlabeled set but not the labeled set, and *vanishing discourse*, where the words, phrases, and their meanings exist in the labeled set but not the unlabeled set. Russian election hacking is an example of emergent discourse, language which did not exist a few years ago, whereas Russian Communism is an example of vanishing discourse, with language that has largely vanished from ongoing conversations over time. However, emergent and vanishing discourse can reverse their meanings if the researcher swaps which set is labeled. For example, in analyzing a large historical data set, a researcher may find it more convenient to read and label a contemporary data set and infer to the historical data sets (say as they are recovered from an archive); to label documents at the start of the period and infer to subsequent periods; or to code a sample spread throughout the period and to infer to the full data set. Either vanishing or emergent discourse can bias readme, but only if the change affects the categories differently. We show how to reduce bias due to vanishing discourse in Section 4.2.

Second is the *lack of textual discrimination*, where the language used in documents falling in different categories is not clearly distinguishable. This problem may arise because the conceptual ideas underlying the chosen categories are not distinct. Hand coding errors can also lead to this problem, which is commonly revealed by low levels of inter-coder reliability. Lack of textual discrimination among categories can also occur because of heterogeneity in how authors express category-related information or a divergence between how authors of the documents express this information and how the analyst conceptualizes the categories.

We have also seen many data sets where the analyst begins with distinct and well-defined conceptual definitions for the set of C categories, with examples of documents that fall unambiguously into each one, but where it turns out upon large-scale coding that large numbers of documents can only be described as falling into multiple categories. Adding categories to represent these more complicated expressions (so that the resulting set is still mutually exclusive and exhaustive) is a logical solution, but this step often leads to a more cognitively demanding hand coding problem that results in even lower levels of inter-coder reliability.

A final problematic situation for readme occurs due to interactions with the other two problems. This issue is *proportion divergence*, when the category proportions in the labeled set (π^{L}) diverge from those in the unlabeled set (π^{U}). To understand why category proportion divergence can be an issue, consider a data set with massive semantic change and no textual discrimination — so the document texts are largely uninformative — but where $E(\pi^{\mathrm{L}}) = \pi^{\mathrm{U}}$, such as occurs when the labeled set is a random sample from the test set. In this situation, readme will return the observed proportion vector in the labeled set, π^{L}, which is an unbiased estimate of π^{U}. This means that we can sometimes protect ourselves from semantic change and the lack of textual discrimination by selecting a labeled set with a similar set of category proportions as the unlabeled set. However, this protective measure is virtually impossible to put into practice, as it requires *a priori* knowledge of category membership.

The rest of this section provides the analytical and simulation evidence to support,

clarify, and further articulate these three situations where readme can be improved.

3.1 Analytical Results

In the classic errors-in-variables linear regression model, with random measurement error only in the explanatory variables, least squares is biased and inconsistent. Under Assumption 2, readme is also a linear regression with random measurement error in the explanatory variables. However, the readme regression is computed in the space of features, the size of which remains constant as the number of observations grows. As such, readme *is* statistically consistent: as we gather and code more documents for the labeled set (and keep W fixed, or at least growing slower than n), its estimator converges to the truth: $\lim_{N^{\mathrm{L}}\to\infty}\widehat{\pi^{\mathrm{U}}} = \lim_{N^{\mathrm{L}}\to\infty}(X^{\mathrm{L}\prime}X^{\mathrm{L}})^{-1}X^{\mathrm{L}\prime}S^{\mathrm{U}} = (X^{\mathrm{U}\prime}X^{\mathrm{U}})^{-1}X^{\mathrm{U}\prime}S^{\mathrm{U}} = \pi^{\mathrm{U}}$. This is a useful result suggesting that, unlike classic errors-in-variables, labeling more observations can reduce bias and variance. It also suggests that, to improve readme, we should focus on finite sample bias rather than consistency results, which are already clear.

To analyze readme's bias in a way that will provide useful intuition, consider a simplified case with only two categories. Because of the simplex constraint, the unlabeled set category proportions can be characterized by a single parameter, π_1^{U}, and the accounting identity for each feature mean w, S_w^{U}, can be written simply as:

$$S_w^{\mathrm{U}} = X_{w2}^{\mathrm{U}} + B_w^{\mathrm{U}}\pi_1^{\mathrm{U}} \tag{3}$$

where $B_w^{\mathrm{U}} = X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}}$. The readme estimator is then the least-squares estimator of π_1^{U}, which we write as follows. (Proofs of all propositions appear in Appendix A.)

Proposition 1. *Two-category readme estimator is*

$$\widehat{\pi_1^U} = \frac{\sum_{w=1}^{W} B_w^L\,(S_w^U - X_{w2}^L)}{\sum_{w=1}^{W}(B_w^L)^2}, \tag{4}$$

where $B_w^L = X_{w1}^L - X_{w2}^L$.

If $X^{\mathrm{L}} = X^{\mathrm{U}}$, the above expression equals π_1^{U} and readme is unbiased. However, due to sampling error, the realized sample value of X^{L} may differ from the unobserved true value X^{U}. By Assumption 2, $X_{wc}^{\mathrm{L}} = X_{wc}^{\mathrm{U}} + \epsilon_{wc}$ where ϵ_{wc} is a random variable with

mean zero and variance inversely proportional to N_c. This enables us to write the readme estimator in terms of X^{U}, the true unlabeled set category proportion π_1^{U}, and the sample category size N_c^{L}. Taking the expectation of this quantity yields:

Proposition 2. *The expected value of the two-category readme estimator is*

$$\mathbb{E}\left[\widehat{\pi_1^U}\right] = E\left[\frac{\sum_{w=1}^{W}\left(B_w^U + \nu_w\right)B_w^U}{\sum_{w=1}^{W}\left(B_w^U + \nu_w\right)^2}\right]\pi_1^U - E\left[\frac{\sum_{w=1}^{W}\left(B_w^U + \nu_w\right)\epsilon_{w2}}{\sum_{w=1}^{W}\left(B_w^U + \nu_w\right)^2}\right]. \tag{5}$$

where $B_w^U = X_{w1}^U - X_{w2}^U$ *and* $\nu_w = \epsilon_{w1} - \epsilon_{w2}$.

The consistency property of readme can be seen here: As the error in measuring X^{U} with X^{L} goes to 0 or N^{L} goes to infinity, the second term in the expectation is 0 (because $\epsilon_{w2} \propto 1/N_C^{\mathrm{L}} \to 0$), while the first converges to π_1^{U}. In the presence of measurement error, the bias is a function of two components of the lack of textual discrimination — the difference in the true category proportions, $B_w^{\mathrm{U}} = X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}}$ and the combined error variance $\nu_w = \epsilon_{w1} - \epsilon_{w2}$.

We obtain further intuition via an approximation using a first-order Taylor polynomial:

Proposition 3. *The approximate bias of the readme estimator is*

$$\textit{Bias}\left(\widehat{\pi_1^U}\right) \approx \frac{\sum_{w=1}^{W}\left[\textit{Var}(\epsilon_{w2}) - \textit{Cov}(\epsilon_{w1},\epsilon_{w2})\right]\left(1-\pi_1^U\right) - \left[\textit{Var}(\epsilon_{w1}) - \textit{Cov}(\epsilon_{w1},\epsilon_{w2})\right]\pi_1^U}{\sum_{w=1}^{W}\left[(B_w^U)^2 + \textit{Var}(\nu_w)\right]}. \tag{6}$$

This expression suggests four insights. First, as the systematic component of textual discrimination, B_w^{U}, increases relative to the variance of the error terms, ϵ_{w1} and ϵ_{w2}, the bias approaches 0. In other words, readme works better when the language of the documents across categories is distinct.

Second, adding more informative numerical representations of the text, so that W increases (but with a fixed n), has an indeterminate impact on the bias. While more informative numerical summaries of the text can increase the sum in the denominator, they may increase the overall bias if the result is an error variance that is high relative to the discriminatory power.

Third, we confirm the intuition that larger labeled sets within each category are better: Since the elements of X^{L} are simple means across documents assumed to be independent,

the variance of the measurement error terms is simply $\mathrm{V}(\epsilon_{wc}) = \sigma^2_{wc}/N^{\mathrm{L}}_c$, which decline as the labeled set category sizes increase.

Finally, we simplify by studying the special case of independence of measurement errors across categories (i.e. $\mathrm{Cov}(\epsilon_{w1}, \epsilon_{w2}) = 0$). In this situation, readme bias is minimized when the following relationship holds between the labeled and unlabeled set category proportions:

Proposition 4. *When measurement errors are independent across categories, the bias of readme is minimized at*

$$\pi^L_1 = \frac{\pi^U_1 \sum_{w=1}^{W} \sigma^2_{w1}}{\pi^U_1 \sum_{w=1}^{W} \sigma^2_{w1} + (1 - \pi^U_1) \sum_{w=1}^{W} \sigma^2_{w2}} \tag{7}$$

What this means is that when measurement error variances are roughly equivalent across categories, the bias of readme is smallest when category proportion divergence is smallest.

