Domain-Specific Sentiment Lexicons Induced from Labeled Documents

SM Mazharul Islam
University of Texas at Arlington
sxi7321@mavs.uta.edu

Xin Dong
Rutgers University–New Brunswick
xin.dongx@rutgers.edu

Gerard de Melo
Hasso Plattner Institute, University of Potsdam
http://sentimentanalysis.org
gdm@demelo.org

Abstract

Sentiment analysis is an area of substantial relevance both in industry and in academia, including for instance in social studies. Although supervised learning algorithms have advanced considerably in recent years, in many settings it remains more practical to apply an unsupervised technique. The latter are oftentimes based on sentiment lexicons. However, existing sentiment lexicons reflect an abstract notion of polarity and do not do justice to the substantial differences of word polarities between different domains. In this work, we draw on a collection of domain-specific data to induce a set of 24 domain-specific sentiment lexicons. We rely on initial linear models to induce initial word intensity scores, and then train new deep models based on word vector representations to overcome the scarcity of the original seed data. Our analysis shows substantial differences between domains, which make domain-specific sentiment lexicons a promising form of lexical resource in downstream tasks, and the predicted lexicons indeed perform effectively on tasks such as review classification and cross-lingual word sentiment prediction.

1 Introduction

Sentiment analysis is among the most prominent forms of natural language processing, with applications such as social media analytics (Rosenthal et al., 2017; Wang et al., 2019; Shoeb et al., 2019), marketing and customer support (Gamon, 2004), as well as recommendation (Yang et al., 2013). Apart from machine learning-driven systems (Pang et al., 2002; Socher et al., 2013; Kalchbrenner et al., 2014, inter alia), which require supervision using labeled training data, there are also lexical resource-driven systems that exploit *sentiment lexicons* and can be run out-of-the-box without the need for any labeled training data. Well-known sentiment lexicons include the Hu and Liu (2004) Opinion Lexicon, SentiWordNet (Baccianella et al., 2010), LIWC (Pennebaker et al., 2001), and VADER (Hutto and Gilbert, 2014). There are numerous techniques for lexicon-driven sentiment analysis (Taboada et al., 2011), SentiStrength (Thelwall et al., 2010) being an example of a more modern lexicon-driven sentiment analysis system. Sentiment lexicons can also be used to bootstrap domain-specific supervised sentiment analysis models (Mudinas et al., 2018).

A sentiment lexicon is a resource that, for a given word (form) w, provides an annotation label l_w describing its overall sentiment polarity. Some lexicons merely provide labels in $\{\texttt{positive}, \texttt{negative}\}$ or $\{\texttt{positive}, \texttt{neutral}, \texttt{negative}\}$. Others offer more informative intensity scores to account for the fact that some words are more negative or positive than others. For example, an emphatic word such as *spectacular* is generally considered stronger than a simple *good* (de Melo and Bansal, 2013). Such scores could be in the range $[-1, 1]$, with -1 denoting the most negative sentiment polarity, whereas $+1$ is the most positive score.

In this paper, we consider two perennial problems with sentiment lexicons:

(i) Sentiment lexicons are based on an abstract domain-independent and context-independent notion of sentiment polarity. In reality, the polarity of a word depends substantially on what one is talking

Proceedings of the 28th International Conference on Computational Linguistics, pages 6576–6587
Barcelona, Spain (Online), December 8-13, 2020

about and how the word is used. For example, when talking about *music*, the word *hot* tends to be positive. When talking about a laptop, the laptop often becoming *hot* would be more negative.

(ii) Sentiment lexicons are typically manually created, and thus have limited coverage. The widely used Hu and Liu (2004) Opinion Lexicon, for instance, consists of around 6,800 words. While this is by no means a small number, the lexicon is still likely to miss important signals.

To mitigate these shortcomings, we induce domain-specific sentiment lexicons using an automated data-driven approach. In our experiments, we consider a corpus of reviews from 24 different domains and first induce seed lexicons using linear predictors. Subsequently, we extend their coverage based on large-scale word vector representations with a deep neural regression model. While our lexicons do not resolve the issues of context and polysemy – these are perhaps best addressed within a full-fledged machine learning architecture – many differences in sentiment polarity for a word stem from divergent uses across different domains. Our experiments confirm that there are substantial differences between domains and that the predicted lexicons prove useful in review classification and cross-lingual word-level sentiment prediction.

2 Method

Our approach proceeds in two steps. First, we rely on labeled documents for a set of different domains to induce seed data for each of the domains using simple linear predictors.

This seed data already accounts for the differences between domains. However, after the first step, the coverage of the resulting seed data is limited to words occurring in the labeled corpora, which may be small. Hence, in a second step, we rely on deep neural models, exploiting vector representations of words to learn sentiment intensity scores for a much larger vocabulary.

2.1 Seed Data Induction

Our approach for seed data induction is simple. Given n domain-specific document sets $\mathcal{D}_i \in \mathcal{X} \times \mathcal{Y}$ $(i = 1, \ldots, n)$ labeled with sentiment polarity labels in $\mathcal{Y} = \{\texttt{positive}, \texttt{negative}\}$, we learn n corresponding linear binary classification models using bag-of-words features. Then, each word present in the vocabulary is assigned a series of domain-specific sentiment polarity scores, by consulting the linear coefficients for the respective word across the n linear models.

