Exploring Sentiment in Social Media: Bootstrapping Subjectivity Clues from Multilingual Twitter Streams

Svitlana Volkova
CLSP
Johns Hopkins University
Baltimore, MD
svitlana@jhu.edu

Theresa Wilson
HLTCOE
Johns Hopkins University
Baltimore, MD
taw@jhu.edu

David Yarowsky
CLSP
Johns Hopkins University
Baltimore, MD
yarowsky@cs.jhu.edu

Abstract

We study subjective language in social media and create Twitter-specific lexicons via bootstrapping sentiment-bearing terms from multilingual Twitter streams. Starting with a domain-independent, high-precision sentiment lexicon and a large pool of unlabeled data, we bootstrap Twitter-specific sentiment lexicons, using a small amount of labeled data to guide the process. Our experiments on English, Spanish and Russian show that the resulting lexicons are effective for sentiment classification for many under-explored languages in social media.

1 Introduction

The language that people use to express opinions and sentiment is extremely diverse. This is true for well-formed data, such as news and reviews, and it is particularly true for data from social media. Communication in social media is informal, abbreviations and misspellings abound, and the person communicating is often trying to be funny, creative, and entertaining. Topics change rapidly, and people invent new words and phrases.

The dynamic nature of social media together with the extreme diversity of subjective language has implications for any system with the goal of analyzing sentiment in this domain. General, domain-independent sentiment lexicons have low coverage. Even models trained specifically on social media data may degrade somewhat over time as topics change and new sentiment-bearing terms crop up. For example, the word "occupy" would not have been indicative of sentiment before 2011.

Most of the previous work on sentiment lexicon construction relies on existing natural language processing tools, e.g., syntactic parsers (Wiebe, 2000), information extraction (IE) tools (Riloff and Wiebe, 2003) or rich lexical resources such as WordNet (Esuli and Sebastiani, 2006). However, such tools and lexical resources are not available for many languages spoken in social media. While English is still the top language in Twitter, it is no longer the majority. Thus, the applicability of these approaches is limited. Any method for analyzing sentiment in microblogs or other social media streams must be easily adapted to (1) many low-resource languages, (2) the dynamic nature of social media, and (3) working in a streaming mode with limited or no supervision.

Although bootstrapping has been used for learning sentiment lexicons in other domains (Turney and Littman, 2002; Banea et al., 2008), it has not yet been applied to learning sentiment lexicons for microblogs. In this paper, we present an approach for bootstrapping subjectivity clues from Twitter data, and evaluate our approach on English, Spanish and Russian Twitter streams. Our approach:
- handles the informality, creativity and the dynamic nature of social media;
- does not rely on language-dependent tools;
- scales to the hundreds of new under-explored languages and dialects in social media;
- classifies sentiment in a streaming mode.

To bootstrap subjectivity clues from Twitter streams we rely on three main assumptions:

i. sentiment-bearing terms of similar orientation tend to co-occur at the tweet level (Turney and Littman, 2002);

ii. sentiment-bearing terms of opposite orientation do not co-occur at the tweet level (Gamon and Aue, 2005);

iii. the co-occurrence of domain-specific and domain-independent subjective terms serves as a signal of subjectivity.

2 Related Work

Mihalcea et.al (2012) classifies methods for bootstrapping subjectivity lexicons into two types: corpus-based and dictionary-based.

Dictionary-based methods rely on existing lexical resources to bootstrap sentiment lexicons. Many researchers have explored using relations in WordNet (Miller, 1995), e.g., Esuli and Sabastiani (2006), Andreevskaia and Bergler (2006) for English, Rao and Ravichandran (2009) for Hindi and French, and Perez-Rosas et al. (2012) for Spanish. Mohammad et al. (2009) use a thesaurus to aid in the construction of a sentiment lexicon for English. Other works (Clematide and Klenner, 2010; Abdul-Mageed et al., 2011) automatically expands and evaluates German and Arabic lexicons. However, the lexical resources that dictionary-based methods need, do not yet exist for the majority of languages in social media. There is also a mismatch between the formality of many language resources, such as WordNet, and the extremely informal language of social media.

