Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora

William L. Hamilton, Kevin Clark, Jure Leskovec, Dan Jurafsky
Department of Computer Science, Stanford University, Stanford CA, 94305
wleif,kevclark,jure,jurafsky@stanford.edu

Abstract

A word's sentiment depends on the domain in which it is used. Computational social science research thus requires sentiment lexicons that are specific to the domains being studied. We combine domain-specific word embeddings with a label propagation framework to induce accurate domain-specific sentiment lexicons using small sets of seed words, achieving state-of-the-art performance competitive with approaches that rely on hand-curated resources. Using our framework we perform two large-scale empirical studies to quantify the extent to which sentiment varies across time and between communities. We induce and release historical sentiment lexicons for 150 years of English and community-specific sentiment lexicons for 250 online communities from the social media forum Reddit. The historical lexicons show that more than 5% of sentiment-bearing (non-neutral) English words completely switched polarity during the last 150 years, and the community-specific lexicons highlight how sentiment varies drastically between different communities.

1 Introduction

Inducing domain-specific sentiment lexicons is crucial to computational social science (CSS) research. Sentiment lexicons allow us to analyze key subjective properties of texts like opinions and attitudes (Taboada et al., 2011). But lexical sentiment is hugely influenced by context. The word *soft* has a very different sentiment in an online sports community than it does in one dedicated to toy animals (Figure 1). *Terrific* once had a highly negative connotation; now it is essentially synonymous with *good* (Figure 2). Without domain-specific lexicons, social scientific analyses can be misled by sentiment assignments biased towards domain-general contexts, neglecting factors like genre, community-specific vernacular, or demographic variation (Deng et al., 2014; Hovy, 2015; Yang and Eisenstein, 2015).

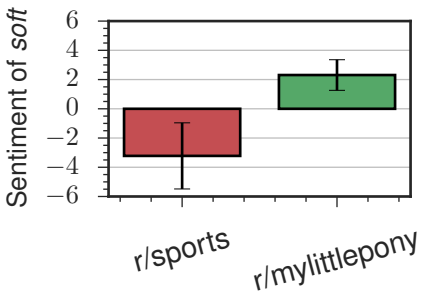


Figure 1: The sentiment of *soft* in different online communities. Sentiment values computed using SENTPROP (Section 3) on comments from Reddit communities illustrate how sentiment depends on social context. Bootstrap-sampled standard deviations provide a measure of confidence with the scores.

Using experts or crowdsourcing to construct domain-specific sentiment lexicons is expensive and often time-consuming (Mohammad and Turney, 2010; Fast et al., 2016), and is especially problematic when non-standard language (as in historical documents or obscure social media forums) prevents annotators from understanding the sociolinguistic context of the data.

Web-scale sentiment lexicons can be automatically induced for large socially-diffuse domains, such as the internet-at-large (Velikovich et al., 2010) or all of Twitter (Tang et al., 2014). However, to study sentiment in domain-specific cases—financial documents, historical texts, or tight-knit social me-

dia forums—such generic lexicons may be inaccurate, and even introduce harmful biases (Loughran and McDonald, 2011).[1] Researchers need a principled and accurate framework for inducing lexicons that are specific to their domain of study.

To meet these needs, we introduce SENTPROP, a framework to learn accurate sentiment lexicons from small sets of seed words and domain-specific corpora. SENTPROP combines the well-known method of label propagation with advances in word embeddings, and unlike previous approaches, is designed to be accurate even when using modestly-sized domain-specific corpora ($\sim 10^7$ tokens). Our framework also provides *confidence scores* along with the learned lexicons, which allows researchers to quantify uncertainty in a principled manner.

The key contributions of this work are:

1. A simple state-of-the-art sentiment induction algorithm, combining high-quality word vector embeddings with a label propagation approach.
2. A novel bootstrap-sampling framework for inferring confidence scores with the sentiment values.
3. Two large-scale studies that reveal how sentiment depends on both social and historical context.
   (a) We induce community-specific sentiment lexicons for the largest 250 "subreddit" communities on the social-media forum Reddit, revealing substantial variation in word sentiment between communities.
   (b) We induce historical sentiment lexicons for 150 years of English, revealing that $>5\%$ of words switched polarity during this time.

To the best of our knowledge, this is the first work to systematically analyze the domain-dependency of sentiment at a large-scale, across hundreds of years and hundreds of user-defined online communities.

All of the inferred lexicons along with code for SENTPROP and all methods evaluated are made available in the SOCIALSENT package released with this paper.[2] The SOCIALSENT package provides a benchmark toolkit for inducing sentiment lexicons, including implementations of previously published algorithms (Velikovich et al., 2010; Rothe et al., 2016), which are not otherwise publicly available.

[1] http://brandsavant.com/brandsavant/the-hidden-bias-of-social-media-sentiment-analysis

[2] http://nlp.stanford.edu/projects/socialsent

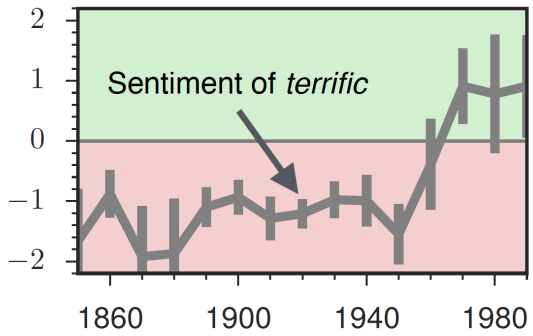


Figure 2: The sentiment of *terrific* changed from negative to positive over the last 150 years. Sentiment values and bootstrapped confidences were computed using SENTPROP on historical data (see Section 6).

