Enhancing General Sentiment Lexicons for Domain-Specific Use

Tim Kreutz   **Walter Daelemans**

CLiPS - Computational Linguistics Group

Department of Linguistics

University of Antwerp

`{tim.kreutz,walter.daelemans}@uantwerpen.be`

Abstract

Lexicon based methods for sentiment analysis rely on high quality polarity lexicons. In recent years, automatic methods for inducing lexicons have increased the viability of lexicon based methods for polarity classification. SentProp is a framework for inducing domain-specific polarities from word embeddings. We elaborate on SentProp by evaluating its use for enhancing DuOMan, a general-purpose lexicon, for use in the political domain. By adding only top sentiment bearing words from the vocabulary and applying small polarity shifts in the general-purpose lexicon, we increase accuracy in an in-domain classification task. The enhanced lexicon performs worse than the original lexicon in an out-domain task, showing that the words we added and the polarity shifts we applied are domain-specific and do not translate well to an out-domain setting.

1 Introduction

Work in sentiment analysis can roughly be divided into two approaches: corpus based techniques and lexicon based techniques (Taboada et al., 2011). Corpus based techniques use supervised learning on large target domain corpora to learn classification models. Currently, implementations using deep learning make up the state of the art in binary polarity classification on Twitter (Rosenthal et al., 2017).

Lexicon based techniques rely on high quality polarity lexicons. Document-level polarity can be predicted by looking at occurrences of lexical units in the document and doing some form of averaging on their respective polarity values. Despite not yielding the state of the art results of corpus based methods, lexicon based methods still see frequent use since they do not rely on big (labelled) data and allow for more interpretability. In recent years, polarity lexicons have been improved by adding new phrases from web data (Velikovich et al., 2010) or by automatically learning domain-specific polarities (Hamilton et al., 2016).

Sentprop is a framework for inducing domain-specific polarities (Hamilton et al., 2016). SentProp has been shown to accurately reproduce domain-specific lexicons from in-domain word embeddings. Additionally, it has been used to examine interesting historical and community-specific polarity shifts.

We elaborate on SentProp by evaluating its use for domain adaptation of a general-purpose lexicon. Starting from a Dutch general polarity lexicon and a corpus of political forum posts, we automatically expand and enhance the general lexicon entries. We show that the enhanced lexicon outperforms the general-purpose lexicon on an in-domain classification task and we further add to the existing body of work by evaluating the enhanced lexicon on an out-domain task. In the out-domain task the general purpose-lexicon yields better accuracy, showing that our proposed method learns domain-specific polarities that do not translate well to other domains.

The paper is organized as follows. Section 2 will embed SentProp in a larger context of semi-supervised lexicon expansion research. In section 3 we describe the corpus used for inducing polarities, our extensions to the SentProp framework and the setup for the classification experiments. In section 4 we describe findings from a qualitative viewpoint and discuss the results of our experiments. We conclude with some final remarks and possible future directions for the research.

2 Related work

In this section we discuss related work to better embed our selected methods for lexicon induction. We first give a broad overview of work done in predicting word polarity. We then address the work that incorporates generated lexicons in polarity classification experiments. The last subsection will discuss research into lexicon adaptation in a domain-specific setting.

2.1 Polarity prediction

Early work in polarity prediction used linguistic rules to extract word relations. Hatzivassiloglou and McKeown (1997) extract conjunctions of adjectives to infer semantic oriëntation (SO; i.e. polarity). The resulting graph modelled similarity links ("and" conjunctions) and dissimilarity links ("but" conjunctions) and clustering was performed to further separate the positive from the negative polarity words. This approach yielded high accuracy (90%) in predicting word polarity, even when using only seven links per adjective.