Specifically, for each $\mathcal{D}_i$, we define the set of features as $\mathcal{F}_i = \mathcal{V}_i \cup \{\overline{w}_j \mid w_j \in \mathcal{V}_i\}$, where $\mathcal{V}_i$ is the term vocabulary of $\mathcal{D}_i$ and $\overline{w}_j$ denotes a negated version of word w_j. In our experiments, we lower-case all terms and simply treat occurrences of "*not* $\langle w_j \rangle$" in the text as negated features, while all other word occurrences are mapped to unnegated features. Of course, one could also invoke much more sophisticated negation detection methods.

Thus, for each $\mathcal{D}_i$, we can map the documents x_j in $\mathcal{D}_i$ to term frequency-based document vectors $\mathbf{x}_j$ in feature space $\mathcal{F}_i$. Along with the labels $y_j \in \mathcal{Y}$ that are given in each $\mathcal{D}_i$, we thus obtain n different labeled feature vector sets $\hat{\mathcal{D}}_i = \{(\mathbf{x}_j, y_j) \mid (x_j, y_j) \in \mathcal{D}_i\}$. These are invoked to train n different linear models

$$f_i(\mathbf{x}) = \mathbf{w}_i^\intercal \mathbf{x} + b_i. \tag{1}$$

Subsequently, for any word $w_j \in \mathcal{V}_i$, we consider its particular score in domain i to be $w_{i,j}$, i.e., the linear coefficient for that word in the weight vector $\mathbf{w}_i$ obtained for the trained model f_i. We disregard the negated features, as their frequency tends to be too low to provide a reliable complementary signal. Rather, the main purpose of the negated features is to eliminate noise that might otherwise affect the primary word features.

2.2 Neural Vector-Based Expansion

The use of supervised learning based on domain-specific datasets $\mathcal{D}_i$ to induce the seed data has two notable drawbacks:

(i) The coverage of words for some domains i may be low, as it is limited to words in the respective labeled training set vocabulary $\mathcal{V}_i$.

(ii) The reliability of induced seed scores may be low if a word was infrequent in the respective domain-specific labeled corpus $\mathcal{D}_i$.

Machine learning based on large-scale distributional semantics as reflected in word vector representations can allow us to overcome the above shortcomings and enable the sentiment scoring of millions of words. Specifically, for each domain i, we train a model $\phi_i(\mathbf{v}_w) \in \mathbb{R}$ to predict a real-valued domain-specific sentiment polarity score for a word w based on its generic vector representation $\mathbf{v}_w$ as input. Word vectors trained on large amounts of data (Mikolov et al., 2013; Pennington et al., 2014) capture important aspects of lexical semantics. Although they are typically trained based on distributional word co-occurrence information, they have also been found to reveal sentiment signals (Rothe et al., 2016).

As the machine learning component, we consider deep neural regression networks as our prediction models $\phi_i(\mathbf{v}_w)$. The architecture is described in Table 1. In particular, we incorporate several hidden layers, but add batch normalization and dropout for regularization. Additionally, we found that initializing the output layer of our model to scale the softmax scores to the sentiment score range observed in the training data proves beneficial. Further training details are given in Section 3.2.

To train these models, we rely on the automatically induced seed data from Section 2.1 as training data for each domain. However, we need to account for the second observation above, i.e., the fact that the reliability of induced seed scores may be low if a word was observed only a few times in the domain-specific corpus $\mathcal{D}_i$. For such words, the predictors $f_i(\mathbf{x})$ (Eq. Section 1) may not have received sufficient signal about their polarity, whereas sentiment scores for words with sufficiently high frequency are expected to be more accurate. To address this, for a given domain i, the corresponding training data is defined as

$$T_i = \{(w_j, w_{i,j}) \mid w_j \in \mathcal{V}_i, \sum_{(x,y)\in\mathcal{D}_i} f(x, w_j) \geq f_{\min}\} \tag{2}$$

where $f(x, w_j)$ denotes the term frequency of word w_j in document x and $f_{\min}$ is a predefined minimal training corpus frequency threshold.

Thus, for each domain i, T_i serves as training data to train a deep neural regression model $\phi_i(\mathbf{v}_w)$ to predict a word w's domain-specific sentiment polarity in that domain, based on w's word vector $\mathbf{v}_w$.

Table 1: Architecture of deep neural regression model for vector-based expanded sentiment prediction

Layers	Dimensionality	Details
Input Layer	300	Word vector of dimensionality 300
FC Layer 1	500	Followed by batch-normalization, ReLU, and dropout layers
FC Layer 2	500	Followed by batch-normalization, ReLU, and dropout layers
FC Layer 3	100	Followed by batch-normalization, ReLU, and dropout layers
FC Layer 4	9	Followed by batch-normalization, soft-max activation
Output Layer	1	Custom-initialized layer to scale output from soft-max layer

3 Results

In the following, we report on a series of experimental results to assess the merits of our proposal. In Section 3.1, we induce seed data based on a large-scale review data set. In Section 3.2, we then proceed with our domain-specific neural expansion approach. We first evaluate it on human-labeled data, and subsequently apply it to the complete vocabulary to induce large-scale domain-specific lexicons with high coverage. Finally, in Section 3.3, we evaluate the effectiveness of these induced domain-specific lexicons on review classification and cross-lingual word-level sentiment prediction.

