Feature Selection as Causal Inference: Experiments with Text Classification

Michael J. Paul
University of Colorado
Boulder, CO 80309, USA
mpaul@colorado.edu

Abstract

This paper proposes a matching technique for learning causal associations between word features and class labels in document classification. The goal is to identify more meaningful and generalizable features than with only correlational approaches. Experiments with sentiment classification show that the proposed method identifies interpretable word associations with sentiment and improves classification performance in a majority of cases. The proposed feature selection method is particularly effective when applied to out-of-domain data.

1 Introduction

A major challenge when building classifiers for high-dimensional data like text is learning to identify features that are not just correlated with the classes in the training data, but associated with classes in a meaningful way that will generalize to new data. Methods for regularization (Hoerl and Kennard, 1970; Chen and Rosenfeld, 2000) and feature selection (Yang and Pedersen, 1997; Forman, 2003) are critical for obtaining good classification performance by removing or minimizing the effects of noisy features. While empirically successful, these techniques can only identify features that are correlated with classes, and these associations can still be caused by factors other than the direct relationship that is assumed.

A more meaningful association is a **causal** one. In the context of document classification using bag-of-words features, we ask the question, which word features "cause" documents to have the class labels that they do? For example, it might be reasonable to claim that adding the word *horrible* to a review would cause its sentiment to become negative, while this is less plausible for a word like *said*. Yet, in one of our experimental datasets of doctor reviews, *said* has a stronger correlation with negative sentiment than *horrible*.

Inspired by methods for causal inference in other domains, we seek to learn causal associations between word features and document classes. We experiment with propensity score matching (Rosenbaum and Rubin, 1985), a technique attempts to mimic the random assignment of subjects to treatment and control groups in a randomized controlled trial by matching subjects with a similar "propensity" to receive treatment. Translating this idea to document classification, we match documents with similar propensity to contain a word, allowing us to compare the effect a word has on the class distribution after controlling for the context in which the word appears. We propose a statistical test for measuring the importance of word features on the matched training data.

We experiment with binary sentiment classification on three review corpora from different domains (doctors, movies, products) using propensity score matching to test for statistical significance of features. Compared to a chi-squared test, the propensity score matching test for feature selection yields superior performance in a majority of comparisons, especially for domain adaptation and for identifying top word associations. After presenting results and analysis in Sections 4–5, we discuss the implications of our findings and make suggestions for areas of language processing that would benefit from causal learning methods.

2 Causal Inference and Confounding

A challenge in statistics and machine learning is identifying causal relationships between variables. Predictive models like classifiers typically learn only correlational relationships between variables, and if spurious correlations are built into a model, then performance will worsen if the underlying distributions change.

A common cause of spurious correlations is **confounding**. A confounding variable is a variable that explains the association between a dependent variable and independent variables. A commonly used example is the positive correlation of ice cream sales and shark attacks, which are correlated because they both increase in warm weather (when more people are swimming). As far as anyone is aware, ice cream does not cause shark attacks; rather, both variables are explained by a confounding variable, the time of year.

There are experimental methods to reduce confounding bias and identify causal relationships. Randomized controlled trials, in which subjects are randomly assigned to a group that receives treatment versus a control group that does not, are the gold standard for experimentation in many domains. However, this type of experiment is not always possible or feasible. (In text processing, we generally work with documents that have already been written: the idea of assigning features to randomly selected documents to measure their effect does not make sense, so we cannot directly translate this idea.)

A variety of methods exist to attempt to infer causality even when direct experiments, like randomized controlled trials, cannot be conducted (Rosenbaum, 2002). In this work, we propose the use of one such method, propensity score matching (Rosenbaum and Rubin, 1985), for reducing the effects of confounding when identifying important features for classification. We describe this method, and its application to text, in Section 3. First, we discuss why causal methods may be important for document classification, and describe previous work in this space.

2.1 Causality in Document Classification

We now discuss where these ideas are relevant to document classification. Our study performs sentiment classification in online reviews using bag-of-words (unigram) features, so we will use examples that apply to this setting.

