# Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora

**William L. Hamilton**, **Kevin Clark**, **Jure Leskovec**, and **Dan Jurafsky**

Department of Computer Science, Stanford University, Stanford CA, 94305

## 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. We show that our approach achieves state-of-the-art performance on inducing sentiment lexicons from domain-specific corpora and that our purely corpus-based approach outperforms methods that rely on hand-curated resources (e.g., WordNet). Using our framework, 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 we induce 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

The sentiment of the word *soft* varies drastically between an online community dedicated to sports and one dedicated to toy animals (Figure 1). *Terrific* once had a highly negative connotation; now it is essentially synonomous with *good* (Figure 2).

Inducing domain-specific sentiment lexicons is crucial to computational social science (CSS) research. Sentiment lexicons allow researchers to analyze key subjective properties of texts, such as user opinions and emotional attitudes (Taboada et al., 2011). However, without domain-specific lexicons, analyses can be misled by sentiment assignments that are biased towards domain-general contexts and that fail to take into account community-specific vernacular or demographic variations in language use (Hovy, 2015; Yang and Eisenstein, 2015).

Experts or crowdsourced annotators can be used to construct sentiment lexicons for a specific domain, but these efforts are expensive and time-consuming (Mohammad and Turney, 2010; Fast et al., 2016). Crowdsourcing is especially problematic when the domain involves very non-standard language (e.g., historical documents or obscure social media forums), since in these cases annotators must understand the sociolinguistic context of the data.

Recent work has shown that 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, in cases where researchers want to analyze the sentiment of domain-specific language—such as in financial documents, historical texts, or tight-knit social media forums—it is not enough to simply use generic crowdsourced or web-scale lexicons. Generic lexicons will not only be inaccurate in specific domains, they may mislead research by introducing 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 this need, we introduce SentProp, a framework to learn accurate sentiment lexicons from small sets of seed words and domain-specific corpora. Unlike previous approaches, SentProp is designed to maintain accurate performance when using modestly-sized domain-specific corpora ($\sim 10^7$ tokens), and it 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 state-of-the-art sentiment induction algorithm, combining high-quality word vector embeddings with an intuitive 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]

## 2 Related work

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

---

[1] http://brandsavant.com/brandsavant/the-hidden-bias-of-social-media-sentiment-analysis
[2] http://nlp.stanford.edu/projects/socialsent

*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; Thelen and Riloff, 2002; Widdows and Dorow, 2002; Jijkoun et al., 2010; Rooth et al., 1999), which can be domain-specific and brittle (e.g., in social media lacking usual grammatical structures). Other models rely on general co-occurrence (Turney and Littman, 2003; Riloff and Shepherd, 1997; Igo and Riloff, 2009). 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-general topic lexicons.

*Dictionary-based* approaches use hand-curated lexical resources—usually WordNet (Fellbaum, 1998)—in order to propagate sentiment from seed 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 (San Vicente et al., 2014; Velikovich et al., 2010; Qiu et al., 2009), 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 this work by incorporating high-quality word vector embeddings, a new graph construction approach, an alternative label propagation algorithm, and a bootstrapping method to obtain confidences. 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 building 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

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

where $\hat{p}$ denotes smoothed empirical probabilities of word (co-)occurrences within fixed-size sliding windows of text.[3] $\mathbf{M}_{i,j}^{PPMI}$ 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 embedding for word $w_i$ is then given by

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

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

We found that this SVD method significantly outperformed word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) on the domain-specific datasets we examined. Our

---

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

results echo the findings of Levy et al. (2015) that the SVD approach performs best on rare word similarity tasks.

**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

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

## 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.

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:

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

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 $\beta$ 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 $\beta$s 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 $\overline{\mathbf{p}}^P(w_i) = \frac{\mathbf{p}^P(w_i)}{\mathbf{p}^P(w_i)+\mathbf{p}^N(w_i)}$ and standardize the final scores to have zero mean and unit variance (within a corpus).

Many variants of this random walk approach and related label propagation techniques exist in the literature (Zhou et al., 2004; Zhu and Ghahramani, 2002; Zhu et al., 2003; Velikovich et al., 2010; San Vicente et al., 2014). We experimented with a number of these approaches and found little difference between their performance, so we present only this random walk approach here. The SocialSent package contains a full suite of these methods.