Turney (2003) used statistical association between seed words and unseen words to induce SO. Specifically, the point-wise mutual information (PMI) of word co-occurrence statistics was used to find similarities of a word to a seed set. The evaluation setup compared induced SO to existing lexicons at different confidence intervals. The highest accuracy (98.20 percent) was found when comparing the top 25 percent most confident SO words induced from the largest corpus (10^{11} tokens). Using lower confidence words and smaller corpora harmed performance significantly.

Another vein in the lexicon expansion literature uses lexical graphs to draw word associations. Rao and Ravichandran (2009) use WordNet to leverage its rich relational structure. They construct a lexical graph by linking words through their synonymy and hypernymy relations. Polarities are propagated over the lexical graph using the label propagation algorithm from Zhu and Ghahramani (2002). Even when reducing the seed set to ten percent of its original size, this yields around 80 percent F-score for nouns and verbs. Adjectives suffer somewhat more from using fewer seed words.

2.2 Classification using induced polarity lexicons

Moving away from the intrinsic evaluation in the papers mentioned above, task-based evaluation is more relevant to our current work.

Velikovich et. al (2010) combine the statistical approach (Turney and Littman, 2003) with simplified label propagation (Zhu and Ghahramani, 2002; Rao and Ravichandran, 2009) to derive a web-based polarity lexicon. Co-occurrence statistics from a large web corpus ($4*10^9$ tokens) were used to construct a lexical graph. Polarities were then propagated over the graph starting from a positive and a negative seed set of respectively 187 and 192 words. The break with predefined lexical resources such as WordNet allows for finding new lexical entries in spelling variations, slang and multi-word expressions (they consider n-grams up to length 10). As such, the purpose of the work is not to compare induced polarities with existing lexicons, but rather to derive a lexicon with high coverage for sentiment analyses.

This results in a 178,104 phrase polarity lexicon which can be applied to a diversity of sentiment analysis subtasks. The web-derived lexicon outperforms two other automatically generated polarity lexicons, including one based on label-propagating WordNet, in a polarity classification and polarity ranking task. Velikovich et. al (2010) show that statistical similarity measures work even when the obtained corpus is very noisy. They also demonstrate the practical usefulness of a web-derived polarity lexicon.

One drawback that remains however, is the neglect of domain-specific information in these general-purpose lexicons. The next subsection will discuss work that has been done in generating lexicons for domain-specific purposes.

2.3 Domain adaptation

There is a large variety of recent work in obtaining domain-specific polarity lexicons. Most are concerned with generating a new lexicon for a target domain (Bross and Ehrig, 2013; Tai and Kao, 2013), while others aim to augment existing lexicons for domain-specific use with a variation of polarity clues.

Choi and Cardie (2009) use linear programming to find the most likely polarity label for words based on their occurrence in opinion expressions. Two types of clues; word-to-word relations within each expression and word-to-expression relations, are exploited to adapt a general-purpose lexicon. The adapted lexicon significantly improves expression classification with regards to the original lexicon.

Du et. al (2010) adapt the information bottleneck (IB) method to take domain-specific knowledge into account. This slightly changes the problem of adapting a general-purpose lexicon to fit a specific domain, because IB is used to adapt an in-domain lexicon to fit out-domain properties. To infer polarity values for words in the unseen domain, the authors use three clues: the relationships between out-of-domain words and out-of-domain documents, the relationships between out-of-domain words and in-domain words and the relationships between out-of-domain words and in-domain documents. To evaluate the domain adaptations, three domain-specific lexicons are manually composed and used as a basis for adaptation. When using IB to adapt an in-domain lexicon for out-domain purposes, polarity classification accuracy of the adapted lexicon is higher than most baselines that use only word occurrence statistics to infer in-domain polarities.

Lu et. al (2011) present an optimization framework that takes a range of signals such as sentiment scores from general lexicons, domain-specific document level sentiment, WordNet and linguistic clues to produce a unified domain-specific lexicon. The framework is tested on hotel reviews and printer reviews and significantly outperforms other lexicon-based methods in a polarity classification task.

