Extending Sparse Text with Induced Domain-Specific Lexicons and Embeddings: A Case Study on Predicting Donations

MeiXing Dong[a], Rada Mihalcea[a], Dragomir Radev[b]

[a]*Department of Computer Science and Engineering, University of Michigan, 2260 Hayward St., Ann Arbor, MI 48105, United States*
[b]*Department of Computer Science, Yale University, 17 Hillhouse Ave., New Haven, CT 06511, United States*

Abstract

This paper addresses the problem of expanding sparse textual content to increase the accuracy of data-driven prediction tasks. We evaluate the use of word embeddings and lexicons within the context of a donation prediction task, where we classify potential donors as either likely or unlikely to donate. We perform several comparative experiments and analyses, and show that our methods to automatically enhance sparse textual data significantly improve the predictive performance on this task.

Keywords: Natural language processing, Text expansion, Sparse text

1. Introduction

Over the past three decades, data-driven learning has made great strides and brought significant progress across many disciplines, ranging from computer science and information sciences, to psychology, astronomy, economics, and many other science or humanities fields. While many of the most recent learning strategies assume the availability of a large amount of data, there are still many applications that only benefit from limited amounts of data. Among these, we often deal with datasets that include only small amounts of textual information that, because of their size and limited vocabulary, end up not contributing as much as they could to the overall learning process.

In this paper, we explore the question of whether we can enrich sparse textual content inside categorical datasets, to bring into the learning framework additional information that is implied by the text but not explicitly stated. As an example, consider a dataset that includes a text field whose value for one of the instances is the word “computer.” Typically, such categorical features are used “as is” and are weighted and used alongside other features, depending on the learning framework. However, aside from being a string of characters, the word “computer” implies “an electronic device for storing and processing data,” has associations with other words such as “data,” “hardware,” “software,” and so forth. In this paper, we present several methods for automatically enriching categorical fields in a dataset where the categorical elements can also be treated as text. Our goal is to improve data-driven predictions, so we perform comparative evaluations that allow us to learn what text expansion techniques work best.

Specifically, we primarily ask our questions in the context of a donation prediction problem, where we use a dataset consisting of the profiles of university alumni who have previously donated, as well as alumni who did not make any donations, and attempt to predict for a new instance whether they are likely to donate or not. We also consider the task of gender prediction on a dataset of blog profiles to determine to what extent our methods can be applied to other datasets.

The amount of textual data available in both datasets is limited in terms of both quantity and variety; each piece of text is a few words at most, and the category definitions restrict the vocabulary. Yet, it can still be quite useful. For instance, a “CEO” is more likely to donate than a “clerk”, or a “senior” employee is more likely to donate than a “recent graduate.”

We explore four different strategies for extending sparse text, including two lexicon generation methods, and two embedding methods that are influenced by domain knowledge. Using features obtained from these methods, we build models that predict whether someone is likely to donate, and compare their performance with baseline models that do not make use of such additional features.

The paper makes two main research contributions. First, we address the question of whether we can effectively augment text fields in a dataset by leveraging information specific to the target domain, and show that with such textual expansion strategies we can significantly improve over a baseline that does not make use of this additional information. Second, we compare several different models for extending sparse text in datasets, including methods that rely on information drawn from (a) the database itself; or (b) external resources, and gain new insights into what methods lead to the highest performance improvements. We seek to answer these questions using the donation prediction task, where we rely on a dataset that has information on previous donors including limited free-form text, and show the role played by different text expansion strategies to improve the effectiveness of our predictive model. We also show that these methods can apply to other cases by evaluating on a second task and dataset.

2. Related Work

Our task is related to the classification of short texts, which is challenging because the text is typically sparse and do not provide much word co-occurrence information. In contrast to standard free-form short-text datasets, such as tweets from Twitter, our categorical text is not only short but also restricted in content. For instance, the set of academic majors available at a particular university only contains text from the the names of the majors.

Unfortunately, the bulk of recent machine learning methods assume the availability of large amounts of varied data, but there exist many ways of tackling machine learning without this. Hand-built lexical resources have been used extensively in natural language processing tasks like word-sense disambiguation (Banerjee and Pedersen (2002)), sentiment analysis (Mohammad et al. (2013)), and short text classification (Jiang et al. (2011)). Text embedding methods allow models trained on one domain to be adapted to new domains that have little data.

We focus on lexical resources and embedding methods as they are two of the most straightforward and commonly used methods for text classification tasks. In this section, we overview the work that has been previously done on these related directions.

2.1. Lexical Resources

Lexicons have been used extensively in sentiment analysis tasks (Taboada et al. (2011)). There are many manually created sentiment lexicons such as the NRC Emotion Lexicon (Mohammad and Turney (2010)), MPQA Lexicon (Wilson et al. (2005)), and Bing Liu Lexicon (Hu and Liu (2004)). General lexical resources have been adapted to the domain of sentiment analysis as well. For instance, SentiWordNet (Esuli and Sebastiani (2007b)) extends WordNet (Miller (1995)) such that each group of synonyms in WordNet, a manually-created lexical database, is tagged with three sentiment scores: positivity, negativity, objectivity. These lexical resources are very useful but manual efforts to create them are costly and time-consuming (Mohammad and Turney (2010)), requiring experts or crowdsourced annotators. This has inspired great interest in automatically inducing sentiment lexicons.