We will use the intuition from each of these results in Section 4.

3.2 Simulation Results

We have found that simulations with large values of C and W generate more complexity without much additional insight, and so for expository purposes in this section we set $C = 2$ and $W = 2$. We then evaluate readme performance as we vary the degree of semantic change, textual discrimination, and proportion divergence. Our improvements to readme exploit the relationship between these three quantities.

For clarity, we divide textual discrimination into two separate concepts that, in the readme regression, are related to the minimal requirement in least squares that X be of full rank and the goal, for reducing variance and model dependence, that X be as far as possible from degeneracy. Since X represents categories as variables and features of the text as rows, we refer to these two variables as *category distinctiveness* and *feature distinctiveness*, respectively. For the purpose of our simulations, define the absolute differences between categories as $b_w = |X_{wc} - X_{wc'}|$ for row $w = 1, 2$. Then, we measure category distinctiveness as $(b_1 + b_2)/2$. Feature distinctiveness, which we measure as $(b_1 - b_2)/2$, is the category distinctiveness between rows. (We show how each definition generalizes with $C > 2$ and $W > 2$ in Section 4.1.)

We create variation in these factors for simulation as follows. First, we control proportion divergence by drawing π^{L} and π^{U} from i.i.d. Dirichlet distributions with concentration parameters set to 2. (In our figures, we measure average proportion divergence as $(|\pi_1^{\mathrm{L}}-\pi_1^{\mathrm{U}}|+|\pi_2^{\mathrm{L}}-\pi_2^{\mathrm{U}}|)/2$.) We sample X_{wc}^{U} from an i.i.d. Normal with mean 0 and variance $1/9$. Then, we generate $X_{wc}^{\mathrm{L}} = X_{wc}^{\mathrm{U}} + \epsilon$, where $\epsilon = 0$ or, to simulate semantic change, from a Normal with a mean at 0 and standard deviation equal to $|X_{wc}^{\mathrm{U}}|$. We then treat these parameters as fixed and, to simulate measurement error in the calculation of X^L, generate 5,000 repeated sample data sets from each set of parameters, apply the readme estimator, and estimate the mean over simulations of the sum of the absolute errors over categories (SAE).[2] To generate each of the 5,000 sampled data sets, we randomly generate document-level features from Normal densities by adding a draw from a standard Normal to each cell value of X_{wc}^{L} and X_{wc}^{U}.

Figure 1 illustrates how the SAE behaves as a function of category distinctiveness (vertically) and proportion divergence (horizontally). SAE is coded in colors from low (red) to high (yellow). The top left of the figure is where readme performance is best: where proportion divergence is low and category distinctiveness is high. When the language is clearly distinguishable among categories, readme can overcome even large divergences between the labeled and unlabeled sets. Without high levels of textual discrimination, readme then becomes vulnerable to high levels of proportion divergence. Category distinctiveness and proportion divergence appear to have roughly the same relative importance, as the contour lines in Figure 1 fall at approximately 45° angles.

Figure 2 studies textual discrimination further by illustrating how category distinctiveness (horizontally) and feature distinctiveness (vertically) jointly impact SAE. If we hold feature distinctiveness fixed, increased category distinctiveness improves performance; if we hold category distinctiveness fixed, greater feature distinctiveness similarly leads to better performance. Of the two, feature distinctiveness is somewhat more predictive of performance, but both can be important.

Finally, Figure 3 illustrates how the relationship between feature distinctiveness (three

[2]We use the SAE to make our analysis consistent across datasets of diverse category number. We have found that simple attempts to normalize, such as dividing by the number of categories, tends to weaken this comparability, especially because in all cases the target quantity is located on the simplex.

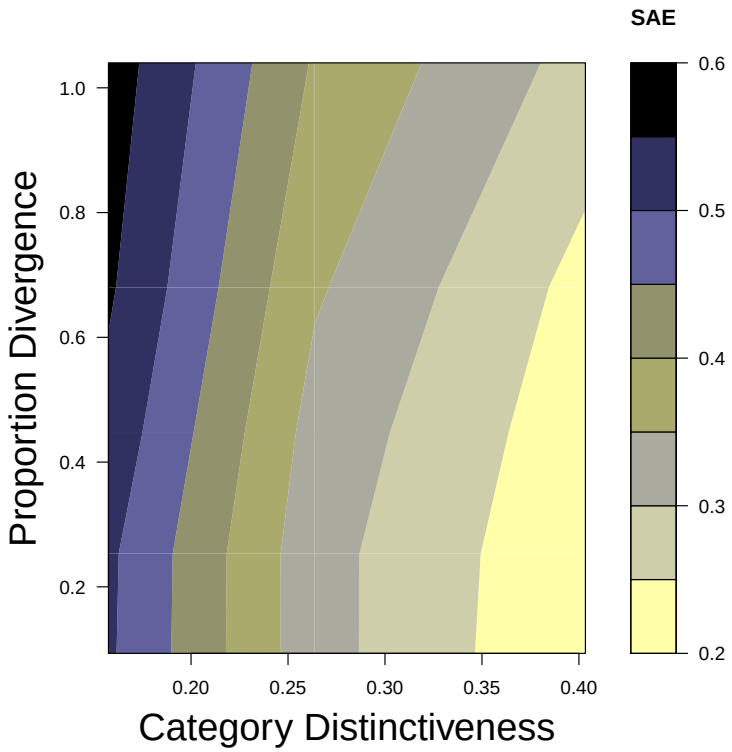

Figure 1: Category distinctiveness is plotted (horizontally) by proportion divergence between the labeled and unlabeled sets is plotted (vertically), with mean absolute sum of errors color coded (with yellow in the lower right corner best).

separate lines in each panel) and proportion divergence (horizontal axis) is mediated by the presence of semantic change (difference between the panels). Without semantic change (left panel), highly distinctive features greatly reduce SAE (which can be seen by the wide separation among the lines on the graph). In contrast, in the presence of semantic change (in this case we moved the mean of $E(X^{\mathrm{L}})$ by a quarter of a standard deviation from X^{U}), more distinctive features still tend to outperform less distinctive features, but the difference is less pronounced. With semantic change, distinctive features in the labeled set may no longer be distinctive in the unlabeled set.

4 Improvements

Section 3 offers analytical and simulation results to show how proportion divergence, textual discrimination (including category and feature distinctiveness), and semantic change

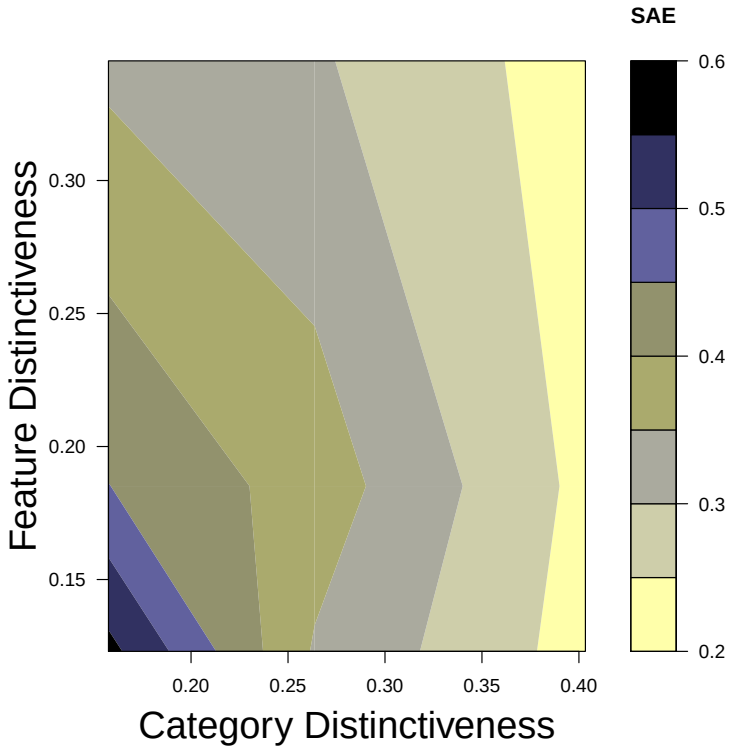

Figure 2: Category distinctiveness is plotted (horizontally) and feature distinctiveness (vertically), with mean absolute sum of errors color coded (with yellow indicating best performance).

impact the performance of readme. We now use these insights to develop an improved method.