2.4 SentProp

The work on domain adaptation for polarity lexicons mirrors the polarity lexicon induction literature in that there are a wide range of possible lexical resources to use as polarity clues and a wide range of induction methods that are all similarly feasible. SentProp implements word-embeddings with label propagation to offer a uniform approach for domain-specific lexicon generation (Hamilton et al., 2016). It further promises accurate performance even when using smaller corpora ((10^7 tokens). Hamilton et. al (2016) demonstrate their method by reproducing existing domain-specific lexicons. They outperform other induction methods including PMI (Turney and Littman, 2003) and graph propagation (Velikovich et al., 2010) in the domain-specific configuration and for standard English with a 1000x reduction of the corpus.

3 Methods

We elaborate on the SentProp method to evaluate it in domain-specific polarity classification tasks. Instead of pursuing state of the art performance in a target domain task, we seek to provide insight into the domain-specificity of polarities derived from the Sentprop method and the applicability of a generated domain-specific lexicon to an in-domain and an out-domain task.

3.1 Word embeddings

The data for the domain-specific input to our lexicon is derived from discussion boards on politics. We use two sources: politics.be and 9lives.be.

Politics.be features a dedicated political discussion board that has seen active use since 2002. It has more than 60,000 registered members and hosts close to 250,000 discussions.

9lives.be is a website originally dedicated to video game news, but also features a discussion board with general topics. We targeted the politics and actuality sub-forum for our data set. The sub-forum hosts close to 2,000 discussions on news and politics since 2004.

Using the Scrapy package for Python, we were able to download close to 20,000 discussion threads with $2x10^6$ posts. Specifications of the collection are reflected in Table 1.

In the choice for a word embedding method we followed Hamilton et. al. (2016), who find that vectors containing positive point-wise mutual information (PPMI) with singular-value decomposition (SVD) outperformed SGNS and GloVe embeddings in lexicon reproduction. This is further supported by Levy et. al. (2015) who also provide a useful overview of hyper parameters in each of the discussed methods. In particular this led us to run 500-dimensional SVD-embeddings with a context-window of

Source	Documents	Types	Tokens
Politics.be	1,778,101	824,978	68,596,562
9lives.be	434,261	315,266	28,006,269
Total	2,212,362	935,292	96,602,831

Table 1: Corpus specifications.

size six (Velikovich et al., 2010) with context-distribution smoothing and eigenvalue weighting (Levy et al., 2015).

3.2 General-purpose lexicon

We used the DuOMAn subjectivity lexicon as our general-purpose lexicon (Jijkoun and Hofmann, 2009). With 8,782 entries, it is the largest polarity lexicon for Dutch. Polarity scores range from -2 (very negative) to 2 (very positive) and contains 3,631 zero (neutral) values.

3.3 Seed sets

SentProp's default pipeline requires the word embeddings and a positive and negative seed word set. We extracted 10 positive and 10 negative seed words by sorting candidates from DuOMan by polarity first and by occurrences in our corpus second. This provided us with a selection of the most positive and negative words that are also used relatively frequently. We then manually filtered words that have any sort of semantic ambiguity. It is important that the seed words are unambiguously positive or negative because they serve as anchors in our graph (Hamilton et al., 2016).

Positive seeds (occurrences)		Negative seeds (occurrences)	
perfect (1.926)	leuk (1.368)	probleem (7.542)	slecht (2.834)
gelukkig (1.306)	effectief (1.304)	onzin (1.805)	onmogelijk (1.600)
correct (1.288)	winnen (887)	jammer (1.321)	dom (1.175)
top (789)	positief (771)	spijtig (919)	hypocriet (771)
succes (667)	prachtig (263)	kwaad (718)	discriminatie (709)

Table 2: Seed words as first sorted by polarity in the DuOMAn lexicon, then sorted by their occurrences in the discussion-board corpus. Some semantically ambiguous words were manually filtered like 're-delijk' which translates to 'reasonable' as an adjective but to 'fairly' as an adverb. All positive seed words have a polarity of 2.0 in the lexicon and all negative seed words have a polarity of -2.0 in the lexicon.