Much work has been focused on Twitter, a microblogging website with hundreds of millions of users from around the world. User-generated text is always short, as tweets are limited to 280 characters. Mohammed et al. (Mohammad et al. (2013)) construct a sentiment lexicon for Twitter based on calculating how closely a word is associated with positive or negative sentiment. A word's association score is calculated using the pointwise mutual information (PMI) between the word and a seed set of hashtags, such as *#good* and *#bad*.

Many other lexicon induction methods use label propagation to build sentiment lexicons from a seed set of words (Rao and Ravichandran (2009); Esuli and Sebastiani (2007a)). Typically, a lexical graph is built, where each word or phrase is a node and edges represent the similarity between two nodes. Then, propagation methods are used to determine the sentiment of each node, given the sentiment of an initial set of nodes.

Most of these lexicons are built for large, general domains like Twitter. However, the sentiment of a word depends on the specific domain in which it is used. Recent work builds domain-specific sentiment lexicons using label propagation methods and domain-specific corpora (Hamilton et al. (2016)).

Lexicons are also used for many tasks outside of sentiment analysis. For instance, LIWC, a general lexicon, is used to quantitatively analyze content in tasks ranging from personality prediction (Schwartz et al. (2013); Pennebaker and Graybeal (2001)) to deception detection (Ott et al. (2011)).

2.2. Text Representations

There are numerous ways of representing text for computational processing, most of which transform text into a numerical vector. These vectors ideally embed important characteristics of the text, such as the semantics.

Classical representations of text include bag-of-words (BOW), where a body of text is represented as the set of words that compose it, and latent semantic analysis (LSA) (Deerwester et al. (1990)), where the representation is derived from the factorization of a term-document occurrence matrix.

Recent text embedding methods such as Word2Vec (Le and Mikolov (2014)) and GloVe (Pennington et al. (2014)) are able to capture semantic relationships such as "man is to woman as brother is to sister." A particular type of the Word2Vec model, skip-gram with negative sampling, has been shown to be implicitly factorizing a word-context matrix (Levy and Goldberg (2014); Levy et al. (2015)). There have been many extensions of these methods that embed larger bodies of text such as sentences, paragraphs and entire documents (Le and Mikolov (2014); Kiros et al. (2015)). A downside of neural embedding models like Word2Vec is the prerequisite of large amounts of training data. For instance, the pre-trained Word2Vec vectors released by Google were trained on part of the Google News dataset, containing about 100 billion words.

Representations for sets of words such as phrases and sentences can be constructed by linearly averaging the embeddings of the constituent words. This has remained a strong feature or baseline across many tasks (Faruqui et al. (2015); Kenter and De Rijke (2015); Yu et al. (2014); Kenter et al. (2016)).

3. Predicting Alumni Donations

We conduct our exploration in the context of a donation prediction task, in which we attempt to determine the likelihood of an alumnus/alumna to donate, based on the limited background data available for that person. This is not a straightforward task. Previous studies on alumni donations (Hoyt (2004); Meer and Rosen (2012); McDearmon (2013)) found that there are many different contributing factors to alumni giving, including having the capacity to give, extracurricular involvement during the time at the university, and the prestige of the university.

We use the dataset described in this section. The ground truth is extracted from the alumni donation history, where those who have donated \$10,000 or more to a single fund are designated as having donated, and those who have not donated anything to any fund are designated as not having donated. The resulting set of alumni has a much greater number of non-donors than donors. There are 31,780 non-donors, as compared to 655 donors, which allow models to achieve 98% donor classification accuracy by simply classifying all samples as the majority class. Sampling methods to balance classes are commonly used when working with imbalanced data. We therefore create a balanced dataset by including all of the 655 alumni who donated more than \$10,000 and randomly sampling an equal number of those who donated nothing.

In all of our experiments, we use 10-fold cross validation, resulting in training and test set sizes of 1179 and 131 respectively for each split. We use

Name	Educational	Professional
Amanda Alamns	MSE in Electrical Engineering - 2000	Electrical Engineer, Senior Project Engineer, Principal Systems Engineer
Bob Beustton	BS in Economics - 2000	Financial Analyst Trainee
Claire Carshter	BS/Teaching Certificate in Elementary Education - 2000	Elementary School Teacher, CEO of EduStartup

Table 1: Fictitious Alumni Examples

a logistic regression model with L2 penalties and a regularization parameter C of 1.0 in all cases.[1]

3.1. What Makes a Donor?

We want to be able to predict whether a person will donate from her personal and professional attributes. Let us consider the fictitious alumni in Table 1 (real examples could not be used due to privacy agreements). Amanda Alamns graduated with a graduate degree in engineering and has steadily climbed the ranks in her professional career. From her position in her career, we can infer that she has the means to donate. Bob Beustton, on the other hand, has somehow remained a trainee for over a decade. It is unlikely that he will make any donations for the time being. Lastly, we have Claire Carshter. If we look solely at her educational history and first job, it appears unlikely that she would donate; the teaching profession is not known for its lucrative opportunities. However, we see that she then proceeded to start her own company. She appears to be a successful individual and is probably more likely to donate because she has the means to do so. Additionally, perhaps her experience at the university helped inspire her to pursue entrepreneurship.