4.1 Improving Textual Discrimination

Since numerical representations of text can have an outsized impact on estimation (see Denny and Spirling, 2016; O. Levy, Goldberg, and Dagan, 2015), we begin by designing a text-to-numbers algorithm optimized for our purposes. Some existing text-to-numbers algorithms reduce the complexity of language without using information in the category labels, such as readme's procedure of *randomly* drawing k word stems (or procedures like principal component analysis). Other text-to-numbers algorithms have been created to improve the performance of classifiers (e.g., Brunzell and Eriksson, 2000; Vincent et al., 2010), but no existing text-to-numbers algorithm we are aware of has been designed to optimize for direct estimation of category proportions. We thus now introduce the first

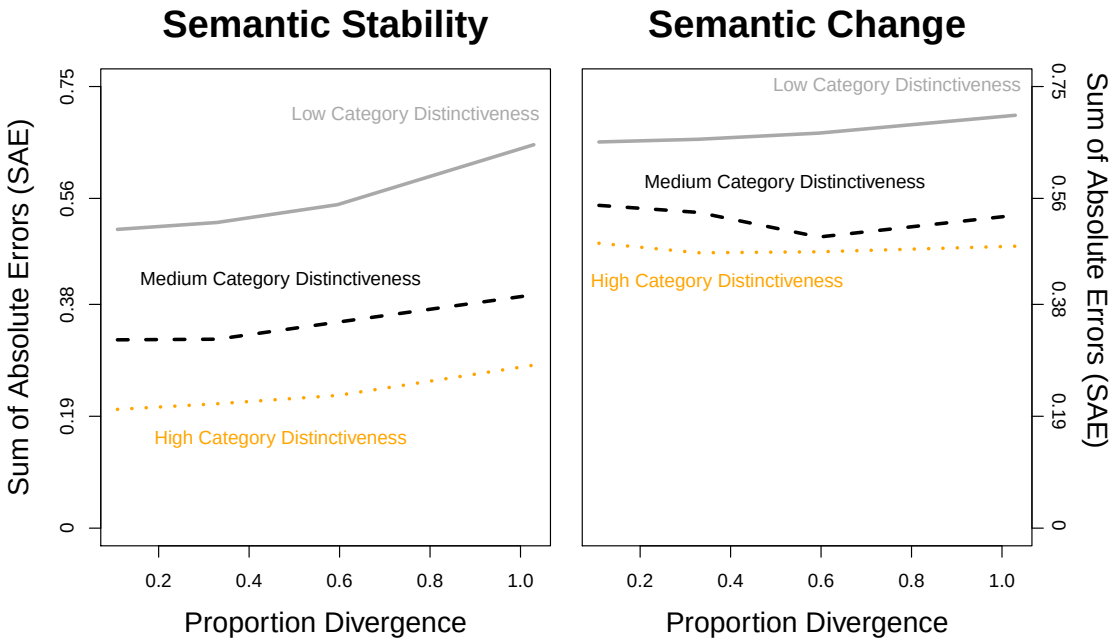

Figure 3: Proportion divergence is plotted (horizontally) by the mean of the sum of the absolute error (vertically) for different levels of category distinctiveness (separate lines) and for the absence (left panel) and presence (right panel) of semantic change.

such algorithm.

First, we replace the discrete feature space in readme with a more informative, continuous feature space, using tools developed since the publication of Hopkins and King (2010). The specific procedure we find that works well is to substitute each term in a document with its global word vector representation ("GloVe") estimated from a pre-trained corpus of 2 billion Twitter posts. This projects each word to a common vector space. We find that 50-dimensional word vectors works well (Pennington, Socher, and Manning, 2014). We first replace each word in each of our documents with its vector representation, and then summarize each document with $W = 150$ features, representing the minimum, maximum, and median values over words of each of the 50-dimensional word vectors. This text-to-numbers summary produces a more informative F matrix, which we now seek to improve further. (The choice of the three summary statistics is useful but also somewhat arbitrary and so could potentially be optimized further. However, we have done experiments incorporating convolutions into our algorithm, but have so far found no

improvements from that approach.)

Second, we project this raw $N \times 150$ document-feature matrix F onto a custom built $N \times W'$ (for $W' \ll 150$) lower dimensional subspace matrix $\underset{\sim}{F}$. We do this through a projection of the form $\underset{\sim}{F}_{N \times W'} = \phi(F \cdot \Gamma)$, with Γ being $150 \times W'$ and ϕ being a (possibly non-linear) transformation function. We can take conditional expectations as before to generate $\underset{\sim}{X}$, which is comparaible to X but now the features that populate the rows of this matrix are now taken from $\underset{\sim}{F}$ instead of F. It is important to note that the objective function for calculating Γ is novel, and no closed-form expression for the optimum is known (unlike in, say, Principle Components Analysis). Instead, we extract the gradients using automatic differentiation and then use a form of stochastic gradient descent optimization (Kingma and Ba, 2015; Martin Abadi et al., 2015) to obtain Γ. If $W' = 10$, there are 1500 parameters to estimate, and gradient-level information is important for obtaining desirable Γ and $\underset{\sim}{F}$ matrices.

While optimizing, it is important to "tie our hands" in order to avoid researcher-induced bias. If we generated X^L by taking into account information about the resulting S, we could generate essentially any $\hat{\pi}^U$ because we would be choosing both the left and right side of the readme regression. Thus, we perform this optimization while adhering to what we will call the *Tied Hands Principle* (THP). This principle, which we formalize here, appears to have many other applications in the statistical literature and has been stated informally in a variety ways.

Tied Hands Principle (THP). *Let t denote a function that transforms data objects A and B, into $A^* = t(A, Z)$ and $B^* = t(B, Z)$, by matching or weighting data subsets (rows), or transforming or selecting features (columns), where Z denotes exogenous information such that $p(A, B|Z) = p(A, B)$. Denote as $T_{A,Z}$ the set of all functions of A and Z (but not B) and $T_{B,Z}$ the set of all functions of B and Z (but not A). Then define $g(j|k)$ as a function to choose an element from set j using information in k. Consider statistical procedures that are functions of both A^* and B^*. Then, for $D = A$ or $D = B$, THP requires that the transformation function t be chosen such that $t = g(T_{D,Z}|D, Z)$ but not $t = g(T_{D,Z}|A, B, Z)$.*

The special case of the THP for what we will call "readme2" prohibits finding Γ^* by minimizing $f(\Gamma,\ L,\ U)$. A different special case of the THP is commonly invoked in causal inference, where the matching of treated and control observations is performed without being allowed to take into account the response variable: the observation weights are calculated explicitly without taking into account the outcome variable at all, or only in the alternative treatment class. Another special case in causal inference allows one, in prospective designs, to select observations conditional on the explanatory variables, but not the outcome variable or, in retrospective (case-control) designs, based on the outcome variable but not on the explanatory variables.

Provided THP is followed, we can safely choose a projection matrix, Γ, that maximizes both category distinctiveness (CD) and feature distinctiveness (FD) in the labeled set, the lack of which were shown in Section 4 to be weaknesses of readme. We define these criteria in their general form as

$$\text{CD}(\Gamma) \propto \sum_{c<c'} \sum_{w=1}^{W'} \left| \underline{X}^{\text{L}}_{wc} - \underline{X}^{\text{L}}_{wc'} \right|. \tag{8}$$

and

$$\text{FD}(\Gamma) \propto \sum_{c<c'} \sum_{w'<w} \left| \left| \underline{X}^{\text{L}}_{wc} - \underline{X}^{\text{L}}_{wc'} \right| - \left| \underline{X}^{\text{L}}_{w'c} - \underline{X}^{\text{L}}_{w'c'} \right| \right|.$$

where the inequalities in the summations prevent double-counting. We then choose Γ through the following optimization:

$$\Gamma^* = \underset{\Gamma \in \mathbb{R}^{W \times W'}}{\arg\max} \ \ \lambda \cdot \text{CD}(\Gamma) + (1-\lambda) \cdot \text{FD}(\Gamma)$$

for some $\lambda \in [0, 1]$. In our experiments, we trade the two forms of distinctiveness equally ($\lambda = 0.5$), although this could be optimized further through cross-validation or other means for some potential additional performance improvement.[3]

Optimizing based on both criteria simultaneously is crucial. Optimizing Γ for category distinctiveness alone would lead to high variance through high collinearity in $\underline{X}$, and

[3]To prevent overfitting, we use dropout and set $\phi(x) = x/(|x| + 1)$. We also normalize the columns of $\underline{F}$ to keep the projections on the same scale (with a mean 0 and standard deviation of 1). In our experiments, we set $W' = 10$. Although this choice could be determined via cross-validation, our experiments indicate that performance does not greatly depend on this parameter.

optimizing Γ for feature distinctiveness alone would lead to higher bias low category distinctiveness. Optimizing both together, as we do, reduces the overall error rate.

We illustrate this point with an analysis of data comprised of 1,426 emails drawn from the broader Enron Corporation email corpus made public during the Federal Energy Regulatory Commission's investigation into the firm's bankruptcy (the data and codebook are available at `j.mp/enronData` and `j.mp/EnronCodeBK`). These emails, also analyzed in Hopkins and King (2010) and below in Section 5, were hand coded by researchers into five broad topics: company business, personal communications, logistics arrangements, employment arrangements, and document editing.