3.4 Lexicon expansion

We ran SentProp with default parameters. Starting from the seed words, positive and negative polarity labels are propagated from a word to their nearest neighbor using random walks. The nearest neighbors for any given word are determined by calculating the cosine-similarities between its embedding vector and the vectors of all other words. This theoretically composes a ranking of words with the most similar meanings to the given word. For our purposes we further imply that words with similar meanings have similar polarities. Although this is likely not plausible for all words, on an aggregate level we hope to at least rely on regularities tying positive words to positive words and negative words to other negative words.

The induced polarity score reflects the probability of a random walk reaching the word from the positive set versus reaching it from the negative set. A score closer to 1 reflects a high likelihood for the polarity to be positive and a score closer to 0 reflects the high likelihood for it be negative (Hamilton et al., 2016).

We considered a vocabulary that is much larger than the size of our general-purpose lexicon. In theory, all words in the vocabulary and their induced polarities could be added to DuOMAn, but we

No.	Configuration rule
#1	Apply all full polarity shifts
#2	Apply all weighted polarity shifts
#3	Apply full polarity shifts for x% largest shifts
#4	Apply full polarity shifts for x% smallest shifts
#5	Apply weighted polarity shifts for x% largest shifts
#6	Apply weighted polarity shifts for x% smallest shifts

Table 3: Polarity shift configuration rules. Each of the percentage-based configurations is applied with 5-percent increments to find the best settings. Full polarity shifts entail applying the induced polarity for any word, weighted polarity shifts take the average of the induced polarity and the polarity from the general-purpose domain and applying the average.

expect words with predominantly neutral polarities to only add noise to the lexicon. To only consider sentiment-bearing words and be able to plot them to the scale the general-purpose domain adheres to, we determined the limits to the amount of positive and negative words with two parameters:

$$num\_neg(X, \beta, \gamma) = \gamma\beta|X| \qquad (1)$$

$$num\_pos(X, \beta, \gamma) = (1 - \gamma)\beta|X| \qquad (2)$$

Where β is the fraction of vocabulary words that will be treated as sentiment-bearing, γ is the proportion of negative polarities to positive polarities in the vocabulary, and $|.|$ is the cardinality operator.

Our algorithm distributes the negative and positive words evenly over the granularity of the general-purpose lexicon. The words with the highest likelihood of being negative hence end up in the -2 group. There is no theoretical rule of thumb for determining the distribution of words over polarity labels and we expect this to be highly domain-dependent. We used a uniform distribution over the labels to make minimal assumptions.

3.5   Enhancing general entries

The algorithm also produces new labels for words already present in the general-purpose lexicon. We do not expect each of these induced polarities to better reflect word polarities than the values in the pre-existing lexicon. Correct label assignment depends on quality of the word embeddings and can change due to the simplified approach to label assignment we described above. We experimented with a few possible configuration rules that combine the information from the general-purpose lexicon and the induced polarities to decide on the size of polarity shifts. We considered applying weighted polarity shifts that average the polarity in the general-purpose lexicon and the induced polarity. We also considered only shifting polarities when the difference between the existing entry and the induced polarity is relatively large or when it is relatively small (Table 3).

3.6   Classification setup

For evaluation, we designed an in-domain and an out-domain binary polarity classification task. The in-domain data consists of over 4,000 polarity annotated tweets that mention the name of a Flemish politician. The out-domain data consists of product reviews from two online web shops: Bol.com and Coolblue.be. The products under review are mostly books, dvd's, games and some consumer electronics. The reviews have a rating that ranges from one to five stars. To match the in-domain binary polarity, we grouped one and two star reviews and four and five star reviews and assigned them negative and positive labels respectively.