3.2. Data Description

The work in this paper is based on a database of alumni information maintained by a large, public Midwest university. We call this dataset Donor

[1] We also obtained results using an SVM classifier, but obtained results and trends similar to those obtained from a logistic regression model. We therefore show results only for the regression model.

Source	Features
DI	age, gender, graduation year, degree level^T, degree type^T, degree major^T
LinkedIn	city, state, country, most recent three job titles^T, most recent three companies^T, NAICS number

Table 2: Dataset features (text fields are marked with ^{T})

Information (DI). In addition, we also have a dataset of public LinkedIn profiles for a subset of the alumni who are in DI. The DI dataset contains each alumna's donation history along with her educational history while at this particular university.

An alumna's educational history contains her major, graduation year, degree level (e.g. Bachelor's level, Master's level, Doctoral Level), and degree type (e.g. BS, MD, PhD). Every record in the LinkedIn dataset contains all job titles and companies listed on the corresponding LinkedIn profile. In our experiments, we only consider the most recent three job titles and companies. We consider the degree level, degree type, degree major, and the most recent three job titles and companies as text fields that are used both as categorical features and as input for the textual feature methods.

There are 56,259 people who appear in both the DI and LinkedIn datasets; we focus on this subset of alumni. Of this set, approximately half have donated some amount. However, many donations are on the order of a few dollars. Therefore, we further hone in on those alumni who have donated more than $10,000 to a single fund.

To represent a person, we extract categorical features such as major, recent job titles, gender, and age, among others. To focus our results on the effects of textual enhancement, we use only the categorical features that can also serve as textual features. Each instance in our dataset is then represented as a feature vector that encodes all of the categorical features by concatenating one-hot embeddings of each feature. Table 2 lists all the features that are available in the dataset.

3.3. Qualitative Analysis

To gain further insight into the data, we conduct several qualitative analyses of the backgrounds of donors. We first look at the percentage of people who donate at different degree levels, shown in Figure 1. Of the different de-

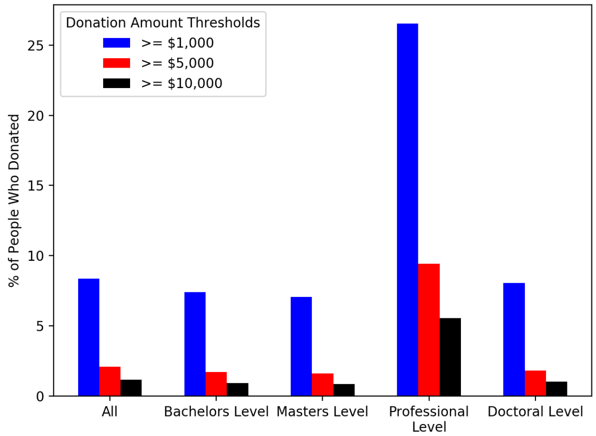

Figure 1: Percentage of population who donated for obtained degree levels at several donation amount thresholds.

gree levels, a much higher percentage of those with professional level degrees are donors. This is consistent across different donation amount thresholds. The donor statistics of the other degree levels are consistent with the overall statistics, across the entire population.

We further look at different types of professional level degrees, which are comprised of various medical and law degrees. The five professional degree types with the highest percentages of donors are shown in Figure 2. We see that Juris Doctor degrees (J.D.) and Doctor of Medicine degrees (M.D.) are among the top five, which is consistent with the correlation lexicons that we automatically generate, as described in the next section.

Medical residencies (Med. Res.) and medical fellowships (Med. Fellowship) occur much less than J.D.s and M.D.s in our dataset, which could have contributed to their lack of representation in the lexicons.

Finally, we look at the number of popular majors across different departments. We see that those who studied law consistently donated more than the others across the different donation thresholds. We also see that education majors have a higher percentage of donors than other popular majors shown in Figure 3. This could be because those who choose to pursue education are more philanthropic by nature, wanting to teach and help others without the promise of a high salary.

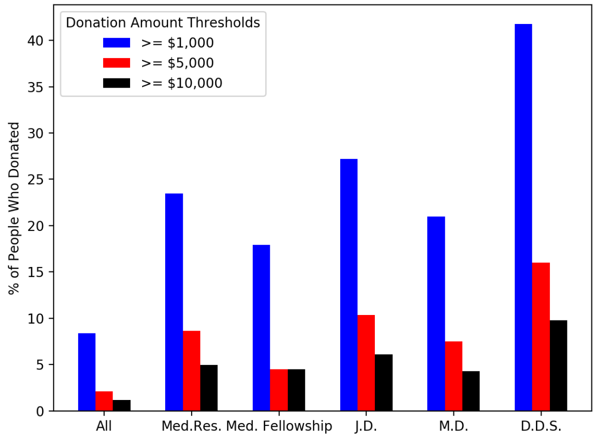

Figure 2: Percentage of population who donated for obtained professional degree types at several donation amount thresholds.

