Reducing confounding bias in observational studies that use text classification

Virgile Landeiro and **Aron Culotta**
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
vlandeir@hawk.iit.edu, aculotta@iit.edu

Abstract

As text classifiers become increasingly used in observational studies, it is critical to consider not only their accuracy but also their robustness to changes in the data distribution. In this paper, we consider the case where there is a confounding variable Z that influences both the text features W and the class variable Y. For example, a classifier trained to predict the health status of a user based on their online communications may be confounded by socioeconomic variables. When the influence of Z changes from training to testing data, we find that classifier accuracy can degrade rapidly. Our approach, based on Pearl's back-door adjustment, estimates the underlying effect of a text variable on the class variable while controlling for the confounding variable. We conduct an observational study to estimate the effect of location on dispositional affect, with gender as a confounder. We find that our adjustment results in more accurate estimates of effect sizes over a range of possible confounding strengths.

1 Introduction

Emerging cross-disciplinary fields like computational social science (Lazer et al. 2009) and computational epidemiology (Marathe and Ramakrishnan 2013) have begun performing observational studies in which variables are inferred using text classification. This includes diverse applications such as public health surveillance (Dredze 2012), political science (Dahllöf 2012), crisis response (Verma et al. 2011), and marketing (Chamlertwat et al. 2012).

To ensure the validity of such studies, one often must control for possible confounding variables (e.g., socioeconomic status). While this is common practice in social sciences, there has been little work studying how to control for confounding variables when the outcome variable is estimated by a classification algorithm. For example, a classifier trained to predict the political affiliation of a Twitter user may be confounded by an unobserved age variable. This may inflate the coefficients of age-related terms in the classifier (a result of *omitted-variable bias* (Clarke 2005)).

While identifying and controlling for confounding variables is central to much of empirical social science, it is mostly overlooked in text classification, presumably because *prediction*, rather than causal inference, is the primary goal. Indeed, if we assume that the confounding variable's influence is consistent from training to testing data, then there should be little harm to prediction accuracy. However, this assumption often does not hold in practice, for at least two reasons. First, due to the cost of annotation, training sets are typically quite small, increasing the chance that the correlation between the confounding variable and target variable varies from training to testing data. Second, and in our view most importantly, in many domains the relationship between the confounder and the target variable is likely to shift over time leading to poor accuracy. For example, diseases may spread to new populations or a new political candidate may attract a unique voter demographic. Without properly controlling for confounding variables, studies based on the output of text classifiers are at risk of reaching erroneous conclusions.

In this paper, we present a text classification algorithm based on Pearl's back-door adjustment (Pearl 2003) to control for confounding variables. The approach conditions on the confounding variable at training time, then sums out the confounding variable at prediction time. We evaluate our approach on an end-to-end observational study conducted with Twitter data that estimates the effect of location on dispositional affect, with gender as a confounder. We compare the relative risk computed with and without back-door adjustment, and find that our approach results in more accurate estimates of effect sizes as the confounding relationship changes from training to testing data.

2 Related Work

In the social sciences, many methods have been developed to control for confounders, including matching, stratification, and regression analysis (Pourhoseingholi, Baghestani, and Vahedi 2012). Pearl (2003) developed tests for causal graphical models to determine which structures allow one to control for confounders using covariate adjustment, also known as the *back-door adjustment*. As far as we know, we are the first to use back-door adjustments to improve the robustness of text classifiers.

In the machine learning community, selection bias has received some attention (Zadrozny 2004; Sugiyama, Krauledat, and Müller 2007; Bareinboim, Tian, and Pearl 2014).

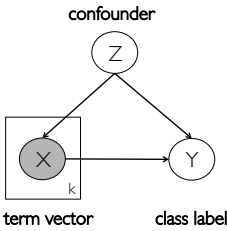


Figure 1: Directed graphical model depicting a confounder variable Z influencing both observed text features X and class variable Y.

Selection bias in text classification occurs when the distribution of text features changes from training to testing; i.e., $P_{train}(X) \neq P_{test}(X)$. Other work has considered the case where the target distribution $P(Y)$ changes from training to testing (Elkan 2001). In the present work, we address the more challenging case of a changing relationship between target labels Y and a confounder Z, i.e., $P_{train}(Y|Z) \neq P_{test}(Y|Z)$.

3 Back-door Adjustment for Text Classifiers

Suppose one wishes to estimate the effect size of a determinant variable X on an outcome variable Y, but a randomized control trial is not possible. In such cases, researchers often conduct observational studies, in which the effect size is estimated from a fixed set of observed X, Y pairs. When doing so, it is often necessary to adjust for potential *confounding* variables Z, which may influence both X and Y. Common approaches include matching, stratification, and regression analysis (Pourhoseingholi, Baghestani, and Vahedi 2012). Without such adjustments, the study validity is threatened by the possibility that an effect of confounder Z on outcome Y is misattributed to determinant X.