There are a number of potentially confounding factors in document classification (Landeiro and Culotta, 2016). Consider a dataset of restaurant reviews, in which fast food restaurants have a much lower average score than other types of restaurants. Word features that are associated with fast food, like *drive-thru*, will be correlated with negative sentiment due to this association, even if the word itself has neutral sentiment. In this case, the type of restaurant is a confounding variable that causes spurious associations. If we had a method for learning causal associations, we would know that *drive-thru* itself does not affect sentiment.

What does it mean for a word to have a causal relationship with a document class? It is difficult to give a natural explanation for a bag-of-words model that ignores pragmatics and discourse, but here is an attempt. Suppose you are someone who understands bag-of-words representations of documents, and you are given a bag of words corresponding to a restaurant review. Suppose someone adds the word *terrible* to the bag. If you previously recognized the sentiment to be neutral or even positive, it is possible that the addition of this new word would cause the sentiment to change to negative. On the other hand, it is hard to imagine a set of words to which adding the word *drive-thru* would change the sentiment in any direction.

In this example, we would say that the word *terrible* “caused” the sentiment to change, while *drive-thru* did not. While most real documents will not have a clean interpretation of a word “causing” a change in sentiment, this may still serve as a useful conceptual model for identifying features that are meaningfully associated with class labels.

2.2 Previous Work

Recent studies have used text data, especially social media, to make causal claims (Cheng et al., 2015; Reis and Culotta, 2015; Pavalanathan and Eisenstein, 2016). The technique we use in this work, propensity score matching, has recently been applied to user-generated text data (Rehman et al., 2016; De Choudhury and Kiciman, 2017).

For the task of document classification specifically, Landeiro and Culotta (2016) experiment with multiple methods to make classifiers robust to confounding variables such as gender in social media and genre in movie reviews. This work requires confounding variables to be identified and included explicitly, whereas our proposed method requires only the features used for classification.

Causal methods have previously been applied to feature selection (Guyon et al., 2007; Cawley, 2008; Aliferis et al., 2010), but not with the match-

People	Text
Subject	Document
Treatment	Word
Outcome	Class label

Table 1: A mapping of standard terminology of randomized controlled trials (left) to our application of these ideas to text classification (right).

ing methods proposed in this work, and not for document classification.

3 Propensity Score Matching for Document Classification

Propensity score matching (PSM) (Rosenbaum and Rubin, 1985) is a technique that attempts to simulate the random assignment of treatment and control groups by matching treated subjects to untreated subjects that were similarly likely to be in the same group. This is centered around the idea of a **propensity score**, which Rosenbaum and Rubin (1983) define as the probability of being assigned to a treatment group based on observed characteristics of the subject, $P(z_i|x_i)$, typically estimated with a logistic regression model. In other words, what is the "propensity" of a subject to obtain treatment? Subjects that did and did not receive treatment are matched based their propensity to receive treatment, and we can then directly compare the outcomes of the treated and untreated groups.

In the case of document classification, we want to measure the effect of each word feature. Using the terminology above, each word is a "treatment" and each document is a "subject". Each word has a treatment group, the documents that contain the word, and a "control" group, the documents that do not. The "outcome" is the document class label.

Each subject has a propensity score for a treatment. In document classification, this means that each document has a propensity score for each word, which is the probability that the word would appear in the document. For a word w, we define this as the probability of the word appearing given all other words in the document: $P(w|\mathbf{d_i} - \{w\})$, where $\mathbf{d_i}$ is the set of words in the ith document. We estimate these probabilities by training a logistic regression model with word features.

Using our example from the previous section, the probability that a document contains the word *drive-thru* is likely to be higher in reviews that describe fast food that those that do not. Matching reviews based on their likelihood of containing this word should adjust for any bias caused by the type of restaurant (fast food) as a confounding variable. This is done without having explicitly included this as a variable, since it will implicitly be learned when estimating the probability of words associated with fast food, like *drive-thru*.

3.1 Creating Matched Samples

Once propensity scores have been calculated, the next step is to match documents containing a word to documents that do not contain the word but have a similar score. There are a number of strategies for matching, summarized by Austin (2011a). For example, matching could be done one-to-one or one-to-many, sampling either with or without replacement. Another approach is to group similar scoring samples into strata (Cochran, 1968).

In this work, we perform one-to-one matching without replacement using a greedy matching algorithm; Gu and Rosenbaum (1993) found no quality difference using greedy versus optimal matching. We also experiment with thresholding how similar two scores must be to match them.

