Supplemental Materials

Reagan Mozer, Luke Miratrix, Aaron Russell Kaufman, and L. Jason Anastasopoulos

A Text representations and distance metrics

In Section 3 we describe a framework for text matching involving choosing both a text representation and a distance metric; we then briefly outline the options for each. Here we expand that discussion.

A.1 Choosing a representation

To operationalize documents for text matching, we must first represent the corpus in a structured, quantitative form. There are two important properties to consider when constructing a representation for text with the goal of matching. First, the chosen representation should be sufficiently low-dimensional such that it is practical to define and calculate distances between documents. If a representation contains thousands of covariates, calculating even a simple measure of distance may be computationally challenging or may suffer from the curse of dimensionality. Second, the chosen representation should be meaningful; that is, it should capture sufficient information about the corpus so that matches obtained based on this representation will be similar in some clear and interpretable way. As discussed in Section 2, text matching is only a useful tool for comparing groups of text documents when the representation defines covariates that contain useful information about systematic differences between the groups.

In this paper, we explore three common types of representations: the term-document matrix (TDM), which favors retaining more information about the text at the cost of dimensionality, statistical topic models, which favor dimension reduction at the potential cost of information, and neural network embeddings, which fall somewhere in between. There are a number of alternative text representations that could also be used to perform matching within our framework, including other representations based on neural networks (Bengio et al., 2003) or those constructed using document embeddings (Le and Mikolov, 2014; Dai et al., 2015), but these are left as a topic for future research.

A.1.1 Representations based on the term-document matrix

Perhaps the simplest way to represent a text corpus is as a TDM. Under the common "bag-of-words" assumption, the TDM considers two documents identical if they use the same terms with the same frequency, regardless of the ordering of the terms (Salton and McGill, 1986). When matching documents, it is intuitive that documents that use the same set of terms at similar rates should be considered similar, so the TDM provides a natural construction for representing text with the goal of matching. However, the dimensionality of a standard TDM may give rise to computational challenges when calculating pairwise distances between documents in some corpora. There are many dimension-reduction strategies that can be applied to help mitigate this issue including techniques based on matrix rescaling using a scheme such as TF-IDF scoring (Salton, 1991), and techniques for bounding the vocabulary to eliminate extremely rare and/or extremely common terms. However, it should be noted that in large corpora, a bounded and rescaled TDM may still have a dimension in the tens of thousands, setting known to be difficult for matching (Roberts et al., 2019).

A.1.2 Representations based on statistical topic models

An alternative representation for text, popular in the text analysis literature, is based on statistical topic models (Blei, 2012), e.g., LDA (Blei et al., 2003) and STM (Roberts et al., 2016a). The main argument for matching using a topic-model-based representation of text is that document similarity can adequately be determined by comparing targeted aspects of the text rather than by comparing the use of specific terms. That is, topic-model-based representations imply that two documents are similar if they cover a fixed number of topics at the same rates. Topic models provide an efficient strategy for considerably reducing the dimension of the covariates while retaining all information that is relevant for matching. In contrast to the tens of thousands of covariates typically defined using a representation based on the TDM, representations built using topic models typically contain no more than a few hundred covariates at most. However, consistent estimation of topic proportions is notoriously difficult due to issues with multimodality of these models, which gives rise to a number of issues for applications of matching in practice (Roberts et al., 2016b).

A.1.3 Representations based on neural network embeddings

Mikolov et al. (Mikolov et al., 2013) introduce a neural network architecture to embed words in an $n-$dimensional space based on its usage and the words which commonly surround it. This architecture has proven remarkably powerful with many intriguing properties. For example, it performs very well in a series of "linguistic algebra" tasks, successfully solving questions like "Japan" − "sushi" + "Germany" = "bratwurst."

A.1.4 Propensity scores

When matching in settings with multiple covariates, a common technique is to first perform dimension reduction to project the multivariate covariates into a univariate space. A popular tool used for this purpose is the propensity score, defined as the probability of receiving treatment given the observed covariates (Rosenbaum and Rubin, 1983). Propensity scores summarize all of the covariates into one scalar, and matching is then performed by identifying groups of units with similar values of this score. In practice, propensity scores are generally not known to the researcher and must be estimated using the observed data. When applied to text, propensity scores can be used to further condense the information within a chosen higher-dimensional representation into a summary of only the information that is relevant for determining treatment assignment. Propensity scores representations can be constructed using a quantitative text representation. For example, using STM-based representations or Word2Vec-based representations where dimension of the covariate space is less than the number of documents, standard techniques such as simple logistic regression can be used to estimate propensity scores. To construct propensity score representations over larger a covariate space, such as those typically spanned by a TDM, we use Multinomial Inverse Regression (MNIR; Taddy, 2013), which provides a novel estimation technique for performing logistic regression of phrase counts from the TDM onto the treatment indicator. After estimating this model, we can calculate a sufficient reduction score that, in principle, will contain all the information from the TDM that is relevant for predicting treatment assignment. Performing a forward regression of the treatment indicator on this sufficient reduction score produces the desired propensity score estimates.

A.2 Design choices for representations

Representations of text data typically involve a number of tuning parameters. When using the bag-of-words representation, researchers often remove very common and very rare words at arbitrary thresholds, as these add little predictive power, or choose to weight terms by their inverse document frequency; these pre-processing decisions can be very important (Denny and Spirling, 2018). Topic models such as the STM are similarly sensitive to these pre-processing decisions (Fan et al., 2017) and also require specification of the number of topics and selecting covariates, which are often unstable. Word2vec values depend on the dimensionality of the word vectors as well as the training data and the architecture of the neural network. Below, we discuss a number of design choices that are required for the different representations considered in our study.

TDM-based representations. Each of the TDM-based representations is characterized by a bounding scheme, which determines the subset of the vocabulary that will be included in X, and a weighting scheme, which determines the numerical rule for how the values of X are measured. We consider standard term-frequency (TF) weighting, TF-IDF weighting, and L2-rescaled TF-IDF weighting. We also consider a number of different screening schemes, including no screening, schemes that eliminate high and low frequency terms, and schemes that consider only high and low frequency terms.

STM-based representations. Each STM-based representation is characterized by a fixed number of topics (K=10, 30, 50, or 100) and takes one of three distinct forms: 1) the vector of K estimated topic proportions ("S1"), 2) the vector of K estimated topic proportions and the SR score ("S2"), or 3) a coarsened version of the vector of K estimated topic proportions ("S3"). This coarsened representation is constructed using the following procedure. For each document, we first identify the three topics with the largest estimated topic proportions. We retain and standardize these three values and set all remaining $K-3$ topic proportions equal to 0, so that the resulting vector of coarsened topic proportions, $\hat{\theta}_i^\star$, contains only three non-zero elements. We then calculate the "focus" of each document, denoted by F_i, a metric we define as the proportion of topical content that is explained by the three most prominent topics. Focus scores close to one indicate content that is highly concentrated on a small number of topics (e.g., a news article covering health care reform may have nearly 100% of its content focused on the topics of *health* and *policy*); conversely, focus scores close to zero indicate more general content covering a wide range of topics (e.g., a news article entitled "The ten events that shaped 2017" may have content spread evenly across ten or more distinct topics). To estimate this score for each document, we take the sum of the raw values of the three non-zero topic proportions identified as above (i.e., $\hat{F}_i = \hat{\theta}_{i[1]} + \hat{\theta}_{i[2]} + \hat{\theta}_{i[3]}$ where $\hat{\theta}_{i[j]}$ is the jth order statistic of the vector $\hat{\theta}$). Appending this estimated focus score to the coarsened topic proportion vector produces the final $(K+1)$-dimensional representation.

TIRM representations. The TIRM procedure of Roberts et al. (2019) uses an STM-based representation with an additional representation based on document-level propensity scores estimated using the STM framework. These separate representations are then combined within the TIRM procedure using a CEM distance. Each variant of the TIRM procedure considered in this paper is characterized by a fixed number of topics and a set coarsening level (2 bins, 3 bins, or 4 bins).

Word Embedding representations. Google and Stanford University have produced a variety of pre-trained word embedding models. Google's GoogleNews model, where each word vector is length 300 using a corpus of 100 billion words, draws from the entire corpus of Google News; this corpus is therefore extremely well-suited to our analysis. As well, we consider several of Stanford's GloVe embeddings (Pennington et al., 2014). In particular, we employ their models with word vectors of length 50, 100, 200, and 300. For each of these five embeddings, we produce document-level vectors by taking the weighted average of all word vectors in a document (Kusner et al., 2015).