2 Related work

Our work builds upon a wealth of previous research on inducing sentiment lexicons, along two threads:

Corpus-based approaches use seed words and patterns in unlabeled corpora to induce domain-specific lexicons. These patterns may rely on syntactic structures (Hatzivassiloglou and McKeown, 1997; Jijkoun et al., 2010; Rooth et al., 1999; Thelen and Riloff, 2002; Widdows and Dorow, 2002), which can be domain-specific and brittle (e.g., in social media lacking usual grammatical structures). Other models rely on general co-occurrence (Igo and Riloff, 2009; Riloff and Shepherd, 1997; Turney and Littman, 2003). Often corpus-based methods exploit distant-supervision signals (e.g., review scores, emoticons) specific to certain domains (Asghar et al., 2015; Blair-Goldensohn et al., 2008; Bravo-Marquez et al., 2015; Choi and Cardie, 2009; Severyn and Moschitti, 2015; Speriosu et al., 2011; Tang et al., 2014). An effective corpus-based approach that does not require distant-supervision—which we adapt here—is to construct lexical graphs using word co-occurrences and then to perform some form of label propagation over these graphs (Huang et al., 2014; Velikovich et al., 2010). Recent work has also learned transformations of word-vector representations in order to induce sentiment lexicons (Rothe et al., 2016). Fast et al. (2016) combine word vectors with crowdsourcing to produce domain-independent topic lexicons.

Dictionary-based approaches use hand-curated lexical resources—usually WordNet (Fellbaum, 1998)—in order to propagate sentiment from seed

Domain	Positive seed words	Negative seed words
Standard English	good, lovely, excellent, fortunate, pleasant, delightful, perfect, loved, love, happy	bad, horrible, poor, unfortunate, unpleasant, disgusting, evil, hated, hate, unhappy
Finance	successful, excellent, profit, beneficial, improving, improved, success, gains, positive	negligent, loss, volatile, wrong, losses, damages, bad, litigation, failure, down, negative
Twitter	love, loved, loves, awesome, nice, amazing, best, fantastic, correct, happy	hate, hated, hates, terrible, nasty, awful, worst, horrible, wrong, sad

Table 1: Seed words. The seed words were manually selected to be context insensitive (without knowledge of the test lexicons).

labels (Esuli and Sebastiani, 2006; Hu and Liu, 2004; Kamps et al., 2004; Rao and Ravichandran, 2009; San Vicente et al., 2014; Takamura et al., 2005; Tai and Kao, 2013). There is an implicit consensus that dictionary-based approaches will generate higher-quality lexicons, due to their use of these clean, hand-curated resources; however, they are not applicable in domains lacking such a resource (e.g., most historical texts).

Most previous work seeks to enrich or enlarge existing lexicons (Qiu et al., 2009; San Vicente et al., 2014; Velikovich et al., 2010), emphasizing recall over precision. This recall-oriented approach is motivated by the need for massive polarity lexicons in tasks like web-advertising (Velikovich et al., 2010). In contrast to these previous efforts, the goal of this work is to induce high-quality lexicons that are accurate to a specific social context.

Algorithmically, our approach is inspired by Velikovich et al. (2010). We extend Velikovich et al. (2010) by incorporating high-quality word vector embeddings, a new graph construction approach, an alternative label propagation algorithm, and a bootstrapping method to obtain confidence values. Together these improvements, especially the high-quality word vectors, allow our corpus-based method to even outperform the state-of-the-art dictionary-based approach.

3 Framework

Our framework, SENTPROP, is designed to meet four key desiderata:

1. **Resource-light:** Accurate performance without massive corpora or hand-curated resources.
2. **Interpretable**: Uses small seed sets of "paradigm" words to maintain interpretability and avoid ambiguity in sentiment values.
3. **Robust**: Bootstrap-sampled standard deviations provide a measure of confidence.
4. **Out-of-the-box**: Does not rely on signals that are specific to only certain domains.

SENTPROP involves two steps: constructing a lexical graph from unlabeled corpora and propagating sentiment labels over this graph.

3.1 Constructing a lexical graph

Lexical graphs are constructed from distributional word embeddings learned on unlabeled corpora.

Distributional word embeddings

The first step in our approach is to build high-quality semantic representations for words using a vector space model (VSM). We embed each word $w_i \in \mathcal{V}$ as a vector $\mathbf{w}_i$ that captures information about its co-occurrence statistics with other words (Landauer and Dumais, 1997; Turney and Pantel, 2010). This VSM approach has a long history in NLP and has been highly successful in recent applications (see Levy et al., 2015 for a survey).