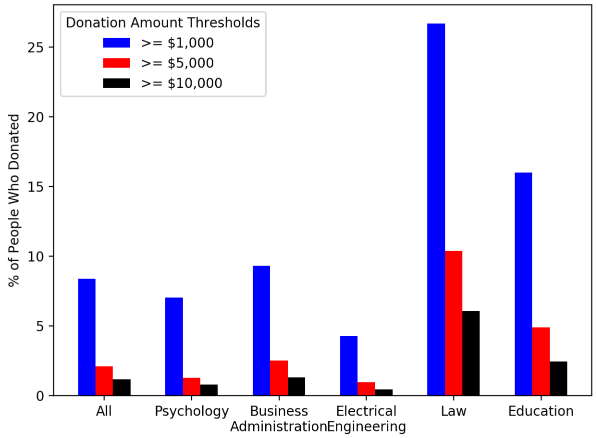

Figure 3: Percentage of population who donated for obtained majors of study at several donation amount thresholds.

4. Text Expansion using Domain-Specific Knowledge

A core hypothesis of our work is that the sparse text that is available in many sources of data, such as our alumni dataset, can still hold much useful information. To make the sparse text useful, we can augment the text with additional information by using natural language processing methods that leverage knowledge about the target domain drawn from within or outside the dataset.

We explore four main methods, described in detail below: (1) word embeddings obtained from a domain-specific corpus; (2) correlation lexicons that aim to identify from within the dataset additional words that are indicative of donations; (3) lexicons induced starting with a few seeds and using external corpora and graph propagation; and (4) domain-specific distance representations, reflecting the semantic similarity between the textual features and a set of domain-specific seeds.

All of these methods are illustrated, and later evaluated, using the donation prediction task and associated dataset described above.

4.1. Domain-specific Embeddings

Unsupervised methods for learning word embeddings represent one of the most recent successes in word representations (Mikolov et al. (2013); Pennington et al. (2014)). As a first method to expand the text fields we thus use word embeddings.

We construct a corpus of articles that discuss philanthropy-related topics from the New York Times that we will refer to as the NYT Philanthropy News corpus. We use their API[2] and collect 8,525 articles dated from January 1981 to March 2017. The final corpus includes 57 million words, with a vocabulary of 94,623 words. Of those, only the words that occur five times or more are considered during the training of the GloVe model; 32,324 such words exist in the corpus.

We create a set of word embeddings using the GloVe embedding model (Pennington et al. (2014)) trained on this philanthropy-focused news corpus. We chose to use GloVe as it was shown to have better performance on several word representation and word similarity tasks (Pennington et al. (2014); Hamilton et al. (2016)). We use 300 dimensions for the embeddings, as is

[2] `https://developer.nytimes.com/`

standard practice[3]. For each text field in the dataset, we take the constituent words. The embeddings for all of the words from every text field are then averaged to form a feature vector.

4.2. Correlation Lexicons

Previous work has shown that domain-specific lexicons can be effectively used to induce features for prediction tasks. Specifically, our method is inspired from previous work on sentiment analysis, where a lexicon of positive and negative words generated specifically for Twitter was found to bring significant improvements (Mohammad et al. (2013)). We adapt their method to our task, and generate a lexicon of words that are specific to the task of donation.

Using pointwise mutual information (PMI), as done in (Mohammad et al. (2013)), we measure the strength of association between each word in the dataset and the labels of donation/no-donation. The words are drawn from all the textual fields, consisting of the degree levels, degree types, degree majors, job titles, and job companies. Note that the correlations are calculated only from the training data. Specifically, given a word W, we calculate its PMI score as:

$$\begin{aligned} PMIScore(W) =& PMI(W, donated) \\ &- PMI(W, nondonated) \end{aligned} \tag{1}$$

where the $PMI(W, class)$ for any of the two classes is calculated as:

$$PMI(W, class) = \log \frac{p(W, class)}{p(W)p(class)}$$

To create the lexicon, we first calculate the $PMIScore$ for each of the words included in the text fields in the dataset, as described in Section 3.2. We then rank the words in decreasing order of their score, and select the top 30 with the assumption that the words that have the highest score are most strongly correlated with the class of donation. Table 3 shows the top 10 words from a generated lexicon.

[3] We use the author-provided code for GloVe at `https://github.com/stanfordnlp/GloVe`. All parameters are left as default other than the embedding size.

	Sample words
Top 10 (donation)	educational, partner, m.d., j.d., ceo, professional, board, law, owner, managing

Table 3: Sample words from the PMI lexicon

Using the PMI lexicon, we generate 30 binary features, one for each entry in the lexicon. We set the value of each feature to 1 (0), reflecting the presence (absence) of the feature in any of the text fields.

4.3. Seed-Induced Lexicons

The third method we consider is to generate a lexicon starting with a few seed words and expanding the set of words using a label propagation algorithm on a lexical graph. We use the SentProp method introduced in (Hamilton et al. (2016)), which was originally proposed for the task of building a lexicon for sentiment analysis.