The in-domain documents contain 2,126 positive and 2,126 negative tweets. The out-domain documents contain 3,959 positive reviews and the same number of negative reviews. We developed our in-domain lexicon on 90 percent of the twitter data and tested on the remaining 10 percent. We also tested the lexicon on the reviews to see if the enhanced lexicon improved classification in general, or mainly in the domain-specific setting.

Positive words			Negative words		
Word	Type	Translation	Word	Type	Translation
Geniaal	Regular addition	Genius	Verzuurd	Regular addition	Soured
Hoera	Regular addition	Hooray	Brainwashen	Regular addition	Brainwashing
Boeiend	Regular addition	Interesting	Oogkleppen	Regular addition	(horse) blinkers
Goei	Spelling variation	Good	Aggressief	Spelling variation	Aggressive
Perfekt	Spelling variation	Perfect	Groffe	Spelling variation	Rough
Intersant	Spelling variation	Interesting	Associaal	Spelling variation	Anti-social
Chapeau	Non-Dutch	Hats off	Debiel	Vulgarity	Imbecile
Style	Non-Dutch	Style	Dement	Vulgarity	Dementing
Zeitgeist	Non-Dutch	Zeitgeist	Pipo	Vulgarity	Idiot

Table 4: Example additions to the enhanced lexicon. The positive additions contain regular word additions, spelling variations and non-dutch words. The negative additions contain a lot of vulgarities.

For classification we implemented a majority-vote algorithm. Since we respect the granularity of the polarity labels in the original lexicon, words carry different weights. Instead of using binary values to vote on a document's polarity, we took the average of the occurring polarity values. If the average is a positive value we judged the document to have a positive label, and vice versa.

4 Results and discussion

We will now discuss our findings with regards to the best configuration for our proposed method, and the eventual classification accuracy this yielded on the proposed tasks. Moreover, in our qualitative evaluation we look at which words were added and which polarities adapted, and in which regard they reflect the domain-specificity of the generated lexicon.

4.1 Lexicon induction

We induced polarities for words with at least 100 occurrences in the corpus (n $\approx$ 1 per million). This resulted in a vocabulary of 26,563 words. We performed grid search for the parameters in equations (1) and (2) and the possible configurations from Table 3, comparing accuracy for each of the settings. The optimal β-value we found was 0.06, which signifies that 6 percent of all words in the vocabulary, which amounts to 1,303 words that were not already present in DuOMAn, were added to the general-purpose lexicon. The optimal γ-value we found was 0.58 which signifies that 58 percent of the added words are negative polarity words. Applying the average of the induced polarity and the lexicon polarity value for words that shift less than 25 percent works best for enhancing general entries (configuration #6 in Table 3). This resulted in 1,165 polarity shifts in the general lexicon. The resulting enhanced lexicon will be used for our qualitative inspection as well as the classification experiments on test data.

4.2 Qualitative evaluation

The 1,303 additions in the enhanced lexicon are a mixture of new words, spelling variations and words borrowed from other languages (Table 4). Some words are not inherently positive or negative but can be expected to occur mostly in a positive or negative context. The negative additions contain a lot of insults and other vulgarities. Generally speaking the additions seem more applicable in the domain of political discussion than in other domains.

The polarity shifts that were applied to further enhance the general-purpose lexicon had a positive effect on the classification performance for the development data, but since we only applied the subtle polarity shifts we cannot easily discern why the shifts improve results. Of the 3,292 entries in the DuOMAn lexicon that also appeared in the vocabulary, 993 had no polarity difference from the induced polarities, 1,165 had only a small polarity difference ($<$25%) and 1,134 had big polarity differences ($>$25%).

	Political tweets					Product Reviews				
	Positive		Negative		Overall	Positive		Negative		Overall
Lexicon	P	R	P	R	Acc.	P	R	P	R	Acc.
Majority baseline	50.0	0.0	50.0	100.0	50.0	50.0	0.0	50.0	100.0	50.0
DuOMan	65.0	54.0	60.6	**70.9**	62.4	**72.9**	79.7	76.5	**66.2**	**72.9**
Enhanced lexicon	**66.0**	**65.7**	**65.9**	66.2	**66.0**	62.6	**89.7**	**81.8**	46.4	68.0

Table 5: Precision, recall and classification accuracy for the original and the enhanced lexicon. The added out-domain task shows that the general-purpose domain was solely enhanced for the in-domain task.

