Deconfounded Lexicon Induction for Interpretable Social Science

Reid Pryzant[*], **Kelly Shen**[*], **Dan Jurafsky**[+†], **Stefan Wager**[‡]
[*+]Department of Computer Science
[†]Department of Linguistics
[‡]Graduate School of Business
Stanford University
{rpryzant,kshen21,jurafsky,swager}@stanford.edu

Abstract

NLP algorithms are increasingly used in computational social science to take linguistic observations and predict outcomes like human preferences or actions. Making these social models transparent and interpretable often requires identifying features in the input that predict outcomes while also controlling for potential confounds. We formalize this need as a new task: inducing a lexicon that is predictive of a set of target variables yet uncorrelated to a set of confounding variables. We introduce two deep learning algorithms for the task. The first uses a bifurcated architecture to separate the explanatory power of the text and confounds. The second uses an adversarial discriminator to force confound-invariant text encodings. Both elicit lexicons from learned weights and attentional scores. We use them to induce lexicons that are predictive of timely responses to consumer complaints (controlling for product), enrollment from course descriptions (controlling for subject), and sales from product descriptions (controlling for seller). In each domain our algorithms pick words that are associated with *narrative persuasion*; more predictive and less confound-related than those of standard feature weighting and lexicon induction techniques like regression and log odds.

1 Introduction

Applications of NLP to computational social science and data science increasingly use *lexical features* (words, prefixes, etc) to help predict non-linguistic outcomes like sales, stock prices, hospital readmissions, and other human actions or preferences. Lexical features are useful beyond predictive performance. They enhance interpretability in machine learning because practitioners know why their system works. Lexical features can also be used to understand the subjective properties of a text.

For social models, we need to be able to select lexical features that predict the desired outcome(s) while also controlling for potential confounders. For example, we might want to know which words in a product description lead to greater sales, regardless of the item's price. Words in a description like "luxury" or "bargain" might increase sales but also interact with our confound (price). Such words don't reflect the *unique* part of text's effect on sales and should not be selected. Similarly, we might want to know which words in a consumer complaint lead to speedy administrative action, regardless of the product being complained about; which words in a course description lead to higher student enrollment, regardless of the course topic. These instances are associated with *narrative persuasion*: language that is responsible for altering cognitive responses or attitudes (Spence, 1983; Van Laer et al., 2013).

In general, we want words which are predictive of their targets yet decorrelated from confounding information. The lexicons constituted by these words are useful in their own right (to develop causal domain theories or for linguistic analysis) but also as interpretable features for down-stream modeling. Such work could help widely in applications of NLP to tasks like linking text to sales figures (Ho and Wu, 1999), to voter preference (Luntz, 2007; Ansolabehere and Iyengar, 1995), to moral belief (Giles et al., 2008; Keele et al., 2009), to police respect (Voigt et al., 2017), to financial outlooks (Grinblatt and Keloharju, 2001; Chatelain and Ralf, 2012), to stock prices (Lee et al., 2014), and even to restaurant health inspections (Kang et al., 2013).

Identifying linguistic features that are indicative of such outcomes and decorrelated with confounds is a common activity among social scientists, data scientists, and other machine learning practitioners. Indeed, it is essential for developing transpar-

Proceedings of NAACL-HLT 2018, pages 1615–1625
New Orleans, Louisiana, June 1 - 6, 2018.

ent and interpretable machine learning NLP models. Yet there is no generally accepted and rigorously evaluated procedure for the activity. Practitioners have conducted it on a largely ad-hoc basis, applying various forms of logistic and linear regression, confound-matching, or association quantifiers like mutual information or log-odds to achieve their aims, all of which have known drawbacks (Imai and Kim, 2016; Gelman and Loken, 2014; Wurm and Fisicaro, 2014; Estévez et al., 2009; Szumilas, 2010).

We propose to overcome these drawbacks via two new algorithms that consider the causal structure of the problem. The first uses its architecture to learn the part of the text's effect which the confounds cannot explain. The second uses an adversarial objective function to match text encoding distributions regardless of confound treatment. Both elicit lexicons by considering learned weights or attentional scores. In summary, we

1. Formalize the problem into a new task.
2. Propose a pair of well-performing neural network based algorithms.
3. Conduct the first systematic comparison of algorithms in the space, spanning three domains: consumer complaints, course enrollments, and e-commerce product descriptions.

The techniques presented in this paper will help scientists (1) better interpret the relationship between words and real-world phenomena, and (2) render their NLP models more interpretable[1].

2 Deconfounded Lexicon Induction

We begin by formalizing this language processing activity into a task. We have access to text(s) T, target variable(s) Y, and confounding variable(s) C. The goal is to pick a lexicon L such that when words in T belonging to L are selected, the resulting set $L(T)$ is related to Y but not C. There are two types of signal at play: the part of Y that T can explain, and that explainable by C. These signals often overlap because language reflects circumstance, but we are interested in the part of T's explanatory power which is *unique* to T, and hope to choose L accordingly.