## 3.3 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[4] 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).[5]

### 4.1 Baselines and state-of-the-art comparisons

We compared SentProp against standard baselines and state-of-the-art approaches. The PMI baseline of Turney and Littman (2003) computes the point-wise mutual information between the seeds and the targets without using propagation. The CountVec baseline, corresponding

---

[4]https://code.google.com/p/word2vec/
[5]The official SemEval task 10E involved fully-supervised learning, so we do not use their evaluation setup.

to the method in 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 which 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 outperforms, are omitted for brevity (e.g., using word-vector cosines in place of PMI in Turney and Littman (2003)'s framework). Code for replicating 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)[6], and (iii) Kendall $\tau$ 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 Sent-Prop 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, SentProp is competitive with the state-of-the-art across all conditions and, unlike previous approaches, it is able to maintain high accuracy even when modestly-sized domain-specific corpora are used. We found that the baseline method of Velikovich et al. (2010), which our method is closely related to, performed very 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.

---

[6]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.

## 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 structures 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.[7] We used all the 2014 comment data to induce the lexicons, with words lower cased and comments from bots and deleted users removed.[8] 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 3 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 4). 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-$\tau$ 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[9], was actually higher than the sentiment correlation between r/TwoX

---

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

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

[9]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 ($\chi_1^2 > 6.8$, $p < 0.01$ for both). r/TheRedPill is dedicated to male empowerment.

and r/sports ($\tau_{25\%} = 0.41$), two communities that are entirely unrelated. This result suggests that conflicting communities may have 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).

The only previous work on automatically inducing historical sentiment lexicons is Cook and Stevenson (2010); they use the PMI method and a full modern sentiment lexicon as their seed set, which problematically assumes that all these words have not changed in sentiment. In contrast, we use a small seed set of words that were manually selected based upon 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).[10] The prevalence of full polarity switches highlights the importance of historical sentiment lexicons for work on diachronic linguistics and cultural change.

Figure 5a 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 5b 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

---

[10]We defined the thresholds for polar vs. neutral using the class-mass normalization method and compared scores averaged over 1850–1880 to those averaged over 1970–2000.

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.

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.

## Acknowledgments

## References

Asghar, Muhammad Zubair, Khan, Aurangzeb, Ahmad, Shakeel, Khan, Imran Ali, Kundi, Fazal Masud. A Unified Framework for Creating Domain Dependent Polarity Lexicons from User Generated Reviews. PLOS ONE. 2015 Oct.10(10):e0140204. [PubMed: 26466101]

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

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

Bullinaria, John A., Levy, Joseph P. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. Behavior Research Methods. 2012 Sep; 44(3):890–907. [PubMed: 22258891]

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

Cook, Paul, Stevenson, Suzanne. Automatically Identifying Changes in the Semantic Orientation of Words. LREC; 2010.

Davies, Mark. The Corpus of Historical American English: 400 million words, 1810–2009. 2010.

Esuli, Andrea, Sebastiani, Fabrizio. Sentiwordnet: A publicly available lexical resource for opinion mining. LREC; Citeseer. 2006.

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

Fellbaum, Christiane. WordNet. Wiley Online Library; 1998.

Hamilton, William L., Leskovec, Jure, Jurafsky, Dan. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. ACL. 2016 arXiv:1605.09096.

Hatzivassiloglou, Vasileios, McKeown, Kathleen R. Predicting the semantic orientation of adjectives. EACL; 1997.

Hendrickx, Iris, Gnreux, Michel, Marquilhas, Rita. Language Technology for Cultural Heritage. Springer; 2011. Automatic pragmatic text segmentation of historical letters; p. 135-153.

Hovy, Dirk. Demographic factors improve classification performance. ACL. 2015

Hu, Minqing, Liu, Bing. KDD. ACM; 2004. Mining and summarizing customer reviews.

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

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

Jijkoun, Valentin, de Rijke, Maarten, Weerkamp, Wouter. Generating focused topic-specific sentiment lexicons. ACL. 2010:585–594.

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

Labov, William. The social stratification of English in New York city. Cambridge University Press; 2006.

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

Lee, Heeyoung, Surdeanu, Mihai, MacCartney, Bill, Jurafsky, Dan. On the Importance of Text Analysis for Stock Price Prediction. LREC; 2014. p. 1170-1175.

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

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

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

Mohammad, Saif M., Turney, Peter D. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. NAACL; 2010. p. 26-34.

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

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

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

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

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

Rao, Delip, Ravichandran, Deepak. Semi-supervised polarity lexicon induction. EACL; 2009. p. 675-682.