For expository clarity, we set $W' = 2$ and choose Γ by first maximizing the category distinctiveness metric alone. We offer a scatterplot of the resulting projections, $\underset{\sim}{F}$, in the left panel of Figure 4, with different colors and symbols to represent the five categories. This panel reveals that these features do indeed discriminate between the categories (which can be seen by separation between the different colored symbols). However, as is also apparent, the two dimensions are highly correlated which, as in linear regression, would lead to higher variance estimates. In linear regression, given a fixed sample size, collinearity is an immutable fact of the fixed data set and specification; in contrast, in our application operating in the space of the features that we construct rather than the data we are given, we can change the projections, the space in which the regression is operating, and therefore the level of collinearity. As a second illustration, we again set $W' = 2$ but now optimize Γ by maximizing only feature distinctiveness. In this case, as can be seen in the middle panel of Figure 4, the columns of $\underset{\sim}{X}^{\mathrm{L}}$ are uncorrelated but unfortunately do not discriminate between categories well (as can be seen by the points with different colors and symbols overlapping).

Thus, we implement our metric, optimizing the sum of both category and feature distinctiveness, which we present in the right panel of Figure 4. This result is well calibrated for estimating category proportions: The dimensions are discriminatory and thus bias reducing, which can be seen by the color separation, but still uncorrelated, and thus variance reducing. After matching on these features and performing the least squares regression

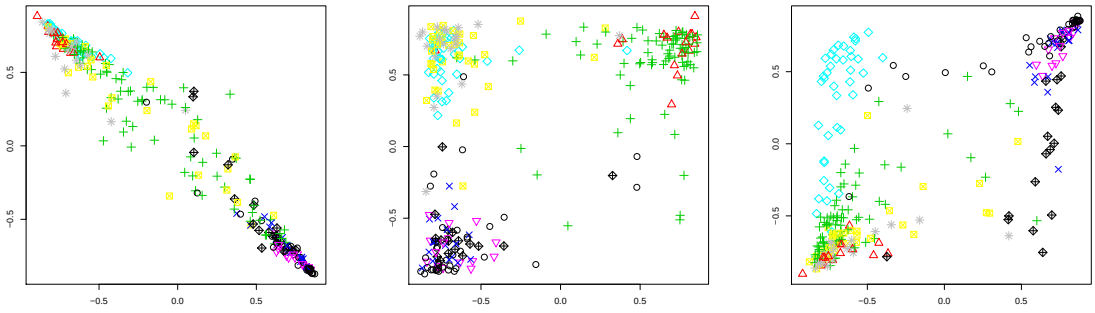

Figure 4: Optimizing Γ by category distinctiveness only (left panel), feature distinctivenss only (middle), and both (right panel). Each point is a document, with categories coded with a unique color and symbol. The axes (or components) are columns of tbe constructed $\phi(F \times \Gamma)$.

in these Enron data, we find the sum of the absolute residuals in estimating π^{U} of 0.30, compared to 0.55 for the original readme, a substantial improvement.

4.2 Overcoming Semantic Change and Proportion Divergence

Given the results of Section 4.1, we could estimate π^{U} by applying the readme regression $(\underline{X}^{\mathrm{L}\prime}\underline{X}^{\mathrm{L}})^{-1}\underline{X}^{\mathrm{L}\prime}\underline{S}^{\mathrm{U}}$, with $\underline{X}^{\mathrm{L}}$ constructed from our newly optimized $\underline{F}$ matrix. However, we have found a way to further alter the space F (and therefore $\underline{X}^{\mathrm{L}}$), in a manner that reduces proportion divergence and any biasing effect of the vanishing discourse component of semantic change. As a result, we depend much less on the veracity of Assumption 2.

To do this, we borrow the idea of matching from the causal inference literature, and attempt to reduce the "imbalance" (a term of art from that literature) between the labeled and unlabeled sets (Ho, Imai, King, and Stuart, 2007; Iacus, King, and Porro, 2012). To keep the quantity of interest the same, while pruning irrelevant observations, matching for causal inference fixes the treated group and estimates the average treatment effect on the treated. In our case, we fix the unlabeled set and our quantity of interest, and selectively prune the labeled set. This enables us to remove much of the biasing effect of vanishing discourse and to simultaneously reduce proportion divergence. We thus use the labeled set to construct a matched (sub)set, $\mathcal{M}$, that more closely resembles the unlabeled set.

For a matching algorithm, we take each document in the unlabeled set, identify the three nearest neighbors in the labeled set (defined in the Euclidean space), and any further

identify other labeled set documents closer than the median nearest neighbor among the sets of the top three. Any labeled set documents not matched by these rules are pruned out and not used further. This act of pruning is what makes matching work in causal inference and, for our problem, reduces semantic change and proportion divergence. We then recompute $\underset{\sim}{F}$ and the matched $\underset{\sim}{X}^{\mathrm{L}}$, which we denote $\underset{\sim}{X}^{L_{\mathcal{M}}}$, and apply the readme regression.

This pruning to the matched set change means that we now do not need to satisfy the assumption in Equation 2 and instead only need to satisfy this substantially weaker version:

$$E[X^{\mathrm{L}_{\mathcal{M}}}] = X^{\mathrm{U}}, \qquad \lim_{N^{\mathrm{L}} \to \infty} X^{\mathrm{L}_{\mathcal{M}}} = X^{\mathrm{U}}. \tag{9}$$

Empirically, we find that matching indeed has the desired effects: in the 72 real-world data sets we introduce in Section 5, matching alone reduces the divergence between $\underset{\sim}{X}^{\mathrm{L}}$ and $\underset{\sim}{X}^{\mathrm{U}}$ in 99.6% of cases and on average by 19.8%. Proportion divergence, which is not observed in real applications but which we can measure because we have coded unlabeled sets for evaluation, has reduced by 83.2% of cases, and by on average 25%. We now turn to the details of these data, and the consequence of using all parts of our new methods, for mean square error in the quantities of interest.

5 Evaluation

Design We performed sixteen large-scale evaluations of our methodology, each following a different protocol. For each protocol, we estimate readme2 and 32 alternative statistical methods that can be used to estimate category proportions (including readme). For each method, we analyzed 7,200 data sets (derived from 72 corpora), comprising a total of $(7,200 \times 16 =)$ 115,200 evaluations. Analyses of each of our our sixteen protocols each yield similar conclusions and so, in this section, we detail the one protocol designed to be as close as possible to real applications (which we refer to as "empirical"). All sixteen protocols are listed in Appendix B, with results summarized here.

The 32 alternative methods of estimating category proportions are of five types. The first four types comprise six classifiers each run within each of the four possible com-

binations of (a) a discrete or continuous feature space and (b) a classification of whole documents and counting or averaging continuous probability estimates to yield estimates of the category proportions. The six classifiers include support vector machines, random forests, Naive Bayes, and L1- and L2-regularized multinomial regression (using standard settings and described in James, Witten, Hastie, and Tibshirani, 2013), and an ensemble of these classifiers based on an average of classifiers within each of the two cells of (b). The fifth type of alternative method includes 8 methods that do not require classification as a first step. Among these, only readme is designed for more than two categories. We adapt the remaining 7 — Friedman, Adjusted Counts, HDX, Median Sweep, Mixture HPMF, Mixture L1, and Mixture L2 (each detailed in Firat 2016) — to multiple categories via estimation of repeated dichtomizations of the set of categories.

Each of the 7,200 datasets we analyze, constructed as a subset of one of 72 corpora, has a labeled out-of-sample test set that plays the role of the unlabeled set, except that we are able to use its labels after estimation to evaluate performance. The 72 corpora include three used in Hopkins and King (2010): The Enron email data set described in Section 4.1; a set of 462 newspaper editorials about immigration (with 3,618 word stems and 5 categories); and a set with 1,938 blog posts about candidate Hillary Clinton from the 2008 presidential election (with 3,623 word stems and 7 categories). Finally, we also include 69 separate Twitter data sets, each created by a different political candidate, private company, nonprofit, or government agency for their own business purposes, covering different time frames and categories (see Firat, 2016); these data cover 150–4,200 word stems, 3–12 categories, and 700–4,000 tweets.

All documents in each of the 72 corpora is labeled with a time stamp. For each, we randomly select a time point and pick the previous 300 documents as the labeled set and the next 300 documents as the out-of-sample evaluation set (wrapping in time if necessary). For each corpora, we repeat this process 100 times, yielding 7,200 data sets in total. This procedure keeps the evaluation highly realistic while also ensuring that we have many types of data sets with variation in proportion divergence, textual discrimination, and semantic change.

Results We present results across our numerous evaluations in three ways.