A.3 Defining a distance metric

After a representation is chosen, applying this representation to the corpus generates a finite set of numerical covariate values associated with each document (i.e., X_i denotes the covariates observed for document i for all $i = 1, \ldots, N$). The next step in the matching procedure concerns how to use these covariate values to quantify the similarity between two documents. There are two main classes of distance metrics. Exact and coarsened exact distances regard distances as binary: the distance between two units is either zero or infinity, and two units are eligible to be matched only if the distance between them is equal to zero. Alternatively, continuous distance metrics define distance on a continuum, and matching typically proceeds by identifying pairs of units for whom the calculated distance is within some allowable threshold ("caliper").

A.3.1 Exact and coarsened exact distances

The exact distance is defined as:

$$D_{ij} = \begin{cases} 0, & \text{if } X_i = X_j \\ \infty, & \text{otherwise.} \end{cases}$$

Matching over this metric (exact matching) generates pairs of documents between treatment and control groups that match exactly on every covariate. Although this is the ideal, exact matching is typically not possible in practice with more than a few covariates. A more flexible metric can be defined by first coarsening the covariate values into "substantively indistinguishable" bins, then using exact distance within these bins (Iacus et al., 2012). For example, using a topic-model-based representation, one might define a coarsening rule such that documents will be matched if they share the same primary topic (i.e., if the topic with the maximum estimated topic proportion among the K topics is the same for both documents). Roberts et al. (2019) advocates using CEM for matching documents based on a representation built using an STM, but, in principle, this technique can also be used with TDM-based representations. For example, one might coarsen the term counts of a TDM into binary values indicating whether each term in the vocabulary is used within each document. Though it is possible in principle, coarsening does not scale well with the dimension of the covariates and so may not be practical for matching with TDM-based representations. This type of distance specification may also create sensitivities in the matching procedure, since even minor changes in the coarsening rules can dramatically impact the resulting matched samples.

A.3.2 Continuous distances

Various continuous distance metrics can be used for matching, including linear distances based on the (estimated) propensity score or best linear discriminant (Rosenbaum and Rubin, 1983), multivariate metrics such as the Mahalanobis metric (Rubin, 1973), or combined metrics, such as methods that match on the Mahalanobis metric within propensity score calipers (Rosenbaum and Rubin, 1985). When matching on covariates defined by text data, care must be taken to define a metric that appropriately captures the complexities of text. For instance, linear distance metrics such as Euclidean distance may often fail to capture information about the relative importance of different covariates. To make this more clear, consider two pairs of documents containing the texts: "obama spoke", "obama wrote" and 'he spoke", "he wrote". Under a TDM-based representation, the Euclidean distances between units in each of these pairs are equal; however, the first pair of documents is intuitively more similar than the second, since the term “obama” contains more information about the content of the documents than the term “he”. Similarly, the Euclidean distance between the pair documents “obama spoke”, “obama obama” is equivalent to the distance between the pair “obama spoke”, “he wrote”, since by this metric distance increases linearly with differences in term frequencies. These issues also arise when using linear distance metrics with topic-model-based representations.

A metric that is less vulnerable to these complications is Mahalanobis distance, which defines the between documents i and j as $D_{ij} = (X_i - X_i)^T \Sigma^{-1} (X_i - X_j)$, where Σ is the variance-covariance matrix of the covariates X. This is essentially a normalized Euclidean distance, which weights covariates according to their relative influence on the total variation across all documents in the corpus. Calculating Mahalanobis distance is practical for lower-dimensional representations, but because the matrix inversion does not scale well with the dimension of X, it may not be computationally feasible for matching using larger, TDM-based representations.

An alternative metric, which can be efficiently computed using representations defined over thousands of covariates, is cosine distance. Cosine distance measures the cosine of the angle between two documents in a vector space:

$$D_{ij} = 1 - \frac{\sum X_i X_j}{\sqrt{\sum X_i^2}\sqrt{\sum X_j^2}}.$$

Cosine distance is commonly used for determining text similarity in fields such as informational retrieval and is an appealing choice for matching because, irrespective of the dimension of the representation, it captures interpretable overall differences in covariate values (e.g., a cosine distance of one corresponds to a 90 degree angle between documents, suggesting no similarity and no shared vocabulary). In general, the utility of a particular continuous distance metric will largely depend on the distribution that is induced on the covariates through the representation.

A.3.3 Calipers and combinations of metrics

When pruning treated units is acceptable, exact and coarsened exact matching methods have the desirable property that the balance that will be achieved between matched samples is established a-priori. Treated units for whom there is at least one exact or coarsened exact match in the control group are matched, and all other treated units are dropped. On the other hand, matching with a continuous distance metric requires tuning after distances have been calculated in order to bound the balance between matched samples. After the distances between all possible pairings of treated and control documents have been calculated, one then chooses a caliper, D_{max}, such that any pair of units i and j with distance $D_{ij} > D_{max}$ cannot be matched. Here, when pruning treated units is acceptable, any treated units without at least one potential match are dropped. Calipers are typically specified according to a "rule of thumb" that asserts that D_{max} be set equal to the value of 0.25 or 0.5 times the standard deviation of the distribution of distance values over all possible pairs of treated and control units, but in some special cases, the caliper can be chosen to reflect a more interpretable restriction. For example, using the cosine distance metric, one might choose a caliper to bound the maximum allowable angle between matched documents.

A.4 Text as covariates and outcomes

The procedure described in Section 3 is relatively straightforward to apply in studies where text enters the problem only through the covariates. However, in more complicated settings where both the covariates and one or more outcomes are defined by features of text, additional steps may be necessary to ensure these components are adequately separated.

In practice it is generally recommended that outcome data be removed from the dataset before beginning the matching process to preclude even the appearance of "fishing," whereby a researcher selects a matching procedure or a particular matched sample that leads to a desirable result (Rubin, 2007). However, this may not be possible when evaluating a text corpus, since both the covariates and outcome may often be latent features of the text (Egami et al., 2017). For instance, suppose we are interested in comparing the level of positive sentiment within articles based on the gender of the authors. One can imagine that news articles that report incidences of crime will typically reflect lower levels of positive sentiment than articles reporting on holiday activities, regardless of the gender of the reporter. Thus, we might like to match articles between male and female reporters based on their topical content and then compare the sentiment expressed within these matched samples. Here, we must extract both the set of covariates that will be used for matching (i.e., topical content) and the outcome (level of positive sentiment) from the same observed text. Because these different components may often be related, measuring both using the same data poses two important challenges for causal inference: first, it requires that the researcher use the observed data to posit a model on the "post-treatment" outcome, and, second, measurement of the covariates creates potential for fishing. In particular, suppose that positive sentiment is defined for each document as the number of times terms such as "happy" are used within that document (standardized by each document's length). Suppose also that we use the entire vocabulary to measure covariate values for each document (e.g., using a statistical topic model). In this scenario, matching on topical content is likely to produce matches that have similar rates of usage of the term "happy" (in addition to having similar rates of usage of other terms), which may actually diminish our ability to detect differences in sentiment.

To address this issue, we recommend that researchers interested in inference in these settings define the covariates and outcome over a particular representation, or set of distinct representations, such that measurement of the outcome can be performed independently of the measurement of covariates. For example, one might measure the covariates using a representation of text defined over only nouns, and separately, measure outcome values using a representation defined over only adjectives. Or, continuing the previous example, one might divide the vocabulary into distinct subsets of terms, where one subset is used to measure topical content and the other is used to measure positive sentiment. In settings where the chosen representation of the text must be inferred from the observed data (e.g., topic-model-based representations), cross-validation techniques can also be employed, as described in Egami et al. (2017). For instance, one might randomly divide the corpus into training set and test set, where the training set is used to build a model for the representation, and this model is then applied to the test set to obtain covariate values that will be used in the matching procedure.