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

Rooth, Mats, Riezler, Stefan, Prescher, Detlef, Carroll, Glenn, Beil, Franz. Inducing a semantically annotated lexicon via EM-based clustering. ACL; 1999. p. 104-111.

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

Rothe, Sascha, Ebert, Sebastian, Schtze, Hinrich. Ultradense Word Embeddings by Orthogonal Transformation. 2016 arXiv preprint arXiv:1602.07572.

Vicente, Inaki San, Agerri, Rodrigo, Rigau, German, Sebastin, Donostia-San. Simple, Robust and (almost) Unsupervised Generation of Polarity Lexicons for Multiple Languages. EACL; 2014. p. 88-97.

Severyn, Aliaksei, Moschitti, Alessandro. On the automatic learning of sentiment lexicons. NAACL-HLT; 2015.

Simpson, John Andrew, Weiner, Edmund SC., et al. The Oxford English Dictionary. Vol. 2. Clarendon Press Oxford; Oxford, UK: 1989.

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

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

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

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

Takamura, Hiroya, Inui, Takashi, Okumura, Manabu. Extracting semantic orientations of words using spin model. ACL; 2005. p. 133-140.

Tang, Duyu, Wei, Furu, Qin, Bing, Zhou, Ming, Liu, Ting. Building Large-Scale Twitter-Specific Sentiment Lexicon: A Representation Learning Approach. COLING; 2014. p. 172-182.

Thelen, Michael, Riloff, Ellen. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. EMNLP; 2002.

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

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

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

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

Velikovich, Leonid, Blair-Goldensohn, Sasha, Hannan, Kerry, McDonald, Ryan. The viability of web-derived polarity lexicons. NAACL-HLT; 2010. p. 777-785.

Warriner, Amy Beth, Kuperman, Victor, Brysbaert, Marc. Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior research methods. 2013; 45(4):1191–1207. [PubMed: 23404613]

Widdows, Dominic, Dorow, Beate. A graph model for unsupervised lexical acquisition. COLING; 2002.

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

Zhou, Dengyong, Bousquet, Olivier, Lal, Thomas Navin, Weston, Jason, Schlkopf, Bernhard. Learning with local and global consistency. NIPS. 2004; 16

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

Zhu, Xiaojin, Ghahramani, Zoubin, Lafferty, John, et al. Semi-supervised learning using gaussian fields and harmonic functions. ICML; 2003. p. 912-919.
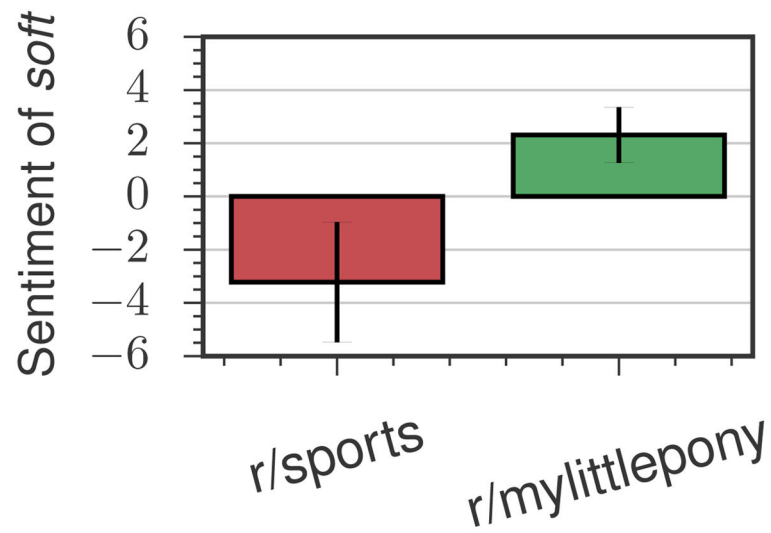
**Figure 1. The sentiment of *soft* in different online communities**
Sentiment values computed using SentProp (Section 3) on comments from Reddit communities illustrate how sentiment depends on social context. Bootstrap-sampled standard deviations provide a measure of confidence with the scores.
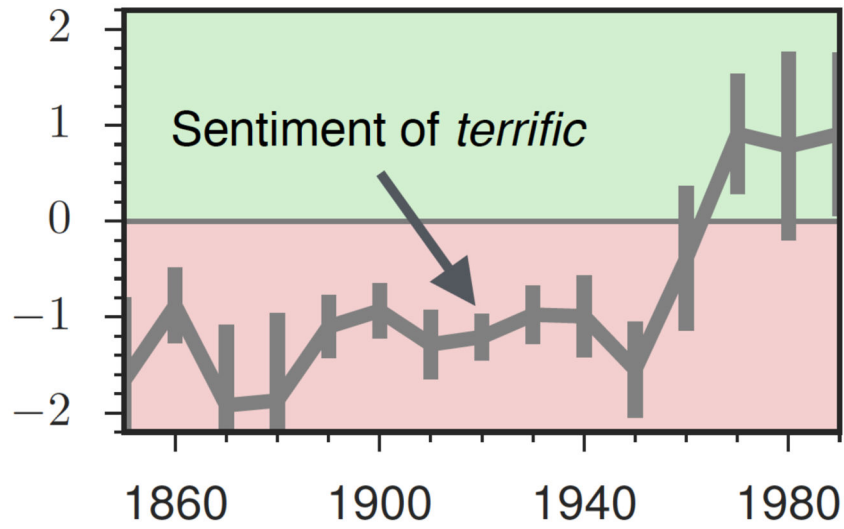
**Figure 2.** *Terrific* **becomes more positive over the last 150 years**
Sentiment values and bootstrapped confidences were computed using SentProp on historical data (Section 6).