Conducting valid observational studies can be challenging even in traditional domains. In this work, we are interested in conducting observational studies from social media, which poses further difficulties. The primary challenge arises from the fact that some or even all of the variables (X, Y, Z) are not observed, but must be inferred from other observed text variables W. For example, in the experiments below, we consider the case where Y is a binary variable representing *dispositional affect* (Cohen and Pressman 2006). I.e., $Y_i = 1$ indicates that user i has positive disposition, and $Y_i = 0$ indicates that user i has a negative disposition. We fit a classifier to a user's tweets, and use the predictions to assign Y variables. Our goal is to train this classifier such that it is robust to changes in the relationship between Y and confounder Z. That is, while traditional observational studies focus on the confounding relationship among Z and (X, Y), here we are concerned with the confounding relationship among Z and (W, Y).

If we let W be a term vector representing a user's tweets, then we wish to fit a classifier $P(Y|W)$ that has been adjusted to reduce the effect of confounder Z. If we have access to a sufficient set of confounder variables Z, then it can be shown that we can estimate the causal effect as follows (Pearl 2003):

$$p(y|\text{do}(w)) = \sum_{z \in Z} p(y|w, z)p(z) \quad (1)$$

This formula is called *covariate adjustment* or *back-door adjustment*. The *back-door criterion* (Pearl 2003) is a graphical test that determines whether Z is a sufficient set of variables to estimate the causal effect. This criterion requires that no node in Z is a descendant of W and that Z blocks every path between W and Y that contains an arrow pointing to W.

While the back-door adjustment is well-studied in causal inference problems, in this paper we consider its application to text classification. We assume we are given a training set $D = \{(\mathbf{w}_i, y_i, z_i)\}_{i=1}^n$, where each instance consists of a term feature vector $\mathbf{w}$, a label y, and a covariate variable z. Our goal is to predict the label y_j for some new instance $\mathbf{w}_j$, while controlling for an unobserved confounder z_j. That is, we assume we observe the confounder at training time, but not at testing time.

Figure 1 displays the directed graphical model for our approach. Omitting the confounder Z, it depicts a standard discriminative approach to text classification, e.g., modeling $P(Y|W)$ with a logistic regression classifier conditioned on the observed term vector $\mathbf{x}$. We assume that the confounder Z influences both the term vector through $P(W|Z)$ as well as the target label through $P(Y|Z)$. For example, in a public health setting, y_i may be health status, $\mathbf{w}_i$ a term vector for online messages, and z_i a demographic variable.

While back-door adjustment is typically presented as a method of identifying the causal effect of W on Y, here we are not attempting any causal interpretation. (Indeed, it would be strange to assert that using a term *causes* one to have a class label.) However, Equation 1 provides a framework for making a prediction for Y given W that controls for Z. In doing so, we can train a classifier that is robust to the case where $P(Y|Z)$ changes from training to testing data.

To compute Equation 1 for a test example $\mathbf{w}$, we need to estimate two quantities from the labeled training data, $p(y|\mathbf{w}, z)$ and $p(z)$. For simplicity, we assume in this paper that $\mathbf{w}_i$ is a vector of binary features and that y_i and z_i are binary scalar variables. For $p(z)$, we use the maximum likelihood estimate $p(z = k) = \frac{\sum_{i \in D} \mathbf{1}[z_i = k]}{|D|}$, where $\mathbf{1}[\cdot]$ is an indicator function. For $p(y|\mathbf{w}, z)$, we use L2-regularized logistic regression. This can be efficiently done by simply appending two additional features $c_{i,0}$ and $c_{i,1}$ to each instance $\mathbf{w}_i$ representing $z = 0$ and $z = 1$. We set the feature values as follows: for training instance i, we set $c_{i,0}$ to v_1 if $z_i = 0$ and v_0 otherwise; we set $c_{i,1}$ to v_1 if $z_i = 1$ and v_0 otherwise. In the default case, we let $v_1 = 1$ and $v_0 = 0$. To predict for a new instance, we compute posteriors using Equation 1.

4 Experiments

We investigate the efficacy of our approach by conducting an observational study of dispositional affect using Twitter data. Dispositional affect indicates the emotional state in which a person typically responds to a situation (Cohen and

Pressman 2006). At a coarse level, we may think of "optimists" as having a positive dispositional affect, and "pessimists" as having a negative dispositional affect. Note that this differs from mood or sentiment in that it is a long-term personality trait, rather than a temporary emotion. Some studies have linked dispositional affect to physical and mental health.

In this study, we consider the effect that location has on dispositional affect, with gender as a potential confounder. Thus, we have three variables: (1) **Outcome variable** (Y): Dispositional affect (Positive or Negative); (2) **Determinant variable** (X): Location (we consider two cities, New York City or Los Angeles); (3) **Confounding variable** (Z): Gender (Male or Female).