So if $\text{Var}\left[\mathbb{E}\left[Y|L(T),C\right]\right]$ is the information in Y explainable by both $L(T)$ and C, then our goal is to choose L such that this variance is maximized *after* C has been fixed. With this in mind, **we formalize the task of deconfounded lexicon induction as finding a lexicon L that maximizes an *informativeness coefficient***,

$$\mathcal{I}(L) = \mathbb{E}\left[\text{Var}\left[\mathbb{E}\left[Y\middle|L(T),\,C\right]\middle|C\right]\right], \qquad (1)$$

which measures the explanatory power of the lexicon beyond the information already contained in the confounders C. Thus, highly informative lexicons cannot simply collect words that reflect the confounds. Importantly, this coefficient is only valid for comparing different lexicons of the same size, because in terms of maximizing this criterion, using the entire text will trivially make for the best possible lexicon.

Our coefficient $\mathcal{I}(L)$ can also be motivated via connections to the causal inference literature: in Section 7, we show that—under assumptions often used to analyze causal effects in observational studies—the coefficient $\mathcal{I}(L)$ can correspond exactly to the strength of T's causal effects on Y.

Finally, note that by expanding out an ANOVA decomposition for Y, we can re-write this criterion as

$$\begin{aligned}\mathcal{I}(L) = \mathbb{E}&\left[\left(Y - \mathbb{E}\left[Y\middle|C,\,L(T)\right]\right)^2\right] \\ &- \mathbb{E}\left[\left(Y - \mathbb{E}\left[Y\middle|C\right]\right)^2\right],\end{aligned} \qquad (2)$$

i.e., $\mathcal{I}(L)$ measures the performance improvement $L(T)$ affords to optimal predictive models that already have access to C. We use this fact for evaluation in Section 4.

3 Proposed Algorithms

We continue by describing the pair of novel algorithms we are proposing for deconfounded lexicon induction problems.

3.1 Deep Residualization (DR)

Motivation. Our first method is directly motivated by the setup from Section 2. Recall that $\mathcal{I}(L)$ measures the amount by which $L(T)$ can improve predictions of Y made from the confounders C. We accordingly build a neural network architecture that first predicts Y directly from C as well as possible, and then seeks to fine-tune those predictions using T.

Description. First we pass the confounds through a feed-forward neural network (FFNN) to obtain

[1] Code, hyperparameters, and instructions for practitioners are online at `https://nlp.stanford.edu/projects/deconfounded-lexicon-induction/`

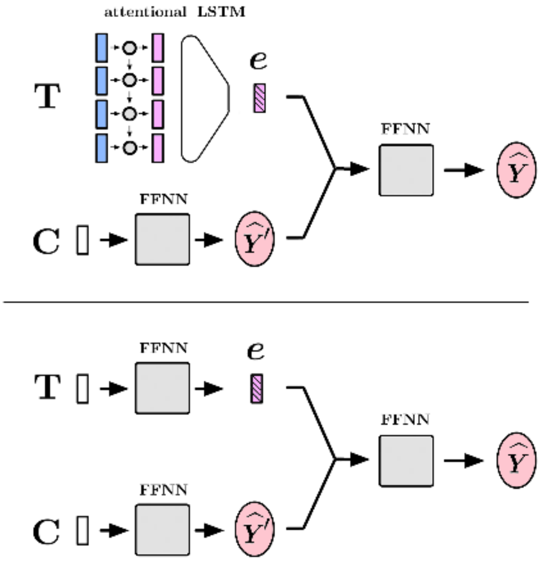

Figure 1: The Deep Residualization (DR) selector. Values which are used to calculate losses are enclosed in red ovals. *Top*: DR+ATTN, which represents text as a sequence of word embeddings. *Bottom*: DR+BOW, which represents text as a vector of word frequencies.

preliminary predictions $\hat{Y}'$. We also encode the text into a continuous vector $\boldsymbol{e} \in \mathcal{R}^d$ via two alternative mechanisms:

1. DR+ATTN: the text is converted into a sequence of embeddings and fed into Long Short-Term Memory (LSTM) cell(s) (Hochreiter and Schmidhuber, 1997) followed by an attention mechanism inspired by Bahdanau et al. (2015). If the words of a text have been embedded as vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_n$ then $\boldsymbol{e}$ is calculated as a weighted average of hidden states, where the weights are decided by a FFNN whose parameters are shared across timesteps:

$$\begin{aligned} \boldsymbol{h}_0 &= \vec{0} \\ \boldsymbol{h}_t &= LSTM(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}) \\ l_t &= ReLU(\boldsymbol{W}^{attn}\boldsymbol{h}_t) \cdot \boldsymbol{v}^{attn} \\ p_t &= \frac{exp(l_t)}{\sum exp(l_i)} \\ \boldsymbol{e} &= \sum p_i \boldsymbol{h}_i \end{aligned}$$

2. DR+BOW: the text is converted into a vector of word frequencies, which is compressed with a two-layer feedforward neural network (FFNN):

$$\begin{aligned} \boldsymbol{t} &= [freq_1, freq_2, ..., freq_k] \\ \boldsymbol{h} &= ReLU(\boldsymbol{W}^{hidden}\boldsymbol{t}) \\ \boldsymbol{e} &= ReLU(\boldsymbol{W}^{output}\boldsymbol{t}) \end{aligned}$$

We then concatenate $\boldsymbol{e}$ with $\hat{Y}'$ and feed the result through another neural network to generate final predictions $\hat{Y}$. If Y is continuous we compute loss with

$$\mathcal{L}_{continuous} = ||\hat{Y} - Y||_2$$

If Y is categorical we compute loss with

$$\mathcal{L}_{categorical} = -p^* \log \widehat{p^*}$$

Where $\widehat{p^*}$ corresponds to the predicted probability of the correct class. The errors from $\hat{Y}$ are propagated through the whole model, but the errors from $\hat{Y}'$ are only used to train its progenitor (Figure 1).