Implementation Even greedy matching is expensive, so we use a fast approximation. We place documents into 100 bins based on their scores (e.g., scores between .50 and .51). For each "treatment" document, we match it to the approximate closest "control" document by pointing to the treatment document's bin and iterating over bins outward until we find the first non-empty bin, and then select a random control document from that bin. Placing documents into bins is related to stratification approaches (Rosenbaum and Rubin, 1984), except that we use finer bins that typical strata and we still return one-to-one pairs.

3.1.1 Comparing Groups

Since our instances are paired (after one-to-one matching), we can use McNemar's test (McNemar, 1947), which tests if there is a significant change in the distribution of a variable in response to a change in the other. The test statistic is:

$$\chi^2 = \frac{(TN - CP)^2}{TN + CP} \quad (1)$$

where TN is the number of treatment instances with a negative outcome (in our case, the number of documents containing the target word with a negative sentiment label) and CP is the number of control instances with a positive outcome (the

	# documents	# tokens	# word types
Doctors	20,000	432,636	2,422
Movies	50,000	9,420,645	3,124
Products	100,000	7,416,381	2,343

Table 2: Corpus summary.

number of documents that do not contain the word with a positive sentiment label).

This test statistic has a chi-squared distribution with 1 degree of freedom. This test is related to a traditional chi-squared test used for feature selection (which we compare to experimentally in Section 4), except that it assumes paired data with a "before" and "after" measurement. In our case, we do not have two outcome measurements for the same subject, but we have two subjects that have been matched in a way that approximates this.

We perform this test for every feature (every word in the vocabulary). The goal of the test is to measure there is a significant difference in the class distribution (positive versus negative, in the case of sentiment) in documents that do and do not contain the word (the "after" and "before" conditions, respectively, when considering words as treatments).

4 Experiments with Feature Selection

To evaluate the ability of propensity score matching to identify meaningful word features, we use it for feature selection (Yang and Pedersen, 1997) in sentiment classification (Pang and Lee, 2004).

4.1 Datasets

We used datasets of reviews from three domains:

- *Doctors:* Doctor reviews from RateMDs.com (Wallace et al., 2014). Doctors are rated on a scale from 1–5 along four different dimensions (knowledgeability, staff, helpfulness, punctuality). We averaged the four ratings for each review and labeled a review positive if the average rating was ≥ 4 and negative if ≤ 2.
- *Movies:* Movie reviews from IMDB (Maas et al., 2011). Movies are rated on a scale from 1–10. Reviews rated ≥ 7 are labeled positive and reviews rated ≤ 4 are labeled negative.
- *Products:* Product reviews from Amazon (Jindal and Liu, 2008). Products are rated on a scale from 1–5, with reviews rated ≥ 4 labeled positive and reviews rated ≤ 2 labeled negative.

All datasets were sampled to have an equal class balance. We used unigram word features. For efficiency reasons (a limitation that is discussed in Section 7), we pruned the long tail of features, removing words appearing in less than 0.5% of each corpus. The sizes of the processed corpora and their vocabularies are summarized in Table 2.

Training Corpus	Test Corpus					
	Doctors		*Movies*		*Products*	
	PSM	χ^2	PSM	χ^2	PSM	χ^2
Doctors	**.8569**	.8560	**.6796**	.6657	**.6670**	.6367
Movies	**.6510**	.5497	**.8094**	.7421	**.6658**	.4917
Products	.7799	**.7853**	**.8299**	.8245	.8234	**.8277**

Table 3: Area under the feature selection curve (see Figure 1) using F1-score as the evaluation metric. All differences between corresponding PSM and χ^2 results are statistically significant with $p \ll 0.01$ except for (*Doctors*, *Doctors*).

4.2 Experimental Details

For each corpus, we randomly selected 50% for training, 25% for development, and 25% for testing. The training set is used for training classifiers as well as calculating all feature selection metrics.

We used the development set to measure classification performance for different hyperparameter values. Our propensity score matching method has two hyperparameters. First, when building logistic regression models to estimate the propensity scores, we adjusted the ℓ_2 regularization strength. Second, when matching documents, we required the difference between scores to be less than $\tau \times SD$ to count as a match, where SD is the standard deviation of the propensity scores. We performed a grid search over different values of τ and different regularization strengths, described more in our analysis in Section 5.2, and used the best combination of hyperparameters for each dataset.

