Analysing domain suitability of a sentiment lexicon by identifying distributionally bipolar words

Lucie Flekova¹, Eugen Ruppert² and Daniel Preoţiu-Pietro³

¹Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt
² FG Language Technology, Technische Universität Darmstadt
³ Computer & Information Science, University of Pennsylvania
flekova@ukp.informatik.tu-darmstadt.de, ruppert@lt.tu-darmstadt.de, danielpr@sas.upenn.edu

Abstract

Contemporary sentiment analysis approaches rely heavily on lexicon based methods. This is mainly due to their simplicity, although the best empirical results can be achieved by more complex techniques. We introduce a method to assess suitability of generic sentiment lexicons for a given domain, namely to identify frequent bigrams where a polar word switches polarity. Our bigrams are scored using Lexicographers Mutual Information and leveraging large automatically obtained corpora. Our score matches human perception of polarity and demonstrates improvements in classification results using our enhanced context-aware method. Our method enhances the assessment of lexicon based sentiment detection algorithms and can be further used to quantify ambiguous words.

1 Introduction

Sentiment prediction from microblogging posts is of the utmost interest for researchers as well as commercial organizations. State-of-the-art sentiment research often focuses on in-depth semantic understanding of emotional constructs (Trivedi and Eisenstein, 2013; Cambria et al., 2013; De Larrabeiti et al., 2010) or neural network models (Socher et al., 2013; Severyn and Moschitti, 2015). However, recent sentiment prediction challenges show that the vast majority of currently used systems is still based on supervised learning techniques with the most important features derived from pre-existing sentiment lexica (Rosenthal et al., 2014; Rosenthal et al., 2015).

Sentiment lexicons were initially developed as general-purpose resources (Pennebaker et al., 2001; Strapparava et al., 2004; Hu and Liu, 2004; Wilson et al., 2005). Recently, there has been an increasing amount of work on platform-specific lexicons such as Twitter (Mohammad, 2012; Mohammad et al., 2013). However, even customized platform- and domain-specific lexica still suffer from ambiguities at a contextual level, e.g. cold beer (+) or cold food (-), dark chocolate (+) or dark soul (-).

In this paper, we propose a method to assess the suitability of an established lexicon for a new platform or domain by leveraging automatically collected data approximating sentiment labels (silver standard). We present a method for creating switched polarity bigram lists to explicitly reveal and address the issues of a lexicon in question (e.g. the positivity of cold beer, dark chocolate or limited edition). Note that the contextual polarity switch does not necessarily happen on sense level, but within one word sense. We demonstrate that the explicit usage of such inverse polarity bigrams and replacement of the words with high ambiguity improves the performance of the classifier on unseen test data and that this improvement exceeds the performance of simply using all in-domain bigrams.

Further, our bigram ranking method is evaluated by human raters, showing high face validity.

2 Related Work

Sentiment research has tremendously expanded in the past decade. Overall, sentiment lexicons are the most popular inputs to polarity classification (Rosenthal et al., 2015; Rosenthal et al., 2014), although the lexicons alone are far from sufficient. Initial studies relied heavily on explicit, manually crafted sentiment lexicons (Kim and Hovy, 2004; Pang and Lee, 2004; Hu and Liu, 2004). There have been efforts to infer the polarity lexicons automatically. Turney and Littman (2003) determined the semantic orientation of a target word \( t \) by comparing its association with two seed sets of manually crafted target words. Others derived the polar-
ity from other lexicons (Baccianella et al., 2010; Mohammad et al., 2009), and adapted lexicons to specific domains, for example using integer linear programming (Choi and Cardie, 2009).

Lexicons are not stable across time and domain. Cook and Stevenson (2010) proposed a method to compare dictionaries for amelioration and pejoration of words over time. Mitra et al. (2014) analyzed changes in senses over time. Dragut et al. (2012) examined inconsistency across lexicons.