Note the similarities between this model and the popular residualizing regression (RR) technique (Jaeger et al., 2009; Baayen et al., 2010, inter alia). Both use the text to improve an estimate generated from the confounds. RR treats this as two separate regression tasks, by regressing the confounds against the variables of interest, and then using the residuals as features, while our model introduces the capacity for nonlinear interactions by backpropagating between RR's steps.

Lexicon Induction. We elicit lexicons from +ATTN style models by (1) running inference on a test set, but rather than saving those predictions, saving the attentional distribution over each source text, and (2) mapping each word to its average attentional score and selecting the k highest-scoring words.

For +BOW style models, we take the matrix that compresses the text's word frequency vector, then score each word by computing the l_1 norm of the column that multiplies it, with the intuition that important words are dotted with big vectors in order to be a large component of $\boldsymbol{e}$.

3.2 Adversarial Selector (A)

Motivation. We begin by observing that a desirable L can explain Y, but is unrelated to C, which implies it should should struggle to predict C. The Adversarial Selector draws inspiration from this.

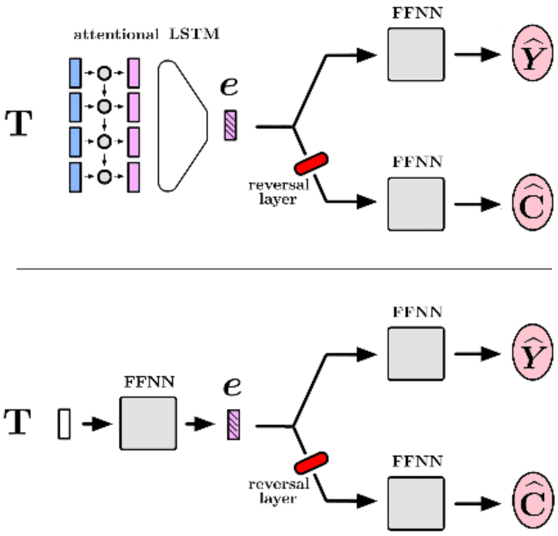

Figure 2: The Adversarial (A) selector. Values which are used to calculate losses are enclosed in red ovals. *Top*: A+ATTN, which represents text as a sequence of word embeddings. *Bottom*: A+BOW, which represents text as a vector of word frequencies.

It learns adversarial encodings of T which are useful for predicting Y, but *not* useful for predicting C. It is depicted in Figure 2.

Description. First, we encode T into $\boldsymbol{e} \in \mathcal{R}^d$ via the same mechanisms as the Deep Residualizer of Section 3.1. $\boldsymbol{e}$ is then passed to a series of FFNNs ("prediction heads") which are trained to predict each target and confound with the same loss functions as that of Section 3.1. As gradients back-propagate from the confound prediction heads to the encoder, we pass them through a *gradient reversal layer* in the style of Ganin et al. (2016) and Britz et al. (2017), which multiplies gradients by -1. If the cumulative loss of the target variables is $\mathcal{L}_t$ and that of the confounds is $\mathcal{L}_c$, then the loss which is implicitly used to train the encoder is $\mathcal{L}_e = \mathcal{L}_t - \mathcal{L}_c$, thereby encouraging the encoder to learn representations of the text which are not useful for predicting the confounds.

Lexicons are elicited from this model via the same mechanism as the Deep Residualizer of Section 3.1.

4 Experiments

We evaluate the approaches described in Sections 3 and 5 by generating and evaluating deconfounded lexicons in three domains: financial complaints, e-commerce product descriptions, and course descriptions. In each case the goal is to find words which can *always* help someone net a positive outcome (fulfillment, sales, enrollment), regardless of their situation. This involves finding words associated with *narrative persuasion*: predictive of human decisions or preferences but decorrelated from non-linguistic information which could also explain things. We analyze the resulting lexicons, especially with respect to the classic Aristotelian modes of persuasion: logos, pathos, and ethos.

We compare the following algorithms: **R**egression (R), **R**egression with **C**onfound features (RC), **M**ixed effects Regression (M), **R**esidualizing **R**egressions (RR), Log-**O**dds **R**atio (OR), **M**utual **I**nformation (MI), and MI/OR with regresssion (R+MI and R+OR). See Section 5 for a discussion of these baselines, and the online supplementary information for implementation details. We also compare the proposed algorithms: **D**eep **R**esidualization using word frequencies (DR+BOW) and embeddings (DR+ATTN), and **A**dversarial Selection using word frequencies (A+BOW) and embeddings (A+ATTN).