3.1 Seed Data Induction Experiments

As our input corpus, we considered a collection of 142.8 million English language reviews from Amazon.com for the time period spanning May 1996 to July 2014, which has been made publicly available

online.[1] The reviews are categorized with respect to an inventory of 24 different classes of products, as listed in Table 2.

The ratings are given on a 5-point scale. We regarded reviews with a rating < 3 as negative, while those with a rating > 3 were deemed positive. Three-star reviews were considered neutral and disregarded for seed model training.

We then followed the approach from Section 2.1 by training 24 linear support vector machine models for binary classification, and extracting the resulting linear coefficients for word features as seed data for those words. The coverage of the resulting data is given in the "Seed (All)" and "Seed (Non-neutral)" columns of Table 2. The non-neutral counts refer to words for which the absolute score is above 0.2, i.e., negative scores <-0.2 as well as positive ones >0.2. We observe that the large corpus gives us orders of magnitude better coverage than existing hard-crafted sentiment lexicons. Still, the coverage differs substantially by domain, and for some we have only limited coverage with high magnitude.

3.2 Neural Vector-Based Expansion Experiments

Our subsequent experiments on the neural vector-based expansion proceeded in two major phases. First, we validated our expansion approach on a smaller dataset, such that the prediction from our system can be verified against human ground truth ratings. After establishing its accuracy, we proceeded to apply this approach on the 24 domains from Section 3.1.

3.2.1 Validation on VADER lexicon

Data. We started off our experiment with a domain-independent, generic sentiment prediction system such that we could draw on ground truth sentiment scores for words solicited from a group of human test subjects. In particular, we relied on the VADER lexicon (Hutto and Gilbert, 2014), a collection of 7,504 unique English words along with mean sentiment rating in $[-4, 4]$, standard deviation, and raw human sentiment ratings from each test subject, as our pilot dataset.[2] As word vectors, we adopted GloVe CommonCrawl embeddings (Pennington et al., 2014), and we eliminated any words in VADER that are not present in GloVe. A random split of 60%/20%/20% with equally diversified sentiment scores (illustrated in Figure 1a) was used to create train/validation/test portions.

Training. To train the model, we relied on a batch size of 32, dropout rate of 20%, and Adam optimization with an initial learning rate of 0.001, dynamic learning rate schedule (halving after 4 epochs of validation loss stagnation), and early stopping.

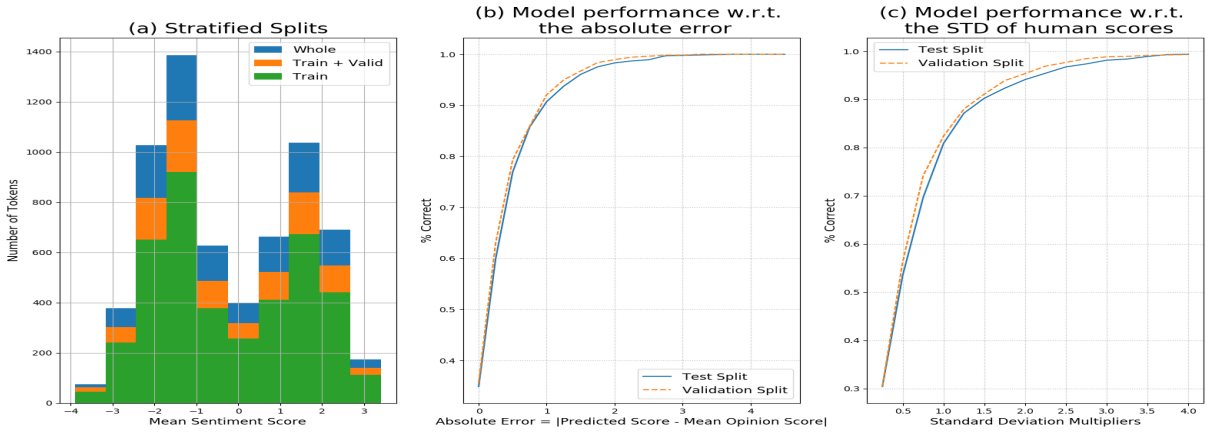

Figure 1: (a) Stratified split; (b) Evaluation on VADER using CDF with respect to absolute prediction error; (c) Evaluation on VADER using CDF with respect to standard deviation multipliers of the human scores.

Results. It is important to note that the original VADER scores were obtained from ten human test subjects and there are discrepancies among these scores, which is to be expected in any such test. The

[1] http://jmcauley.ucsd.edu/data/amazon/

[2] VADER contains a few instances of duplicate lexical entries, for which we use the first provided scores.

highest standard deviation of the mean sentiment rating scores was found to be 2.5, while the lowest was 0. Hence, a simple prediction accuracy is not sufficient to capture the performance of any sentiment prediction model. Thus, three different evaluation methods are presented here to assess the performance of our neural regression model.

First, we evaluated the effectiveness of the proposed model in terms of the raw accuracy across different absolute error tolerances with respect to the human mean sentiment rating. For different absolute prediction error thresholds, we obtain a different percentage of correct predictions, as plotted in Figure 1b. We observe that for 76.23% of cases, the absolute prediction error falls within 0.5 of the mean sentiment rating scores and for around 91% of cases, it falls within unity difference to the human ground truth.