Negation and its scope has been studied extensively (Moilanen and Pulman, 2008; Pang and Lee, 2004; Choi and Cardie, 2009). Polar words can even carry an opposite sentiment in a new domain (Blitzer et al., 2007; Andreevskaia and Bergler, 2006; Schwartz et al., 2013; Wilson et al., 2005). Wilson et al. (2005) identified polarity shifter words to adjust the sentiment on phrase level. Choi and Cardie (2009) validated that topic-specific features would enhance existing sentiment classifiers. Ikeda et al. (2008) first proposed a machine learning approach to detect polarity shifting for sentence-level sentiment classification. Taboada et al. (2011) presented a polarity lexicon with negation words and intensifiers, which they refer to as contextual valence shifters (Polanyi and Zaenen, 2006). Research by Kennedy and Inken (2006) dealt with negation and intensity by creating a discrete modifier scale, namely, the occurrence of good might be either good, not good, intensified good, or diminished good. A similar approach was taken by Steinberger et al. (2012). Polarity modifiers, however, do not distinguish cases such as cannot be bad from cannot be worse.

Further experiments revealed that some nouns can carry sentiment per se (e.g. chocolate, injury). Recently, several noun connotation lexicons have been built (Feng et al., 2013; Klenner et al., 2014) based on a set of seed adjectives. One of the biggest disadvantages of polarity lexicons, however, is that they rely on either positive or negative score of a word, while in reality it can be used in both contexts even within the same domain (Volkova et al., 2013).

3 Method

This section describes our methodology for identifying ambiguous sentiment bearing lexicon words based on the contexts they appear in. We demonstrate our approach on two polarity lexicons consisting of single words, namely the lexicon of Hu and Liu (Hu and Liu, 2004), further denoted HL, and the MPQA lexicon (Wilson et al., 2005). First we use a corpus of automatically collected Twitter sentiment data set of over one million tweets (detailed in section 3.2) to compute bigram polarities for the lexicon words and determine contexts which alter the polarity of the original lexicon word. Using the JoBimText framework (Biemann and Riedl, 2013), we build a large Twitter bigram thesaurus which serves as a background frequency distribution which aids in ranking the bigrams (see section 3.1). For each lexicon word, we then replace the most ambiguous words with bigrams. We compare this on sentiment prediction with a straightforward usage of all bigrams.

3.1 Twitter Bigram Thesaurus

Methods based on word co-occurrence have a long tradition in NLP research, being used in tasks such as collocation extraction or sentiment analysis. Turney and Littman (2003) used polarity seeds to measure words which co-occur with positive/negative contexts. However, the PMI is known to be sensitive to low count words and bigrams, overemphasising them over high frequency words. To account for this, we express the mutual information of a word bigram by means of Lexicographer’s Mutual Information (LMI).

The LMI, introduced by Kilgarriff et al. (2004), offers an advantage to Pointwise Mutual Information (PMI), as the scores are multiplied over the bigram frequency, boosting more frequent combinations of word (w) and context (c).

\[
\text{PMI}(w, c) = \log_2 \left( \frac{f(w, c)}{f(w) \cdot f(c)} \right)
\]

\[
\text{LMI}(w, c) = \text{PMI}(w, c) \cdot f(w, c)
\]

3.2 Bigram Sentiment Scores

We compute the LMI over a corpus of positive, respectively negative tweets, in order to obtain positive (LMI\textsubscript{pos}) and negative (LMI\textsubscript{neg}) bigram scores. We combine the following freely available data, leading to a large corpus of positive and negative tweets:

- 1.6 million automatically labeled tweets from the Sentiment140 data set (Go et al., 2009), collected by searching for positive and negative emoticons;

\footnote{An online demo illustrating the score values and distributional term similarities in this Twitter space can be found at the LT website http://maggie.lt.informatik.tu-darmstadt.de/jobimviz/}
– 7,000 manually labeled tweets from University of Michigan;¹
– 5,500 manually labeled tweets from Niek J. Sanders;²
– 2,000 manually labeled tweets from the STS-Gold data set (Saif et al., 2013).

We filtered out fully duplicate messages, as these appear to bring more noise than realistic frequency information. The resulting corpus contains 794,000 positive and 791,000 negative tweets. In pursuance of comparability between the positive and negative LMI scores, we weight the bigrams by their relative frequency in the respective data set, thus discounting rare or evenly distributed bigrams, as illustrated for negative score in:

\[
LMI_{negREL}(w, c) = \frac{LMI_{neg}(w, c)}{f_{neg}(w, c) + f_{pos}(w, c)}
\]

Since the LMI scores from a limited sized data set are not the most reliable, we further boost them by incorporating scores from a background corpus (LMI\textsubscript{GLOB}) – described below. This approach emphasizes significant bigrams, even when their score in one polarity data set is low:

\[
LMI_{negGLOB}(w, c) = \frac{LMI_{negREL}(w, c) \cdot LMI_{GLOB}(w, c)}{f_{neg}(w, c)}
\]

As background data we use a Twitter corpus of 1% of all tweets from the year 2013, obtained through the Twitter Spritzer API. We filtered this corpus with a language filter,³ resulting in 460 million English tweets.