We used logistic regression classifiers for sentiment classification. While we experimented with ℓ_2 regularization for constructing propensity scores, we used no regularization for the sentiment classifiers. Since regularization and feature selection are both used to avoid overfitting, we did not want to conflate the effects of the two, so by using unregularized classifiers we can directly assess the efficacy of our feature selection methods on held-out data. All models were implemented with `scikit-learn` (Pedregosa et al., 2011).

Baseline We compare propensity score matching with McNemar's test **(PSM)** to a standard chi-squared test **(χ^2)** for feature selection, one of the

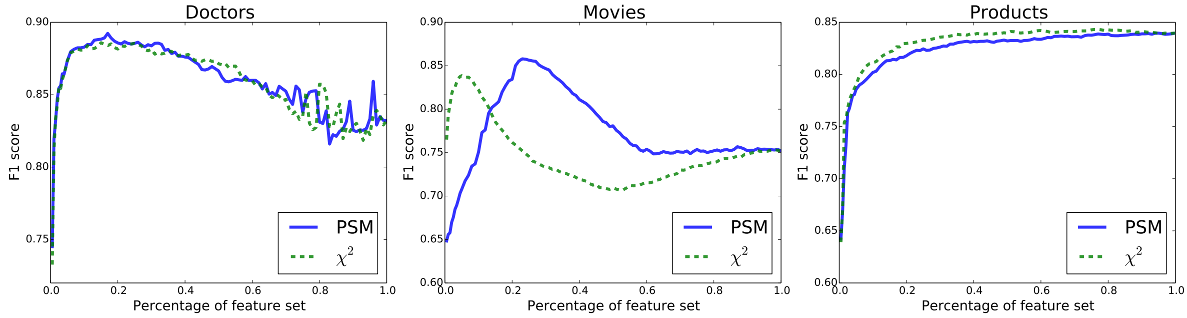

Figure 1: F1 scores when using a varying numbers of features ranked by two feature selection tests.

most common statistical tests for features in document classification (Manning et al., 2008). Since both tests follow a chi-squared distribution, and since McNemar's test is loosely like a chi-squared test for paired data, we believe this baseline offers the most direct comparison.

4.3 Results

We calculated the F1 scores of the sentiment classifiers when using different numbers of features ranked by significance. For example, when training a classifier with 1% of the feature set, this is the most significant 1% (with the lowest p-values). Results for varying feature set sizes on the three test datasets are shown in Figure 1.

To summarize the curves with a concise metric, we calculated the area under these curves (AUC). AUC scores for each dataset can be found along the diagonal of Table 3. We find that PSM gives higher AUC scores than χ^2 in two out of three datasets, though one is not statistically significant based on a paired t-test of the F1 scores.

PSM gives a large improvement over χ^2 on the *Movies* corpus, though the feature selection curve is unusual in that it rises gradually and peaks much later than χ^2. This appears to be because the highest ranking words with PSM have mostly positive sentiment. There is a worse balance of class associations in the top features with PSM than χ^2, so the classifier has a harder time discriminating with few features. However, PSM eventually achieves a higher score than the peak from χ^2 and the performance does not drop as quickly after peaking.

In the next two subsections, we examine additional settings in which PSM offers larger advantages over the χ^2 baseline.

4.3.1 Generalizability

A motivation for learning features with causal associations with document classes is to learn robust features that can generalize to changes in the data distribution. To test this, we evaluated each of the three classifiers on the other two datasets (for example, testing the classifier trained on *Doctors* on the *Products* dataset). The AUC scores for all pairs of datasets are shown in Table 3.

On average, PSM improves the AUC over χ^2 by an average of .021 when testing on the same domain as training, while the improvement increases to an average of .053 when testing on out-of-domain data. In thus seems that PSM may be particularly effective at identifying features that can be applied across domains.