First, Figure 5 compares the performance of readme2 to the 32 alternative methods (for our "empirical" evaluation). For each method, we compute the proportion of data sets with higher error than readme2 vertically by the proportion divergence in quantiles horizontally. Our new approach outperforms the best classifier (in these data, the regularized multinomial regression run in the continuous feature space) in 84% of data sets and the average classifier in the continuous space in more than 90% of corpora. Our approach outperforms the best discrete classifier in over 90% of corpora as well. Most of the 32 methods are outperformed by readme2 in 100% of the cases, as indicated by appearing at the top of the graph. Relative performance remains excellent across the different levels of category proportion divergence between labeled and unlabeled sets. The new method's relative performance improves when proportion divergence is high (at the right), with even more substantial changes between labeled and unlabeled sets, which makes sense since ours is the only approach to directly address semantic change. Compared to baseline readme, the new algorithm achieves better average performance in 96% of our sample corpora, with an average corpus-wise improvement of 38.4%.

Second, we present a more detailed analysis in Figure 6 of estimation error (vertically) for readme compared to our new approach (ordered horizontally by size of the improvement). The length of each arrow represents the average improvement over the 100 separate analyses of subsets of each of the 72 data sets, with one arrow for each data set. In all but seven cases (at the left), the arrows face downward, meaning that on average our new method usually outperforms readme, and the seven exceptions are all narrow misses. The three colored arrows refer to the data sets used in Hopkins and King (2010); we improve performance of all three, with levels of improvement ranging from low (immigration in green), to medium (Clinton posts in orange), to high (Enron emails in red). Overall, across all our data sets, we find a 38.4% average corpus-wide improvement over readme, which in terms of SAE is a substantial nine percentage points.

Finally, we show that our results are robust across sixteen diverse simulation designs (described in Appendix B). In the left panel of Figure 7 we compare readme to readme2

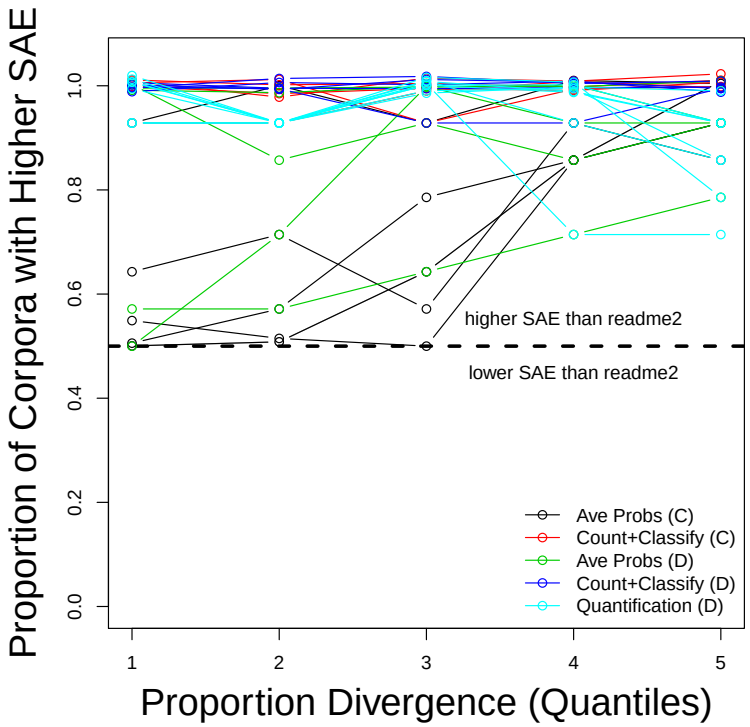

Figure 5: The proportion of corpora with higher error than our approach plotted (vertically) by quantile in proportion divergence (horizontally), with 32 different methods color coded by type (described in the text, with feature space "D" for discrete and "C"for continuous). Each point is an average over 7,200 data sets.

in each design. On average, every design shows that readme2 outperforms readme on average (as indicated by being above the dotted horizontal line). The empirical result, noted in red, is the design described above. Then, the right panel of Figure 7 illustrates how, across the 16 simulation designs, readme2 outperforms not only readme, but also the other 32 alternative methods most of the time. Moreover, there does not appear to be a consistent best alternative, even in second place to readme2: the best performing alternative method is different in 13 of the 16 designs. In sum, our new methodology would seem to be preferably to readme and other existing methods across a wide array of data sets and evaluation protocols.

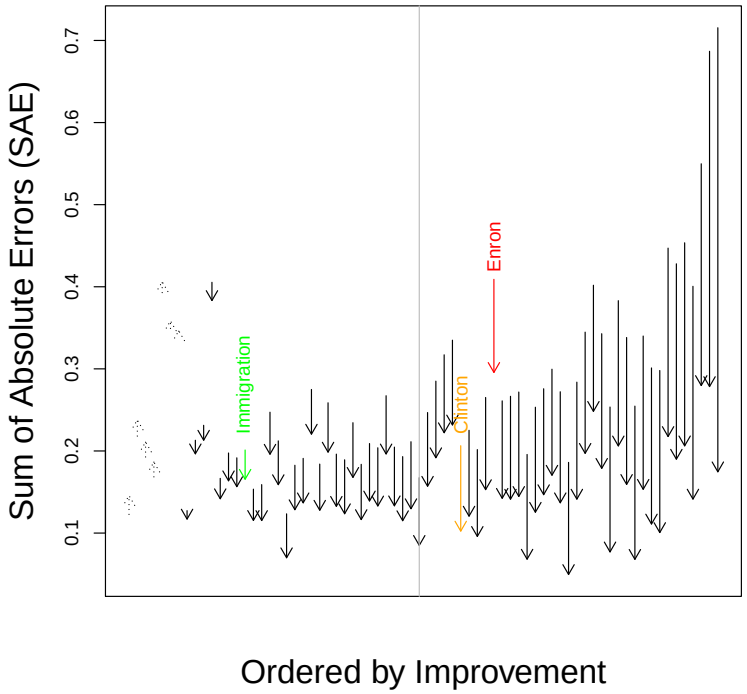

Figure 6: Average Error Reduction: readme to readme2, in analyses of 100 subsets of each of the 72 data sets. The length of each arrow indicates the change in performance, with downward arrows indicating how SAE is reduced by our new methodology. Colored arrows refer to the three data sets in Hopkins and King (2010).

6 What Can Go Wrong?

In practice, we find that the methods described here are robust in a wide variety of circumstances, category types, and data sets. However, no inferential method works all the time, and so we focus on the three situations to be aware of, where our approach may not help.

First, when we use matching in continuous space, we generally reduce proportion divergence and the effects of vanishing discourse. However, emerging discourse can not only cause bias in any method, but this bias can sometimes come about from the process of dealing with vanishing discourse. In addition, although readme2 is the only method that can reduce the effects of vanishing discourse, the method is of no help if all the relevant discourse vanishes within a category. This is akin to a violation of the common

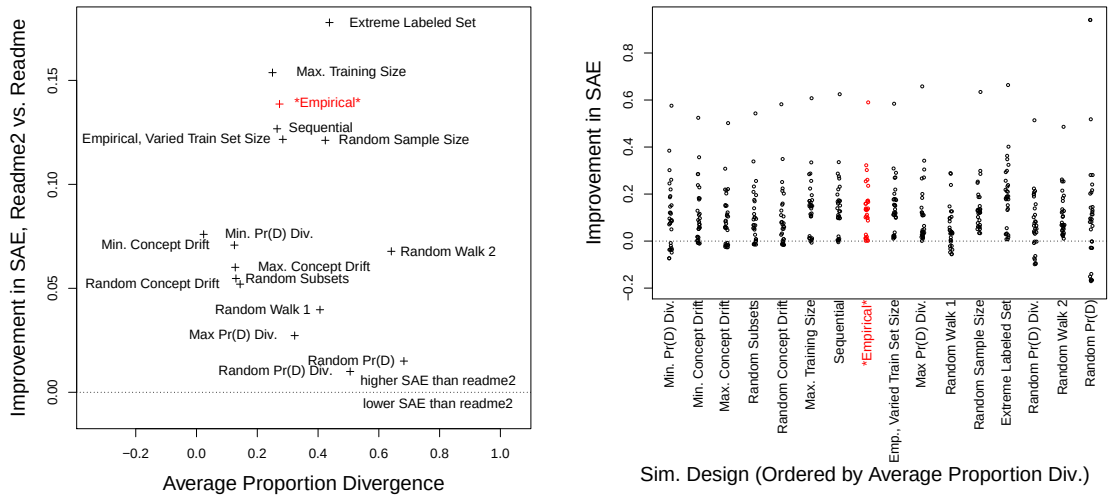

Figure 7: Average Error Reduction Across Simulation Designs, relative to readme (left panel) and 32 alternative methods (right panel). The simulation marked Empirical in both is the one described in the text; for the others, see Appendix B. Items above the dotted horizontal line in both panels indicate that readme2 reduced SAE compared to a competing method.

support assumption in uses of matching for causal inference and so must rely on risky extrapolations. Unlike with classifiers, our methodology does not need to assume that the labeled and unlabeled sets are drawn from the same distribution, but it must assume that the distributions have some overlap. If one suspects that meaning or language is changing dramatically, one should consider coding additional observations from later points in time.