For each bigram, we then compute its semantic orientation:

\[
LMI_{SO} = LMI_{posGLOB} - LMI_{negGLOB}
\]

These two large bigram lists, which at this point still contain all bigrams from the Twitter sentiment corpus, are then filtered by sentiment lexica, as we are only interested in bigrams with at least one word from the original sentiment lexicon (containing single words). We chose two sentiment polarity lexica for our experiments:

– the HL lexicon (Hu and Liu, 2004) having 4,782 negative and 2,004 positive words (e.g. happy, good, bad);
– the MPQA sentiment lexicon (Wilson et al., 2005), with 1,751 positive and 2,693 negative words.⁴

The most interesting candidates for a novel bigram sentiment lexicon are:

– bigrams containing a word from a negative lexicon, which has a positive semantic orientation LMI\textsubscript{SO}, i.e. having higher global LMI in the positive data set than in the negative;
– bigrams containing a word from a positive lexicon with negative semantic orientation LMI\textsubscript{SO}

The top ranked bigrams, where local contextualization reverts the original lexicon score, are listed for both lexicons in Table 1. We can observe that the polarity shifting occurs in a broad range of situations, e.g. by using polar word as an intensity expression (super tired), by using polar word in names (desperate housewives, frank iero), by using multiword expressions, idioms and collocations (cloud computing, sincere condolences, light bulbs), but also by adding a polar nominal context to the adjective (cold beer/person, dark chocolate/thoughts, stress reliever/management, guilty pleasure/feeling).

3.3 Quantifying Polarity

We have shown how to identify words which switch to the opposite polarity based on their word context. Our next goal is to identify words which occur in many contexts with both the original and the switched polarity and therefore are, without further disambiguation, harmful in either of the lexicons. With this aim we calculate a polarity score POL\textsubscript{word} for each word \((w)\) in the polarity lexicon, using the number of its positive and negative contexts determined by their semantic orientation LMI\textsubscript{SO} as previously computed:

\[
POL(w) = p_{pos}(w) - p_{neg}(w)
\]

where we define \(p_{pos}(w)\) and \(p_{neg}(w)\), as the count of positive and negative bigrams respectively, of a

¹This lexicon also contains neutral words, which might be interesting for some applications. Since the HL lexicon does not feature neutral words, we chose to omit those entries for comparable results. The words in MPQA are further distinguished as ‘strong’ or ‘weak’ by POS tag. Since we do not maintain POS information in our distributional LMI lists, we chose to utilize all indicators equally.

²http://inclass.kaggle.com/c/s1650winter11/data
³http://www.sananalytics.com/lab/twitter-sentiment/
⁴https://github.com/shuyo/language-detection
Negative to Positive

<table>
<thead>
<tr>
<th>HL</th>
<th>Word</th>
<th>Context</th>
<th>MPQA</th>
<th>Word</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>limit</td>
<td>why</td>
<td>vice</td>
<td>-versa</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sneak</td>
<td>-peek</td>
<td>stress</td>
<td>-reliever</td>
<td></td>
</tr>
<tr>
<td></td>
<td>impossible</td>
<td>mission-</td>
<td>down</td>
<td>calmed-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>lazy</td>
<td>-sunday</td>
<td>deep</td>
<td>-breath</td>
<td></td>
</tr>
<tr>
<td></td>
<td>desperate</td>
<td>-housewives</td>
<td>long</td>
<td>-awaited</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cold</td>
<td>-beer</td>
<td>cloud</td>
<td>-computing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>guilty</td>
<td>-pleasure</td>
<td>dark</td>
<td>-haired</td>
<td></td>
</tr>
<tr>
<td></td>
<td>belated</td>
<td>-birthday</td>
<td>bloody</td>
<td>-mary</td>
<td></td>
</tr>
</tbody>
</table>