4.3.2 Top Features

Having measured performance across the entire feature set, we now focus on only the most highly associated features. The top features are important because these can give insights into the classification task, revealing which features are most associated with the target classes. Having top features that are meaningful and interpretable will lead to more trust in these models (Paul, 2016), and iden-

Doctors		*Movies*		*Products*	
PSM	χ^2	PSM	χ^2	PSM	χ^2
great	told	great	worst	excellent	waste
caring	great	excellent	bad	wonderful	money
rude	rude	wonderful	and	great	great
best	best	best	great	waste	worst
excellent	said	love	waste	bad	best

Table 4: The highest scoring words from the two feature selection methods.

	$M = 5$		$M = 10$		$M = 20$	
	PSM	χ^2	PSM	χ^2	PSM	χ^2
Doctors	**.5573**	.4806	**.6318**	.5520	**.6999**	.6503
Movies	**.5211**	.4962	.5841	**.6196**	.6171	**.6921**
Products	**.5388**	.3478	**.5514**	.4696	**.6031**	.5622

Table 5: Area under the feature selection curve when using only a small number of features, M.

tifying meaningful features can itself be the goal of a study (Eisenstein et al., 2011b).

We experimented with a small number of features $M \in \{5, 10, 20\}$. Under the assumption that optimal hyperparameters may be different when using such a small number of features, we retuned the PSM parameters again for the experiments in this subsection, using $M{=}10$.

Table 4 shows the five words with the lowest p-values with both methods. At a glance, the top words from PSM seem to have strong sentiment associations; for example, *excellent* is a top five feature in all three datasets using PSM, and none of the datasets using χ^2. Words without obvious sentiment associations seem to appear more often in the top χ^2 features, like *and*.

To quantify if there is a difference in quality, we again calculated the area under the feature selection F1 curves, where the number of features ranged from 1 to M. Results are shown in Table 5. For M of 10 and 20, PSM does worse on *Movies*, which is not surprising based on our finding above that the top features in this dataset are not balanced across the two labels, so PSM does worse for smaller numbers of features. For the other two datasets, PSM substantially outperforms χ^2. PSM appears to be an effective method for identifying strong feature associations.

5 Empirical Analysis

We now perform additional analyses to gain a deeper understanding of the behavior of propensity score matching applied to feature selection.

5.1 An Example

To better understand what happens during matching, we examined the word *said* on the *Doctors* corpus. This word does not have an obvious sentiment association, but is the fifth-highest scoring word with χ^2. It is still highly ranked when using propensity score matching, but this approach reduces its rank to ten.

Upon closer inspection, we find that reviews tend to use this word when discussing logistical issues, like interactions with office staff. These issues seem to be discussed primarily in a negative context, giving *said* a strong association with negative sentiment. If, however, reviews that discussed these logistical issues were matched, then within these matched reviews, those containing *said* are probably not more negative than those that do not. With propensity score matching, documents are matched based on how likely they are to contain the word *said*, which is meant to control for the negative context that this word has a tendency (or propensity) to appear in.

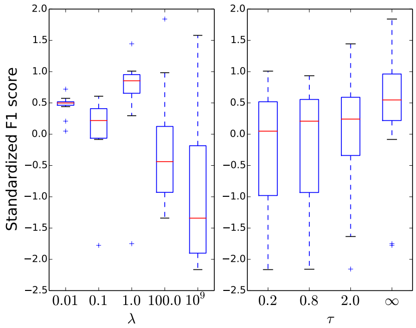

Figure 2: The distribution of the area under the feature selection curve scores when using different hyperparameter settings (propensity inverse regularization strength λ and matching threshold τ).

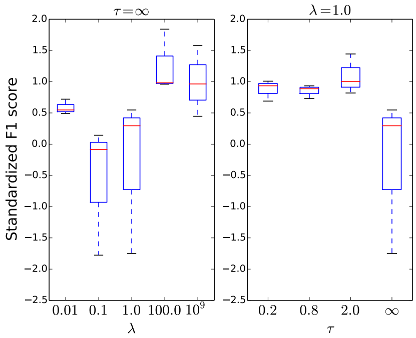

Figure 3: The distribution of scores when using different hyperparameter settings, restricted to the best performing setting for each independent parameter as shown in Figure 2 (varying λ with the optimal τ, and varying τ with the optimal λ).