B Index of representations evaluated

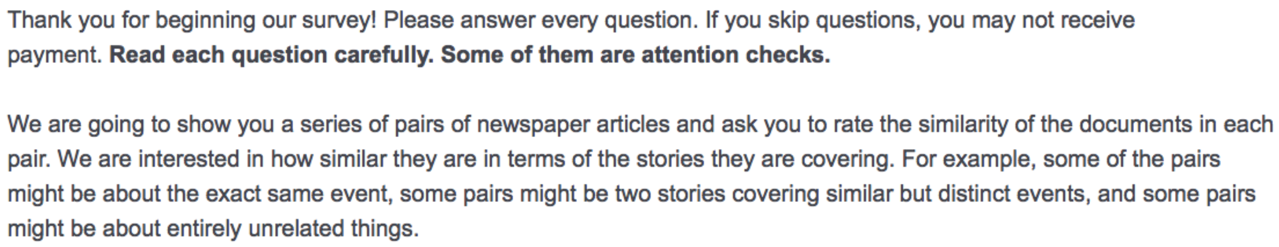
Table 1: Specification of the 26 representations considered

Type	Name	Description	Dimension
TDM	T1	TF Bounded from 4-1000	10726
	T2	TF-IDF Bounded from 4-1000	10726
	T3	TF-IDF Bounded from 4-100	9413
	T4	TF-IDF Bounded from 4-10	4879
	T5	TF-IDF Bounded from 10-500	6000
	T6	TF-IDF Bounded from 500-1000	154
	T7	L2 Rescaled TF-IDF Bounded from 4-1000	10726
	T8	TF on unbounded TDM	34397
	T9	TF-IDF on unbounded TDM	34397
STM	S1-10	STM on 10 Topics	10
	S2-10	10 Topics + estimated sufficient reduction	11
	S3-10	10 Topics, top 3 topics + focus	11
	S1-30	30 Topics	30
	S2-30	30 Topics + estimated sufficient reduction	31
	S3-30	30 Topics, top 3 topics + focus	31
	S1-50	50 Topics	50
	S2-50	50 Topics + estimated sufficient reduction	51
	S3-50	50 Topics, top 3 topics + focus	51
	S1-100	100 Topics	100
	S2-100	100 Topics + estimated sufficient reduction	101
	S3-100	100 Topics, top 3 topics + focus	101
Word2Vec	W1	Word embedding of dimension 50 (Google)	50
	W2	Word embedding of dimension 100 (Google)	100
	W3	Word embedding of dimension 200 (Google)	200
	W4	Word embedding of dimension 300 (Google)	300
	W5	Word embedding of dimension 300	300

C Survey used in human evaluation experiment

The figures below show snapshots of different components of the survey as they were presented to participants in each of our human evaluation experiments. In particular, Figure 1 shows the survey landing page, where participants were informed about the nature of the task. Participants were then presented with the scoring rubric shown in Figure 2 and were informed to use this rubric as "a guide to help [them] determine the similarity of a pair of articles." In the final component of training, participants completed a series of three training tasks, as depicted in Figure 3, where each task required them to read and score one pre-selected pairs of articles. The articles presented in each task were chosen to represent pairings that we believe have match quality scores of zero, five, and ten, respectively. After scoring each training pair, participants were informed about the anticipated score for that pair and provided with an explanation for how that determination was made.

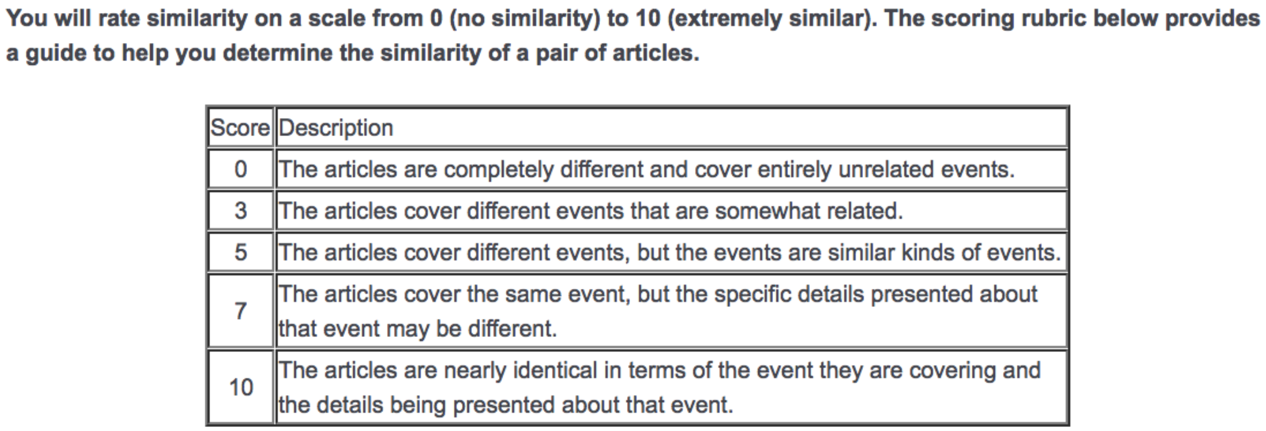
Figure 1: The survey landing page informed participants about the nature of the task.



Thank you for beginning our survey! Please answer every question. If you skip questions, you may not receive payment. **Read each question carefully. Some of them are attention checks.**

We are going to show you a series of pairs of newspaper articles and ask you to rate the similarity of the documents in each pair. We are interested in how similar they are in terms of the stories they are covering. For example, some of the pairs might be about the exact same event, some pairs might be two stories covering similar but distinct events, and some pairs might be about entirely unrelated things.

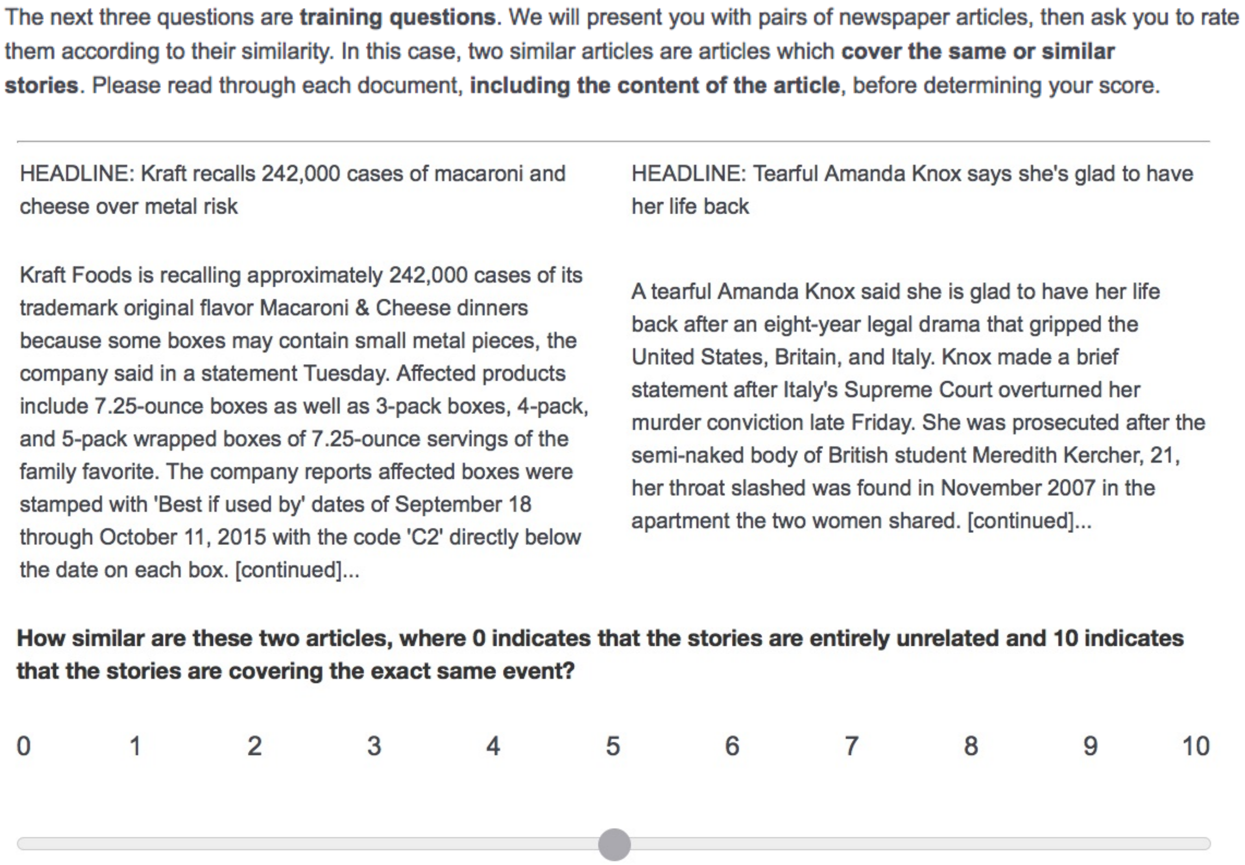
Figure 2: After enrolling in the experiment, participants were presented with a scoring rubric to use as a guide for determining the similarity of a pair of documents.



You will rate similarity on a scale from 0 (no similarity) to 10 (extremely similar). The scoring rubric below provides a guide to help you determine the similarity of a pair of articles.

Score	Description
0	The articles are completely different and cover entirely unrelated events.
3	The articles cover different events that are somewhat related.
5	The articles cover different events, but the events are similar kinds of events.
7	The articles cover the same event, but the specific details presented about that event may be different.
10	The articles are nearly identical in terms of the event they are covering and the details being presented about that event.

Figure 3: In the first training task of the survey, participants were ask to read and score a pair of articles and were then informed that the anticipated score for this pair was zero. Specifically, they were told “We think these articles’ similarity is 0 out of 10. The first article is related to macaroni and cheese, while the second article is about a murder trial.”