We first manually build two sets of seed words, associated with philanthropic tendencies and the lack thereof, respectively. Table 4 shows these seed words.

	Seed words
Donation	donation, endowment, investment, charity, generosity, benefaction, giver, grantor, donor, donator, benefactor, benefactress, endow, sponsor, backer
Non-donation	miserly, stingy, uncharitable, ungenerous, frugal, selfish, skimping, scrimping, tightfisted, closefisted, parsimonious, inhospitable, greedy, cheap

Table 4: Seed words used to generate the SentProp lexicon

We then build a weighted lexical graph using the words from the text fields in the dataset, as well as all of the seed words. Each word is connected to its nearest *10* neighbors by using a measure of cosine similarity applied on word embedding representations for each word that is present in both the dataset vocabulary and the trained word embeddings. We use GloVe embeddings, following the original SentProp implementation.

The donation and non-donation labels are then propagated through the graph using a random walk method. Finally, a word's donation score is calculated as the probability of a random walk from the corresponding seed set hitting that word. In our experiments, we try lexicon generation using both generic pre-trained GloVe embeddings[4] as well as GloVe embeddings that we train on the NYT Philanthropy News corpus. They perform comparably; we only show results using the latter embeddings.

To create the final lexicon, we take only the words that have a donation association score higher than 0.7. We chose this threshold heuristically; lower thresholds introduced noisy words and higher thresholds excluded many words that appear in the dataset. Sample words from the resulting lexicon are shown in Table 5. As with the PMI lexicon, we create a feature for each of the lexicon words, and set its value as 1 (0) depending on whether the feature is present (absent) among the words in the text fields.

	Sample words
NYT GloVe based	contributor, giving, investor, management, banking, mutual, venture, institutional, profit, corporate, philanthropy, market, cash, asset, hedge, managed

Table 5: Seed-induced lexicon entries using label propagation on graphs

4.4. Seed-Similarity Embeddings

Finally, as an alternative to the previous seed-induced lexicon method, we also consider a method that measures the semantic distance between the

[4] https://nlp.stanford.edu/projects/glove/

Source	Features
Baselines	
DI	degree level, degree type, degree major
LinkedIn	most recent three job titles, most recent three companies
Text expansion features	
DomainEmbed	300-dimension GloVe embeddings trained on the donation corpus, averaged over all the words in the text fields
CorrelLex	30-word correlation lexicon generated from training data (text fields in both DI and LinkedIn); one feature for each lexicon word, reflecting presence/absence among words from text fields
SeedProp	Seed-induced donation lexicon using label propagation on a lexical graph formed by using pretrained GloVe embeddings; one feature for each lexicon word, reflecting presence/absence among words from text fields
SeedSim	Semantic similarity between the text fields and the 15 donation seeds, using cosine similarity between pretrained GloVe embeddings; one feature for each of the 15 donation seeds

Table 6: Summary of features

words in the text fields and the donation seed words. The hypothesis behind this method is that we can circumvent the need for a domain-specific corpus by measuring the distance between a small set of domain words and the text fields.

We use the same seed set as listed in Table 4 (row Donation). We use the pre-trained GloVe embeddings with 300 dimensions. For each seed word, we find the maximum cosine similarity score between that word's embedding and each of the word embeddings from the text fields. The result is a feature vector that reflects these maximum similarity scores, and is the same length as the seed set.

5. Results and Discussion

We evaluate the performance of the donation prediction task described in Section 3 using the original features available in the dataset, as well as expanded feature sets obtained with the four text expansion methods described above. Table 6 summarizes the features we use, described in the previous sections.

The top part of Table 7 shows the results obtained with the two baselines (DI features, and DI combined with LinkedIn features), while the bottom part of the table shows the results obtained when augmenting the top performing baseline with the various text-expansion features. We combine features by concatenating their feature vectors. This combination method has been shown to work well in many applications (Argamon et al. (2007); Le and Mikolov (2014); Maas et al. (2011)).

Statistical significance over the DI+LinkedIn baseline is calculated using the McNemar two-tailed test. We used an alpha value of 0.05.

Among the four text expansion methods, the correlation lexicons, domain-specific embeddings, and seed-induced lexicon result in significant improvements over the baselines as seen from Table 7. The seed-similarity embedding features also bring small improvements, but they are not found to be significant.

To gain further insight into the performance of these models, we perform several additional analyses and evaluations, which we describe next.

5.1. Model Correlation

First, we measure the correlation between the output produced by the top baseline model (DI+LinkedIn) and by the four different methods considered.

Source	Accuracy
BASELINES	
DI	68.8%
DI+LinkedIn	76.8%
TEXT EXPANSION FEATURES	
DI+LinkedIn+DomainEmbed	81.3%*
DI+LinkedIn+CorrelLex	80.1%*
DI+LinkedIn+SeedProp	78.5%*
DI+LinkedIn+SeedSim	77.2%

Table 7: Donation prediction results using text expansion methods. Results with * are statistically significant compared to the DI+LinkedIn baseline system.