Table 6 shows example reviews that do (the "treatment" group) and do not (the "control" group) contain *said*. We see that the higher propensity reviews do tend to discuss issues like receptionists and records, and controlling for this context may explain why this method produced a lower ranking for this word.

5.2 Hyperparameter Settings

We investigate the effect of different hyperparameter settings. To do this, we first standardized the results across the three development datasets by converting them to z-scores so that they can be directly compared. The distribution of scores (specifically, the area under the F1 curve scores from Table 3) is summarized in Figure 2.

"Treatment"		"Control"	
High Propensity			
.8040 $-$	She repeatedly said, "I don't care how you feel" when my wife told her the medication (birth control) was causing issues. She failed to mention a positive test result, giving a clean bill of health.	.7880 $-$	After a long, long conversation during which I tried to explain that I did not have records as I was only looked at by a sport trainer, they still would not see me without previous records.
.6320 $-$	I went for a checkup and he ended up waiting for over 2 hours just to get into the room. Then I waited some more until he eventually came in and dedicated the whole 10 minutes of his time. When I asked what exactly is going to take place, the assistant said, no big deal, just a little scrape.	.5047 $+$	The receptionist was able to get me in the next day and really worked around my busy schedule. I downloaded my paperwork off the website and had it ready at my appointment. I waited maybe 10 minutes and was in the exam room. The doctor was really nice and took the time to talk to me.
Low Propensity			
.2012 $+$	I said he was on time but usually you have to wait because he does procedures in all hospitals in town, has emergencies and runs a little late. No matter how busy he is, he greets you warmly and chats with you.	.1959 $-$	For over a week I was going to the pharmacy every day after being told by her staff that it had been called in. Finally after a week then told she would not call it in, I had to come in to see her!
.0597 $-$	This doctor did not do what he said he would, was massively late, unwilling to talk to us about the condition we were facing.	.0598 $-$	DR.Taylor is usually not around. Staff is rude and antagonistic. They do not care about you as a person or your children.

Table 6: Examples of reviews that were matched based on the word *said*. Reviews on the left contain the word *said* while those on the right do not. Each row corresponds to a pair of matched documents (edited for length). The propensity score and sentiment label ($+$ or $-$) is shown for each document.

Regularization When training the logistic regression model to create propensity scores, we experimented with the following values of the inverse regularization parameter: $\lambda \in \{0.01, 0.1, 1.0, 100.0, 10^9\}$, where $\lambda{=}10^9$ is essentially no regularization other than to keep the optimal parameter values finite. We make two observations. First, high λ values (less regularization) generally result in worse scores. Second, small λ values lead to more consistent results, with less variance in the score distribution. Based on these results, we recommend a value of $\lambda{=}1.0$ based on its high median score, competitive maximum score, and low variance.

Matching We required that the scores of two documents were within $\tau{\times}SD$ of each other, and experimented with the following thresholds: $\tau \in \{0.2, 0.8, 2.0, \infty\}$. Austin (2011b) found that $\tau{=}0.2$ was optimal for continuous features and $\tau{=}0.8$ was optimal for binary features. Based on these guidelines, 0.8 would be appropriate for our scenario, but we also compared to a larger threshold (2.0) and no threshold (∞). We find that scores consistently increase as τ increases.

Coupling Looking at the two hyperparameters independently does not tell the whole story, due to interactions between the two. In particular, we observe that lower thresholds (lower τ) work better when using heavier regularization (lower λ), and vice versa. It turns out that it is ill-advised to use $\tau{=}\infty$, as Figure 2 would suggest, when using our recommendation of $\lambda{=}1.0$. Figure 3 shows the λ distribution when set to $\tau{=}\infty$ and the τ distribution when set to $\lambda{=}1.0$. This shows that when $\lambda{=}1.0$, scores are much worse when $\tau{=}\infty$. When $\tau{=}\infty$, scores are better with higher λ values.

The best combinations of hyperparameters are $(\lambda = 100.0, \tau = \infty)$ and $(\lambda = 1.0, \tau = 2.0)$. Between these, we recommend $(\lambda = 1.0, \tau = 2.0)$ due to its higher median and lower variance.