Our goal is to estimate the effect of location (X) on disposition (Y); that is, are New Yorkers more likely to have a positive disposition than Los Angelenos? We measure the effect size using *relative risk*:

$$RR = \frac{P(Y = \text{Positive}|X = \text{NYC})}{P(Y = \text{Positive}|X = \text{LA})}$$

To build this dataset, we use the Twitter streaming API to collect tweets with geocoordinates from New York City (NYC) and Los Angeles (LA). We gather a total of 246,930 tweets for NYC and 218,945 for LA over a four-day period (June 15th to June 18th, 2015). We attempt to filter bots, celebrities, and marketing accounts by removing users with fewer than 10 followers or friends, more than 1,000 followers or friends, or more than 5,000 posts. We then label unique users with their gender using U.S. census name data, removing ambiguous names. We then collect all the available tweets (up to 3,200) for each user and represent each user as a binary unigram vector, using standard tokenization. Finally, we subsample this collection and keep the tweets from 6,000 users such that gender and location are uniformly distributed over the users.

We fit a binary classifier to predict the dispositional affect of a user based on their tweets. Collecting labeled data for this task is difficult, so for the purposes of generating the many labeled instances required for this study, we use a heuristic to annotate users as having positive or negative affect. To do so, we annotate individual tweets as positive if they contain one of 21 positive emoticons/emojis, and negative if they contain one of 15 sad emoticons/emojis. Emoticons have been used in prior work as a way of generating weakly labeled sentiment classification data (Agarwal et al. 2011). (We remove emoticons from the feature vector used by the classifier.)

To annotate a user's dispositional affect, we compute the number of negative tweets divided by the number of positive and negative tweets; that is, the proportion of sentiment-bearing tweets that are negative. To obtain a nearly equal number of positive and negative affect users, we labeled users with more than 20% negative tweets as having negative affect, and the remainder were labeled as positive affect. (Positive emoticons are used much more frequently than negative emoticons overall.)

We conducted experiments in which the relationship between the confounder Z and the class variable Y varies between the training and testing set. To control this relationship, we sample train/test sets with different $P(Y|Z)$ distributions. We assume we have labeled training datasets D_{train}, D_{test}, with elements $\{(\mathbf{w}_i, y_i, z_i)\}$, where y_i and z_i are binary variables. We introduce a bias parameter $P(y = 1|z = 1) = b$; by definition, $P(y = 0|z = 1) = 1 - b$. For each experiment, we sample without replacement from each set $D'_{train} \subseteq D_{train}$, $D'_{test} \subseteq D_{test}$. To simulate a change in $P(Y|Z)$, we use different bias terms for training and testing, b_{train}, b_{test}. We thus sample according to the following constraints: $P_{train}(y = 1|z = 1) = b_{train}$; $P_{test}(y = 1|z = 1) = b_{test}$; $P_{train}(Y) = P_{test}(Y)$. The latter constraint attempts to isolate the effect of $P(Y|Z)$ on results, while controlling for the class prior $P(Y)$. We emphasize that we do not alter any of the actual labels in the data; we merely sample instances to meet these constraints.

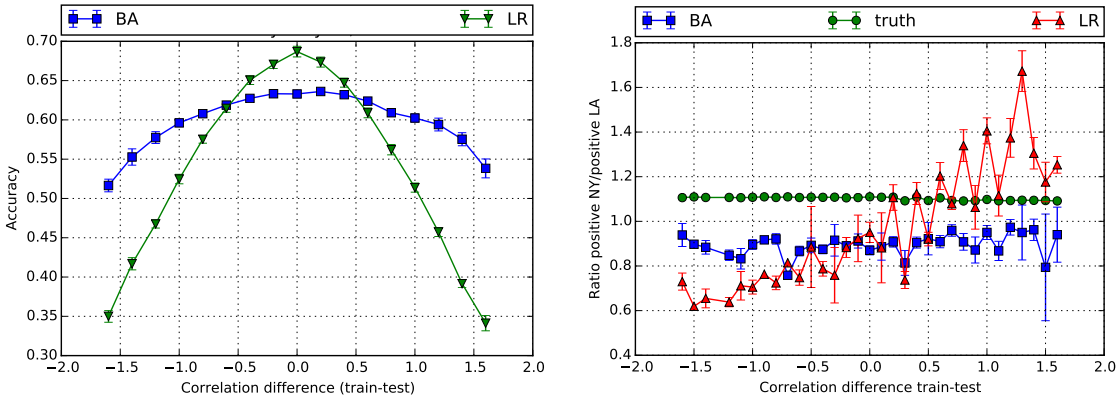
We make the bias value b vary from 0.1 to 0.9 (i.e. from 10% to 90% of bias) for both the training and the testing sets and we compare the accuracy of several classification models. For each b_{train}, b_{test} pair, we sample 5 train/test splits and report the average accuracy.