Second, if the original feature space is extraordinarily sparse (as with a regression with a large number of irrelevant covariates), then our optimization algorithm may have difficulty arriving at a stable solution for Γ. This can happen with highly uninformative text, categories with labels that may be more meaningful to investigators than the authors of the text, or error-ridden hand coding.

Finally, we emphasize that our approach relies on meaningful text in each document, conceptually coherent and mutually exclusive and exhaustive categores, and a labeling effort that validly and reliably codes documents into the right categories. These may seem like obvious criteria, but they always constitute the most important steps in any automated text analysis method, including ours. In our experience most of the effort in getting an analysis right involves, or should involve, getting these preliminary steps right.

7 Concluding Remarks

We improve on a popular method of estimating category proportions (Hopkins and King, 2010; King and Lu, 2008), a task of central interest to social scientists among others. We do this without having to tune or even use the often model dependent methods of individual classification developed in other fields for different quantities of interest. We prove properties and provide intuition about readme and then build our alternative. We have tested our analysis in 72 separate data sets, 7,200 subsets, and 16 evaluation protocols, all with encouraging results. Overall, our approach weakens the key assumptions of readme while creating new, more meaningful numerical representations of each of the documents specifically tuned to reduce the mean square error of multi-category, nonparametric quantification.

We can identify several ways of building on our work to further improve performance. These include methods for optimizing the raw continuous textual representations used in readme2. In this analysis, we use document-level summaries of word vectors for the raw features, but there is no quantitative principle implying that this choice is optimal and so could be improved. Indeed, our results suggest that the quantitative features used in readme greatly influence the performance of the estimator. It is natural, then, to explore continuous document-level representations directly from the labeled (and unlabeled) sets, possibly using category-wise information from the labeled set. We could also optimize over λ, rather than setting it to 0.5 as we do now, among other related small changes. With these additions, the estimation process could be more fully optimized for quantification.

Finally, we note that some of the innovations described here may be profitably applied in other domains. For example, our dimension reduction method could be used for data visualization. For example, in visualizing data on partisanship, we could use our methods, find the 2-dimensional projection that maximally discriminates between Democrats, Republicans, and Independents and simultaneously contains minimal redundancy. The relevant clusters would then become more visible, and could even be paired with a data clustering algorithm on the 2-dimensional projection for additional visualization or analysis purposes. Our dimensionality reduction approach could also be applied in the study of

causality, where we borrowed some of our techniques. In causal inference, investigators often use nonparametric methods such as matching, but there is considerable interest in performing this matching in an optimal feature space, such as in the space of predicted values for the outcome under the control intervention (such as in "predictive mean matching"). Matching could also be performed on the features we derive in this paper. The resulting causal estimator may have attractive properties, since it would take into account the relationship between the covariates and the outcome (leading to lower bias) while incorporating several independent sources of information (leading to lower variance).

Appendix A Bias on the Simplex in Two Categories

Proof of Proposition 1

Start with the least-squares minimization problem

$$\widehat{\pi_1^{\mathrm{U}}} = \arg\min_{\pi_1^U} \sum_{w=1}^{W} \left(S_w^{\mathrm{U}} - \widehat{S_w^{\mathrm{U}}} \right)^2 .$$

Write $\widehat{S_w^{\mathrm{U}}}$ in terms of X^{L} and π_1^{U}

$$\widehat{\pi_1^{\mathrm{U}}} = \arg\min_{\pi_1^U} \sum_{w=1}^{W} \left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}} - \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right) \pi_1^{\mathrm{U}} \right)^2 .$$

Take the derivative and set equal to 0

$$\begin{aligned}
0 &= \frac{\partial}{\partial \pi_1^{\mathrm{U}}} \sum_{w=1}^{W} \left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}} - \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right) \pi_1^{\mathrm{U}} \right)^2 \\
&= \sum_{w=1}^{W} \left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}} \right) \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right) - \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right)^2 \pi_1^{\mathrm{U}} \\
\sum_{w=1}^{W} \left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}} \right) \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right) &= \sum_{w=1}^{W} \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right)^2 \pi_1^{\mathrm{U}} \\
\frac{\sum_{w=1}^{W} \left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}} \right) \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right)}{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}} \right)^2} &= \pi_1^{\mathrm{U}} .
\end{aligned}$$

Since the expression being optimized is quadratic, this is a global optimum. Therefore the readme estimator in two categories has the closed-form expression

$$\widehat{\pi_1^{\mathrm{U}}} = \frac{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}}\right)\left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}}\right)}{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}}\right)^2}.$$

Proof of Proposition 2

Start with the expression for $\widehat{\pi_1^{\mathrm{U}}}$.

$$\widehat{\pi_1^{\mathrm{U}}} = \frac{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}}\right)\left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{L}}\right)}{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{L}} - X_{w2}^{\mathrm{L}}\right)^2}.$$

Write X_{wc}^{L} in terms of X_{wc}^{U} and ϵ_{wc}.

$$\widehat{\pi_1^{\mathrm{U}}} = \frac{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}} + \epsilon_{w1} - \epsilon_{w2}\right)\left(S_w^{\mathrm{U}} - X_{w2}^{\mathrm{U}} - \epsilon_{w2}\right)}{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}} + \epsilon_{w1} - \epsilon_{w2}\right)^2}.$$

Substitute the accounting identity for S_w^{U}

$$\widehat{\pi_1^{\mathrm{U}}} = \frac{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}} + \epsilon_{w1} - \epsilon_{w2}\right)\left((X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}})\pi_1^{\mathrm{U}} - \epsilon_{w2}\right)}{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}} + \epsilon_{w1} - \epsilon_{w2}\right)^2}.$$

Expanding the numerator

$$\widehat{\pi_1^{\mathrm{U}}} = \frac{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}} + \epsilon_{w1} - \epsilon_{w2}\right)\left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}}\right)}{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}} + \epsilon_{w1} - \epsilon_{w2}\right)^2}\pi_1^{\mathrm{U}} - \frac{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}} + \epsilon_{w1} - \epsilon_{w2}\right)\left(\epsilon_{w2}\right)}{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}} + \epsilon_{w1} - \epsilon_{w2}\right)^2}.$$

Taking the expectation, we find

$$E\left[\widehat{\pi_1^{\mathrm{U}}}\right] = E\left[\frac{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}} + \epsilon_{w1} - \epsilon_{w2}\right)\left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}}\right)}{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}} + \epsilon_{w1} - \epsilon_{w2}\right)^2}\right]\pi_1^{\mathrm{U}} - E\left[\frac{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}} + \epsilon_{w1} - \epsilon_{w2}\right)\left(\epsilon_{w2}\right)}{\sum_{w=1}^{W} \left(X_{w1}^{\mathrm{U}} - X_{w2}^{\mathrm{U}} + \epsilon_{w1} - \epsilon_{w2}\right)^2}\right].$$

Proof of Proposition 3

Using the first-order Taylor approximation $E\left[\frac{X}{Y}\right] \approx \frac{E[X]}{E[Y]}$, we have

$$
\begin{aligned}
E\left[\widehat{\pi_1^{\mathrm{U}}}\right] &\approx \frac{E\left[\sum_{w=1}^{W}\left(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}}+\epsilon_{w1}-\epsilon_{w2}\right)\left(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}}\right)\right]}{E\left[\sum_{w=1}^{W}\left(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}}+\epsilon_{w1}-\epsilon_{w2}\right)^2\right]}\pi_1^{\mathrm{U}} - \frac{E\left[\sum_{w=1}^{W}\left(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}}+\epsilon_{w1}-\epsilon_{w2}\right)\left(\epsilon_{w2}\right)\right]}{E\left[\sum_{w=1}^{W}\left(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}}+\epsilon_{w1}-\epsilon_{w2}\right)^2\right]} \\
&= \frac{\sum_{w=1}^{W}\left(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}}\right)^2}{\sum_{w=1}^{W}(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}})^2+E[(\epsilon_{w1}-\epsilon_{w2})^2]}\pi_1^{\mathrm{U}} - \frac{\sum_{w=1}^{W}E[\epsilon_{w1}\epsilon_{w2}]-E[(\epsilon_{w2})^2]}{\sum_{w=1}^{W}(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}})^2+E[(\epsilon_{w1}-\epsilon_{w2})^2]} \\
&= \frac{\sum_{w=1}^{W}\left(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}}\right)^2\pi_1^{\mathrm{U}}-E[\epsilon_{w1}\epsilon_{w2}]+E[(\epsilon_{w2})^2]}{\sum_{w=1}^{W}(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}})^2+E[(\epsilon_{w1}-\epsilon_{w2})^2]} \\
&= \frac{\sum_{w=1}^{W}\left(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}}\right)^2\pi_1^{\mathrm{U}}+\mathrm{Var}(\epsilon_{w2})-\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})}{\sum_{w=1}^{W}(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}})^2+\mathrm{Var}(\epsilon_{w1}-\epsilon_{w2})} \\
&= \frac{\sum_{w=1}^{W}\left(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}}\right)^2\pi_1^{\mathrm{U}}+\mathrm{Var}(\epsilon_{w2})-\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})}{\sum_{w=1}^{W}(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}})^2+\mathrm{Var}(\epsilon_{w1})+\mathrm{Var}(\epsilon_{w2})-2\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})},
\end{aligned}
$$