In Section 2 we observed that $\mathcal{I}(L)$ measures the improvement in predictive power that $L(T)$ affords a model already having access to C. Thus, we evaluate each algorithm by (1) regressing C on Y, (2) drawing a lexicon L, (3) regressing $C + L(T)$ on Y, and (4) measuring the size of gap in test prediction error between the models of step (1) and (3). For classification problems, we measured error with cross-entropy (XE):

$$XE = -\sum_i p_i \log \hat{p}_i$$

$$\text{performance} = XE_C - XE_{L(T),C}$$

And for regression, we computed the mean squared error (MSE):

$$MSE = \frac{1}{n}\sum_i (\hat{Y}_i - Y_i)^2$$

$$\text{performance} = MSE_C - MSE_{L(T),C}$$

Because we fix lexicon size but vary lexicon content, lexicons with good words will score highly under this metric, yielding the large performance improvements when combined with C.

We also report the average strength of association between words in L and C. For categorical confounds, we measure Cramer's V (V) (Cramér, 2016), and for continuous confounds, we use the point-biserial correlation coefficient (r_{pb}) (Glass and Hopkins, 1970). Note that r_{pb} is mathematically equivalent to Pearson correlation in bivariate settings. Here the best lexicons will score the lowest.

We implemented neural models with the Tensorflow framework (Abadi et al., 2016) and optimized using Adam (Kingma and Ba, 2014). We implemented linear models with the scikit learn package (Pedregosa et al., 2011). We implemented mixed models with the lme4 R package (Bates et al., 2014). We refer to the online supplementary materials for per-experiment hyperparameters.

For each dataset, we constructed vocabularies from the 10,000 most frequently occurring tokens, and randomly selected 2,000 examples for evaluation. We then conducted a wide hyperparameter search and used lexicon performance on the evaluation set to select final model parameters. We then used these parameters to induce lexicons from 500 random train/test splits. Significance is estimated with a bootstrap procedure: we counted the number of trials each algorithm "won" (i.e. had the largest $error_C - error_{L(T),C}$). We also report the average performance and correlation of all the lexicons generated from each split. We ran these experiments using lexicon sizes of k = 50, 150, 250, and 500 and observed similar behavior. The results reported in the following sections are for k = 150, and the words in Tables 1, and 2, 3 are from randomly selected lexicons (other lexicons had similar characteristics).

4.1 Consumer Financial Protection Bureau (CFPB) Complaints

Setup. We consider 189,486 financial complaints publicly filed with the Consumer Financial Protection Bureau (CFPB)[2]. The CFPB is a product of Dodd-Frank legislation which solicits and addresses complaints from consumers regarding a variety of financial products: mortgages, credit reports, etc. Some submissions are handled on a timely basis (< 15 days) while others languish.

We are interested in identifying salient words which help push submissions through the bureaucracy and obtain timely responses, regardless of the specific nature of the complaint. Thus, our target variable is a binary indicator of whether the complaint obtained a timely response. Our confounds are twofold, (1) a categorical variable tracking the type of issue (131 categories), and (2) a categorical variable tracking the financial product (18 categories). For the proposed DR+BOW, DR+ATTN, A+BOW, and A+ATTN models, we set $|\boldsymbol{e}|$ to 1, 64, 1, and 256, respectively.

Results. In general, this seems to be a tractable classification problem, and the confounds alone are moderately predictive of timely response (XE_C = 1.06). The proposed methods appear to perform the best, and DR+BOW achieved the largest performance/correlation ratio (Figure 3).

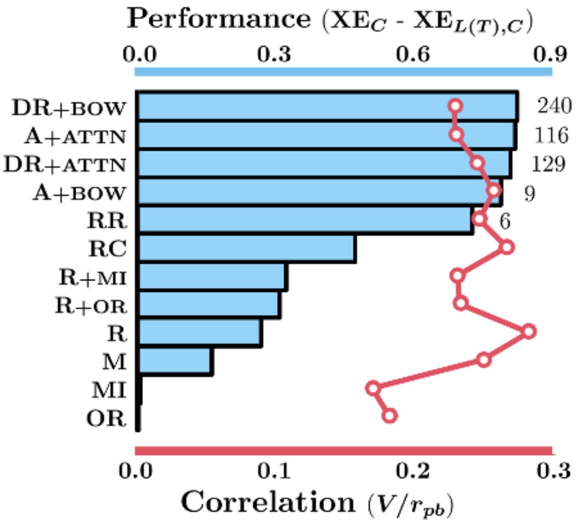

Figure 3: Predictive performance ($XE_C - XE_{L(T),C}$) and average confound correlation (V/r_{pb}) of lexicons generated via our proposed algorithms and a variety of methods in current use. The numbers to the right of each bar indicate the number of winning bootstrap trials.

DR+BOW	MI	RR	R
.	secondly	being	100
ma'am	forget	6	fargo
multiple	focus	issued	wells
guide	questions	agreement	.
submitted	battle	starting	fdcpa
'nt	vs	150.00	angry
honor	certainly	question	owe
,	contained	in	hipaa
xx/xx/xxxx	the	.	file
ago	be	agreement	across

Table 1: The ten highest-scoring words in lexicons generated by Deep Residualization + BOW (DR+BOW), Mutual Information (MI), Residialized Regression (RR), and regression (R).