	DI+LinkedIn	+SeedSim	+SeedProp	+DomainEmbed	+CorrelLex
DI+LinkedIn	1.0	0.90	0.83	0.56	0.63
+SeedSim		1.0	0.83	0.57	0.63
+SeedProp			1.0	0.56	0.66
+DomainEmbed				1.0	0.68
+CorrelLex					1.0

Table 8: Pearson correlation coefficients among the output of the four models and the baseline. Each model includes the DI and LinkedIn categorical features with the specified additional single feature type.

Source	Accuracy
DI	66.5%
DI+LinkedIn	72.6%
DI+LinkedIn+DomainEmbed	75.3%*
DI+LinkedIn+CorrelLex	75.6%*
DI+LinkedIn+SeedProp	73.3%
DI+LinkedIn+SeedSim	73.9%

Table 9: Classification results when non-donors include alumni who donated any amount below $10,000. Results with * are statistically significant compared to the DI+LinkedIn baseline.

Table 8 shows the Pearson correlation between all the pairs of two models. As seen in this table, most of the models are medium correlated, which indicates there is some overlap between the predictions they make. The models that are most divergent from the baseline are the correlated lexicons (CorrelLex) and the domain specific embeddings (DomainEmbed), which is also reflected in the higher performance of these models (see Table 7). The highest correlation is found between the model that measures the similarity with the seed set (SeedSim) and the model that performs label propagation on a lexical graph starting with the seed set (SeedProp); their high correlation is likely a reflection of the dependence of these two models on the same seed set.

5.2. Classification with Less Distinguishable Classes

We also perform an evaluation for a classification task where the division between donors and non-donors is less clear. Specifically, we again consider all the alumni who donated $10,000 and above as donors, but now we consider alumni who donated any amount below $10,000 as a non-donor. The non-donors are randomly sampled from the instances corresponding to people who donated less than $10,000. Table 9 shows the results obtained during these evaluations. As expected, all the results are lower than the ones obtained during the earlier evaluations. In this more difficult setup, the use of correlation lexicons and domain embeddings continues to bring consistent improvements over the baseline.

5.3. Influence of Seeds on Sentprop Lexicon

We analyze the effects of changing the seed words used for SentProp by looking at the overlap between our original lexicon and the new lexicons generated using different seed words. To highlight the effects of different donation words, we change the donation words to be entirely different from those used in our experiments, but maintain the topic of donation among the words. For each of the three different donation word sets, the non-donation words remain the same as those used in our experiments. To understand the influence of the non-donation words, we also choose a set of random words as non-donation words. For this, we retain the same set of donation words as used in our experiments. The chosen words are shown in Table 10.

For these different sets of seed words, we generate lexicons with SentProp on the NYT Donation corpus at two different association score thresholds. We measure the overlap between the new lexicons and the original one used in our experiments by calculating their Jaccard similarity and overlap coefficient. For two sets of words, X and Y, we have

$$Jaccard(X,Y) = |X \cap Y|/|X \cup Y|$$

and

$$Overlap(X,Y) = |X \cap Y|/min(|X|,|Y|).$$

The new lexicons corresponding to altered donation words (Set 1, Set 2, Set 3), generated at an association score threshold of 0.7, contain many more words that are not related to philanthropy. The large size of the lexicon is indicative of this. This could be a result of the seed words not being as unambiguously tied in topic as the original set of seeds. However, the overlap coefficients are close to 1, showing that the original lexicon words are present in the new lexicons. This implies that the donation topic was still captured, but with much more noise.

We raise the association score threshold to filter out the less relevant words. Overall, the lexicons resulting from Set 1, Set 2, and Set 3 still maintain much overlap with the original lexicon. Set 2's lexicon has a much lower overlap coefficient than Set 1 or Set 3. This is likely because Set 2's donation seeds contain words like "kind" and "charitable" that have more ambiguous meanings.

Interestingly, having random non-donation words does not greatly perturb the captured topic of the lexicon. All words generated also appear in the original lexicon. Additionally, the number of words is actually smaller than

Seed Words	Donation	Non-donation
Set1	contribution, gift, funding, foundation	miserly, stingy, uncharitable, ungenerous, frugal, selfish, skimping, scrimping, tightfisted, closefisted, parsimonious, inhospitable, greedy, cheap
Set2	kind, supporter, charitable, patron	
Set3	contribution, gift, funding, foundation, kind, supporter, charitable, patron, compassionate	
NegRand	donation, endowment, investment, charity, generosity, benefaction, giver, grantor, donor, donation, benefactor, benefactress, endow, sponsor, backer	cattle, evanescent, vague, jittery, trade, grade, excited, signify, clear, toad

Table 10: Different sets of seed words used for SentProp. Sets 1-3 retain the same set of non-donation words as used in the experiments, but with different donation words. NegRand retains the same set of experiment donation words, but with random non-donation words.

the original set. This may be because having random non-donation words encourages SentProp to choose words that are unambiguously related to the donation words. There is a separation of the donation topic from effectively all others, rather than from just the non-donation topic. This is desirable in applications where we are primarily interested in generating a lexicon related to one theme, rather than two polar themes, as is the case here.