Corpus-based methods extract subjectivity and sentiment lexicons from large amounts of unlabeled data using different similarity metrics to measure the relatedness between words. Hatzivassiloglou and McKeown (1997) were the first to explore automatically learning the polarity of words from corpora. Early work by Wiebe (2000) identifies clusters of subjectivity clues based on their distributional similarity, using a small amount of data to bootstrap the process. Turney (2002) and Velikovich et al. (2010) bootstrap sentiment lexicons for English from the web by using Pointwise Mutual Information (PMI) and graph propagation approach, respectively. Kaji and Kitsuregawa (2007) propose a method for building sentiment lexicon for Japanese from HTML pages. Banea et al. (2008) experiment with Lexical Semantic Analysis (LSA) (Dumais et al., 1988) to bootstrap a subjectivity lexicon for Romanian. Kanayama and Nasukawa (2006) bootstrap subjectivity lexicons for Japanese by generating subjectivity candidates based on word co-occurrence patterns.

In contrast to other corpus-based bootstrapping methods, we evaluate our approach on multiple languages, specifically English, Spanish, and Russian. Also, as our approach relies only on the availability of a bilingual dictionary for translating an English subjectivity lexicon and crowdsourcing for help in selecting seeds, it is more scalable and

better able to handle the informality and the dynamic nature of social media. It also can be effectively used to bootstrap sentiment lexicons for any language for which a bilingual dictionary is available or can be automatically induced from parallel corpora.

3 Data

For the experiments in this paper, we use three sets of data for each language: 1M *unlabeled* tweets (BOOT) for bootstrapping Twitter-specific lexicons, 2K labeled tweets for development data (DEV), and 2K labeled tweets for evaluation (TEST). DEV is used for parameter tuning while bootstrapping, and TEST is used to evaluating the quality of the bootstrapped lexicons.

We take English tweets from the corpus constructed by Burger et al. (2011) which contains 2.9M tweets (excluding retweets) from 184K users.[1] English tweets are identified automatically using a compression-based language identification (LID) tool (Bergsma et al., 2012). According to LID, there are 1.8M (63.6%) English tweets, which we randomly sample to create BOOT, DEV and TEST sets for English. Unfortunately, Burger's corpus does not include Russian and Spanish data on the same scale as English. Therefore, for other languages we construct a new Twitter corpus by downloading tweets from followers of region-specific news and media feeds.

Sentiment labels for tweets in DEV and TEST sets for all languages are obtained using Amazon Mechanical Turk. For each tweet we collect annotations from five workers and use majority vote to determine the final label for the tweet. Snow et al. (2008) show that for a similar task, labeling emotion and valence, on average four non-expert labelers are needed to achieve an expert level of annotation. Table 1 gives the distribution of tweets over sentiment labels for the development and test sets for English (E-DEV, E-TEST), Spanish (S-DEV, S-TEST), and Russian (R-DEV, R-TEST). Below are examples of tweets in Russian with English translations labeled with sentiment:

- Positive: В планах вкусный завтрак и куча фильмов (Planning for delicious breakfast and lots of movies);
- Negative: Хочу сдохнуть, и я это сделаю (I want to die and I will do that);

Data	Positive	Neg	Both	Neutral
E-DEV	617	357	202	824
E-TEST	596	347	195	862
S-DEV	358	354	86	1,202
S-TEST	317	387	93	1203
R-DEV	452	463	156	929
R-TEST	488	380	149	983

Table 1: Sentiment label distribution in development DEV and test TEST datasets across languages.

- **Both:** Хочется написать грубее про фильм но не буду. Хотя актеры хороши (I want to write about the movie rougher but I will not. Although the actors are good);
- **Neutral:** Почему умные мысли приходят только ночью? (Why clever thoughts come only at night?).

4 Lexicon Bootstrapping

To create a Twitter-specific sentiment lexicon for a given language, we start with a general-purpose, high-precision sentiment lexicon[2] and bootstrap from the unlabeled data (BOOT) using the labeled development data (DEV) to guide the process.