**Figure 3. 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*.
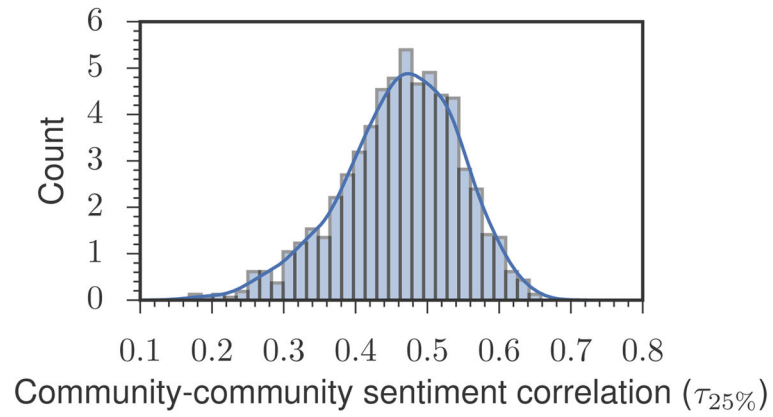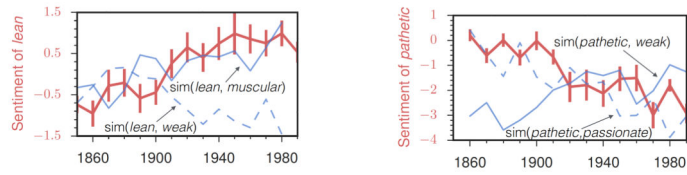
**Figure 4. There is a long tail of communities with very different word sentiments**

Some communities have very similar sentiment (e.g., r/sports and r/hockey), while other other community pairs differ drastically (e.g., r/sports and r/TwoX).

(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 5. Examples of amelioration and pejoration**
(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*.

**Table 1**

**Seed words**

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

| 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 2**

Results on recreating known lexicons.

**(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 | $\tau$ |
|---|---|---|---|
| SentProp | 90.6 | 58.6 | 0.44 |
| Densifier | **93.3** | **62.1** | **0.50** |
| WordNet | 89.5 | 58.7 | 0.34 |
| Majority | – | 24.8 | – |

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

| Method | AUC | Ternary F1 | $\tau$ |
|---|---|---|---|
| SentProp | 86.0 | **60.1** | 0.50 |
| Densifier | **90.1** | 59.4 | **0.57** |
| Sentiment140 | 86.2 | 57.7 | 0.51 |
| Majority | – | 24.9 | – |

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

| Method | AUC | Ternary F1 |
|---|---|---|
| SentProp | **91.6** | **63.1** |
| Densifier | 80.2 | 50.3 |
| PMI | 86.1 | 49.8 |
| CountVecs | 81.6 | 51.1 |
| Majority | – | 23.6 |

**(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 $2 \times$ ~$10^7$ tokens).**

| Method | AUC | Ternary F1 | $\tau$ |
|---|---|---|---|
| SentProp | **83.8** | **53.0** | **0.28** |
| Densifier | 77.4 | 46.6 | 0.19 |
| PMI | 70.6 | 41.9 | 0.16 |
| CountVecs | 52.7 | 32.9 | 0.01 |
| Majority | – | 24.3 | – |