Next, we consider our model as just another opinion along with the 10 original human responses. We then compute the standard deviation among the human scorers and evaluated our predicted scores against it. Figure 1c shows the percentage correct when evaluating the predictions using different *standard deviation multiplier* thresholds. It is observed that 80% of the model predictions fall within unity standard deviation σ of the ground truth scores, whereas 94% of the predictions fall within just two standard deviations, 2σ, of the mean sentiment rating scores.

Finally, the Pearson correlation coefficient between the predicted scores and the mean sentiment rating scores was found to be 0.903. We can conclude from these three results that our deep model succeeds at learning to recreate scores for held-out data.

3.2.2 Domain-Specific Sentiment Scores

Subsequently, we proceeded to apply the technique on our larger seed data set from Section 3.1, which provides domain-specific sentiment scores. Recall that this seed data was obtained from a corpus of domain-specific reviews and hence sentiment scores obtained through the previously described automated seed data induction method served as the training data for our prediction model. A separate model was trained on each domain, resulting in 24 domain-specific predictors. Hence, each word may obtain 24 different sentiment scores corresponding to the 24 domains.

In this section, we shall denote our neural model's predictions as *predicted scores*, while sentiment ratings from the automatic seed induction are referred to as *seed scores*, which here can be regarded as silver standard ground truth targets.

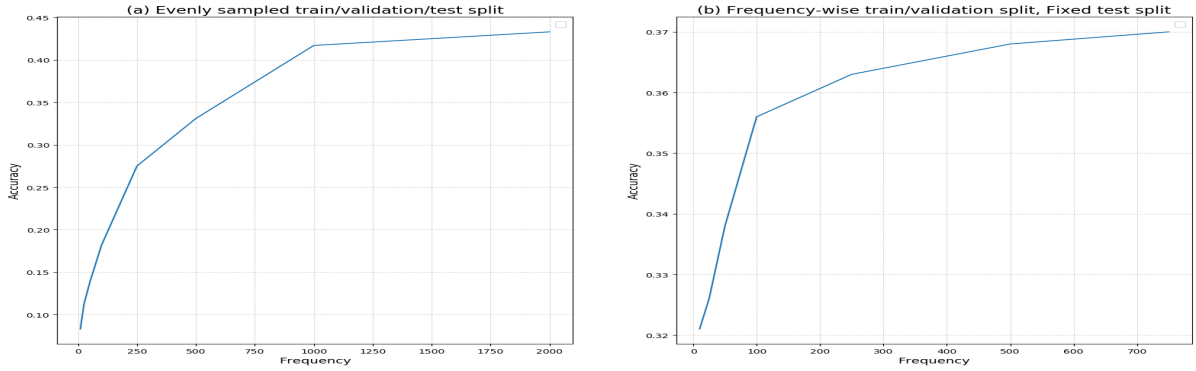

Figure 2: (a) Frequency-wise accuracy indicators (Pearson correlation averaged over all the domains) on evenly sampled train/validation/test splits, (b) Frequency-wise accuracy indicators (Pearson correlation averaged over all domains) on train/validation sampled from different frequency thresholds with consistent test split.

Given that we consider the frequency of a word in the original labeled data as a factor that affects the accuracy of our seed data induction, we generated train/validation/test splits with words that have a frequency equal or above different predefined frequency thresholds $f_{\min}$ in a given domain.

For each considered frequency threshold $f_{\min}$, we computed the Pearson correlation coefficients between predicted scores and seed data scores on each of the domains, and consider the average of such

Table 2: Word counts in seed data vs. predicted data. We compared two settings. All: All words with non-zero sentiment score are considered. Non-neutral: Only words with absolute sentiment score > 0.2.

Domains	Seed (All)	Seed (Non-neutral)	Predicted (All)	Predicted (Non-neutral)
Electronics	241,942	99,451	2,195,999	88,733
Beauty	42,023	19,435	2,195,999	241,661
Apps for Android	99,146	48,893	2,195,999	49,818
Sports and Outdoors	56,401	24,195	2,195,999	227,235
Toys and Games	38,024	15,304	2,195,999	639,685
Home and Kitchen	79,405	34,363	2,195,999	346,856
CDs and Vinyl	254,127	91,142	2,195,999	222,706
Health and Personal Care	73,640	32,032	2,195,999	130,098
Kindle Store	102,229	32,032	2,195,999	630,767
Patio Lawn and Garden	16,299	1,546	2,195,999	4,179
Tools and Home Improvement	41,277	16,542	2,195,999	740,818
Movies and TV	374,774	134,182	2,195,999	129,003
Pet Supplies	40,205	19,142	2,195,999	32,861
Amazon Instant Video	28,546	6,461	2,195,999	260,202
Cell Phones and Accessories	42,567	19,229	2,195,999	186,594
Books	326,268	116,300	2,195,999	214,745
Automotive	13,661	2,519	2,195,999	122,792
Office Products	24,200	5,981	2,195,999	336,457
Baby	34,874	16,326	2,195,999	204,700
Grocery and Gourmet Food	39,715	17,088	2,195,999	143,637
Clothing Shoes and Jewelry	41,522	19,623	2,195,999	134,652
Musical Instruments	9,429	754	2,195,999	23,590
Video Games	96,245	34,278	2,195,999	1,015,866
Digital Music	45,145	9,585	2,195,999	883,777

Pearson correlation coefficients across different domains as the overall accuracy indicator for that $f_{\min}$. Figure 2a plots the outcome of this experiment.