4.1 High-Precision Subjectivity Lexicons

For English we seed the bootstrapping process with the strongly subjective terms from the MPQA lexicon[3] (Wilson et al., 2005). These terms have been previously shown to be high-precision for recognizing subjective sentences (Riloff and Wiebe, 2003).

For the other languages, the subjective seed terms are obtained by translating English seed terms using a bilingual dictionary, and then collecting judgments about term subjectivity from Mechanical Turk. Terms that truly are strongly subjective in translation are used for seed terms in the new language, with term polarity projected from the English. Finally, we expand the lexicons with plurals and inflectional forms for adverbs, adjectives and verbs.

4.2 Bootstrapping Approach

To bootstrap, first the new lexicon $L_{B(0)}$ is seeded with the strongly subjective terms from the original lexicon L_I. On each iteration $i \geq 1$, tweets in the unlabeled data are labeled using the lexicon

from the previous iteration, $L_{B(i-1)}$. If a tweet contains one or more terms from $L_{B(i-1)}$ it is considered subjective, otherwise objective. The polarity of subjective tweets is determined in a similar way: if the tweet contains ≥ 1 positive terms, taking into account the negation, it is considered negative; if it contains ≥ 1 negative terms, taking into account the negation, it is considered positive.[4] If it contains both positive and negative terms, it is considered to be both. Then, for every term not in $L_{B(i-1)}$ that has a frequency $\geq \theta_{freq}$, the probability of that term being subjective is calculated as shown in Algorithm 1 line 10. The top θ_k terms with a subjective probability $\geq \theta_{pr}$ are then added to $L_{B(i)}$. The polarity of new terms is determined based on the probability of the term appearing in positive or negative tweets as shown in line 18.[5] The bootstrapping process terminates when there are no more new terms meeting the criteria to add.

Algorithm 1 BOOTSTRAP $(\sigma, \theta_{pr}, \theta_{freq}, \theta_{topK})$

1: $iter = 0, \sigma = 0.5, L_B(\vec{\theta}) \leftarrow L_I(\sigma)$
2: **while** $(stop \neq true)$ **do**
3:     $L_B^{iter}(\vec{\theta}) \leftarrow \emptyset, \Delta L_B^{iter}(\vec{\theta}) \leftarrow \emptyset$
4:     **for** each new term $w \in \{V \setminus L_B(\vec{\theta})\}$ **do**
5:         **for** each tweet $t \in T$ **do**
6:             **if** $w \in t$ **then**
7:                 UPDATE $c(w, L_B(\vec{\theta})), c(w, L_B^{pos}(\vec{\theta})), c(w)$
8:             **end if**
9:         **end for**
10:         $p^{subj}(w) \leftarrow \frac{c(w, L_B(\vec{\theta}))}{c(w)}$
11:         $p^{pos}(w) \leftarrow \frac{c(w, L_B^{pos}(\vec{\theta}))}{c(w, L_B(\vec{\theta}))}$
12:         $L_B^{iter}(\vec{\theta}) \leftarrow w, p^{subj}(w), p^{pol}(w)$
13:     **end for**
14:     SORT $L_B^{iter}(\vec{\theta})$ by $p^{subj}(w)$
15:     **while** $(K \leq \theta_{topK})$ **do**
16:         **for** each new term $w \in L_B^{iter}(\vec{\theta})$ **do**
17:             **if** $[p^{subj}(w) \geq \theta_{pr}$ **and** $c_w \geq \theta_{freq}$ **then**
18:                 **if** $[p^{pos}(w) \geq 0.5]$ **then**
19:                     $w^{pol} \leftarrow positive$
20:                 **else**
21:                     $w^{pol} \leftarrow negative$
22:                 **end if**
23:                 $\Delta L_B^{iter}(\vec{\theta}) \leftarrow \Delta L_B^{iter}(\vec{\theta}) + w^{pol}$
24:             **end if**
25:         **end for**
26:         $K = K + 1$
27:     **end while**
28:     **if** $[\Delta L_B^{iter}(\vec{\theta}) == 0]$ **then**
29:         $stop \leftarrow true$
30:     **end if**
31:     $L_B(\vec{\theta}) \leftarrow L_B(\vec{\theta}) + \Delta L_B^{iter}(\vec{\theta})$
32:     $iter = iter + 1$
33: **end while**

---

[4]If there is a negation in the two words before a sentiment term, we flip its polarity.