where the last two lines follow from the definition of variance and the assumption that $E[\epsilon_{wc}] = 0$. Subtracting π_1^{U} to get the bias:

$$
\begin{aligned}
\mathrm{Bias}(\widehat{\pi_1^{\mathrm{U}}}) &\approx \frac{\sum_{w=1}^{W}\left(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}}\right)^2\pi_1^{\mathrm{U}}+\mathrm{Var}(\epsilon_{w2})-\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})}{\sum_{w=1}^{W}(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}})^2+\mathrm{Var}(\epsilon_{w1})+\mathrm{Var}(\epsilon_{w2})-2\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})} - \pi_1^{\mathrm{U}} \\
&= \frac{\sum_{w=1}^{W}\mathrm{Var}(\epsilon_{w2})-\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})-\mathrm{Var}(\epsilon_{w1})\pi_1^{\mathrm{U}}-\mathrm{Var}(\epsilon_{w2})\pi_1^{\mathrm{U}}+2\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})\pi_1^{\mathrm{U}}}{\sum_{w=1}^{W}(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}})^2+\mathrm{Var}(\epsilon_{w1})+\mathrm{Var}(\epsilon_{w2})-2\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})} \\
&= \frac{\sum_{w=1}^{W}\mathrm{Var}(\epsilon_{w2})(1-\pi_1^{\mathrm{U}})-\mathrm{Var}(\epsilon_{w1})\pi_1^{\mathrm{U}}-\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})+2\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})\pi_1^{\mathrm{U}}}{\sum_{w=1}^{W}(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}})^2+\mathrm{Var}(\epsilon_{w1})+\mathrm{Var}(\epsilon_{w2})-2\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})} \\
&= \frac{\sum_{w=1}^{W}\mathrm{Var}(\epsilon_{w2})(1-\pi_1^{\mathrm{U}})-\mathrm{Var}(\epsilon_{w1})\pi_1^{\mathrm{U}}-\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})(1-\pi_1^{\mathrm{U}})+\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})\pi_1^{\mathrm{U}}}{\sum_{w=1}^{W}(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}})^2+\mathrm{Var}(\epsilon_{w1})+\mathrm{Var}(\epsilon_{w2})-2\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})} \\
&= \frac{\sum_{w=1}^{W}\left(\mathrm{Var}(\epsilon_{w2})-\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})\right)(1-\pi_1^{\mathrm{U}})-\left(\mathrm{Var}(\epsilon_{w1})-\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})\right)\pi_1^{\mathrm{U}}}{\sum_{w=1}^{W}(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}})^2+\mathrm{Var}(\epsilon_{w1})+\mathrm{Var}(\epsilon_{w2})-2\mathrm{Cov}(\epsilon_{w1},\epsilon_{w2})}.
\end{aligned}
$$

Proof of Proposition 4

Substituting in the known measurement error variances and assuming independence in measurement errors across categories yields:

$$
\mathrm{Bias}(\widehat{\pi_1^{\mathrm{U}}}) \approx \frac{\sum_{w=1}^{W}\left(\frac{\sigma_{w2}^2}{N_2^{\mathrm{L}}}\right)(1-\pi_1^{\mathrm{U}})-\left(\frac{\sigma_{w1}^2}{N_1^{\mathrm{L}}}\right)\pi_1^{\mathrm{U}}}{\sum_{w=1}^{W}(X_{w1}^{\mathrm{U}}-X_{w2}^{\mathrm{U}})^2+\left(\frac{\sigma_{w2}^2}{N_2^{\mathrm{L}}}\right)+\left(\frac{\sigma_{w1}^2}{N_1^{\mathrm{L}}}\right)}
$$

Using the fact that $N_c^{\mathrm{L}} = N^{\mathrm{L}}\pi_c^{\mathrm{L}}$

$$
\begin{aligned}
\text{Bias}(\widehat{\pi_1^{\text{U}}}) &\approx \frac{(1-\pi_1^{\text{U}})\sum_{w=1}^{W}\left(\frac{\sigma_{w2}^2}{N^{\text{L}}\pi_2^{\text{L}}}\right) - \pi_1^{\text{U}}\sum_{w=1}^{W}\left(\frac{\sigma_{w1}^2}{N^{\text{L}}\pi_1^{\text{L}}}\right)}{\sum_{w=1}^{W}(X_{w1}^{\text{U}} - X_{w2}^{\text{U}})^2 + \left(\frac{\sigma_{w2}^2}{N^{\text{L}}\pi_2^{\text{L}}}\right) + \left(\frac{\sigma_{w1}^2}{N^{\text{L}}\pi_1^{\text{L}}}\right)} \\
&\approx \frac{\frac{1}{N^{\text{L}}}\left[\frac{(1-\pi_1^{\text{U}})}{(1-\pi_1^{\text{L}})}\sum_{w=1}^{W}\sigma_{w2}^2 - \frac{\pi_1^{\text{U}}}{\pi_1^{\text{L}}}\sum_{w=1}^{W}\sigma_{w1}^2\right]}{\sum_{w=1}^{W}(X_{w1}^{\text{U}} - X_{w2}^{\text{U}})^2 + \frac{1}{N^{L}(1-\pi_1^{\text{L}})}\sum_{w=1}^{W}\sigma_{w2}^2 + \frac{1}{N^{L}\pi_1^{\text{L}}}\sum_{w=1}^{W}\sigma_{w1}^2} \\
&\approx \frac{\frac{(1-\pi_1^{\text{U}})}{(1-\pi_1^{\text{L}})}\sum_{w=1}^{W}\sigma_{w2}^2 - \frac{\pi_1^{\text{U}}}{\pi_1^{\text{L}}}\sum_{w=1}^{W}\sigma_{w1}^2}{N^{\text{L}}\sum_{w=1}^{W}(X_{w1}^{\text{U}} - X_{w2}^{\text{U}})^2 + \frac{1}{(1-\pi_1^{\text{L}})}\sum_{w=1}^{W}\sigma_{w2}^2 + \frac{1}{\pi_1^{\text{L}}}\sum_{w=1}^{W}\sigma_{w1}^2}
\end{aligned}
$$

The denominator is strictly positive. Therefore, bias is minimized when the numerator is equal to 0. Solving for π_1^{U} in terms of π_1^{L} yields

$$
\begin{aligned}
0 &= \frac{(1-\pi_1^{\text{U}})}{(1-\pi_1^{\text{L}})}\sum_{w=1}^{W}\sigma_{w2}^2 - \frac{\pi_1^{\text{U}}}{\pi_1^{\text{L}}}\sum_{w=1}^{W}\sigma_{w1}^2 \\
\frac{\pi_1^{\text{U}}}{\pi_1^{\text{L}}}\sum_{w=1}^{W}\sigma_{w1}^2 &= \frac{(1-\pi_1^{\text{U}})}{(1-\pi_1^{\text{L}})}\sum_{w=1}^{W}\sigma_{w2}^2 \\
\pi_1^{\text{U}}(1-\pi_1^{\text{L}})\sum_{w=1}^{W}\sigma_{w1}^2 &= (1-\pi_1^{\text{U}})\pi_1^{\text{L}}\sum_{w=1}^{W}\sigma_{w2}^2 \\
\left(\pi_1^{\text{U}} - \pi_1^{\text{U}}\pi_1^{\text{L}}\right)\sum_{w=1}^{W}\sigma_{w1}^2 &= \left(\pi_1^{\text{L}} - \pi_1^{\text{U}}\pi_1^{\text{L}}\right)\sum_{w=1}^{W}\sigma_{w2}^2 \\
\pi_1^{\text{U}}\sum_{w=1}^{W}\sigma_{w1}^2 &= \pi_1^{\text{L}}\sum_{w=1}^{W}\sigma_{w2}^2 + \pi_1^{\text{U}}\pi_1^{\text{L}}\left(\sum_{w=1}^{W}\sigma_{w1}^2 - \sigma_{w2}^2\right) \\
\frac{\pi_1^{\text{U}}\sum_{w=1}^{W}\sigma_{w1}^2}{\sum_{w=1}^{W}\sigma_{w2}^2 + \pi_1^{\text{U}}\left(\sum_{w=1}^{W}\sigma_{w1}^2 - \sigma_{w2}^2\right)} &= \pi_1^{\text{L}} \\
\frac{\pi_1^{\text{U}}\sum_{w=1}^{W}\sigma_{w1}^2}{\pi_1^{\text{U}}\sum_{w=1}^{W}\sigma_{w1}^2 + (1-\pi_1^{\text{U}})\sum_{w=1}^{W}\sigma_{w2}^2} &= \pi_1^{\text{L}}
\end{aligned}
$$

When the measurement error variances are generally constant across categories, the bias is zero when the labeled set proportions are equal to the unlabeled set proportions.