In order to find an optimum training frequency threshold to filter out training data with ambiguous sentiment scores, we ran a separate additional experiment, creating a fixed dev./test set by sampling 1,000 tokens from each domain with frequency over 1,000, while generating training data with varied frequency thresholds. Again, we computed the Pearson correlation coefficients of predicted sentiment scores and ground truth seed ones for each domain and took their average as the overall score for a given frequency threshold. The corresponding results are plotted in Figure 2b. Based on the observed scores, we adopted a frequency threshold of 500 for all subsequent experiments.

3.2.3 Extension to Very Large Vocabulary

Finally, in this section, we describe our extension of the sentiment prediction on different domains to all tokens in the word vector vocabulary. At this point, the weights and hyperparameters of our neural regression models were all frozen. We used models trained with frequency threshold $f_{\min} = 500$ from the last section and generated 24 domain-specified sentiment scores for each word in GloVe.

Table 2 compares the low coverage of the original seed data with the coverage of the predicted data. Due to the network architecture of the prediction model, it virtually always predicts a non-zero value. However, many words obtained a low score very close to 0. Hence, it is more informative to again consider the filtered higher-intensity words with absolute score above 0.2 as non-neutral. From this, we can observe that our deep prediction helps filled the gaps in domains for which we had smaller amounts of training data. It achieved this in part by exploiting semantic relatedness between new words and words for which we had known scores in our seed data, as revealed by the embeddings.

No ground truth scores are available for the large GloVe vocabulary. However, we confirmed in Section 3.2.1 that our deep model succeeds at learning to predict very high-quality sentiment scores. Figure 3 considers the Pearson correlation of the different domain-specific lexicons with the polarity scores given by the complete VADER lexicon. Any words not covered by our lexicons were assumed to have 0.0 as our polarity score. Obviously, an overly strong correlation with VADER is not desirable, as we seek

domain-specific lexicons precisely for their ability to capture domain-specific polarities that differ from generic ones. For example, in the movie domain, a word such as *twist* typically indicates a plot twist, which is often regarded as positive. In general, however, a word such as *twist* does not inherently convey anything positive. Still, the fact that our predicted lexicons correlate vastly better with VADER than the initial seed data suggests that they are more reliable. This mainly stems from their better coverage.

For additional analysis, we studied the cross-correlation matrix of sentiment scores obtained from the 24 domains, illustrated in Figure 4(a) as a heat-map. We further applied classical MDS based on the cross-correlation matrix for dimensionality reduction in order to render a 2D representation of the inter-relationships among the sentiment scores from 24 domains, shown in Figure 4(b). We found that the sentiment scores across different domains reflect intuitive connections. For example, entertainment-related domains such as Digital Music, Books, CDs and Vinyl, Toys and Games, and Video Games bear clear connections in light of their similarity. Likewise, categories related to household usage such as Pet Supplies, Grocery and Gourmet Food, Tools and Home Improvement, Home and Kitchen, etc. reside in similar locations, in light of the similarity of reviews in such domains.

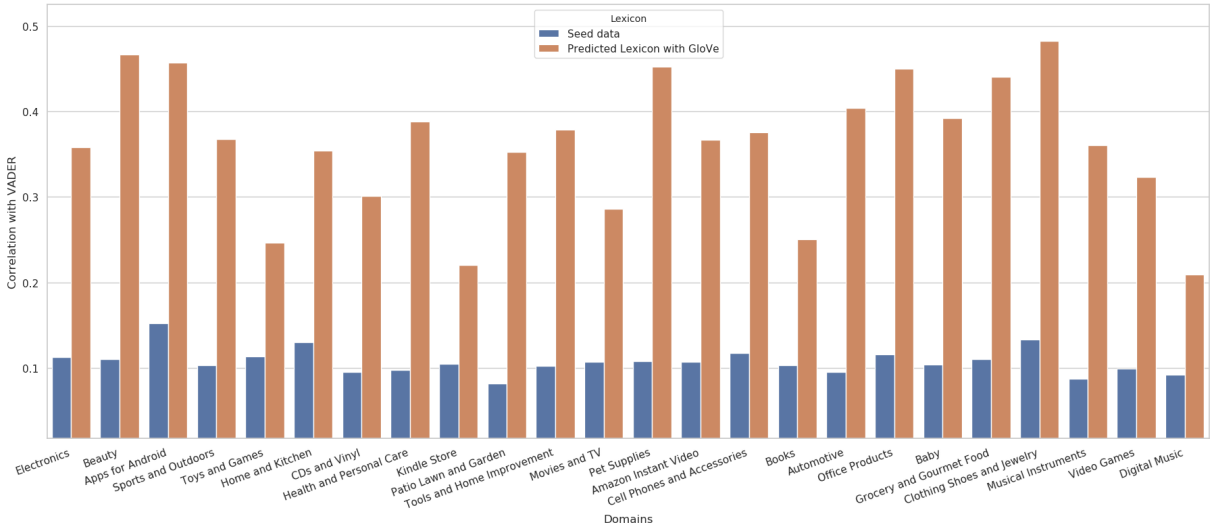

Figure 3: Sentiment score comparison of seed data and predicted lexicon in terms of correlation with domain-independent VADER scores.

These results, along with the high correlation of the predictions in Section 3.2.2, corroborate that our domain-specific lexicons capture human-like sentiment toward different domains.