[5]Polarity association probabilities should sum up to 1 $p^{pos}(w|L_B(\vec{\theta})) + p^{neg}(w|L_B(\vec{\theta})) = 1$.

	English		Spanish		Russian	
	L_I^E	L_B^E	L_I^S	L_B^S	L_I^R	L_B^R
Pos	2.3	16.8	2.9	7.7	1.4	5.3
Neg	2.8	4.7	5.2	14.6	2.3	5.5
Total	**5.1**	**21.5**	**8.1**	**22.3**	**3.7**	**10.8**

Table 2: The original and the bootstrapped (high-lighted) lexicon term count ($L_I \subset L_B$) with polarity across languages (thousands).

The set of parameters $\vec{\theta}$ is optimized using a grid search on the development data using F-measure for subjectivity classification. As a result, for English $\vec{\theta} = [0.7, 5, 50]$ meaning that on each iteration the top 50 new terms with a frequency ≥ 5 and probability ≥ 0.7 are added to the lexicon. For Spanish, the set of optimal parameters $\vec{\theta} = [0.65, 3, 50]$ and for Russian - $\vec{\theta} = [0.65, 3, 50]$. In Table 2 we report size and term polarity from the original L_I and the bootstrapped L_B lexicons.

5 Lexicon Evaluations

We evaluate our bootstrapped sentiment lexicons English L_B^E, Spanish L_B^S and Russian L_B^R by comparing them with existing *dictionary-expanded* lexicons that have been previously shown to be effective for subjectivity and polarity classification (Esuli and Sebastiani, 2006; Perez-Rosas et al., 2012; Chetviorkin and Loukachevitch, 2012). For that we perform subjectivity and polarity classification using rule-based classifiers[6] on the test data E-TEST, S-TEST and R-TEST.

We consider how the various lexicons perform for rule-based classifiers for both subjectivity and polarity. The subjectivity classifier predicts that a tweet is subjective if it contains a) at least one, or b) at least two subjective terms from the lexicon. For the polarity classifier, we predict a tweet to be positive (negative) if it contains at least one positive (negative) term taking into account negation. If the tweet contains both positive and negative terms, we take the majority label.

For English we compare our bootstrapped lexicon L_B^E against the original lexicon L_I^E and strongly subjective terms from SentiWordNet 3.0 (Esuli and Sebastiani, 2006). To make a fair comparison, we automatically expand SentiWord-Net with noun plural forms and verb inflectional forms. In Figure 1 we report precision, recall

[6]Similar approach to a rule-based classification using terms from he MPQA lexicon (Riloff and Wiebe, 2003).

and F-measure results. They show that our boot-strapped lexicon significantly outperforms Senti-WordNet for subjectivity classification. For polarity classification we get comparable F-measure but much higher recall for L_B^E compared to S_{WN}.

(a) Subj $\geq$ 1   (b) Subj $\geq$ 2   (c) Polarity
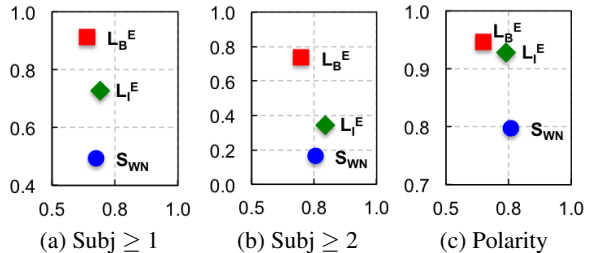
Lexicon	$F_{subj \geq 1}$	$F_{subj \geq 2}$	$F_{polarity}$
S_{WN}	0.57	0.27	0.78
L_I^E	0.71	0.48	**0.82**
L_B^E	**0.75**	**0.72**	0.78

Figure 1: Precision (x-axis), recall (y-axis) and F-measure (in the table) for English: L_I^E = initial lexicon, L_B^E = bootstrapped lexicon, S_{WN} = strongly subjective terms from SentiWordNet.