[2] These data can be obtained from https://www.consumerfinance.gov/data-research/consumer-complaints/

We obtain further evidence upon examining the lexicons selected by four representative algorithms: proposed (DR+BOW), a well-performing baseline (RR), and two naive baselines (R, MI) (Table 1). MI's words appear unrelated to the confounds, but don't seem very persuasive, and our results corroborate this: these words failed to add predictive power over the confounds (Figure 3). On the opposite end of the spectrum, R's words appear somewhat predictive of the timely response, but are confound-related: they include the FDCPA (Fair Debt Collection Practices Act) and HIPAA (Health Insurance Portability and Accountability Act), which are directly related to the confound of financial product.

The top-scoring words in RR's lexicon include numbers ("6", "150.00") and words that suggest that the issue is ongoing ("being", "starting"). On the other hand, the words of DR+BOW draw on the rhetorical devices of ethos by respecting the reader's authority ("ma'am", "honor"), and logos by suggesting that the writer has been proactive about solving the issue ("multiple", "submitted", "xx/xx/xxx", "ago"). These are narrative qualities that align with two of the persuasion literature's "weapons of influence": reciprocation and commitment (Kenrick et al., 2005). Several algorithms implicitly favored longer (presumably more detailed) complaints by selecting common punctuation.

4.2 University Course Descriptions

Setup. We consider 141,753 undergraduate and graduate course offerings over a 6-year period (2010 - 2016) at Stanford University. We are interested in how the writing style of a description convinces students to enroll. We therefore choose log(enrollment) as our target variable and control for non-linguistic information which students also use when making enrollment decisions: course subject (227 categories), course level (26), number of requirements satisfied (7), whether there is a final (3), the start time, and the combination of days the class meets (26). All except start time are modeled as categorical variables. For the proposed DR+BOW, DR+ATTN, A+BOW, and A+ATTN models, we set $|e|$ to 1, 100, 16, and 64, respectively.

Results. This appears to be a tractable regression problem; the confounds alone are highly predictive of course enrollment (MSE_C = 3.67). (Figure 4). A+ATTN performed the best, and in general, the proposed techniques produced the most-predictive and least-correlated lexicons. Interestingly, Residualization (RR) and Regression with Confounds (RC) appear to outperform the Deep Residualization selector.

A+ATTN	R	OR
future	programming	summer
instructor	required	interpretation
eating	prerequisites	stability
or	computer	attitude
doing	management	optimization
guest	introduction	completion
sexual	chemical	during
culture	applications	labor
research	you	production
project	clinical	background

Table 2: The ten highest-scoring words in lexicons generated by Adversarial + ATTN (A+ATTN), Regression (R), and Log-Odds Ratio (OR).

In Table 2 we observe stark differences between the highest-scoring words of a proposed technique (A+ATTN) and two baselines with opposing characteristics (R, OR) (Table 2). Words chosen via Regression (R) appear predictive of enrollment, but also related to the confounds of subject ("programming", "computer", "management", "chemical", "clinical") and level ("required", "prerequisites", "introduction").

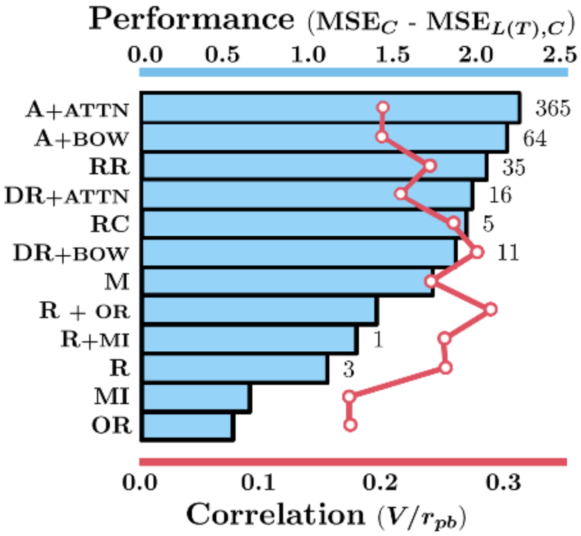

Figure 4: Course description comparative performance.

Log-Odds Ratio (OR) selected words which

A+BOW			RR		
word	transliteration	translation	word	transliteration	translation
ます	*masu*	polite suffix	７５		
プロテイン	*purotein*	protein	５		
お	*oh*	polite prefix	ニチバン	*nichiban*	adhesive company
粒	*tsubu*	grain	４		
栄養	*eiyō*	nutrition	群	*gun*	group
ご	*go*	polite prefix	サイズ	*saizu*	size
配合	*haigō*	formulation	摂取	*sesshu*	intake
デザート	*dezāto*	dessert	枚	*mai*	sheet
錠	*jō*	tablet	化学	*kagaku*	chemical
大豆	*daizu*	soy	ミニ	*mini*	mini

Table 3: The ten highest-scoring words in lexicons generated by Adversarial Selection + BOW (A+BOW) and Residualization (RR).

appear unrelated to both the confounds and enrollment. The Adversarial Selector (A+ATTN) selected words which are both confound-decorrelated and predictive of enrollment. Its words appeal to the concept of variety ("or", "guest"), and to pathos, in the form of universal student interests ("future", "eating", "sexual"). Notably, the A+ATTN words are also shorter (mean length of 6.2) than those of R (9.3) and OR (9.0), which coincides with intuition (students often skim descriptions) and prior research (short words are known to be more persuasive in some settings (Pratkanis et al., 1988)). The lexicon also suggests that students prefer courses with research project components ("research", "project").