Positive to Negative

<table>
<thead>
<tr>
<th>HL</th>
<th>Word</th>
<th>Context</th>
<th>MPQA</th>
<th>Word</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>luck</td>
<td>good-</td>
<td>super</td>
<td>-duper</td>
<td></td>
</tr>
<tr>
<td></td>
<td>wisdom</td>
<td>-tooth</td>
<td>happy</td>
<td>-camper</td>
<td></td>
</tr>
<tr>
<td></td>
<td>well</td>
<td>oh-</td>
<td>just</td>
<td>-puked</td>
<td></td>
</tr>
<tr>
<td></td>
<td>work</td>
<td>gotta-</td>
<td>heart</td>
<td>-breaker</td>
<td></td>
</tr>
<tr>
<td></td>
<td>hot</td>
<td>-outside</td>
<td>gold</td>
<td>-digger</td>
<td></td>
</tr>
<tr>
<td></td>
<td>better</td>
<td>feels-</td>
<td>light</td>
<td>-bulbs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>super</td>
<td>-tired</td>
<td>sincere</td>
<td>-condolences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>enough</td>
<td>-money</td>
<td>frank</td>
<td>-iero</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Bigrams with opposite LMI sentiment orientation than the original lexicon word. Note that the polarity rarely changes on sense level i.e., same sense can have different polar contexts.

\[
p_{neg}(w) = \frac{\sum (w, c) \mathbb{1}(LMI_{SO} \leq 0)}{\sum (w, c)}
\]

Lexicon words with the lowest absolute polarity score and the highest number of different contexts \((w, c)\) are listed in Table 2.

4 Experiments

To evaluate the quality of our bigrams, we perform two studies. First, we rate our inverted polarity bigrams intrinsically using crowdsourced annotations. Second, we assess the performance of the original and adjusted lexicons on a distinct expert-constructed data set of 1,600 Facebook messages annotated for sentiment. The disambiguated bigram lexicons are available on author’s website.

4.1 Intrinsic Evaluation

We crowdsource ratings for the inverted polarity bigrams found using both the HL and MPQA lexicon. The raters were presented a list of 100 bigrams of each lexicon, with 25% having the same positive polarity as in the original lexicon, 25% the same negative polarity, 25% switching polarity from positive unigram to negative bigram and the remaining quarter vice versa. They had to answer the question ‘Which polarity does this word pair have?’, given positive, negative and also neutral as options. Each bigram is rated by three annotators and the majority vote is selected. The inter-annotator agreement is measured using weighted Cohen’s kappa (Cohen, 1968), which is especially useful for ordered annotations, as it accounts not only for chance, but also for the seriousness of a disagreement between annotators. \(\kappa\) can range from -1 to 1, where the value of 0 represents an agreement equal to chance while 1 equals to a perfect agreement, i.e. identical annotation values. We obtained an agreement of weighted Cohen’s \(\kappa = 0.55\), which represents a “moderate agreement” (Landis and Koch, 1977). The confusion matrix of average human judgement compared to our computed bigram polarity is shown in Table 3. Some of the bigrams, especially for the MPQA lexicon, were assessed as objective, which our LMI method unfortunately does not reflect beyond the score value (neutral words are less polar). However, the confusion between negatively and positively labeled bigrams was very low.
We evaluate our method on a data set of Facebook posts annotated for positive and negative sentiment by two psychologists. The posts are annotated on a scale from 1 to 9, with 1 indicating strong negative sentiment and 9 strong positive sentiment. An average rating between annotators is considered to be the final message score. Ratings follow a normal distribution, i.e. with more messages having less polar score. An inter-annotator agreement of weighted Cohen’s ω = 0.61 on exact score was reached, representing a “substantial agreement”(Landis and Koch, 1977). Given our task, in which we attempt to improve on misleading bipolar words, we removed the posts annotated as neutral (rating 5.0). This left us with 2,087 posts, of which we use only those containing at least one word from the polarity lexicons of our interest, i.e., 1,601 posts for MPQA and 1,526 posts for HL. We then compute a sentiment score of a post as a difference of positive and negative word counts present in the post. If a bigram containing the lexicon word is found, its LMI_{SO} score is used instead of the lexicon word polarity score. For the two lexicons and their modifications, we employ two evaluation measures - Pearson correlation of the sentiment score of a post with the affect score, and classification accuracy on binary label, i.e., distinguishing if the affect is negative (1–4) or positive (6–9). Table 4 presents our results of four experiments using the following features:

- using the original unigram lexicon only (1);
- using original lexicon corrected by polarity score of lexicon bigrams when they appear (2–4);
- using pruned unigram lexicon, removing words that exceed entropy threshold of 0.99 or appear in more contexts of the opposite polarity than of the assumed one (5);
- using pruned unigram lexicon corrected by polarity score of (unpruned) lexicon bigrams when they appear (6–8);
- all bigrams (9).