In 1,078 of the 1,165 applied shifts, the polarity of a word is reduced. This leads us to believe the induced polarities had a lot of neutral values for words that were sentiment-bearing in the general-purpose lexicon. This weakening of sentiment in part of the lexicon did have a positive effect on classification.

4.3 Quantitative evaluation

We tested the enhanced lexicon on the held out tweets and the review data. The results are presented in Table 5. The improvement of the enhanced lexicon over DuOMan is restricted to the political tweets. In fact, with the additions and shifts applied for the political domain, classification on product reviews performs a lot worse overall. Each polarity class shows a precision-recall trade-off in the results. For classification of the political tweets, the additions and shifts in the lexicon yield a more optimal balance between the two classes, while for classification of the product reviews the balance is upset, incorrectly predicting far more positive labels than DuOMan.

Our interpretation is that for the political domain specifically, the negative sentiment additions contain human traits rather than words that may describe a product. The positive additions seem more general and reusable, which explains the increased tendency of the enhanced lexicon to predict positive labels in the product reviews. When solely using a polarity lexicon to classify documents, the balance between the generality of positive versus negative words is very important.
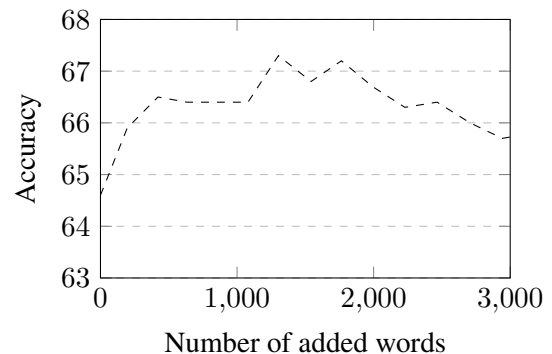

Figure 1: Adding more words from the vocabulary to the enhanced lexicon improves accuracy on the development data initially, but as more neutral words are added, the effect is diminished.

4.4 Error analysis

Figure 2 shows some of the typical corrections that were made when applying the domain-specific lexicon as opposed to the general purpose lexicon. The first sentence gets the default negative label because none of its words are present in DuOMAn. By expanding the general purpose lexicon with more words, the correct positive label is assigned. This first type of correction accounted for 21 out of 28 total corrections on the test data (75%). The second sentence demonstrates a similar correction in which words were found in DuOMAn but the addition of new words corrected the label. This type of correction was present in 5 out of 28 corrections on the test data (about 18 %). The last type of correction is uncommon, but is caused by a reduction of polarity in the original lexicon where the default label is actually the correct label. This explained the remaining 2 corrections.

We look at errors caused by the adapted lexicon on the out-domain test set because they were far more numerous than the errors on the in-domain data. Our interpretation in section 4.3 is right: in 88 percent of the 656 additional misclassifications the (correct) negative label was flipped because of new positive

D: boeiend nieuwtje gehoord op lezing kris peeters : #mer zou behoorlijk wat procedurefouten bevatten
   -       -   -   -   -   -   -   - -   -   -   -   -   -

A: boeiend nieuwtje gehoord op lezing kris peeters : #mer zou behoorlijk wat procedurefouten bevatten
   +2.0   -   -   -   +1.0 -   -   - -   -   -   -   -   -

D: zeker niet tegenover bart de wever , die het overigens schitterend doet op televisie .
   +1.0 -1.0 -   -   - -   - -   -   -   -   -   - -   -