For Spanish we compare our bootstrapped lexicon L_B^S against the original L_I^S lexicon, and the full and medium strength terms from the Spanish sentiment lexicon constructed by Perez-Rosas et el. (2012). We report precision, recall and F-measure in Figure 2. We observe that our boot-strapped lexicon yields significantly better performance for subjectivity classification compared to both full and medium strength terms. However, our bootstrapped lexicon yields lower recall and similar precision for polarity classification.

(a) Subj $\geq$ 1   (b) Subj $\geq$ 2   (c) Polarity
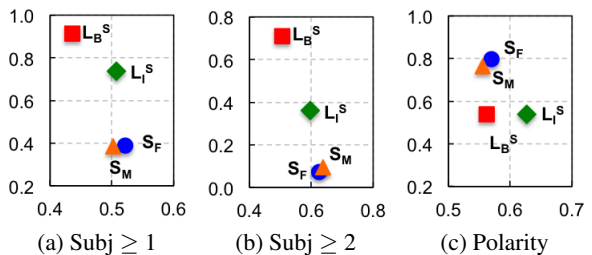
Lexicon	$F_{subj \geq 1}$	$F_{subj \geq 2}$	$F_{polarity}$
S_M	0.44	0.17	0.64
S_F	0.47	0.13	**0.66**
L_I^S	0.59	0.45	0.58
L_B^S	**0.59**	**0.59**	0.55

Figure 2: Precision (x-axis), recall (y-axis) and F-measure (in the table) for Spanish: L_I^S = initial lexicon, L_B^S = bootstrapped lexicon, S_F = full strength terms; S_M = medium strength terms.

508

For Russian we compare our bootstrapped lexicon L_B^R against the original L_I^R lexicon, and the Russian sentiment lexicon constructed by Chetviorkin and Loukachevitchet (2012). The external lexicon in Russian P was built for the domain of product reviews and does not include polarity judgments for subjective terms. As before, we expand the external lexicon with the inflectional forms for adverbs, adjectives and verbs. We report results for Russian in Figure 3. We find that for subjectivity our bootstrapped lexicon shows better performance compared to the external lexicon (5k terms). However, the expanded external lexicon (17k terms) yields higher recall with a significant drop in precision. Note that for Russian, we report polarity classification results for L_B^R and L_I^R lexicons only because P does not have polarity labels.
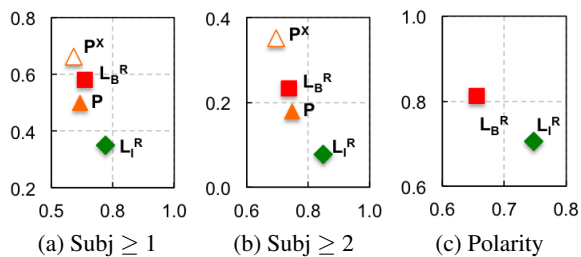
Lexicon	$F_{subj \geq 1}$	$F_{subj \geq 2}$	$F_{polarity}$
P	0.55	0.29	–
P_X	0.62	0.47	–
L_I^R	0.46	0.13	0.73
L_B^R	**0.61**	**0.35**	**0.73**

Figure 3: Precision (x-axis), recall (y-axis) and F-measure for Russian: L_I^R = initial lexicon, L_B^R = bootstrapped lexicon, P = external sentiment lexicon, P^X = expanded external lexicon.

We next perform error analysis for subjectivity and polarity classification for all languages and identify common errors to address them in future.