Table 4 shows that adding contextual bigrams brings a consistent improvement (1 vs. 2, 5 vs. 6). Especially the negative part of the bigram lexica, including bigrams of negative words which have positive orientation, consistently improves results (1 vs. 4, 5 vs. 8). Likewise, pruning of the lexicon with the polar entropy score (1 vs. 5) enhances the sentiment prediction performance. For both polarity lexicons the best performance is achieved by combining the two effects (8).

In case of the first lexicon, the performance is even higher than in case of applying for the same data a fully in-domain bigram lexicon, generated from the same large public Twitter corpus (Mohammad et al., 2013).

The correction of negative unigrams to positive bigrams does not improve the prediction as much as its counterpart. The main cause appears to be the fact that those expressions with shifted polarity shall be rather neutral - as discussed in section 4.1 and by some recent research (Zhu et al., 2014).

4.3 Discussion

Usage of bigrams does not always bring improvement, but sometimes also introduces new errors. One of the frequent sources of errors appears to be the remaining ambiguity of the bigrams due to more complex phrase structure. While the bigrams are tremendously helpful in message chunks such as ‘holy shit, tech support...’, where the holy (+1) and support (+1) is replaced by its appropriately polar contexts (-0.35, -0.85), the same replacement is harmful in a post ‘holy shit monday night was amazing’. Same applies for bigrams such as work ahead (-0.89) in ‘new house...yeah!! lots of work ahead of us!!’ or nice outside (-0.65) in ‘it’s nice outside today!’.
Additionally, the performance suffers when a longer negation window is applied, such as feeling sick in the post ‘Isn’t feeling sick woohoo!’ In our setup we did not employ explicit polarity switchers commonly used with word lexicons (Wilson et al., 2005; Pang and Lee, 2008; Steinberger et al., 2012) since the context captured by the bigrams often incorporated subtle negation hints per se, including their misspelled variations. This would make the combination of bigrams with more sophisticated syntactic features challenging.

Another very interesting issue are the bigrams which are explicitly positive but have learnt their negative connotation from a broader context, such as happy camper or looking good, which are more often used jointly with negations. Posts that use these bigrams without negation (‘someone is a happy camper!’) then lead to errors, and similarly a manual human assessment without a longer context fails. This issue concerns distributional approaches in general.

Lastly, several errors arise from the non-standard, slang and misspelled words which are not present often enough in our silver standard corpus. For example, while love you is clearly positive, love ya has a negative score. On corpora such as Twitter, further optimization of word frequency thresholds in lexical methods requires special attention.

5 Conclusion

Lexicon based methods currently remain, due to their simplicity, the most prevalent sentiment analysis approaches. While it is taken for granted that using more in-domain training data is always helpful, a little attention has been given to determining how much and why a given general-purpose lexicon can help in a specific target domain or platform. We introduced a method to identify frequent bigrams where a word switches polarity, and to find out which words are bipolar to the extent that it is better to have them removed from the polarity lexica. We demonstrated that our scores match human perception of polarity and bring improvement in the classification results using our enhanced context-aware method. Our method enhances the assessment of lexicon based sentiment detection algorithms and can be further used to quantify ambiguous words.

Acknowledgements

LF and DP-P acknowledge the support from Templeton Religion Trust, grant TRT-0048. The work of ER has been supported by the German Federal Ministry of Education and Research (BMBF) within the context of the Software Campus project LiCoRes under grant No. 01IS12054

References


Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin, 70(4).


Marie-Catherine De Marneffe, Christopher D Manning, and Christopher Potts. 2010. Was it good? It was provocative. Learning the meaning of scalar adjectives. In Proceedings of the 48th Annual Meeting of