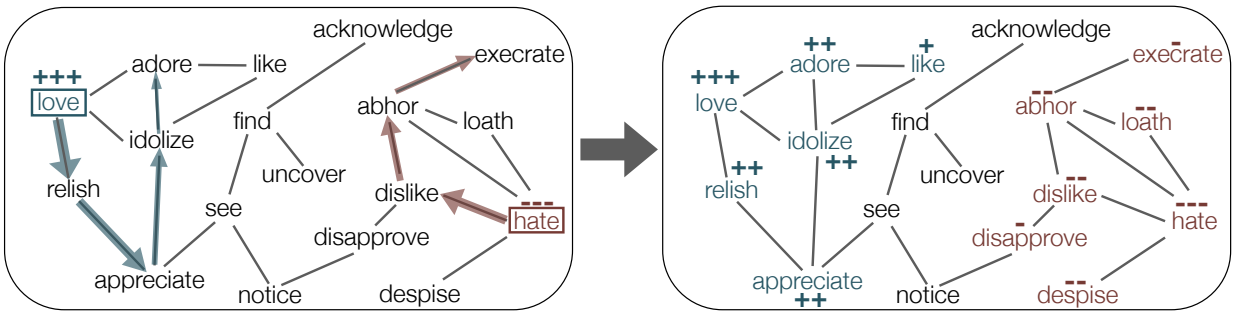
When recreating known lexicons, we used a number of publicly available embeddings (Section 4).

In the cases where we learned embeddings ourselves, we employed an SVD-based method to construct the word-vectors. First, we construct a matrix $\mathbf{M}^{PPMI} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ with entries given by

$$\mathbf{M}^{PPMI}_{i,j} = \max\left\{\log\left(\frac{\hat{p}(w_i, w_j)}{\hat{p}(w)\hat{p}(w_j)}\right), 0\right\}, \quad (1)$$

where $\hat{p}$ denotes smoothed empirical probabilities of word (co-)occurrences within fixed-size sliding windows of text.[3] $\mathbf{M}^{PPMI}_{i,j}$ is equal to a smoothed variant of the positive pointwise mutual information between words w_i and w_j (Levy et al., 2015). Next, we compute $\mathbf{M}^{PPMI} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, the truncated singular value decomposition of $\mathbf{M}^{PPMI}$. The vector

[3] We use contexts of size four on each side and context-distribution smoothing with $c = 0.75$ (Levy et al., 2015).



a. Run random walks from seed words. **b. Assign polarity scores based on frequency of random walk visits.**

Figure 3: Visual summary of the SENTPROP algorithm.

embedding for word w_i is then given by

$$\mathbf{w}_i^{\text{SVD}} = (\mathbf{U})_i \, . \tag{2}$$

Excluding the singular value weights, Σ, has been shown known to dramatically improve embedding quality (Turney and Pantel, 2010; Bullinaria and Levy, 2012). Following standard practices, we learn embeddings of dimension 300.

We found that this SVD-based method significantly outperformed word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) on preliminary experiments with the domain-specific data we used (see Section 4).

Defining the graph edges

Given a set of word embeddings, a weighted lexical graph is constructed by connecting each word with its nearest k neighbors within the semantic space (according to cosine-similarity). The weights of the edges are set as

$$\mathbf{E}_{i,j} = \arccos\left(-\frac{\mathbf{w_i}^\top \mathbf{w_j}}{\|\mathbf{w_i}\|\|\mathbf{w_j}\|}\right). \tag{3}$$

3.2 Propagating polarities from a seed set

Once a weighted lexical graph is constructed, we propagate sentiment labels over this graph using a random walk method (Zhou et al., 2004). A word's polarity score for a seed set is proportional to the probability of a random walk from the seed set hitting that word (Figure 3).

Let $\mathbf{p} \in \mathbb{R}^{|\mathcal{V}|}$ be a vector of word-sentiment scores constructed using seed set $\mathcal{S}$ (e.g., ten negative words); $\mathbf{p}$ is initialized to have $\frac{1}{|\mathcal{V}|}$ in all entries. And let $\mathbf{E}$ be the matrix of edge weights given by equation (3). First, we construct a symmetric transition matrix from $\mathbf{E}$ by computing $\mathbf{T} = \mathbf{D}^{\frac{1}{2}}\mathbf{E}\mathbf{D}^{\frac{1}{2}}$, where $\mathbf{D}$ is a matrix with the column sums of $\mathbf{E}$ on the diagonal. Next, using $\mathbf{T}$ we iteratively update $\mathbf{p}$ until numerical convergence:

$$\mathbf{p}^{(t+1)} = \beta\mathbf{T}\mathbf{p}^{(t)} + (1-\beta)\mathbf{s}, \tag{4}$$

where $\mathbf{s}$ is a vector with values set to $\frac{1}{|\mathcal{S}|}$ in the entries corresponding to the seed set $\mathcal{S}$ and zeros elsewhere. The β term controls the extent to which the algorithm favors local consistency (similar labels for neighbors) vs. global consistency (correct labels on seed words), with lower βs emphasizing the latter.

To obtain a final polarity score for a word w_i, we run the walk using both positive and negative seed sets, obtaining positive ($\mathbf{p}^P(w_i)$) and negative ($\mathbf{p}^N(w_i)$) label scores. We then combine these values into a positive-polarity score as $\bar{\mathbf{p}}^P(w_i) = \frac{\mathbf{p}^P(w_i)}{\mathbf{p}^P(w_i)+\mathbf{p}^N(w_i)}$ and standardize the final scores to have zero mean and unit variance (within a corpus).