The next three questions are **training questions**. We will present you with pairs of newspaper articles, then ask you to rate them according to their similarity. In this case, two similar articles are articles which **cover the same or similar stories**. Please read through each document, **including the content of the article**, before determining your score.

HEADLINE: Kraft recalls 242,000 cases of macaroni and cheese over metal risk

Kraft Foods is recalling approximately 242,000 cases of its trademark original flavor Macaroni & Cheese dinners because some boxes may contain small metal pieces, the company said in a statement Tuesday. Affected products include 7.25-ounce boxes as well as 3-pack boxes, 4-pack, and 5-pack wrapped boxes of 7.25-ounce servings of the family favorite. The company reports affected boxes were stamped with 'Best if used by' dates of September 18 through October 11, 2015 with the code 'C2' directly below the date on each box. [continued]...

HEADLINE: Tearful Amanda Knox says she's glad to have her life back

A tearful Amanda Knox said she is glad to have her life back after an eight-year legal drama that gripped the United States, Britain, and Italy. Knox made a brief statement after Italy's Supreme Court overturned her murder conviction late Friday. She was prosecuted after the semi-naked body of British student Meredith Kercher, 21, her throat slashed was found in November 2007 in the apartment the two women shared. [continued]...

How similar are these two articles, where 0 indicates that the stories are entirely unrelated and 10 indicates that the stories are covering the exact same event?

D Supplemental results from the human evaluation experiment

D.1 Sensitivity of match quality scores to the population of respondents

To determine the generalizability of the match quality ratings obtained from our survey experiment, we compare two identical pilot surveys using respondents from two distinct populations. The first pilot survey was administered through Mechanical Turk, and the second pilot was administered through the Digital Laboratory for the Social Sciences (Enos et al., 2016). For each survey, respondents were asked to read and evaluate ten paired articles, including one attention check and one anchoring question. Each respondent was randomly assigned to evaluate eight matched pairs from a sample of 200, where this pilot sample was generated using the same weighted sampling scheme described above. Figure 4 shows the average match quality scores for each of the 200 matched pairs evaluated based on sample of 337 respondents from Mechanical Turk and 226 respondents from DLABSS. The large correlation between average matched quality scores across samples (ρ=0.88) suggests that our survey is a useful instrument for generating consistent average ratings of match quality across diverse populations of respondents. In particular, even though individual conceptions of match quality may differ across respondents, the average of these conceptions both appears to meaningfully separate the pairs of documents and to be stable across at least two different populations.

D.2 Performance of the predictive model

Figure 5 shows the out-of-sample predictive performance of the model for a distinct sample of 472 pairs of documents evaluated in a separate survey experiment. The correlation of predictions to measured quality for this sample was approximately 94%. In sample correlation

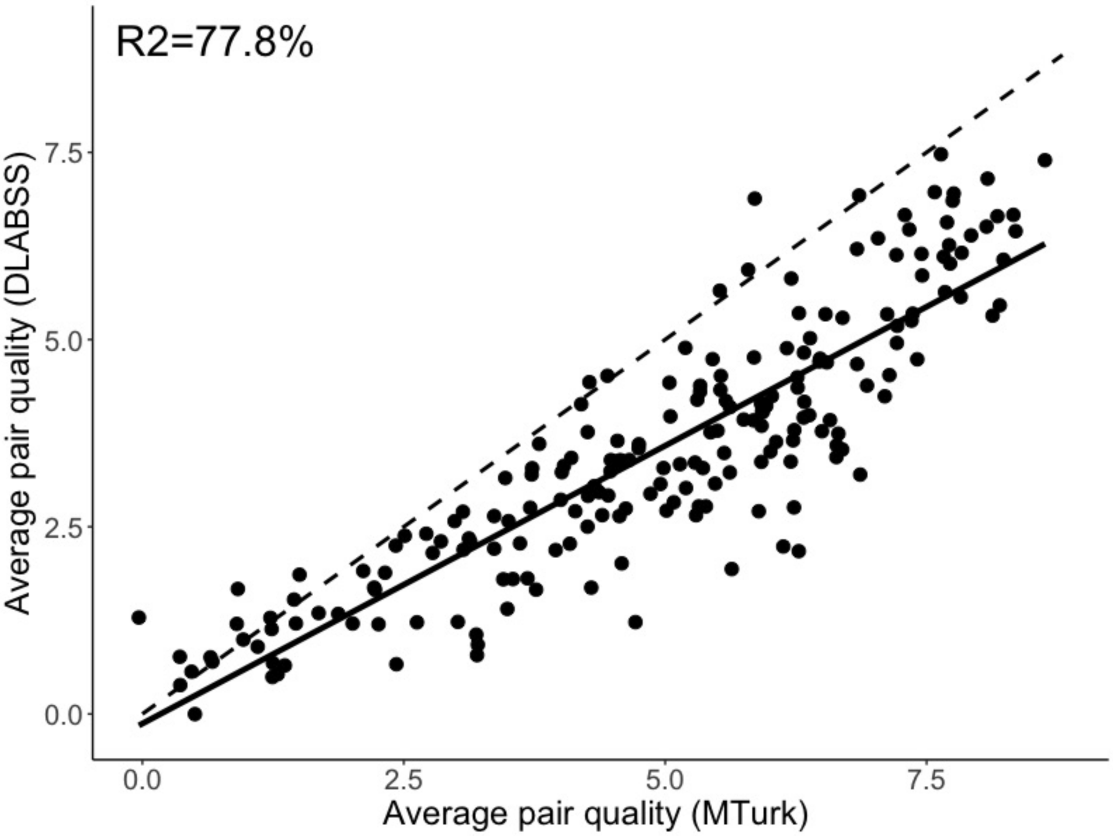


Figure 4: The strong linear relationship between the average match quality scores for 200 pairs of articles evaluated in two separate pilot studies (solid line) compared to a perfect fit (dotted line) suggests that the survey produces consistent results across samples, when averaged across multiple respondents.

was 88% (the stronger out-of-sample correlation is likely driven by a different distribution of matched pairs evaluated).To evaluate the sensitivity of this model to the chosen regularization scheme, we performed a similar analysis using ridge regression and found only a negligible difference in predictive performance.

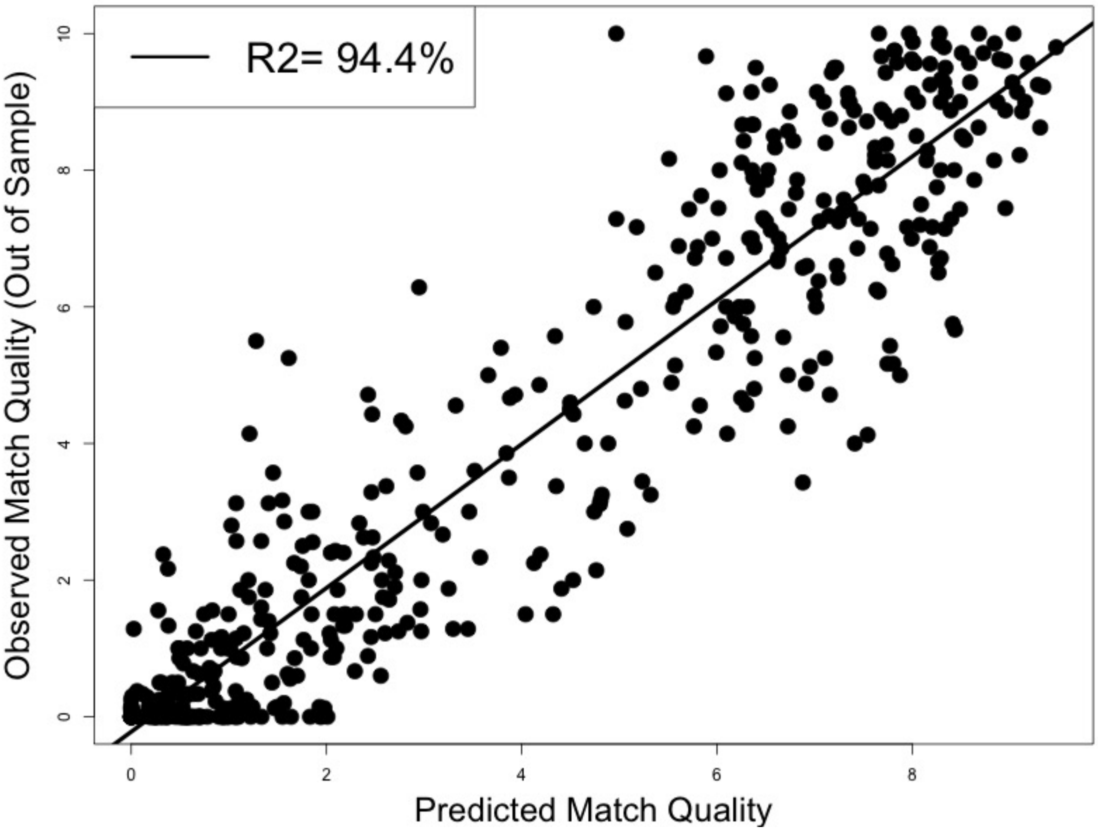


Figure 5: Predictive model for match quality trained on human evaluations has a correlation of 0.944 with observed quality scores obtained in a separate human evaluation experiment on a different set of pairs, indicating high out-of-sample predictive accuracy.