4.3 eCommerce Descriptions

Setup. We consider 59,487 health product listings on the Japanese e-commerce website Rakuten[3]. These data originate from a December 2012 snapshot of the Rakuten marketplace. They were tokenized with the JUMAN morphological analyzer (Kurohashi and Nagao, 1999).

We are interested in identifying words which advertisers could use to increase their sales, regardless of the nature of the product. Therefore, we set log(sales) as our target variable, and control for an item's price (continuous) and seller (207 categories). The category of an item (i.e. toothbrush vs. supplement) is not included in these data. In practice, sellers specialize in particular product types, so this may be indirectly accounted for. For the proposed DR+BOW, DR+ATTN, A+BOW, and A+ATTN models, we set $|e|$ to 4, 64, 4, and 30, respectively.

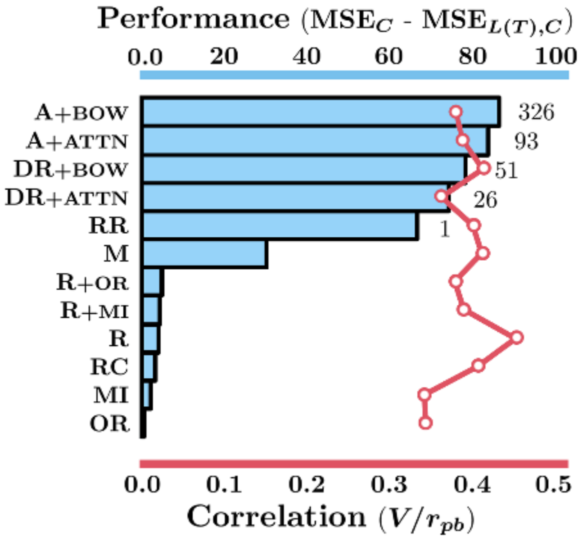

Figure 5: E-commerce comparative performance.

Results. This appears to be a more difficult prediction task, and the confounds are only slightly predictive of sales (MSE_C = 116.34) (Figure 5). Again, lexicons obtained via the proposed methods were the most successful, achieving the highest performance with the lowest correlation (Table 3). When comparing the words selected by A+BOW (proposed) and RR (widely used and well performing), we find that both draw on the rhetorical element of logos and demonstrate informativeness ("nutrition", "size", etc.). A+BOW also draws on ethos by identifying word stems associated with politeness. This quality draws on the authority of shared cultural values, and has been shown to appeal to Japanese shoppers (Pryzant et al., 2017). On the other hand, RR selected sev-

[3] These data can be obtained from `https://rit.rakuten.co.jp/data_release/`

eral numbers and failed to avoid brand indicators: "nichiban", a large company which specializes in medical adhesives, is one of the highest-scoring words.

5 Related Work

There are three areas of related work which we draw on. We address these in turn.

Lexicon induction. Some work in lexicon induction is intended to help interpret the subjective properties of a text or make make machine learning models more interpretable, i.e. so that practitioners can know why their system works. For example, Taboada et al. (2011); Hamilton et al. (2016) induce sentiment lexicons, and Mohammad and Turney (2010); Hu et al. (2009) induce emotion lexicons. Practitioners often get these words by considering the high-scoring features of regressions trained to predict an outcome (McFarland et al., 2013; Chahuneau et al., 2012; Ranganath et al., 2013; Kang et al., 2013). They account for confounds through manual inspection, residualizing (Jaeger et al., 2009; Baayen et al., 2010), hierarchical modeling (Bates, 2010; Gustarini, 2016; Schillebeeckx et al., 2016), log-odds (Szumilas, 2010; Monroe et al., 2008), mutual information (Berg, 2004), or matching (Tan et al., 2014; DiNardo, 2010). Many of these methods are manual processes or have known limitations, mostly due to multicollinearity (Imai and Kim, 2016; Chatelain and Ralf, 2012; Wurm and Fisicaro, 2014). Furthermore, these methods have not been tested in a comparative setting: this work is the first to offer an experimental analysis of their abilities.

Causal inference. Our methods for lexicon induction have connections to recent advances in the causal inference literature. In particular, Johansson et al. (2016) and Shalit et al. (2016) propose an algorithm for counterfactual inference which bear similarities to our Adversarial Selector (Section 3.2), Imai et al. (2013) advocate a lasso-based method related to our Deep Residualization (DR) method (Section 3.1), and Egami et al. (2017) explore how to make causal inferences from text through careful data splitting. Unlike us, these papers are largely unconcerned with the underlying features and algorithmic interpretability. Athey (2017) has a recent survey of machine learning problems where causal modeling is important.

Persuasion. Our experiments touch on the mechanism of *persuasion*, which has been widely studied. Most of this prior work uses lexical, syntactic, discourse, and dialog interactive features (Stab and Gurevych, 2014; Habernal and Gurevych, 2016; Wei et al., 2016), power dynamics (Rosenthal and Mckeown, 2017; Moore, 2012), or diction (Wei et al., 2016) to study *discourse* persuasion as manifested in argument. We study *narrative* persuasion as manifested in everyday decisions. This important mode of persuasion is understudied because researchers have struggled to isolate the "active ingredient" of persuasive narratives (Green, 2008; De Graaf et al., 2012), a problem that the formal framework of deconfounded lexicon induction (Section 2) may help alleviate.