A: zeker niet tegenover bart de wever , die het overigens schitterend doet op televisie .
   +1.0 -1.0 -0.5   -   - -   - -   -   -   +1.5   -   - -   -

D: als rik torfs zonder werk valt , kan ie nog steeds proberen te solliciteren bij bond zonder naam .
   -   -   -   -   0.0 -   - 0.0 - -   -   0.0   - -   -   -   -   1.0 -

A: als rik torfs zonder werk valt , kan ie nog steeds proberen te solliciteren bij bond zonder naam .
   -   -   -   -   0.0 -   - 0.0 - -   -   0.0   - -   -   -   -   0.0 -

Figure 2: Classifications from the DuOMAn lexicon (D) and respective corrections made by the adapted lexicon (A) on the political tweets test set.

polarity word in the review. Similarly, 79 percent of the 268 corrections on the review data was made this way.

There were specific additions to the lexicon that caused many misclassifications, for example 'hoofd-stuk' (chapter), *literatuur* (literature) and 'uitgelezen' (finished reading; but can also mean excellent) all have positive polarities in the domain-specific lexicon but cause many book reviews to be wrongfully classified as positive. Personal traits such as 'asociale' (anti-social), *egocentrische* (selfish) and 'houding' (attitude) caused problems when book or movie plots were described in the review. The same is true for *mannetje* (little man), *goal* and *pass* but for video game reviews.

With regards to domain-specificity these specific word examples show us that there are many (sub)domains within the product review data to be considered. The lexicon adaptation method we propose in this paper can generate a lexicon for use on any level, given there is enough in-domain textual data available. Beyond polarity classification, generated lexicons may be used to study sentiment differences between domains.

5 Conclusion

We have shown that enhancing a lexicon with SentProp for domain-specific use improves its accuracy in an in-domain classification task. Not all induced polarities should be considered when using majority voting for polarity classification. The words with the most extreme polarities make for the best additions to the lexicon (see Figure 1).

When applying polarity shifts, only applying small shifts that average the general-purpose lexicon polarity with the induced polarity had a positive effect on classification accuracy. A large difference between an induced polarity and the general-purpose domain is likely due to some very specific context in the word embeddings. We ignored these large differences and used the general-purpose polarity instead.

We added 1,303 new words to the general-purpose lexicon and applied 1,165 small polarity shifts to its entries. This gave us the best enhanced lexicon for classification on political tweets. We also tested the enhanced lexicon on out-domain product reviews and found that this yielded worse accuracy than the general-purpose domain, showing that the additions and adaptations were domain-specific and do not translate well to other domains.

In future work, we would like to compare lexicons enhanced on different corpora. This should provide more insight into the word additions and shifts that are particular for a domain. We also expect the balance between positive and negative words to shift depending on the domain. Another important addition to the current work would be a comparison of SentProp to other polarity induction methods for domain-specific classification.

SentProp has been shown to accurately reproduce domain-specific lexicon (Hamilton et al., 2016). We have elaborated on the existing body of work by demonstrating its capability to enhance a general-purpose domain for domain-specific classification.

References

Juergen Bross and Heiko Ehrig. 2013. Automatic construction of domain and aspect specific sentiment lexicons for customer review mining. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1077–1086. ACM.

Yejin Choi and Claire Cardie. 2009. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 590–598. Association for Computational Linguistics.

Weifu Du, Songbo Tan, Xueqi Cheng, and Xiaochun Yun. 2010. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 111–120. ACM.

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

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics.

Valentin Jijkoun and Katja Hofmann. 2009. Generating a non-english subjectivity lexicon: Relations that matter. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–405. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. 2011. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM.

Delip Rao and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682. Association for Computational Linguistics.

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-2017)*, pages 502–518.

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

Yen-Jen Tai and Hung-Yu Kao. 2013. Automatic domain-specific sentiment lexicon generation with label propagation. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services*, page 53. ACM.

Peter D Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 777–785. Association for Computational Linguistics.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. *CMU-CALD-02-107*.