E Technical details of the evaluation of match quality of pairs of news articles

In this section we more fully describe the design and analysis of the human evaluation experiment for the newspaper matching example. We start by discussing how we generated our sampling strategy and weights, and then discuss how we used model-assisted survey sampling to estimate average match quality for the different methods along with associated uncertainty.

E.1 Details of the sampling design

The study presented in this paper is in fact a replication study as our initial study did not directly assess all procedures considered (in particular, we did not initially evaluate the Word2Vec procedures). We therefore designed our second study to both directly extend our findings, verify the prior results, and further investigate the predictive accuracy of our models to out-of-sample pairs. In order to achieve this, we designed a sampling scheme that has three components: (1) we sampled 4 pairs from each procedure considered, (2) we directly sampled pairs that were previously evaluated to assess the stability of the evaluation process, and (3) we sampled pairs not selected by any method to examine differences between selected and non-selected pairs. The first stage sampled pairs with weights based on the predicted quality of the pairs in order to sample predicted high-quality pairs more heavily. We used the prior study's fit predictive model to generate these predictions. The second and third stage sampled a fixed number of pairs within each tier of quality (from 0 to 8+) to see the full range of pair qualities in our sample (simple random sampling would not work since the vast majority of pairs are scored as quality 1 or lower). This overall process resulted in a sample of 505 pairs that fully represents all possible pairs (selected and not). For each pair we have an initial predicted quality score, a sampling probability π_i, and an associated sampling weight $w_i \propto 1/\pi_i$.

Because many of the procedures generally select the same high-quality pairs, the sequential sampling of 4 pairs for each procedure tends to give many of the same pairs back. This is by design, and means that our sample primarily consists of pairs shared by multiple procedures which gives greater precision in estimating these procedures' average quality. We simply take the unique set of pairs sampled as our final evaluation sample.

We calculate the actual sampling weights of each pair for this scheme using simulation. In particular, we conduct our sampling scheme 100,000 times and calculate how often each pair is selected into the sample. These provide (up to monte carlo error) the true selection probabilities π_i; inverting them provides the true sampling weights w_i. For the out-of-matched pairs sampling stage (3), we averaged these final weights across groups of pairs that all have the same probability of selection to increase precision.

The stage (1) sampling scheme intentionally induces selection bias into the sample by discouraging rare pairs, especially singleton pairs, which are expected to be low quality with little variability, in favor of pairs that are identified by multiple matching procedures. Regardless, because the sampling probabilities are fixed a-priori, weighted averages of the pairs' match quality gives good estimates of the average quality of the pairs selected by each procedures; this approach is simply classic survey sampling as described in, e.g.,Sarndal et al. (2003). All this complexity in the sampling design is to ensure that the sample evaluated is targeted to give information on as many procedures as possible, a difficult task when evaluating 130 procedures with a sample size of about 500.

E.2 Estimating pair and procedure quality.

Let $u_{t,c}$ denote a potential pairing of treatment and control documents, where t is the index of the treated unit and c is the index of the control unit. In our evaluation study, $t = 1, \ldots, 1565$ and $c = 1, \ldots, 1796$. For matching procedure j, let $\mathcal{R}_j$ denote the set of n_j matched pairs of articles identified using procedure j. The set of all unique pairs selected by any of the J procedures considered in the evaluation experiment, denoted $\mathcal{R}$, is defined by the union of these subsets:

$$\mathcal{R} = \cup_{j=1}^{J} \mathcal{R}_j.$$

We index the pairs with $i = 1, \ldots, N$.

The frequency of how often each pair u_i in $\mathcal{R}$ was selected by a procedure is:

$$F_i = \sum_{j=1}^{J} 1\{u_i \in \mathcal{R}_j\},$$

where $1\{i \in \mathcal{R}_j\}$ is an indicator variable taking value 1 if pair u_i is identified using matching procedure j and 0 otherwise.

From the human evaluation, we, for each element i of $\mathcal{S}$, where $\mathcal{S}$ is the set of all sampled pairs, observe m_i similarity ratings, $q_{i,1}^{obs}, \ldots, q_{i,m_i}^{obs}$ where $q_{i,\cdot}^{obs} \in [0, 10]$. We estimate the match quality for each evaluated pair i using the average of observed ratings for that pair, $\bar{q}_i^{obs}$.[1]

We wish to estimate, for each procedure, the finite-population quantities of the average true quality of the pairs selected. In particular, if we let q_i be the average quality score score we would see if we had an arbitrarily large number of human respondents evaluate that pair, our targets of inference are, for each procedure j,

$$Q_j = \frac{1}{N_j} \sum_{u_i \in \mathcal{R}_j} q_i.$$

The Q_j are population quantities of how the matching procedure did in the specific context considered. This estimand does not necessarily take into account how the methods would perform on other corpora, even ones similar to this one.

To estimate Q_j for any matching procedure j in our evaluation we use a weighted average of the match quality estimates across the pairs contained in $\mathcal{R}_j \cap \mathcal{S}$, where weights for each

---

[1] We also explored modeling these ratings to account for rater effects and variable number of ratings per question, but as the results were essentially unchanged, elected to use the simple averages.

pair are equal to the inverse probability of being sampled:

$$\hat{Q}_{samp,j} = \frac{1}{Z_j} \sum_{u_i \in \mathcal{R}_j} \frac{1}{\pi_i} S_i \bar{q}_i^{obs} \text{ with } Z_j = \sum_{u_i \in \mathcal{R}_j} \frac{1}{\pi_i} S_i. \tag{1}$$

with S_i an indicator of whether pair i was sampled for evaluation, with sampling probability π_i, and Z_j a normalizing constant. This is a simple Hajék estimator and is known to have good properties.

Unfortunately, despite the sampling scheme, some of our methods only had a small number of pairs sampled for evaluation. Estimating the average match quality for such procedures could therefore be fairly imprecise. We address this by using our model for predicting the match quality of a pair of documents based on different machine measures of similarity to construct model-assisted survey sampling estimators that use the predicted qualities to adjust these estimated average quality scores. We describe this analysis approach next.

E.3 Improving the estimates of procedure quality.

To enhance our predictions of match quality for our procedures, we use a model trained on the pairs in $\mathcal{S}_{pre}$, the sample collected in our initial study, to calculate the predicted match quality, $\hat{q}_i$ for all pairs $i = 1, \dots, N$. These $\hat{q}_i$ are fixed, and do not depend on the analyzed (i.e., second) random sample. We can use these predictions to adjust our estimates of the average quality of all pairs for each procedure using survey sampling methods.

In particular, our model adjusted quality for procedure j is

$$\hat{Q}_{adj,j} = \frac{1}{n_j} \sum_{u_i \in \mathcal{R}_j} \hat{q}_i + \frac{1}{Z_j} \sum_{u_i \in \mathcal{R}_j} S_i \frac{1}{\pi_i} \left(\bar{q}_i^{obs} - \hat{q}_i \right)$$

Here $\hat{q}_i$ is the predicted quality based on the initial sample. Note the first term in the above is a fixed constant, not dependent on the sample. The second term is random, depending on the sample, and, ignoring the small bias induced by Z_j being random, we see the expected value is

$$\begin{aligned}
\mathbb{E}\left[\hat{Q}_{adj,j}\right] &\approx \frac{1}{n_j}\sum_{u_i\in\mathcal{R}_j}\hat{q}_i + \frac{1}{\mathbb{E}\left[Z_j\right]}\sum_{u_i\in\mathcal{R}_j}\mathbb{E}\left[S_i\frac{1}{\pi_i}\left(\bar{q}_i^{obs} - \hat{q}_i\right)\right] \\
&= \frac{1}{n_j}\sum_{u_i\in\mathcal{R}_j}\hat{q}_i + \frac{1}{\mathbb{E}\left[Z_j\right]}\sum_{u_i\in\mathcal{R}_j}\mathbb{E}\left[S_i\right]\frac{1}{\pi_i}\left(\mathbb{E}\left[\bar{q}_i^{obs}\right] - \hat{q}_i\right) \\
&= \frac{1}{n_j}\sum_{u_i\in\mathcal{R}_j} q_i = Q_j.
\end{aligned}$$

This is a *model-adjusted estimate*; the first summation gives the predicted average quality of the method. The second summation adds an adjustment based on the residuals for the actually sampled and evaluated pairs; this adjustment makes the overall estimate effectively unbiased[2] regardless of whether the predictive model is useful, predictive, or even correct. The more the predictive model aligns with the actual measured values, however, the more precise our estimates will be (as the residuals and adjustment part will get smaller and smaller as predictive accuracy grows).