5.3 P-Values

Lastly, we examine the p-values produced by McNemar's test on propensity score matched data compared to the standard chi-squared test. Figure 4 shows the distribution of the log of the p-values from both methods, using the same hyperparameters as in Section 4.3. We find that χ^2 tends to assign lower p-values, with more extreme values. This suggests that propensity score matching yields more conservative estimates of the statistical significance of features.

6 Related Work

In addition to the prior work already discussed, we wish to draw attention to work in related areas with respect to text classification.

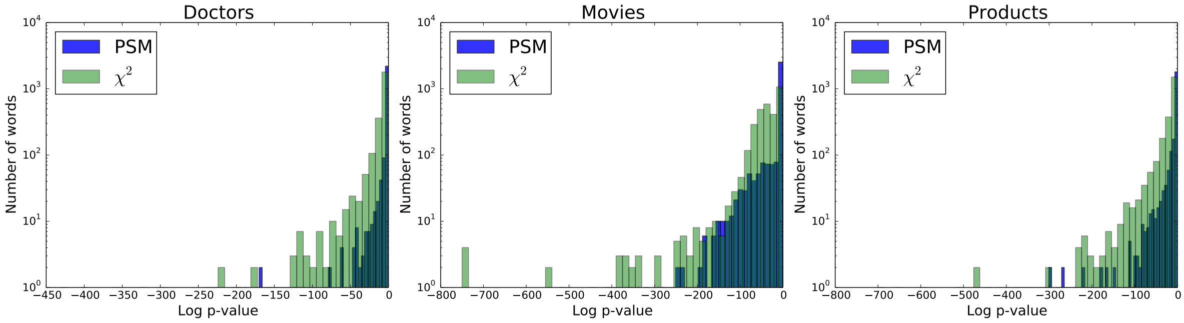

Figure 4: Distribution of p-values of features from the two methods of testing. Counts are on a log scale.

Matching There have been instances of using matching techniques to improve text training data. Tan et al. (2014) built models to estimate the number of retweets of Twitter messages and addressed confounding factors by matching tweets of the same author and topic (based on posting the same link). Zhang et al. (2016) built classifiers to predict media coverage of journal articles used matching sampling to select negative training examples, choosing articles from the same journal issue. While motivated differently, contrastive estimation (Smith and Eisner, 2005) is also related to matching. In contrastive estimation, negative training examples are synthesized by perturbing positive instances. This strategy essentially matches instances that have the same semantics but different syntax.

Annotation Perhaps the work that most closely gets at the concept of causality in document classification is work that asks for annotators to identify which features are important. There are branches of active learning which ask annotators to label not only documents, but to label features for importances or relevance (Raghavan et al., 2006; Druck et al., 2009). Work on annotator rationales (Zaidan et al., 2007; Zaidan and Eisner, 2008) seeks to model **why** annotators labeled a document a certain way—in other words, what "caused" the document to have its label? These ideas could potentially be integrated with causal inference methods for document classification.

7 Future Work

Efficiency is a drawback of the current work. The standard way of defining propensity scores with logistic regression models is not designed to scale to the large number of variables used in text classification. Our proposed method is slow because it requires training a logistic regression model for every word in the vocabulary. Perhaps documents could instead be matched based on another metric, like cosine similarity. This would match documents with similar context, which is what the PSM method appears to be doing based on our analysis.

We emphasize that the results of the PSM statistical analysis could be used in ways other than using it to select features ahead of training, which is less common today than doing feature selection directly through the training process, for example with sparse regularization (Tibshirani, 1994; Eisenstein et al., 2011a; Yogatama and Smith, 2014). One way to integrate PSM with regularization would be to use each feature's test statistic to weight its regularization penalty, discouraging features with high p-values from having large coefficients in a classifier.

In general, we believe this work shows the utility of controlling for the context in which features appear in documents when learning associations between features and classes, which has not been widely considered in text processing. Prior work that used matching and related techniques for text classification was generally motivated by specific factors that needed to be controlled for, but our study found that a general-purpose matching approach can also lead to better feature discovery. We want this work to be seen not necessarily as a specific prescription for one method of feature selection, but as a general framework for improving learning of text categories.