3.3 SENTPROP variants

Many variants of the random walk approach and related label propagation techniques exist in the literature (San Vicente et al., 2014; Velikovich et al., 2010; Zhou et al., 2004; Zhu and Ghahramani, 2002; Zhu et al., 2003). For example, there are differences in how to normalize the transition matrix in the random walks (Zhou et al., 2004) and variants of label propagation, e.g. where the labeled seeds are clamped to the correct values (Zhu and Ghahramani, 2002) or where only shortest-paths through the graph are used for propagation (Velikovich et al., 2010).

We experimented with a number of these approaches and found little difference in their performance. We opted to use the random walk method because it had a slight edge in terms of performance

in preliminary experiments[4] and because it produces well-behaved distributions over label scores, whereas Zhu and Ghahramani (2002)'s method and its variants produce extremely peaked distributions. We do note report in detail on all the label propagation variants here, but the SOCIALSENT package contains a full suite of these methods.

3.4 Bootstrap-sampling for robustness

Propagated sentiment scores are inevitably influenced by the seed set, and it is important for researchers to know the extent to which polarity values are simply the result of corpus artifacts that are correlated with these seeds words. We address this issue by using a bootstrap-sampling approach to obtain confidence regions over our sentiment scores. We bootstrap by running our propagation over B random equally-sized subsets of the positive and negative seed sets. Computing the standard deviation of the bootstrap-sampled polarity scores provides a measure of confidence and allows the researcher to evaluate the robustness of the assigned polarities. We set $B = 50$ and used 7 words per random subset (full seed sets are size 10; see Table 1).

4 Recreating known lexicons

We validate our approach by recreating known sentiment lexicons in the three domains: Standard English, Twitter, and Finance. Table 1 lists the seed words used in each domain.

Standard English: To facilitate comparison with previous work, we focus on the well-known General Inquirer lexicon (Stone et al., 1966). We also use the continuous valence (i.e., polarity) scores collected by Warriner et al. (2013) in order to evaluate the fine-grained performance of our framework. We test our framework's performance using two different embeddings: off-the-shelf Google news embeddings constructed from 10^{11} tokens[5] and embeddings we constructed from the 2000s decade of the Corpus of Historical American English (COHA), which contains $\sim 2 \times 10^7$ words in each decade, from 1850 to 2000 (Davies, 2010). The COHA corpus allows us to test how the algorithms deal with this smaller historical corpus, which is important since we will use the COHA corpus to infer historical sentiment lexicons (Section 6).

Finance: Previous work found that general purpose sentiment lexicons performed very poorly on financial text (Loughran and McDonald, 2011), so a finance-specific sentiment lexicon (containing binary labels) was hand-constructed for this domain (ibid.). To test against this lexicon, we constructed embeddings using a dataset of $\sim 2 \times 10^7$ tokens from financial 8K documents (Lee et al., 2014).

Twitter: Numerous works attempt to induce Twitter-specific sentiment lexicons using supervised approaches and features unique to that domain (e.g., follower graphs; Speriosu et al., 2011). Here, we emphasize that we can induce an accurate lexicon using a simple domain-independent and resource-light approach, with the implication that lexicons can easily be induced for related social media domains without resorting to complex supervised frameworks. We evaluate our approach using the test set from the 2015 SemEval task 10E competition (Rosenthal et al., 2015), and we use the embeddings constructed by Rothe et al. (2016).[6]

4.1 Baselines and state-of-the-art comparisons

We compare SENTPROP against standard baselines and state-of-the-art approaches. The PMI baseline of Turney and Littman (2003) computes the pointwise mutual information between the seeds and the targets without using propagation. The baseline method of Velikovich et al. (2010) is similar to our method but uses an alternative propagation approach and raw co-occurrence vectors instead of learned embeddings. Both these methods require raw corpora, so they function as baselines in cases where we do not use off-the-shelf embeddings. We also compare against DENSIFIER, a state-of-the-art method that learns orthogonal transformations of word vectors instead of propagating labels (Rothe et al., 2016). Lastly, on standard English we compare against a state-of-the-art WordNet-based method, which performs label propagation over a WordNet-derived graph (San Vicente et al., 2014). Several variant baselines, all of which SENTPROP

[4] >2% improvement across metrics on the standard and historical English datasets described in Section 4.

[5] https://code.google.com/p/word2vec/

[6] The official SemEval task 10E involved fully-supervised learning, so we do not use their evaluation setup.

Method	AUC	Ternary F1	τ
SENTPROP	90.6	58.6	0.44
DENSIFIER	**93.3**	**62.1**	**0.50**
WordNet	89.5	58.7	0.34
Majority	–	24.8	–

(a) Corpus methods outperform WordNet on standard English. Using word-vector embeddings learned on a massive corpus (10^{11} tokens), we see that both corpus-based methods outperform the WordNet-based approach overall.

Method	AUC	Ternary F1	τ
SENTPROP	86.0	**60.1**	0.50
DENSIFIER	**90.1**	59.4	**0.57**
Sentiment140	86.2	57.7	0.51
Majority	–	24.9	–

(b) Corpus approaches are competitive with a distantly supervised method on Twitter. Using Twitter embeddings learned from $\sim 10^9$ tokens, we see that the semi-supervised corpus approaches using small seed sets perform very well.