E.4 Uncertainty estimation

Classic survey sampling results allowed us to estimate each procedure's average quality with the estimated qualities of our sampled pairs. We can also increase the precision of these estimates using model adjustment, using the predicted quality scores to adjust the same by population averaged characteristics. In both cases, the next step is to obtain appropriate uncertainty estimates (standard errors) for these point estimates. Unfortunately, the task of appropriately calculating uncertainty in this context for both the raw estimates and the model-adjusted estimates is a surprisingly difficult and subtle problem. In particular, while there are classic survey sampling formula that can be used to calculate uncertainty, they

---

[2] The bias is purely from using a Hajék rather than Horvitz-Thompson estimator, and comes from the normalizing Z_j being a random quantity. It is *not* a function of model misfit or misspecification.

are asymptotic and are sensitive to extreme weights (which we have). This creates some perverse results (i.e. near zero standard errors) for some of the procedures that only had a few pairs sampled. To avoid this we, by instituting a homoscedastic assumption for the error terms, did a parametric simulation to calculate uncertainty in order to work around this problem. This procedure captures the variability induced by the varying sample weights and the measurement error due to the human evaluation. We describe this next.

Uncertainty estimates for the raw quality estimates. For the unadjusted quality measures, we estimate uncertainty using the principles of a case-wise bootstrap with some modifications. In particular, especially for those methods with very few (e.g. 4) sampled pairs, estimating the variability of quality of the pairs via case-wise bootstrap is unreliable unless we pool or partially pool estimates of variability across the different methods.

To see this consider a hypothetical method with 4 of its pairs sampled, 1 with very high weight due to being a rare pair and 3 with a low weight due to being selected by most methods. Any bootstrap sample that includes the high weight unit will essentially give an average quality score close to that of the high weight unit. Even bootstrap samples with multiple draws of the high weight unit will still get nearly that same average quality score since the values of these large elements will all be the same. Across bootstrap samples, this will give low variability, i.e., seemingly high precision. It does not take into account the variability of scores we might have actually seen across other units of similar weight. We address this with the a parametric approach that we describe next.

We first assess the typical variability of the quality scores of pairs within the procedures. For the unadjusted quality scores of the individual pairs we first calculated an estimate of the standard deviation of scores within a given match method (we did this by calculating the weighted standard deviation of scores). We then took the median of these values as our measure of within-method variation of pair quality. We use the median to avoid the impact of the extreme standard deviations due to the methods with small samples of pairs.[3]

---

[3] We actually calculated this (pooled) standard deviation a variety of ways and took the largest to be

To calculate standard errors for our methods, we then simulated the pair sampling step followed by the scoring of sampled pairs step by first selecting pairs using the original sampling strategy, and then generating pseudo-quality scores with the same variance as we generally saw for pairs selected by a method. We then calculated the overall pseudo-quality for each of our methods based on these scores and associated sample weights. Our standard errors are then the standard deviation of these generated overall pseduo-quality scores.

To compare, we also conducted a simple case-wise bootstrap. Here we sampled the evaluated pairs with replacement and calculated each methods' quality score using the bootstrap sample, finally obtaining standard errors using the standard deviation of the resulting values. This approach works well for those methods with 10 or more sampled pairs. Overall, our parametric approach generally produced larger standard errors, which is a mixture of the overall conservatism of our approach and of the aforementioned issue of the naíve approach giving small standard errors those methods with few pairs and a few high-weight pairs that dominate the overall quality measure. We thus report our parametric simulation-based standard errors.

Uncertainty for the model-adjusted approach. For the model-adjusted case, we again worried about those methods with few samples having less variability due to small numbers of high weight units giving nearly the same model adjustment with each step. We therefore follow the above process, but instead of generating synthetic outcomes we generated synthetic residuals by generating normally distributed noise with variance equal to the variance of the original residuals from our predictive model. These simulated residual-based standard errors were again conservative when compared to the naïve case-wise approach for those procedures with enough selected pairs to make this comparison.

Remarks. All our uncertainty estimation methods capture the uncertainty in the pair quality evaluation process as the variability of the pairs' quality scores captures both the

---

maximally conservative.

measurement error and the structural variation of the pairs themselves. In our plots, we report the simulation-based standard errors for the model adjusted estimates. As noted in the text, the model-adjusted quality scores themselves were generally similar to unadjusted (for the directly evaluated methods where we had both scores), and the differences between the two had no impact on our overall findings.

For methods that we did not initially identify for our human evaluation, we could calculate a predicted quality based on our model of

$$\hat{Q}_{pred,j} = \frac{1}{n_j} \sum_{u_i \in \mathcal{R}_j} \hat{q}_i.$$

This is extrapolation, however. If the new procedure was selecting pairs that systematically were better than predicted, for example, this extrapolation would be biased. Even if such a new method happened to use some pairs randomly selected for evaluation, we cannot use the survey adjusted $\hat{Q}_{adj,j}$ or raw estimate $\hat{Q}_{samp,j}$ since the pairs *not* in the sampling frame had no chance of selection. One could create a hybrid estimator by splitting the sample into potentially sampled, but we do not explore that further here.

E.5 Prior evaluation study details

As mentioned above, we performed an initial full study on an initial subset of the matching procedures considered (in particular, we did not initially evaluate the Word2Vec procedures). Overall, this study produced the same results as our final study.

We sampled pairs differently for our initial study. In particular, we did not have baseline predicted quality scores to calculate sampling weights from. We therefore, to produce a representative sample of matched articles for evaluation, did not take a sample from each procedure's pair list but instead took a weighted random sample of 500 pairs from the union of these lists, $\mathcal{R}$, with sampling weights roughly proportional to F_i, where F_i is the number of times pair i was selected by a procedure. Because singleton pairs comprised over 75% of the pairs in $\mathcal{R}$, we further downweighted pairs with $F_i = 1$ by a factor of 5. Our overall sampling probabilities for pair i were then

$$w_i \propto \begin{cases} 0.20 \text{ if } F_i = 1, \\ F_i \text{ otherwise.} \end{cases}$$

We then calculated true sampling probabilities and weights via simulation as described above (due to high weights for some pairs and the sampling without replacement these initially weights are not truly proportional to inverse probability of selection).

F Notes on the sample and unadjusted human experiment results

The final evaluation sample consisted of 33 pairs that were originally evaluated in the initial evaluation, 50 pairs that were not identified by any matching method considered, and 422 pairs that were used by at least one matching method evaluated. The sampling weights for those pairs that were selected by at least one method ranged from 0.02 to 10.7, with a median of 0.23. This corresponded to selection probabilities ranging from 1 in 1000 to 77%. 25% of the pairs had less than a 1% chance of being selected. The very rare pairs tend to come from the propensity score methods that had a large number of low-quality matches. Across procedures, some had only 4 pairs sampled, and some had up to 100. The average was 28 pairs.

The standard deviation of quality scores did depend on the sampling weight, with a standard deviation of around 2.5 for low pi_i and 1 for the highest pi_i. On the other hand, the standard deviation of scores for very low and very high predicted qualities was less than 0.5, rising to around 1.6 for pairs predicted to have a quality of 5. Within a given procedure, scores tended to have a standard deviation of around 2.37, for those procedures with 10 or more pairs sampled. If we look across all procedures the median decreases markedly due to poor estimates for small sample sizes. We used 2.37 in our simulation.

For the residual scores, residuals had a lower standard deviation near the endpoints (due to truncation) and peaked at around 1.6 for the middle scores. We therefore use a residual standard deviation of 1.6 in our simulations to calculate our standard errors, which will be generally conservative. Even with this conservative approximation, we are explaining 55% of our variation with our predictive score.

Figure 6 shows the simple weighted average match quality of the directly evaluated pairs sampled for each of the 130 procedures considered in the evaluation experiment. The nominal 95% confidence intervals are from standard errors calculated from the parametric bootstrap described above.

The standard errors seem small, but some mild calculations suggest they are reasonable. In particular, with 28 pairs, if the pairs have a standard deviation of about 2, we would expect, roughly a standard error of $2/\sqrt{28} = 0.38$, which is what we tended to see. We also point out that we are considering the population of pairs selected by a method as fixed: this is a finite sample inference problem.