6 Conclusion

Computational social scientists frequently develop algorithms to find words that are related to some information but not other information. We encoded this problem into a formal task, proposed two novel methods for it, and conducted the first principled comparison of algorithms in the space. Our results suggest the proposed algorithms offer better performance than those which are currently in use. Upon linguistic analysis, we also find the proposed algorithms' words better reflect the classic Aristotelian modes of persuasion: logos, pathos, and ethos.

This is a promising new direction for NLP research, one that we hope will help computational (and non-computational!) social scientists better interpret linguistic variables and their relation to outcomes. There are many directions for future work. This includes algorithmic innovation, theoretical bounds for performance, and investigating rich social questions with these powerful new techniques.

7 Appendix: Causal Interpretation of the Informativeness Coefficient

Recall the definition of $\mathcal{I}(L)$:

$$\mathcal{I}(L) = \mathbb{E}\left[\text{Var}\left[\mathbb{E}\left[Y \middle| L(T), C\right] \middle| C\right]\right]$$

Here, we discuss how under standard (albeit strong) assumptions that are often made to identify causal effects in observational studies, we can interpret $\mathcal{I}(L)$ with $L(T) = T$ as a measure of the strength of the text's causal effect on Y.

Following the potential outcomes model of Rubin (1974) we start by imagining potential outcomes $Y(t)$ corresponding to the outcome we would have observed given text t for any possible text $t \in \mathcal{T}$; then we actually observe $Y = Y(T)$. With this formalism, the causal effect of the text is clear, e.g., the effect of using text t' versus t is simply $Y(t') - Y(t)$.

Suppose that T, our observed text, takes on values in $\mathcal{T}$ with a distribution that depends on C. Let's also assume that the observed text T is independent of the potential outcomes $\{Y(t)\}_{t\in\mathcal{T}}$, conditioned on the confounders C (Rosenbaum and Rubin, 1983). So we know what would happen with any given text, but don't yet know which text will get selected (because T is a random variable). Now if we fix C and there is any variance remaining in $Y(T)$ (i.e. $\mathbb{E}\left[\text{Var}\left[Y(T)\middle|C, \{Y(t)\}_{t\in\mathcal{T}}\right]\right] > 0$) then the text has a causal effect on Y.

Now we assume that $Y(t) = f_c(t) + \epsilon$, meaning that the difference in effects of one text t relative to another text t' is always the same given fixed confounders. For example, in a bag of words model, this would imply that switching from using the word "eating" versus "homework" in a course description would always have the same impact on enrollment (conditionally on confounders). With this assumption in hand, then the causal effects of T, $\mathbb{E}\left[\text{Var}\left[Y(T)\middle|C, \{Y(t)\}_{t\in\mathcal{T}}\right]\right]$, matches $\mathcal{I}(L)$ as described in equation (1) (Imbens and Rubin, 2015). In other words, given the same assumptions often made in observational studies, the informativeness coefficient of the full, uncompressed text in fact corresponds to the amount of variation in Y due to the causal effects of T.

8 Acknowledgements

We gratefully acknowledge support from NSF Award IIS-1514268. We thank Youngjoo Chung for her invaluable assistance, advice, and the Rakuten data. We also thank Will Hamilton for his advice and direction while writing.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* .

Stephen Ansolabehere and Shanto Iyengar. 1995. *Going Negative: How Attack Ads Shrinks and Polarize the Electorate*. New York: Free Press.

Susan Athey. 2017. Beyond prediction: Using big data for policy problems. *Science* 355(6324):483–485.

R. Harald Baayen, Victor Kuperman, and Raymond Bertram. 2010. Frequency effects in compound processing. In *Compounding*, Benjamins, pages 257–270.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *4th International Conference on Learning Representations (ICLR)* .

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. 2014. lme4: Linear mixed-effects models using eigen and s4. *R package version* 1(7):1–23.

Douglas M Bates. 2010. lme4: Mixed-effects modeling with r.

Bruce L. Berg. 2004. *Methods for the social sciences*. Pearson Education Inc, United States of America.

Denny Britz, Reid Pryzant, and Quoc V. Le. 2017. Effective domain mixing for neural machine translation. In *Second Conference on Machine Translation (WMT)*.

V. Chahuneau, K. Gimpel, B. R. Routledge, L. Scherlis, and N. A. Smith. 2012. Word salad: Relating food prices and descriptions. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

Jean-Bernard Chatelain and Kirsten Ralf. 2012. Fallacious liaisons: Near multicolinearity and "classical suppressors," aid policies, and growth. *Revue économique* 63(3):557–567.

Harald Cramér. 2016. *Mathematical Methods of Statistics (PMS-9)*, volume 9. Princeton university press.

Anneke De Graaf, Hans Hoeken, José Sanders, and Johannes WJ Beentjes. 2012. Identification as a mechanism of narrative persuasion. *Communication Research* 39(6):802–823.

John DiNardo. 2010. Natural experiments and quasi-natural experiments. In *Microeconometrics*, Springer, pages 139–153.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2017. How to make causal inferences using texts .

Pablo A. Estévez, Michel Tesmer, Claudio A. Perez, and Jacek M. Zurada. 2009. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks* 20(2):189–201.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* 17(59):1–35.