3.3 Applications of Induced Lexicons

Finally, we assessed the performance of the induced domain-specific sentiment lexicons on downstream tasks such as review sentiment classification and cross-lingual word-level sentiment prediction.

3.3.1 Unsupervised Review Sentiment Classification

Here, we used our predicted domain-specific lexicon to perform sentiment classification on the IMDB movie review dataset compiled by Maas et al. (2011). The *test* portion of movie review data set has 25,000 reviews in total, among which 12,500 are positive and 12,500 are negative.

As for the word embeddings, in this evaluation, along with GloVe, we also used fastText (Mikolov et al., 2018) to obtain a second set of domain-specific lexicons for comparison. As baselines, along with the raw VADER lexicon, two further domain-independent lexicons were derived by using the VADER lexicon as seed data and invoking GloVe and fastText to expand their coverage using our neural expansion approach.

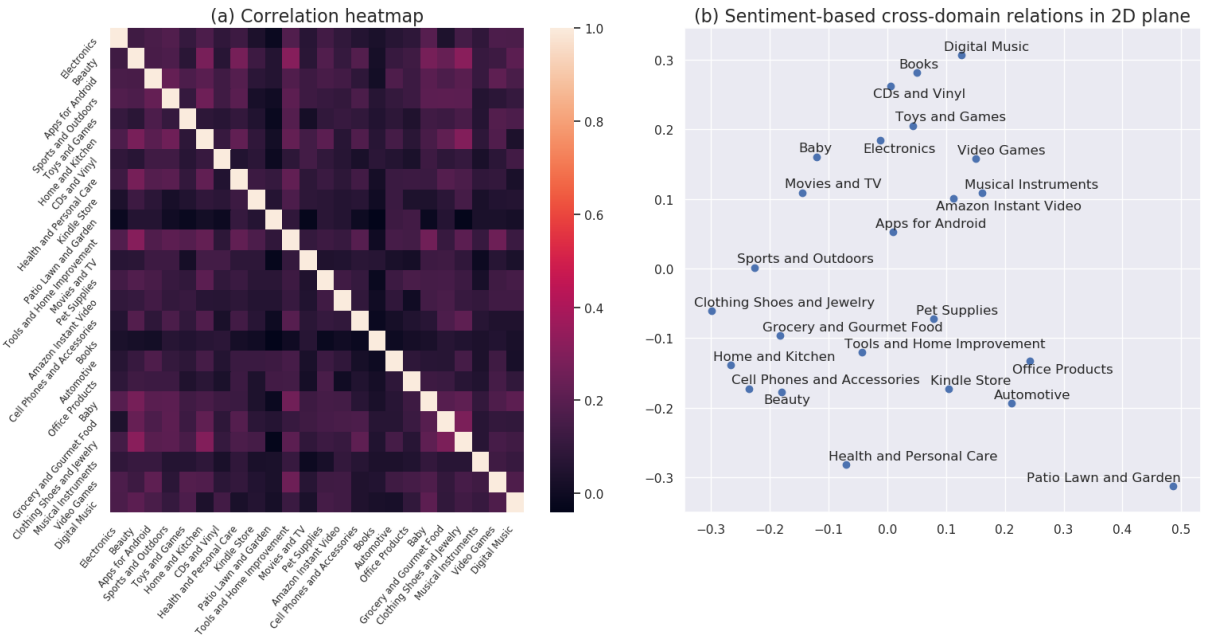

Figure 4: (a) Heat-map of cross-correlation values for predicted sentiment among all domains, (b) 2D representation of 24 domains according to the predicted domain-specific sentiment lexicons.

For unsupervised prediction given a document x in the test set, we simply compute a prediction score

$$f(x) = \frac{1}{|x|} \sum_{i=0}^{|x|} \phi(\vec{v}_{x_i}), \tag{3}$$

where $|x|$ denotes the document length and x_i denotes the i-th word in x. Recall that $\phi(\vec{v}_w)$ is the neural prediction score, given the word vector for w. Subsequently, we predict the polarity by setting the average of all such prediction scores in the corpus as a binary threshold.

Figure 5 plots the results of this evaluation. We observe that in almost all the domains, the domain-specific lexicons (plotted as bars) outperformed the domain-independent lexicons (horizontal lines). As expected, the results are particularly strong for the domains that are closest to the movie domain.

3.3.2 Cross-Lingual Word-Level Sentiment Prediction

Finally, we evaluated the performance of predicted domain-specific lexicons on cross-lingual word-level sentiment score prediction. For this, cross-lingually aligned fastText word vectors (Bojanowski et al., 2017; Joulin et al., 2018) for four languages (English, Spanish, French, and Polish) were used as word embeddings. As the ground truth, we considered the mean sentiment scores of 7,504 English tokens from VADER, as well as the mean human ratings of valence for 875 Spanish words (Hinojosa et al., 2015), 1,031 French words (Monnier and Syssau, 2013), and 1,586 Polish words (Imbir, 2014). Any words from the ground truth data that are missing in the aligned fastText word vectors are eliminated.

The sentiment prediction model was trained on 24 different domains separately, as described in Section 3.2.2, except that we here used the aligned word vectors for English during training. After the training stage, the same models could then be invoked to cross-lingually predict sentiment scores for words from the ground truth data sets using aligned word vectors for non-English words. Correlations between the predicted scores and ground truth datasets are plotted in Figure 6. Although the cross-lingual results did not attain the level of the monolingual English correlation, we obtained a promising degree of cross-lingual generalization across languages. Note again that we do not desire a perfect correlation, as the domain-specific scores are expected to diverge from the generic domain-independent valence ratings.