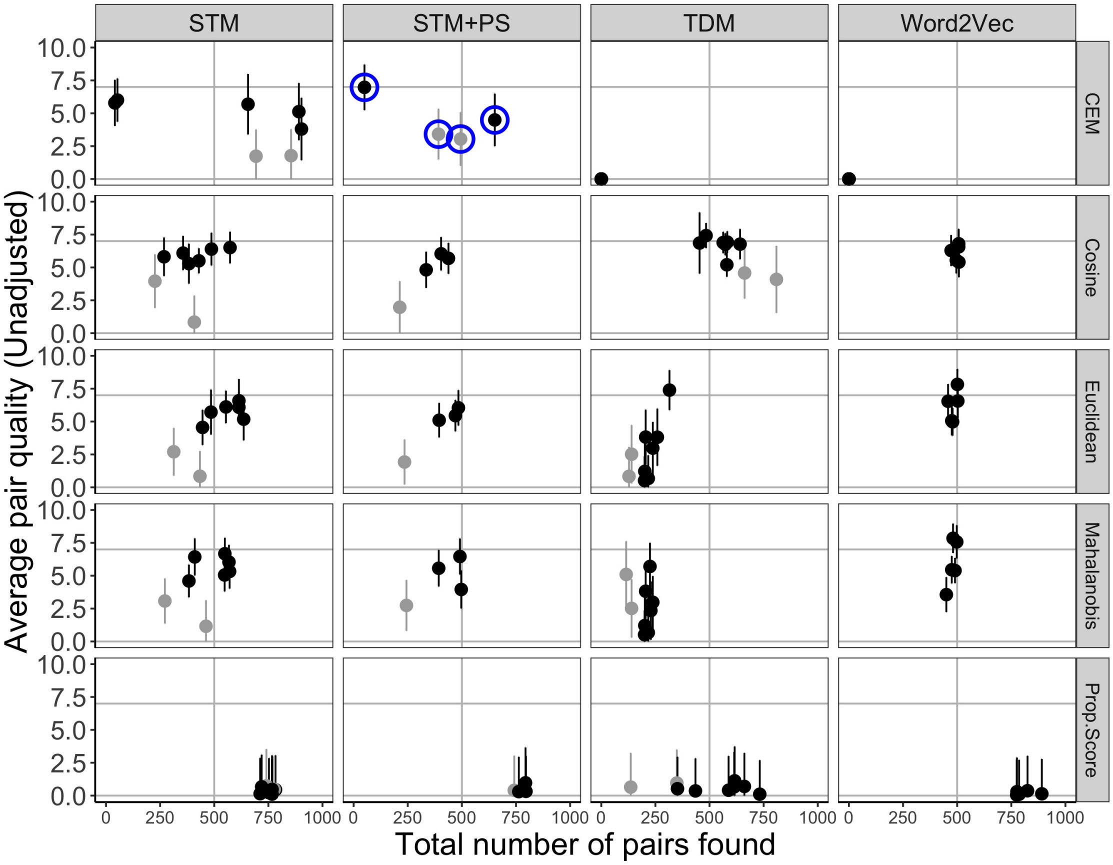


Figure 6: Number of matches found versus estimated (unadjusted) average match quality scores for each combination of matching methods. Grey points indicate procedures with extreme reduction in information (e.g., procedures that match on only stop words). Blue circles highlight procedures that use existing state-of-the-art methods for text matching.

G Template matching and sensitivity analyses for the media bias application

To evaluate the robustness of our findings, we performed a series of sensitivity checks to assess how our results and subsequent conclusions change when using different specifications of the matching procedure. Figure 7 shows the results produced by three alternative text matching methods. These robustness checks highlight the importance of the specification of the matching procedure: weaker methods (i.e., methods that produce low quality matches) typically lead to weaker inferences. For example, the results produced from template matching using the Mahalanobis distance metric on a vector of 100 topic proportions show generally smaller changes in average favorability within each source before and after matching than the results shown in Figure 3 in the main text. The null results in this case provide further evidence in support of the claim that text matching is an effective strategy for reducing differences in the observed biases across news sources that are due to topic selection.

As a final robustness check of the results based on our template-matched sample, we performed the following consistency test. First, we randomly generated 10,000 pairs of documents containing 150 randomly selected articles from each news source. In each iteration of random sampling and for each news source, we then calculated the average favorability scores towards Democrats and Republicans within the matched sample. Figure 8 shows the distributions of these favorability scores for each news source after 100 iterations of random matching. Finally, we calculated the total *change* in favorability observed after matching in each iteration, averaged across all 13 sources. More formally, for each iteration $i = 1, \ldots, 10000$ we calculated the test statistic:

$$T_i = \frac{1}{13} \sum_{j=1}^{13} \left(|\hat{Y}_j^{dem} - \hat{Y}_{j,M_i}^{dem}| + |\hat{Y}_j^{rep} - \hat{Y}_{j,M_i}^{rep}| \right),$$

where $\hat{Y}_j^{dem}$ and $\hat{Y}_j^{rep}$ denote the average favorability scores toward democrats and republicans, respectively, for all articles corresponding to source j in the original, unmatched sample. Quantities $\hat{Y}^{dem}_{j,M_i}$ and $\hat{Y}^{rep}_{j,M_i}$ denote the partisan favorability scores averaged across the set of 150 articles from source j that were selected by random matching in iteration i. The sampling distribution of this test statistic provides a reference for values of the test statistic that may occur when comparing randomly selected sets of 150 articles across these 13 sources. Therefore, by comparing the value of our observed test statistic based on the results of our template-matching procedure described in Section 5 to the randomization distribution defined by T, we can estimate the probability that our template-matched results are due to random chance. Our results from this randomization test indicated that template matching on text removes a significant amount of the bias observed across sources that remains after adjusting for differences in topic selection (p=0.004).

H Results of the systematic evaluation applied to the medical data

Figure 9 shows the average pairwise Jaccard similarity achieved after matching (within propensity score calipers based on the numerical covariates) using each of the 130 text matching specifications described in Section 3. The best identified specification for maximizing the average Jaccard similarity achieved between matched pairs of medical documents, a metric believed to mimic manual evaluation by medical experts, uses a bounded TDM together with the cosine distance metric. Specifically, the best-performing representation is a TDM that is bounded to exclude extremely rare and extremely frequent terms, defined operationally as terms that appear in less than four or more than 1000 documents within this corpus,

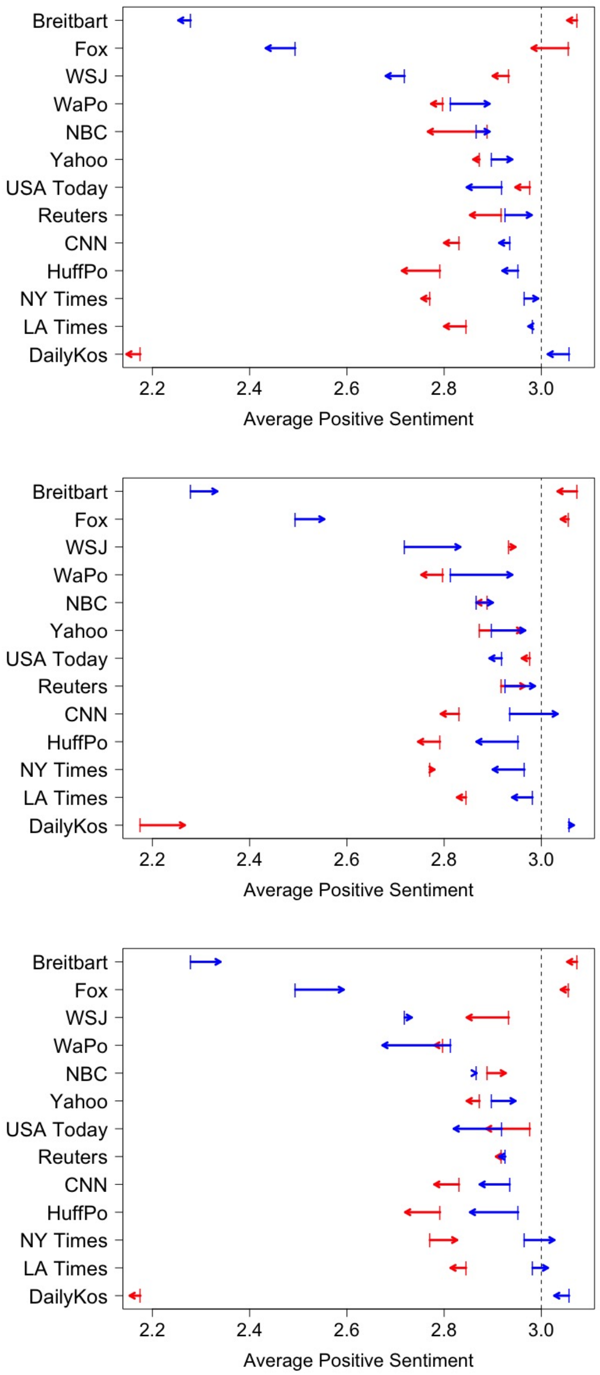


Figure 7: Estimates of average favorability toward Democrats (blue) and Republicans (red) for each source both before and after matching using Mahalanobis matching on an STM with 100 topics (top), propensity score matching on an STM with 100 topics (center) and propensity score matching on a TDM (bottom).