5.4. Error Analyses

To better understand the performance of our methods, and where and when they fail, we perform several error analyses. Specifically, since the job related information from LinkedIn was the most varied, and therefore the area that could benefit the most from our text expansion methods, we mainly focus our analyses on how well our features understood LinkedIn information. We have anonymized the examples below by excluding names and modifying job titles to be generic.

Categorical features were not able to understand complex or non-standard job titles such as “Director of Major, Planned, and Special Gifts”, “Senior Director of Major Gifts”, or “CEO of A Philanthropic Trust”. These particular job titles are highly indicative of philanthropic tendencies, but

Seed Words (Threshold)	Lexicon Size	Jaccard Sim.	Overlap Coef.
Set1 (0.7)	897	0.14	0.95
Set2 (0.7)	1667	0.08	1.00
Set3 (0.7)	929	0.13	0.95
NegRand (0.7)	49	0.37	1.00
Set1 (0.8)	37	0.17	0.65
Set2 (0.8)	126	0.21	0.35
Set3 (0.8)	48	0.21	0.65
NegRand (0.8)	0	0.00	0.00

Table 11: Number of words, Jaccard similarity, and overlap coefficients for different sets of seed words at different association score thresholds. Jaccard similarity and overlap coefficient are calculated with respect to the generated lexicon used in the experiments. The original generated lexicon has 132 words.

the categorical-only DI+LinkedIn model classified these individuals as non-donors. The categorical model also was not able to correctly detect people working in known high-pay fields because of non-standard titles. For instance, one donor is a "Pulmonary Specialist", which is a type of doctor. From our data (Figure 4), we can see that health care professionals are the most charitable individuals. However, the categorical model was unable to make the association between "Pulmonary Specialist" and the health care profession.

The embedding features helped find such associations. The "Pulmonary Specialist" was found to be a donor by the model that incorporated embedding features. It was also much better at detecting individuals with advanced career positions such as "Senior Vice President", "Strategic Advisor", and "Executive Director". While these titles may seem obvious, there exist many variations on advanced titles, such as "Managing Director", "Principal Advisor", and "Creative Director". Embedding features implicitly help the model understand that positions like these are indicative of donors, without explicitly having a list of such titles. However, the embeddings were not good at distinguishing between those who had a single position indicative of a donor and those who had a history of such positions.

The correlation lexicon features focused on finding individuals that held multiple indicative positions. For example, some of the donors that were correctly identified only by including CorrelLex each had at least three advanced career positions. One was a “Senior Counselor”, “CEO”, and “Founder”; another was a “Senior Development Officer”, “Consultant”, and “President”; and yet another was a “Senior Manager”, “Chief Operating Officer”, and “Senior Clinical Manager”. These results follow the fact that the generated correlation lexicons from Table 3 seem to mainly focus on high income or advanced titles. However, this misses people who are philanthropic but do not necessarily hold traditional advanced positions.

Some of the donors that were correctly identified only by using SeedProp had titles such as “Head of Police Board”, “Workplace Learning Specialist”, “Program Director/Scholarship Manager”, or worked at foundations. These careers involve public service, interacting with people, and being in environments that are geared towards philanthropy. SeedSim produced similar results, though the detected associations were limited to very explicit indicators, such as someone being an “Evangelist”.

5.5. Evaluation on Other Datasets

We also want to determine to what extent our methods can be applied to other datasets. Although there are public donation records available at crowdfunding sites such as Kickstarter.com and DonorsChoose.org, there is usually little information revealed about the donors themselves beyond what they have donated to.

We therefore evaluate our proposed text expansion methods on a different task: gender classification on a dataset of blog profiles collected from Blogger.com. Previous work has shown that it is possible to detect demographic information such as gender from writings and social media content Farnadi et al. (2018); Mukherjee and Liu (2010); Schler et al. (2006); Sarawgi et al. (2011).

Bloggers can choose to fill in information, such as gender, occupation, and interests, on their profile page. We use a set of 76,971 profiles that have both gender and interests listed and are in the USA. The full set of features is listed in Table 12. We classify each blogger as male or female based on the information available on their profile.

Our text expansion features are replicated for gender in this setting. The domain embeddings are trained on the blog dataset. The correlation lexicon is generated from the training set of blog data. Gender-based seed words are used for SeedProp and SeedSim. The results are shown in Table 13. All of the text expansion methods improve significantly over the baseline categorical method, with the domain embeddings (DomainEmbed) yielding

Source	Features
Blogs	gender, interests,^{T}, occupation^T, city, state, country, introduction^T, movies^T, music^T, books^T

Table 12: Blog dataset features (text fields are marked with ^{T})

Source	Accuracy
Blogs	70.6%
Blogs+DomainEmbed	83.3%*
Blogs+CorrelLex	75.6%*
Blogs+SeedProp	72.0%*
Blogs+SeedSim	74.5%*

Table 13: Gender prediction results on blog profiles. Results with * are statistically significant compared to the Blog baseline.

the highest performance. These results demonstrate that our methods can be successfully applied to other datasets.

6. Conclusions

In this paper, we explored whether we can enhance sparse textual content to improve data-driven predictions using the task of alumni donation prediction.