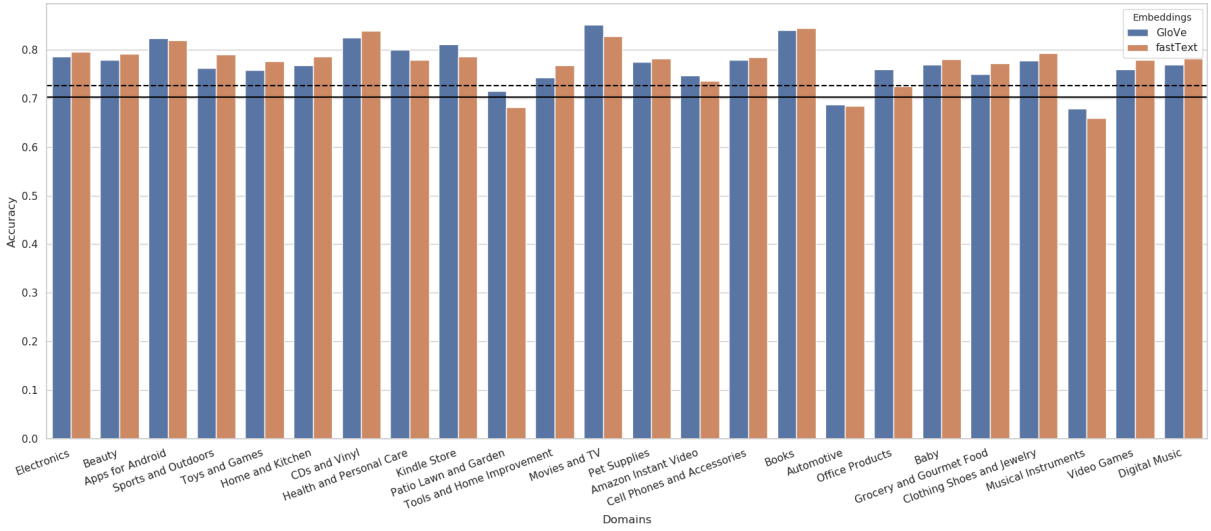

Figure 5: Accuracy of unsupervised review sentiment classification. The solid and dotted horizontal lines represent the baseline from GloVe-based induction from VADER and fast-based induction from VADER respectively.

4 Related Work

The traditional way of obtaining sentiment lexicons has been to build them manually, relying either on experts or invoking crowd-sourcing. A prominent example is the Hu and Liu (2004) Opinion Lexicon. There are numerous algorithms that aim to increase the coverage of an individual sentiment lexicon. Often, these start from seeds and then rely on graph-based algorithms to gather additional data, as for instance explored by Kim and Hovy (2004) and in the approach used to induce SentiWordNet (Baccianella et al., 2010). The extension can also be based on vector representations of words, as proposed in the Densifier approach (Rothe et al., 2016). Such work has shown that dense word vectors trained on large amounts of data harbour signals that are useful for sentiment analysis. Instead of a regular supervised setup, Castellucci et al. (2016) used distant supervision based on emoticons to obtain sentiment labels for entire sentences. They then trained a sentiment model on sentence vector representations sharing a common representation space with word vectors, which allowed them to apply the trained model to predict word-level scores. However, techniques such as the above mostly have not targeted domain-specific sentiment lexicons.

The SocialSent project (Hamilton et al., 2016) induced Reddit community-specific sentiment lexicons without labeled corpora. Their SentProp approach constructs a graph of words and then considers random walks emanating from a small set of seed words with known sentiment polarity. The polarity scores are based on the frequency of random walk visits and the polarity of the seed word from which those random walks started. While Reddit communities provide substantial diversity, the language used in Reddit posts differs quite substantially from the kinds of language one encounters in reviews. Kreutz and Daelemans (2018) adopted SentProp to customize an existing general-purpose sentiment lexicon for use in one specific domain.

We instead focus on inducing a number of domain-specific lexicons to obtain a lexical resource that is more suitable for typical sentiment analysis use cases. The approach by Labille et al. (2017) also starts from labeled data for consumer products. It infers word polarity scores directly based on posterior probabilities and inverse document frequencies. However, such scores are limited to words that occur in the labeled training data.

Instead, in our work, we draw on word vectors to greatly enhance the coverage of the lexicons beyond the words present in the category-specific labeled data. Our initial seed data approach is based on linear models optimized for maximum margin discrimination between the positive and negative classes, in line with the observations by Mudinas et al. (2018), who found that linear models outperformed more