Method	AUC	Ternary F1
SENTPROP	**91.6**	**63.1**
DENSIFIER	80.2	50.3
PMI	86.1	49.8
Velikovich et al. (2010)	81.6	51.1
Majority	–	23.6

(c) SENTPROP performs best with domain-specific finance embeddings. Using embeddings learned from financial corpus ($\sim 2 \times 10^7$ tokens) , SENTPROP significantly outperforms the other methods.

Method	AUC	Ternary F1	τ
SENTPROP	**83.8**	**53.0**	**0.28**
DENSIFIER	77.4	46.6	0.19
PMI	70.6	41.9	0.16
Velikovich et al. (2010)	52.7	32.9	0.01
Majority	–	24.3	–

(d) SENTPROP performs well on standard English even with 1000x reduction in corpus size. SENTPROP maintains strong performance even when using embeddings learned from the 2000s decade of COHA (only $2 \times \sim 10^7$ tokens).

Table 2: Results on recreating known lexicons.

outperforms, are omitted for brevity (e.g., using word-vector cosines in place of PMI in Turney and Littman (2003)'s framework). Code for all these variants is available in the SOCIALSENT package.

4.2 Evaluation setup

We evaluate the approaches according to (i) their binary classification accuracy (ignoring the neutral class, as is common in previous work), (ii) ternary classification performance (positive vs. neutral vs. negative)[7], and (iii) Kendall τ rank-correlation with continuous human-annotated polarity scores.

For all methods in the ternary-classification condition, we use the class-mass normalization method (Zhu et al., 2003) to label words as positive, neutral, or negative. This method assumes knowledge of the label distribution—i.e., how many positive/negative vs. neutral words there are—and simply assigns labels to best match this distribution.

4.3 Evaluation results

Tables 2a-2d summarize the performance of our framework along with baselines and other state-of-the-art approaches. Our framework significantly outperforms the baselines on all tasks, outperforms a state-of-the-art approach that uses WordNet on standard English (Table 2a), and is competitive with Sentiment140 on Twitter (Table 2b), a distantly-supervised approach that uses signals from emoticons (Mohammad and Turney, 2010). DENSIFIER also performs extremely well, outperforming SENTPROP when off-the-shelf embeddings are used (Tables 2a and 2b). However, SENTPROP significantly outperforms all other approaches when using the domain-specific embeddings (Tables 2c and 2d).

Overall our results show that SENTPROP— a relatively simple method, which combines high-quality word vectors embeddings with standard label propagation — can perform at a state-of-the-art level, even performing competitively with methods relying on hand-curated lexical graphs. Unlike previous published approaches, SENTPROP is able to maintain high accuracy even when modest-sized domain-specific corpora are used. In cases where very large corpora are available and where there is an abundance of training data, DENSIFIER performs extremely well, since it was designed for this sort of setting (Rothe et al., 2016).

[7]Only GI contains words explicitly marked neutral, so for ternary evaluations in Twitter and Finance we sample neutral words from GI to match its neutral-vs-not distribution.

We found that the baseline method of Velikovich et al. (2010), which our method is closely related to, performed relatively poorly with these domain-specific corpora. This indicates that using high-quality word-vector embeddings can have a drastic impact on performance. However, it is worth noting that Velikovich et al. (2010)'s method was designed for high recall with massive corpora, so its poor performance in our regime is not surprising.

Lastly, we found that the choice of embedding method could have a drastic impact. Preliminary experiments on the COHA data showed that using word2vec SGNS vectors (with default settings) instead of our SVD-based embeddings led to a $>40\%$ performance drop for SENTPROP across all measures and a $>10\%$ performance drop for DENSIFIER. It is possible that certain settings of word2vec could perform better, but previous work has shown that SVD-based methods have superior results on smaller datasets and rare-word similarity tasks (Levy et al., 2015; Hamilton et al., 2016), so this result is not surprising.

5 Inducing community-specific lexicons

As a first large-scale study, we investigate how sentiment depends on the social context in which a word is used. It is well known that there is substantial sociolinguistic variation between different communities, whether these communities are defined geographically (Trudgill, 1974) or via underlying sociocultural differences (Labov, 2006). However, no previous work has systematically investigated community-specific variation in word sentiment at a large scale. Yang and Eisenstein (2015) exploit social network structure in Twitter to infer a small number (1-10) of communities and analyzed sentiment variation via a supervised framework. Our analysis extends this line of work by analyzing the sentiment across hundreds of user-defined communities using only unlabeled corpora and a small set of "paradigm" seed words (the Twitter seed words outlined in Table 1).

In our study, we induced sentiment lexicons for the top-250 (by comment-count) subreddits from the social media forum Reddit.[8] We used all the 2014 comment data to induce the lexicons, with words lower cased and comments from bots and deleted users removed.[9] Sentiment was induced for the top-5000 non-stop words in each subreddit (again, by comment-frequency).