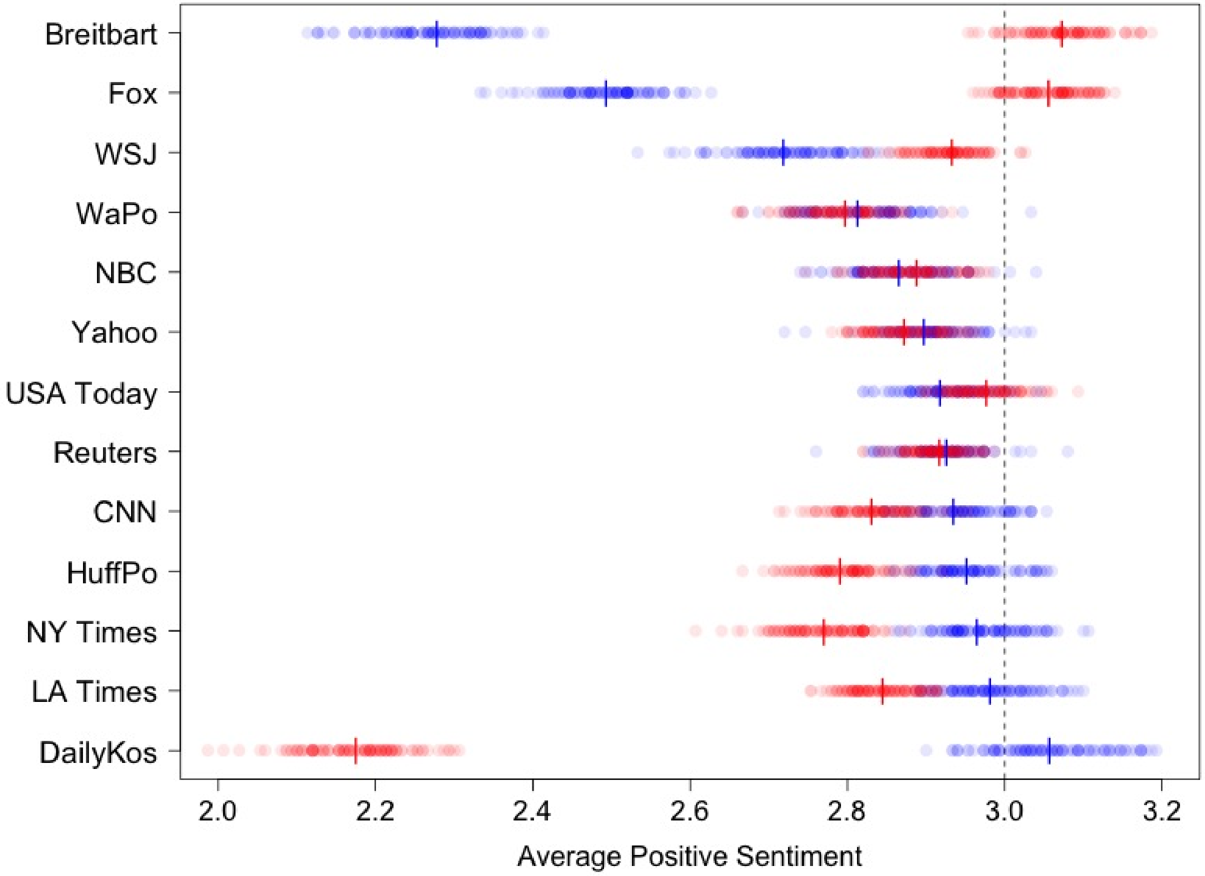


Figure 8: Estimates of average favorability toward Democrats (blue) and Republicans (red) for each source for 100 iterations of random matching. Blue and red lines represent the average favorability scores before matching for Democrats and Republicans, respectively.

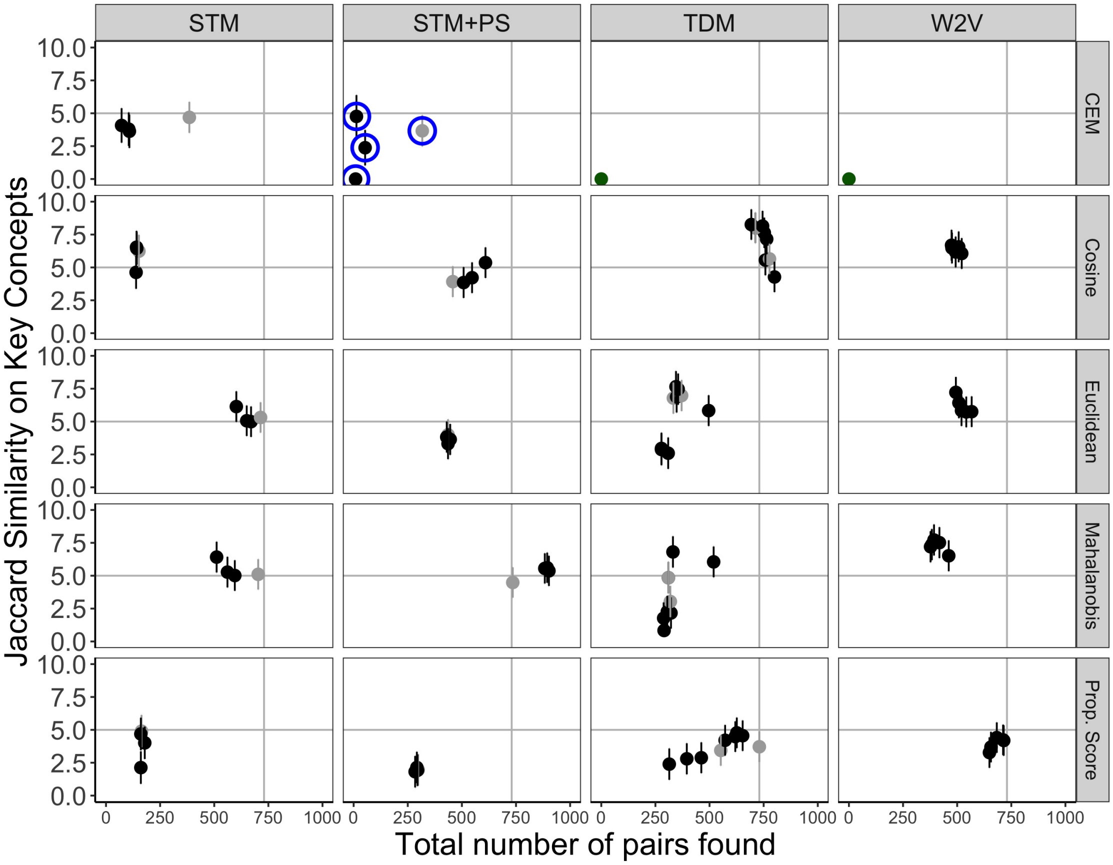


Figure 9: Number of matches found versus average pairwise Jaccard similarity for each combination of matching methods. Grey points indicate procedures with extreme reduction in information (e.g., procedures that match on only stop words). Blue circles highlight procedures that use existing state-of-the-art methods for text matching.

References

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Dai, A. M., Olah, C., and Le, Q. V. (2015). Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.

Denny, M. J. and Spirling, A. (2018). Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis*, pages 1–22.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., and Stewart, B. M. (2017). How to make causal inferences using texts. *arXiv preprint*.

Enos, R. D., Hill, M., and Strange, A. M. (2016). Voluntary digital laboratories for experimental social science: The harvard digital lab for the social sciences. *Working Paper*.

Fan, A., Doshi-Velez, F., and Miratrix, L. (2017). Promoting domain-specific terms in topic models with informative priors. *arXiv preprint arXiv:1701.03227*.

Iacus, S. M., King, G., Porro, G., and Katz, J. N. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, pages 1–24.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016a). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Roberts, M. E., Stewart, B. M., and Nielsen, R. A. (2019). Adjusting for confounding with text matching. *arXiv preprint*.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2016b). Navigating the local modes of big data. *Computational Social Science*, 51.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, pages 159–183.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36.

Salton, G. (1991). Developments in automatic text retrieval. *Science*, pages 974–980.

Salton, G. and McGill, M. J. (1986). *Introduction to modern information retrieval.* McGraw-Hill, Inc.

Sarndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling.* Springer.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770.