Andrew Gelman and Eric Loken. 2014. The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American Scientist* 102(6):460.

Micheal W. Giles, Bethany Blackstone, and Richard L. Vining Jr. 2008. The Supreme Court in American democracy: Unraveling the linkages between public opinion and judicial decision making. *The Journal of Politics* 70(2):293–306.

Gene V. Glass and Kenneth D. Hopkins. 1970. *Statistical methods in education and psychology*. Prentice-Hall Englewood Cliffs, NJ.

Melanie C. Green. 2008. Research challenges: Research challenges in narrative persuasion. *Information Design Journal* 16(1):47–52.

Mark Grinblatt and Matti Keloharju. 2001. How distance, language, and culture influence stockholdings and trades. *The Journal of Finance* 56(3):1053–1073.

Mattia Gustarini. 2016. *Analysing smartphone users "inner-self": the perception of intimacy and smartphone usage changes*. Ph.D. thesis, University of Geneva.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional lstm. In *54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. *2016 Conference on Empirical Methods on Natural Language Processing (EMNLP)* .

Chin-Fu Ho and Wen-Hsiung Wu. 1999. Antecedents of customer satisfaction on the internet: An empirical study of online shopping. In *Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on*. IEEE, pages 9–pp.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural computation* 9(8):1735–1780.

Yajie Hu, Xiaoou Chen, and Deshun Yang. 2009. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *The International Society of Music Information Retrieval (ISMIR)*.

Kosuke Imai and In Song Kim. 2016. *When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data?*. Ph.D. thesis, Working paper, Princeton University, Princeton, NJ.

Kosuke Imai, Marc Ratkovic, et al. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7(1):443–470.

Guido W. Imbens and Donald B. Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

T. Florian Jaeger, Victor Kuperman, and Austin Frank. 2009. Issues and solutions in fitting, evaluating, and interpreting regression models. In *Talk given at WOMM presession to the 22nd CUNY Conference on Sentence Processing*.

Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International Conference on Machine Learning (ICLR)*.

Jun Seok Kang, Polina Kuznetsova, Michael Luca, and Yejin Choi. 2013. Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. In *Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

Denise M. Keele, Robert W. Malmsheimer, Donald W. Floyd, and Lianjun Zhang. 2009. An analysis of ideological effects in published versus unpublished judicial opinions. *Journal of Empirical Legal Studies* 6(1):213–239.

Douglas T. Kenrick, Steven L. Neuberg, and Robert B. Cialdini. 2005. *Social psychology: Unraveling the mystery*. Pearson Education New Zealand.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *3rd International Conference for Learning Representations (ICLR)* .

Sadao Kurohashi and Makoto Nagao. 1999. Japanese morphological analysis system juman version 3.61. *Department of Informatics, Kyoto University* .

Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. 2014. On the importance of text analysis for stock price prediction. In *International Conference on Language Resources and Evaluation (LREC)*.

Frank Luntz. 2007. *Words that work: It's not what you say, it's what people hear*. Hachette Books.

Daniel A. McFarland, Dan Jurafsky, and Craig Rawlings. 2013. Making the connection: Social bonding in courtship situations. *American journal of sociology* 118(6):1596–1649.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. Association for Computational Linguistics, pages 26–34.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16(4):372–403.

Brian C.J. Moore. 2012. *An introduction to the psychology of hearing*. Brill.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12(Oct):2825–2830.

Anthony R. Pratkanis, Anthony G. Greenwald, Michael R. Leippe, and Michael H. Baumgardner. 1988. In search of reliable persuasion effects: III. The sleeper effect is dead: Long live the sleeper effect. *Journal of personality and social psychology* 54(2):203.

Reid Pryzant, Young-joo Chung, and Dan Jurafsky. 2017. Predicting sales from the language of product descriptions. In *Special Interest Group on Information Retrieval (SIGR) eCommerce Workshop*.

Rajesh Ranganath, Dan Jurafsky, and Daniel A. McFarland. 2013. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech & Language* 27(1):89–115.

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

Sara Rosenthal and Kathleen Mckeown. 2017. Detecting influencers in multiple online genres. *ACM Transactions on Internet Technology (TOIT)* 17(2):12.

Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688.

Simon J.D. Schillebeeckx, Sankalp Chaturvedi, Gerard George, and Zella King. 2016. What do I want? The effects of individual aspiration and relational capability on collaboration preferences. *Strategic Management Journal* 37(7):1493–1506.

Uri Shalit, Fredrik Johansson, and David Sontag. 2016. Estimating individual treatment effect: Generalization bounds and algorithms. *34th International Conference on Machine Learning (ICML)* .

Donald P. Spence. 1983. Narrative persuasion. *Psychoanalysis & Contemporary Thought* .

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Magdalena Szumilas. 2010. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry* 19(3):227.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37(2):267–307.

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

Tom Van Laer, Ko De Ruyter, Luca M. Visconti, and Martin Wetzels. 2013. The extended transportation-imagery model: A meta-analysis of the antecedents and consequences of consumers' narrative transportation. *Journal of Consumer research* 40(5):797–817.

Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences* .

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? Ranking argumentative comments in the online forum. In *The 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lee H. Wurm and Sebastiano A. Fisicaro. 2014. What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language* 72:37–48.