For subjectivity classification we observe that applying part-of-speech tagging during the bootstrapping could improve results for all languages. We could further improve the quality of the lexicon and reduce false negative errors (subjective tweets classified as neutral) by focusing on sentiment-bearing terms such as adjective, adverbs and verbs. However, POS taggers for Twitter are only available for a limited number of languages such as English (Gimpel et al., 2011). Other false negative errors are often caused by misspellings.[7]

---

[7]For morphologically-rich languages, our approach covers different linguistic forms of terms but not their misspellings. However, it can be fixed by an edit-distance check.

We also find subjective tweets with philosophical thoughts and opinions misclassified, especially in Russian, e.g., Иногда мы бываем не готовы к исполнению заветной мечты но все равно так не хочется ее спугнуть (Sometimes we are not ready to fulfill our dreams yet but, at the same time, we do not want to scare them). Such tweets are difficult to classify using lexicon-based approaches and require deeper linguistic analysis.

False positive errors for subjectivity classification happen because some terms are weakly subjective and can be used in both subjective and neutral tweets e.g., the Russian term хвастаться (brag) is often used as subjective, but in a tweet никогда не стоит хвастаться будущим (never brag about your future) it is used as neutral. Similarly, the Spanish term *buenas* (good) is often used subjectively but it is used as neutral in the following tweet "@Diveke me falto el buenas! jaja que onda que ha pasado" (I miss the good times we had, haha that wave has passed!).

For polarity classification, most errors happen because our approach relies on either positive or negative polarity scores for a term but not both.[8] However, in the real world terms may sometimes have both usages. Thus, some tweets are misclassified (e.g., "It is too warm outside"). We can fix this by summing over weighted probabilities rather than over term counts. Additional errors happen because tweets are very short and convey multiple messages (e.g., "What do you mean by unconventional? Sounds exciting!") Thus, our approach can be further improved by adding word sense disambiguation and anaphora resolution.

6 Conclusions

We propose a scalable and language independent bootstrapping approach for learning subjectivity clues from Twitter streams. We demonstrate the effectiveness of the bootstrapping procedure by comparing the resulting subjectivity lexicons with state-of the-art sentiment lexicons. We perform error analysis to address the most common error types in the future. The results confirm that the approach can be effectively exploited and further improved for subjectivity classification for many under-explored languages in social media.

---

[8]During the bootstrapping we calculate probability for a term to be positive and negative, e.g., $p(warm|+) = 0.74$ and $p(warm|-) = 0.26$. But during polarity classification we rely on the highest probability score and consider it to be "the polarity" for the term e.g., positive for *warm*.

References

Muhammad Abdul-Mageed, Mona T. Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of ACL/HLT*.

Alina Andreevskaia and Sabine Bergler. 2006. Mining wordnet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of EACL*.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of LREC*.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific Twitter collections. In *Proceedings of 2nd Workshop on Language in Social Media*.

John D. Burger, John C. Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twittier. In *Proceedings of EMNLP*.

Ilia Chetviorkin and Natalia V. Loukachevitch. 2012. Extraction of Russian sentiment lexicon for product meta-domain. In *Proceedings of COLING*.

Simon Clematide and Manfred Klenner. 2010. Evaluation and extension of a polarity lexicon for German. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*.

Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Scott Deerwester, and Richard Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *Proceedings of SIGCHI*.

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

Michael Gamon and Anthony Aue. 2005. Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twittier: annotation, features, and experiments. In *Proceedings of ACL*.

Vasileios Hatzivassiloglou and Kathy McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of ACL*.

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

Nobuhiro Kaji and Masaru Kitsuregawa. 2007. Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of EMNLP*.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of EMNLP*.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2012. Multilingual subjectivity and sentiment analysis. In *Proceedings of ACL*.

George A. Miller. 1995. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11).

Saif Mohammad, Cody Dunne, and Bonnie Dorr. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of EMNLP*.

Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. 2012. Learning sentiment lexicons in Spanish. In *Proceedings of LREC*.

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

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of EMNLP*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast – but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*.

Peter D. Turney and Michael L. Littman. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *Computing Research Repository*.

Peter D. Turney. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*.

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

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of AAAI*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of EMNLP*.