5.1 Examining the lexicons

Analysis of the learned lexicons reveals the extent to which sentiment can differ across communities. Figure 4 highlights some words with opposing sentiment in two communities: in `r/TwoXChromosomes` (`r/TwoX`), a community dedicated to female perspectives and gender issues, the words *crazy* and *insane* have negative polarity, which is not true in the `r/sports` community, and, vice-versa, words like *soft* are positive in `r/TwoX` but negative in `r/sports`.

To get a sense of how much sentiment differs across communities in general, we selected a random subset of 1000 community pairs and examined the correlation in their sentiment values for highly sentiment-bearing words (Figure 5). We see that the distribution is noticeably skewed, with many community pairs having highly uncorrelated sentiment values. The 1000 random pairs were selected such that each member of the pair overlapped in at least half of their top-5000 word vocabulary. We then computed the correlation between the sentiments in these community-pairs. Since sentiment is noisy and relatively uninteresting for neutral words, we compute $\tau_{25\%}$, the Kendall-τ correlation over the top-25% most sentiment bearing words shared between the two communities.

Analysis of individual pairs reveals some interesting insights about sentiment and inter-community dynamics. For example, we found that the sentiment correlation between `r/TwoX` and `r/TheRedPill` ($\tau_{25\%} = 0.58$), two communities that hold conflicting views and often attack each other[10], was actually higher than the sentiment correlation between `r/TwoX` and `r/sports` ($\tau_{25\%} = 0.41$), two communities that are entirely unrelated. This result suggests that conflicting communities may have

[8] Subreddits are user-created topic-specific forums.

[9] https://archive.org/details/2015_reddit_comments_corpus

[10] This conflict is well-known on Reddit; for example, both communities mention each others' names along with *fuck*-based profanity in the same comment far more than one would expect by chance ($\chi^2_1 > 6.8$, $p < 0.01$ for both). `r/TheRedPill` is dedicated to male empowerment.

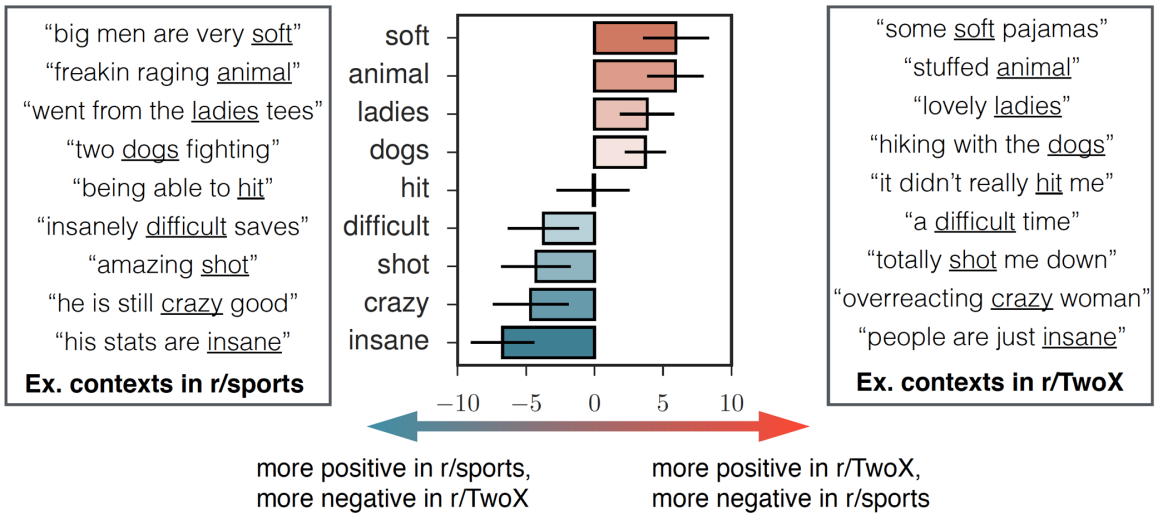


Figure 4: Word sentiment differs drastically between a community dedicated to sports (`r/sports`) and one dedicated to female perspectives and gender issues (`r/TwoX`). Words like *soft* and *animal* have positive sentiment in `r/TwoX` but negative sentiment in `r/sports`, while the opposite holds for words like *crazy* and *insane*.

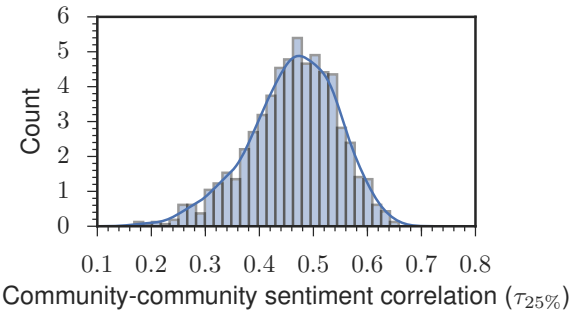


Figure 5: There is a long tail of communities with very different word sentiments. Some communities have very similar sentiment (e.g., `r/sports` and `r/hockey`), while other community pairs differ drastically (e.g., `r/sports` and `r/TwoX`).

more similar sentiment in their language compared to communities that are entirely unrelated.