8 Conclusion

We have introduced and experimented with the idea of using propensity score matching for document classification. This method matches documents of similar propensity to contain a word as a way to simulate the random assignment to treatment and control groups, allowing us to more reliably learn if a feature has a significant, causal effect on document classes. While the concept of causality does not apply to document classification as naturally as in other tasks, the methods used for causal inference may still lead to more interpretable and generalizable features. This was evidenced by our experiments with feature selection using corpora from three domains, in which our proposed approach resulted in better performance than a comparable baseline in a majority of cases, particularly when testing on out-of-domain data. In future work, we hope to consider other metrics for matching to improve the efficiency, and to consider other ways of integrating the proposed feature test into training methods for text classifiers.

References

C.F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X.D. Koutsoukos. 2010. Local causal and markov blanket induction for causal discovery and feature selection for classification. *Journal of Machine Learning Research* 11:171–234.

P.C. Austin. 2011a. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 46(3):399–424.

P.C. Austin. 2011b. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat* 10(2):150–161.

G.C. Cawley. 2008. Causal & non-causal feature selection for ridge regression. In *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*.

S.F. Chen and R. Rosenfeld. 2000. A survey of smoothing techniques for maximum entropy models. *IEEE Transactions on Speech and Audio Processing* 8(1):37–50.

J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. 2015. Antisocial behavior in online discussion communities. In *International Conference on Web and Social Media (ICWSM)*.

W.G. Cochran. 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24:295–313.

M. De Choudhury and E. Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *International Conference on Web and Social Media (ICWSM)*.

G. Druck, B. Settles, and A. McCallum. 2009. Active learning by labeling features. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

J. Eisenstein, A. Ahmed, and E.P. Xing. 2011a. Sparse additive generative models of text. In *International Conference on Machine Learning (ICML)*.

J. Eisenstein, N.A. Smith, and E.P. Xing. 2011b. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the Association for Computational Linguistics (ACL)*.

G. Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research* 3:1289–1305.

X.S. Gu and P.R. Rosenbaum. 1993. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* 2:405–420.

I. Guyon, C. Aliferis, and A. Elisseeff. 2007. Causal feature selection. In H. Liu and H. Motoda, editors, *Computational Methods of Feature Selection*, Chapman and Hall/CRC Press.

A.E. Hoerl and R.W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12:55–67.

N. Jindal and B. Liu. 2008. Opinion spam and analysis. In *International Conference on Web Search and Data Mining (WSDM)*.

V. Landeiro and A. Culotta. 2016. Robust text classification in the presence of confounding bias. In *AAAI*.

A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts. 2011. Learning word vectors for sentiment analysis. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

C.D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Q. McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2):153–157.

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL)*.

M.J. Paul. 2016. Interpretable machine learning: lessons from topic modeling. In *CHI Workshop on Human-Centered Machine Learning*.

U. Pavalanathan and J. Eisenstein. 2016. Emoticons vs. emojis on Twitter: A causal inference approach. In *AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and

E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

H. Raghavan, O. Madani, and R. Jones. 2006. Active learning with feedback on features and instances. *J. Mach. Learn. Res.* 7:1655–1686.

N.A. Rehman, J. Liu, and R. Chunara. 2016. Using propensity score matching to understand the relationship between online health information sources and vaccination sentiment. In *AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content*.

V.L.D. Reis and A. Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health from Twitter. In *AAAI*.

P.R. Rosenbaum. 2002. *Observational Studies*. Springer-Verlag.

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

P.R. Rosenbaum and D.B. Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79:516–524.

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

N.A. Smith and J. Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the Association for Computational Linguistics (ACL)*.

C. Tan, L. Lee, and B. Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

R. Tibshirani. 1994. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58:267–288.

B.C. Wallace, M.J. Paul, U. Sarkar, T.A. Trikalinos, and M. Dredze. 2014. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association* 21(6):1098–1103.

Y. Yang and J.O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*.

D. Yogatama and N.A. Smith. 2014. Linguistic structured sparsity in text categorization. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

O.F. Zaidan and J. Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of EMNLP 2008*. pages 31–40.

O.F. Zaidan, J. Eisner, and C. Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *NAACL HLT 2007; Proceedings of the Main Conference*. pages 260–267.

Y. Zhang, E. Willis, M.J. Paul, N. Elhadad, and B.C. Wallace. 2016. Characterizing the (perceived) newsworthiness of health science articles: A data-driven approach. *JMIR Med Inform* 4(3):e27.