In addition to accuracy, we also compare the estimated relative risk using back-door adjustment with that of traditional logistic regression. We use the labeled data in the testing set to compute the "true" relative risk, then compare the quality of the results of each method. For all experiments, we fix the true relative risk to be 1.1 (that is, New Yorkers are somewhat more likely than Los Angelenos to have a positive disposition). We then classify all users in the test set to compute the estimated relative risk.

5 Results

Figure 2(a) shows testing accuracy of disposition classification as the difference between training and testing bias varies. To determine the x-axis, we compute the Pearson correlation between Z and Y, and report the difference between the testing and training correlations. This figure indicates that back-door adjustment results in more robust classification in the presence of confounding bias. This is most pronounced in extreme shifts in the confounding relationship, where the classification accuracy of LR becomes much worse than random. Note, however, that this comes at a cost of slightly lower classification accuracy when there is no change in confounding relationship between training and testing sets.

Figure 2(b) shows how the estimated relative risk values of the resulting classifiers compare with the "true" values set by the experiment. That is, this figure shows the effect that the confounding bias has on the final conclusions of the study. We can see that across a wide range of values, back-door adjustment produces more reliable estimates of the effect of location on disposition. (The periodic nature of the **LR** curve is in part due to shifts between positive and negative correlations between Y and Z, which we have not separated since we only consider the correlation difference.) It is notable that even when overall accuracy is fairly low, the relative risk can be estimated fairly well if the confounding variables have been accounted for properly.



(a) (b)

Figure 2: Experimental results: (a) dispositional affect classification accuracy as the strength of the confounder (gender) varies between training and testing set; (b) true and estimated relative risk values of the effect of location on disposition. In both cases, the classifier using back-door adjustment (**BA**) results in more robust classification and effect size estimates than standard logistic regression (**LR**).

6 Conclusion

In this paper, we have proposed an efficient and effective method of using back-door adjustment to control for confounders in text classification. We have found such adjustment to result in more accurate estimates effect sizes in web-based observational studies, thereby improving the validity of studies conducted using such noisy data sources. In our experiments, we have assumed that we observe the confounding variable at training time, and that the confounder is a single binary variable. In future work, we will consider the case where we only have a noisy estimate of Z at training time (Kuroki and Pearl 2014), as well as the case where Z is a vector of variables.

References

Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; and Passonneau, R. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, 30–38. Association for Computational Linguistics.

Bareinboim, E.; Tian, J.; and Pearl, J. 2014. Recovering from selection bias in causal and statistical inference. In *Proceedings of The Twenty-Eighth Conference on Artificial Intelligence (CE Brodley and P. Stone, eds.). AAAI Press, Menlo Park, CA*.

Chamlertwat, W.; Bhattarakosol, P.; Rungkasiri, T.; and Haruechaiyasak, C. 2012. Discovering consumer insight from twitter via sentiment analysis. *J. UCS* 18(8):973–992.

Clarke, K. A. 2005. The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science* 22(4):341–352.

Cohen, S., and Pressman, S. D. 2006. Positive affect and health. *Current Directions in Psychological Science* 15(3):122–125.

Dahllöf, M. 2012. Automatic prediction of gender, political affiliation, and age in swedish politicians from the wording of their speechesa comparative study of classifiability. *Literary and linguistic computing* fqs010.

Dredze, M. 2012. How social media will change public health. *IEEE Intelligent Systems* 27(4):81–84.

Elkan, C. 2001. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, 973–978.

Kuroki, M., and Pearl, J. 2014. Measurement bias and effect restoration in causal inference. *Biometrika* 101(2):423–437.

Lazer, D.; Pentland, A. S.; Adamic, L.; Aral, S.; Barabasi, A. L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; et al. 2009. Life in the network: the coming age of computational social science. *Science (New York, NY)* 323(5915):721.

Marathe, M., and Ramakrishnan, N. 2013. Recent advances in computational epidemiology. *IEEE intelligent systems* 28(4):96.

Pearl, J. 2003. Causality: models, reasoning, and inference. *Econometric Theory* 19:675–685.

Pourhoseingholi, M. A.; Baghestani, A. R.; and Vahedi, M. 2012. How to control confounding effects by statistical analysis. *Gastroenterology and Hepatology from bed to bench* 5(2):79.

Sugiyama, M.; Krauledat, M.; and Müller, K.-R. 2007. Covariate shift adaptation by importance weighted cross validation. *The Journal of Machine Learning Research* 8:985–1005.

Verma, S.; Vieweg, S.; Corvey, W. J.; Palen, L.; Martin, J. H.; Palmer, M.; Schram, A.; and Anderson, K. M. 2011. Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency. In *ICWSM*.

Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, 114. ACM.