6 Inducing diachronic sentiment lexicons

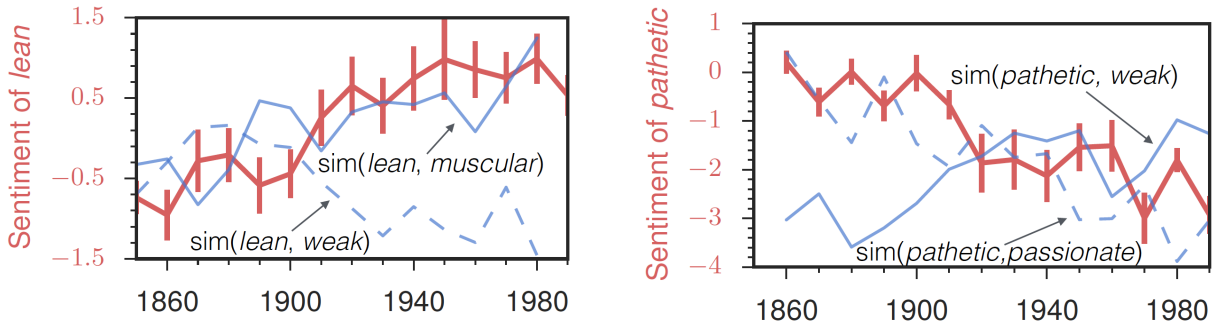
Sentiment also depends on the historical time-period in which a word is used. To investigate this dependency, we use our framework to analyze how word polarities have shifted over the last 150 years. The phenomena of *amelioration* (words becoming more positive) and *pejoration* (words becoming more negative) are well-discussed in the linguistic literature (Traugott and Dasher, 2001); however, no comprehensive polarity lexicons exist for historical data (Cook and Stevenson, 2010). Such lexicons are crucial to the growing body of work on NLP analyses of historical text (Piotrowski, 2012) which are informing diachronic linguistics (Hamilton et al., 2016), the digital humanities (Muralidharan and Hearst, 2012), and history (Hendrickx et al., 2011).

Our work is inspired by the only previous work on automatically inducing historical sentiment lexicons, Cook and Stevenson (2010); they use the PMI method and a full modern sentiment lexicon as their seed set, which relies on the assumption that all these words have not changed in sentiment. In contrast, in addition to our different algorithm, we use a small seed set of words that were manually selected based on having strong and stable sentiment over the last 150 years (Table 1; confirmed via historical entries in the Oxford English Dictionary).

6.1 Examining the lexicons

We constructed lexicons from COHA, since it was carefully constructed to be genre balanced (e.g., compared to the Google N-Grams; Pechenick et al., 2015). We built lexicons for all adjectives with counts above 100 in a given decade and also for the top-5000 non-stop words within each year. In both these cases we found that $>5\%$ of sentiment-bearing (positive/negative) words completely switched polarity during this 150-year time-period and $>25\%$ of all words changed their sentiment label (including switches to/from neutral).[11] The prevalence of

[11] We defined the thresholds for polar vs. neutral using the class-mass normalization method and compared scores aver-



(a) *Lean* becomes more positive. *Lean* underwent amelioration, becoming more similar to *muscular* and less similar to *weak*.

(b) *Pathetic* becomes more negative. *Pathetic* underwent pejoration, becoming similar to *weak* and less similar to *passionate*.

Figure 6: Examples of amelioration and pejoration.

full polarity switches highlights the importance of historical sentiment lexicons for work on diachronic linguistics and cultural change.

Figure 6a shows an example amelioration detected by this method: the word *lean* lost its negative connotations associated with "weakness" and instead became positively associated with concepts like "muscularity" and "fitness". Figure 6b shows an example pejoration, where *pathetic*, which used to be more synonymous with *passionate*, gained stronger negative associations with the concepts of "weakness" and "inadequacy" (Simpson et al., 1989). In both these cases, semantic similarities computed using our learned historical word vectors were used to contextualize the shifts.

Some other well-known examples of sentiment changes captured by our framework include the semantic bleaching of *sorry*, which shifted from negative and serious ("he was in a sorry state") to uses as a neutral discourse marker ("sorry about that") and *worldly*, which used to have negative connotations related to materialism and religious impurity ("sinful worldly pursuits") but now is frequently used to indicate sophistication ("a cultured, worldly woman") (Simpson et al., 1989). Our hope is that the full lexicons released with this work will spur further examinations of such historical shifts in sentiment, while also facilitating CSS applications that require sentiment ratings for historical text.

7 Conclusion

SENTPROP allows researchers to easily induce robust and accurate sentiment lexicons that are relevant to their particular domain of study. Such lexicons are crucial to CSS research, as evidenced by our two studies showing that sentiment depends strongly on both social and historical context.

Our methodological comparisons show that simply combining label propagation with high-quality word vector embeddings can achieve state-of-the-art performance competitive with methods that rely on hand-curated dictionaries, and the code package released with this work contains a full benchmark toolkit for this area, including implementations of several variants of SENTPROP. We hope these tools will facilitate future quantitative studies on the domain-dependency of sentiment.

Of course, the sentiment lexicons induced by SENTPROP are not perfect, which is reflected in the uncertainty associated with our bootstrap-sampled estimates. However, we believe that these user-constructed, domain-specific lexicons, which quantify uncertainty, provide a more principled foundation for CSS research compared to domain-general sentiment lexicons that contain unknown biases. In the future our method could also be integrated with supervised domain-adaption (e.g.,Yang and Eisenstein, 2015) to further improve these domain-specific results.