We introduced a dataset of alumni donations, and we qualitatively analyzed the donations and the backgrounds of the donors to highlight the differences between the backgrounds of donors and non-donors as well as the patterns of donations attracted by different academic departments.

We used four different methods of expanding sparse text, including lexicon generation methods and text embedding methods. We evaluated these methods on the task of predicting whether someone is likely to donate, and compared with baseline models that do not make use of any textual features.

We showed that we can classify large donors from non-donors with an accuracy of up to 80%. We also showed that the enrichment of sparse text through the extraction and use of textual features does benefit model performance. Our domain-specific embeddings and correlation-based lexicon consistently improved over the baseline models that only use categorical features. We also showed that our methods can be successfully applied to other sparse-text datasets.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., Levitan, S., 2007. Stylistic text classification using functional lexical features. Journal of the American Society for Information Science and Technology 58 (6), 802–822.

Banerjee, S., Pedersen, T., 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In: International conference on intelligent text processing and computational linguistics. Springer, pp. 136–145.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R., 1990. Indexing by latent semantic analysis. Journal of the American society for information science 41 (6), 391.

Esuli, A., Sebastiani, F., 2007a. Pageranking wordnet synsets: An application to opinion mining. In: ACL. Vol. 7. pp. 442–431.

Esuli, A., Sebastiani, F., 2007b. Sentiwordnet: A high-coverage lexical resource for opinion mining. Evaluation, 1–26.

Farnadi, G., Tang, J., De Cock, M., Moens, M.-F., 2018. User profiling through deep multimodal fusion. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM, pp. 171–179.

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., Smith, N. A., 2015. Retrofitting word vectors to semantic lexicons. In: Proc. of NAACL.

Hamilton, W. L., Clark, K., Leskovec, J., Jurafsky, D., 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. arXiv preprint arXiv:1606.02820.

Hoyt, J. E., 2004. Understanding alumni giving: Theory and predictors of donor status. Online Submission.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 168–177.

Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T., 2011. Target-dependent twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, pp. 151–160.

Kenter, T., Borisov, A., de Rijke, M., 2016. Siamese cbow: Optimizing word embeddings for sentence representations. CoRR abs/1606.04640.

Kenter, T., De Rijke, M., 2015. Short text similarity with word embeddings. In: Proceedings of the 24th ACM international on conference on information and knowledge management. ACM, pp. 1411–1420.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Skip-thought vectors. In: Advances in neural information processing systems. pp. 3294–3302.

Le, Q. V., Mikolov, T., 2014. Distributed representations of sentences and documents. In: ICML. Vol. 14. pp. 1188–1196.

Levy, O., Goldberg, Y., 2014. Neural word embedding as implicit matrix factorization. In: Advances in neural information processing systems. pp. 2177–2185.

Levy, O., Goldberg, Y., Dagan, I., 2015. Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics 3, 211–225.

Maas, A., Daly, R., Pham, P., Huang, D., Ng, A., Potts, C., 2011. Learning word vectors for sentiment analysis. In: Proceedings of the Association for Computational Linguistics (ACL 2011). Portland, OR.

McDearmon, J. T., 2013. Hail to thee, our alma mater: Alumni role identity and the relationship to institutional support behaviors. Research in Higher Education 54 (3), 283–302.

Meer, J., Rosen, H. S., 2012. Does generosity beget generosity? alumni giving and undergraduate financial aid. Economics of Education Review 31 (6), 890–907.

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

Miller, G. A., 1995. Wordnet: a lexical database for English. Communications of the ACM 38 (11), 39–41.

Mohammad, S., Kiritchenko, S., Zhu, X., 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of Semeval.

Mohammad, S. M., Turney, P. D., 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, pp. 26–34.

Mukherjee, A., Liu, B., 2010. Improving gender classification of blog authors. In: Proceedings of the Conference on Empirical Methods in natural Language Processing. pp. 207–217.

Ott, M., Choi, Y., Cardie, C., Hancock, J. T., 2011. Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, pp. 309–319.

Pennebaker, J. W., Graybeal, A., 2001. Patterns of natural language use: Disclosure, personality, and social integration. Current Directions in Psychological Science 10 (3), 90–93.

Pennington, J., Socher, R., Manning, C. D., 2014. Glove: Global vectors for word representation. In: EMNLP. Vol. 14. pp. 1532–1543.

Rao, D., Ravichandran, D., 2009. Semi-supervised polarity lexicon induction. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 675–682.

Sarawgi, R., Gajulapalli, K., Choi, Y., 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In: Proceedings of the Fifteenth

Conference on Computational Natural Language Learning. Association for Computational Linguistics, pp. 78–86.

Schler, J., Koppel, M., Argamon, S., Pennebaker, J., 2006. Effects of age and gender on blogging. In: Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. Stanford, pp. 199–204.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al., 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one 8 (9), e73791.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M., 2011. Lexicon-based methods for sentiment analysis. Computational linguistics 37 (2), 267–307.

Wilson, T., Wiebe, J., Hoffmann, P., 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, pp. 347–354.

Yu, L., Hermann, K. M., Blunsom, P., Pulman, S. G., 2014. Deep learning for answer sentence selection. CoRR abs/1412.1632.