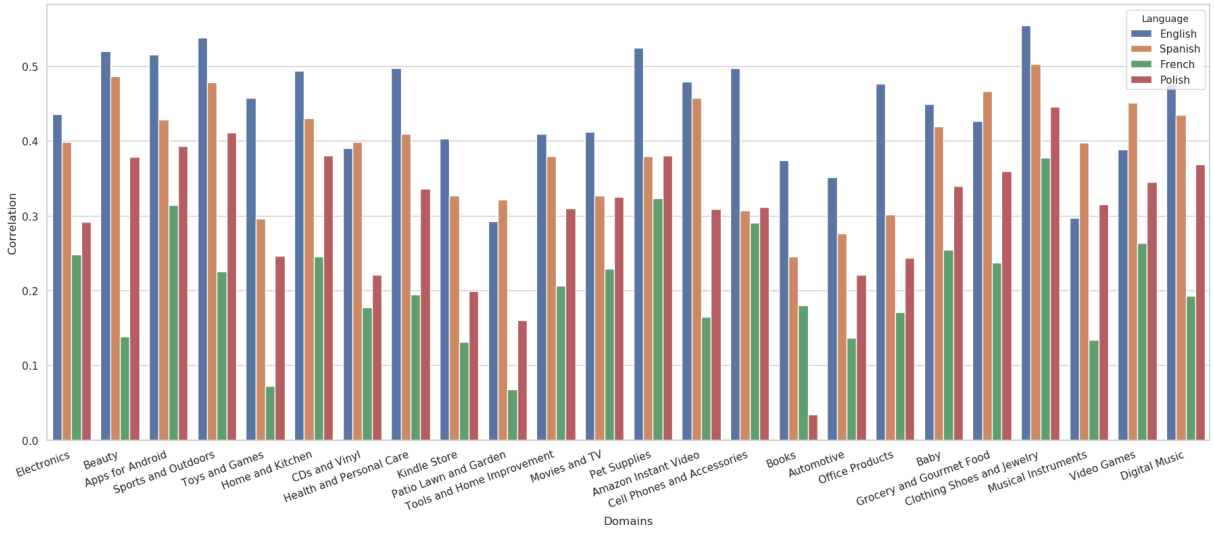

Figure 6: Correlation of cross-lingually predicted word-level sentiment scores with human ratings.

sophisticated semi-supervised or transductive learning algorithms in their experiments.

Cross-lingual propagation of sentiment lexicons has been studied in a number of previous approaches. For example, Dong and de Melo (2018a) and Dong and de Melo (2018b) induced sentiment embeddings using translation graphs. In our experiments, we considered cross-lingual word embeddings for cross-lingual transfer.

5 Conclusion

In this paper, we present new domain-specific sentiment lexicons for a number of domains. We bootstrap this data from a large-scale review corpus covering 24 domains and then rely on a neural model to substantially extend its coverage. Our analysis shows that there are substantial differences between domains, which make domain-specific sentiment lexicons an important form of lexical resource in downstream tasks. Further experiments show that the predicted lexicons outperform domain-independent lexicons on unsupervised review classification and can also be used for cross-lingual word-level sentiment prediction. Our data is freely available under an open source license from `http://sentimentanalysis.org`.

Acknowledgments

We thank Yupeng Zhang (Rutgers University) for conducting helpful additional analyses.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2016. A language independent method for generating large scale polarity lexicons. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 38–45, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Gerard de Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1(July 2013):279–290.

Xin Dong and Gerard de Melo. 2018a. Cross-lingual propagation for deep sentiment analysis. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*, pages 5771–5778. AAAI Press.

Xin Dong and Gerard de Melo. 2018b. A helping hand: Transfer learning for deep sentiment analysis. In *Proceedings of ACL 2018*, pages 2524–2534.

Michael Gamon. 2004. Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605.

J. A. Hinojosa, N. Martínez-García, C. Villalba-García, U. Fernández-Folgueiras, A. Sánchez-Carmona, M. A. Pozo, and P. R. Montoro. 2015. Affective norms of 875 spanish words for five discrete emotional categories and two emotional dimensions. *Behavior Research Methods*, 48(1):272–284, March.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD 2004: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA. ACM.

C.J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc. ICWSM-14*.

Kamil K. Imbir. 2014. Affective norms for 1, 586 polish words (ANPW): Duality-of-mind approach. *Behavior Research Methods*, 47(3):860–870, October.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of COLING 2004*, pages 1367–1373, Geneva, Switzerland. COLING.

Tim Kreutz and Walter Daelemans. 2018. Enhancing general sentiment lexicons for domain-specific use. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1056–1064.

Kevin Labille, Susan Gauch, and Sultan Alfarhood. 2017. Creating domain-specific sentiment lexicons via text mining. In *Proceedings of the 6th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM 2017)*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Catherine Monnier and Arielle Syssau. 2013. Affective norms for french words (FAN). *Behavior Research Methods*, 46(4):1128–1137, December.

Andrius Mudinas, Dell Zhang, and Mark Levene. 2018. Bootstrap domain-specific sentiment classifiers from unlabeled corpora. *Transactions of the Association for Computational Linguistics*, 6:269–285.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*. Lawerence Erlbaum Associates, Mahwah, NJ.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada. Association for Computational Linguistics.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June.

Abu Awal Md Shoeb, Shahab Raji, and Gerard de Melo. 2019. EmoTag – Towards an emotion-based analysis of emojis. In *Proceedings of RANLP 2019*, pages 1094–1103, sep.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*.

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

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment in short strength detection informal text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558.

Liqiang Wang, Yafang Wang, Gerard de Melo, and Gerhard Weikum. 2019. Understanding archetypes of fake news by fine-grained classification. *Social Network Analysis and Mining (SNAM)*, 9(1):37.

Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhu Wang. 2013. A sentiment-enhanced personalized location recommendation system. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, pages 119–128, New York, NY, USA. ACM.