Appendix B Alternative Evaluation Designs

Each of the sixteen evaluation designs summarized in Table 1 offers a different way of generating 7,200 data sets as subsets of the 72 corpora described in Section 5. Each data set is divided into a labeled set as well as a test set that serves the purpose of the unlabeled set during estimation, but can also be used for evaluation since all its document labels are observed.

Table 1: Alternative Evaluation Designs

Design Name	**Description**
Empirical	Sample consecutive chronological slices of the data to form labeled and unlabeled sets, wrapping when necessary (detailed in Section 5).
Empirical, Varied Training Set Size	Sample consecutive chronological slices of the data to form labeled and unlabeled sets, wrapping when necessary (detailed in Section 5). Randomly sample the labeled set size from $\{100, 300, 500, 1000\}$.
Sequential	Sample one time point, with the first half of the data for the labeled set and second half for the test set. Wrap when necessary.
Random Subsets	Sample documents randomly without replacement for labeled and unlabeled sets.
Min. Proportion Divergence	Sample documents randomly without replacement to form 100,000 candidate labeled and unlabeled sets. Select the pair which minimizes $\lvert\pi^{\mathrm{L}} - \pi^{\mathrm{U}}\rvert$ divergence.
Uniform Random Proportion Divergence	Draw random uniform on the interval $[0, 0.75]$, the target $\lvert\pi^{\mathrm{L}} - \pi^{\mathrm{U}}\rvert$ divergence. Draw 10,000 candidate π^{L} and π^{U} uniform from the simplex. Select the pair closest to the target.
Max. Proportion Divergence	Sample documents randomly without replacement to form 100,000 candidate labeled and unlabeled sets. Select pair that maximizes $\lvert\pi^{\mathrm{L}} - \pi^{\mathrm{U}}\rvert$ divergence.
Min. Semantic Change	Sample documents randomly without replacement to form 100,000 candidate labeled and unlabeled sets. Select the pair that minimizes semantic change.
Uniform Random Semantic Change	Sample documents randomly without replacement to form 100,000 candidate labeled and unlabeled sets. Select a uniform random target amount of semantic change. Select the pair closest to the target.
Max. Semantic Change	Sample documents randomly without replacement to form 100,000 candidate labeled and unlabeled sets. Select the pair which maximizes semantic change.
Random Walk 1	Draw π^{L} from a uniform density on the simplex. For iteration i, draw π^{U} from a Dirichlet with parameter $\alpha \propto 1_{C\times 1}$ for the first iteration and $\alpha \propto (\pi^{\mathrm{U}})_{i-1}$ for subsequent iterations.
Random Walk 2	Draw the labeled set chronologically. Then, draw π^{U} by selecting a random point on the simplex.
Random Sample Size	Draw N^{L} uniform on the integers 100,...,500. Generate 1,000 chronological labeled/unlabeled set divisions. Select division that best approximates one of the categories having $< 5\%$ of the labeled set, but $> 25\%$ of the unlabeled set.
Uniform Random Proportions	Draw π^{L} and π^{U} from independent uniform distributions on the simplex.

Table 1 . . . Continued

Design Name	Description
Uniform Divergence	Draw random uniform value from the interval $[0, 0.75]$, the value of $\lvert\pi^{\mathrm{L}} - \pi^{\mathrm{U}}\rvert$. Draw 10,000 candidate π^{L} and π^{U} values uniform from the simplex. Select the pair closest to target.
Extreme Labeled Set	Calculate document-level word vector features. Form labeled set from documents falling furthest from the average document. Form the unlabeled set from a random selection of the remaining documents.

References

Brunzell, H. and J. Eriksson (2000): "Feature reduction for classification of multidimensional data". In: *Pattern Recognition*, no. 10, vol. 33, pp. 1741–1748. DOI: 10.1016/S0031-3203(99)00142-9. URL: bit.ly/2ihoYdl.

Buck, Alfred A and John J Gart (1966): "Comparison of a Screening Test and a Reference Test in Epidemiologic Studies. I. Indices of Agreements and their Relation to Prevalence." In: *American Journal of Epidemiology*, no. 3, vol. 83, pp. 586–92.

Denny, Matthew James and Arthur Spirling (2016): "Assessing the Consequences of Text Preprocessing Decisions". In: https://ssrn.com/abstract=2849145.

Esuli, Andrea and Fabrizio Sebastiani (2015): "Optimizing text quantifiers for multivariate loss functions". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)*, no. 4, vol. 9, p. 27.

Firat, Aykut (2016): "Unified Framework for Quantification". In: *arXiv:1606.00868*.

Forman, George (2007): *Quantifying counts, costs, and trends accurately via machine learning*. Tech. rep. Technical report, HP Laboratories, Palo Alto, CA.

Hand, David J. (2006): "Classifier Technology and the Illusion of Progress". In: *Statistical Science*, no. 1, vol. 21, pp. 1–14.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart (2007): "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference". In: *Political Analysis*, vol. 15, pp. 199–236. URL: j.mp/matchP.

Hoadley, Bruce (2001): "[Statistical Modeling: The Two Cultures]: Comment". In: *Statistical Science*, no. 3, vol. 16, pp. 220–224.

Hopkins, Daniel and Gary King (Jan. 2010): "A Method of Automated Nonparametric Content Analysis for Social Science". In: *American Journal of Political Science*, no. 1, vol. 54, pp. 229–247. URL: j.mp/jNFDgI.

Hopkins, Daniel, Gary King, Matthew Knowles, and Steven Melendez (2013): *Readme: Software for Automated Content Analysis*. Versions 2007–2013. URL: GaryKing.org/readme.

Iacus, Stefano M., Gary King, and Giuseppe Porro (2012): "Causal Inference Without Balance Checking: Coarsened Exact Matching". In: *Political Analysis*, no. 1, vol. 20, pp. 1–24. URL: j.mp/woCheck.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013): *An introduction to statistical learning*. Vol. 112. Springer.

Kar, Purushottam et al. (2016): "Online Optimization Methods for the Quantification Problem". In: *arXiv preprint arXiv:1605.04135*.

King, Gary, Daniel Hopkins, and Ying Lu (2012): "System for estimating a distribution of message content categories in source data". US Patent 8,180,717. URL: j.mp/VApatent.

King, Gary and Ying Lu (2008): "Verbal Autopsy Methods with Multiple Causes of Death". In: *Statistical Science*, no. 1, vol. 23, pp. 78–91. URL: j.mp/2AuA8aN.

King, Gary, Ying Lu, and Kenji Shibuya (2010): "Designing Verbal Autopsy Studies". In: *Population Health Metrics*, no. 19, vol. 8. URL: j.mp/DAutopsy.

King, Gary, Jennifer Pan, and Margaret E. Roberts (2013): "How Censorship in China Allows Government Criticism but Silences Collective Expression". In: *American Political Science Review*, vol. 107, pp. 1–18. URL: j.mp/LdVXqN.

Kingma, Diederik and Jimmy Lei Ba (2015): "Adam: A Method for Stochastic Optimization". In: *Proceedings of the International Conference on Learning Representations*. International Union for the Scientific Study of Population.

Levy, Omer, Yoav Goldberg, and Ido Dagan (2015): "Improving distributional similarity with lessons learned from word embeddings". In: *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 211–225.

Levy, P.S. and E. H. Kass (1970): "A three population model for sequential screening for Bacteriuria". In: *American Journal of Epidemiology*, vol. 91, pp. 148–154.

Martin Abadi et al. (2015): *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. URL: tensorflow.org.

Milli, Letizia et al. (2013): "Quantification trees". In: *2013 IEEE 13th International Conference on Data Mining*, pp. 528–536.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014): "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Tasche, Dirk (2016): "Does quantification without adjustments work?" In: *arXiv preprint arXiv:1602.08780*.

Vincent, Pascal et al. (2010): "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion". In: *Journal of Machine Learning Research*, no. Dec, vol. 11, pp. 3371–3408. URL: bit.ly/2gPcedw.