Acknowledgements

The authors thank P. Liang for his helpful comments. This research has been supported in part by NSF CNS-1010921, IIS-1149837, IIS-1514268 NIH BD2K, ARO MURI, DARPA XDATA, DARPA SIMPLEX, Stanford Data Science Initiative, SAP Stanford Graduate Fellowship, NSERC PGS-D, Boeing, Lightspeed, and Volkswagen.

aged over 1850-1880 to those averaged over 1970-2000.

References

Muhammad Zubair Asghar, Aurangzeb Khan, Shakeel Ahmad, Imran Ali Khan, and Fazal Masud Kundi. 2015. A Unified Framework for Creating Domain Dependent Polarity Lexicons from User Generated Reviews. *PLOS ONE*, 10(10):e0140204, October.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A. Reis, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*.

Felipe Bravo-Marquez, Eibe Frank, and Bernhard Pfahringer. 2015. From Unlabelled Tweets to Twitter-specific Opinion Words. In *SIGIR*.

John A. Bullinaria and Joseph P. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, 44(3):890–907, September.

Yejin Choi and Claire Cardie. 2009. Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *EMNLP*.

Paul Cook and Suzanne Stevenson. 2010. Automatically Identifying Changes in the Semantic Orientation of Words. In *LREC*.

Mark Davies. 2010. The Corpus of Historical American English: 400 million words, 1810-2009. http://corpus.byu.edu/coha/.

Lingjia Deng, Janyce Wiebe, and Yoonjung Choi. 2014. Joint inference and disambiguation of implicit sentiments via implicature constraints. In *COLING*.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding Topic Signals in Large-Scale Text. In *CHI*.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *ACL*.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *EACL*.

Iris Hendrickx, Michel Gnreux, and Rita Marquilhas. 2011. Automatic pragmatic text segmentation of historical letters. In *Language Technology for Cultural Heritage*, pages 135–153. Springer.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *ACL*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD*.

Sheng Huang, Zhendong Niu, and Chongyang Shi. 2014. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. *Knowledge-Based Systems*, 56:191–200.

Sean P. Igo and Ellen Riloff. 2009. Corpus-based semantic lexicon induction with web-based corroboration. In *ACL Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics*.

Valentin Jijkoun, Maarten de Rijke, and Wouter Weerkamp. 2010. Generating focused topic-specific sentiment lexicons. In *ACL*.

Jaap Kamps, M. J. Marx, Robert J. Mokken, and M. de Rijke. 2004. Using wordnet to measure semantic orientations of adjectives. In *LREC*.

William Labov. 2006. *The social stratification of English in New York City*. Cambridge University Press.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.*, 104(2):211.

Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. 2014. On the Importance of Text Analysis for Stock Price Prediction. In *LREC*.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Trans. Assoc. Comput. Ling.*, 3.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *NAACL*.

Aditi Muralidharan and Marti A Hearst. 2012. Supporting exploratory text analysis in literature study. *Literary and Linguistic Computing*.

Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE*, 10(10).

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Michael Piotrowski. 2012. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2009. Expanding Domain Sentiment Lexicon through Double Propagation. In *IJCAI*.

Delip Rao and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *EACL*.

Ellen Riloff and Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. *arXiv preprint cmp-lg/9706013*.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *ACL*.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 task 10: Sentiment analysis in Twitter. *SemEval-2015*.

Sascha Rothe, Sebastian Ebert, and Hinrich Schutze. 2016. Ultradense Word Embeddings by Orthogonal Transformation. In *NAACL-HLT*.

Inaki San Vicente, Rodrigo Agerri, German Rigau, and Donostia-San Sebastin. 2014. Simple, Robust and (almost) Unsupervised Generation of Polarity Lexicons for Multiple Languages. In *EACL*.

Aliaksei Severyn and Alessandro Moschitti. 2015. On the automatic learning of sentiment lexicons. In *NAACL-HLT*.

John Andrew Simpson, Edmund SC Weiner, et al. 1989. *The Oxford English Dictionary*, volume 2. Clarendon Press Oxford, Oxford, UK.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *ACL Workshop on Unsupervised Learning in NLP*.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The General Inquirer: A Computer Approach to Content Analysis.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Comput. Ling.*, 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.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *ACL*.

Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu. 2014. Building Large-Scale Twitter-Specific Sentiment Lexicon: A Representation Learning Approach. In *COLING*.

Michael Thelen and Ellen Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *EMNLP*.

Elizabeth Closs Traugott and Richard B Dasher. 2001. *Regularity in Semantic Change*. Cambridge University Press, Cambridge, UK.

Peter Trudgill. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, 3(2):215–246.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Sys.*, 21(4):315–346.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.*, 37(1):141–188.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *NAACL-HLT*.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4):1191–1207.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *COLING*.

Yi Yang and Jacob Eisenstein. 2015. Putting Things in Context: Community-specific Embedding Projections for Sentiment Analysis. *arXiv preprint arXiv:1511.06052*.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Scholkopf. 2004. Learning with local and global consistency. In *NIPS*.

Xiaojin Zhu and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. Technical report.

Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, and others. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*.