Expanding Opinion Lexicon with Domain Specific Opinion Words Using Semi-Supervised Approach

Abstract
Opinion words as well as opinion phrases and idioms are very useful in sentiment analysis. All these terms together build opinion or sentiment lexicons. Therefore, opinion lexicons are large lists of terms that encode the sentiment of each phrase within it. Generally, to create such a lexicon automatically, high-precision classifiers use known sentiment vocabulary, e.g. the prior polarity of an adjective at word-level, to separate corresponding phrases from a non-annotated text collection. Most unsupervised approaches try to determine prior polarity, also called semantic orientation, of adjectives. However, adjective phrases or verb phrases are useful indicators of sentiment as well. To build domain independent opinion lexicons classifiers need to be applied to a high number of corpora regarding different text categories. This introduces the challenge of ambiguity, as opinion terms or phrases often show different sentiment when used in various sorts of texts. Therefore, a tradeoff which takes the most applicable sentiment in regards of a general domain has to be developed in such a case. In this paper we show a novel approach to extract domain specific adjectives from the Twitter corpus and expand the general lexicon. We build an undirected weighted graph of the adjective pairs, and use the weighted adjacency matrix as input of the clustering algorithm.

Keywords: sentiment analysis, lexicon-based classification, opinion lexicon expansion

1 Introduction
The common place in the sentiment analysis is a group of (machine learning) algorithms, which are used to build corresponding classifiers. Besides these algorithms the topic of many papers as well as commercial systems is the construction of opinion lexicons, which are lists of words and phrases that encode the polarity (positive or negative) of each term within it. Opinion lexicons can be constructed manually or using one of automatic approaches (dictionary-based or corpus-based). Manual approach is very costly in relation to time. Hence, it is not usually used alone, but as the final method, combined with one of automated techniques.

The dictionary-based approach starts with a small set of opinion words and an online dictionary. After that the bootstrapping process is applied to extend the source set with newly found words. The idea behind this approach is to collect manually several opinion words with known polarity and then to extend this set in several iteration processes. The iteration ends when no more new opinion words are found.

The corpus-based approach depends on syntactic patterns as well as an initial list of opinion words to find other opinion words in a selected corpus. It starts with a list of seed adjectives and uses them, together with a set of constraints, to identify additional adjectives and their orientations. For instance, “and” and “or” are constraints, which specify that a paired adjectives combined with one of them usually have the same orientation. On the other hand, the constraint “but” indicates the different orientation of combined adjectives.

In this paper we show a novel approach to extract domain specific adjectives from the Twitter corpus and expand the general lexicon. We build an undirected weighted graph of the adjective pairs, and use the weighted adjacency matrix as input of the clustering algorithm. The clustering task separates all nodes in two polarity classes. This approach can easily be transferred to other domains. The paper is a part of our work, which studies the possibility of aspect-based sentiment analysis on Twitter in the domain of politics. The data was collected in context of the 2012 Republican presidential primaries in
the United States. The extracted aspects and their sentiment reveal what people like or dislike about the political candidates.

1.1 Related Work

As we already stated, to gather opinion words two main approaches have been investigated: manual and automated. Although the former approach is not effective, there are several research papers concerning this issue. In their work [1], Morinaga et al. generate syntactic and semantic rules for product reputations to determine whether any statement is an opinion and if so, whether the opinion is positive or negative. After that they search for statistically meaningful information in the stored data using different data mining techniques. The paper [2] introduces the system called Sentiment Analyzer (SA), which among other things, generates and uses an opinion lexicon and a sentiment pattern database, after applying different natural processing techniques to detect references and determines sentiment in each of them. In [3], the authors develop the methodology to extract sentiment from stock message boards. For this task, they built a lexicon which is a collection of manually gathered finance words. These words form the variables for statistical inference undertaken by the given algorithms (naïve classifier, vector-distance classifier and discriminant-based classifier).

All articles which use dictionary-based approach usually access information from WordNet [4]. Hu and Liu in their paper [5] proposed several techniques based on data mining and natural language processing to provide a feature-based summary of a large number of product reviews. First, they identify a set of adjectives. After that, for each opinion word, semantic orientation is determined and a bootstrapping technique is used to extend the original set of adjectives. The authors of paper [6] use the similar approach. They evaluate several various models of classifying and combining sentiment at word and sentence levels, and show that rather good results can be obtained with simple models and only a small set of seed opinion words. Additional techniques for refinement of list of opinion words are studied in [7, 8]. The former article develops WordNet-based measures for semantic orientation of opinion words. Measures, based on distances in the synonymy-graph are specified and results are evaluated against a collection of words generated by humans. The latter article uses three different forms of already applied semi-supervised methods for orientation detection, to decide whether a given term has a positive, negative or no subjective orientation. The results of this paper show that determining subjectivity and orientation is a much harder problem than determining orientation per se. Riloff and Wiebe first collected subjectivity clues as a part of their work. The clues were then used in [9] to detect semantic orientation. In this paper, a bootstrapping process was developed, where high-precision classifiers use known subjective vocabulary to separate subjective and objective sentences from a non-annotated text collection.

One of the early works in relation to corpus-based approach is the work of Hatzivassiloglou and McKeown [10]. A technique used in this paper starts with a set of adjectives and uses them, together with the set of linguistic constraints to identify additional opinion words and their orientations. For instance, one of the constraints specifies that conjoined adjectives usually have the same orientation. After that, a learning process is applied to a corpus to specify whether two conjoined adjectives have the same or different orientation and lists of positive and negative words are produced. Kanayama and Nasukawa in [11] expand the approach of the previous paper by introducing the idea of intra-sentential and inter-sentential sentiment consistency, where the former is similar to the considerations expressed in the work of Hatzivassiloglou and McKeown. The inter-sentential sentiment consistency, i.e. sentiment consistency between neighbouring sentences follows the idea that the same opinion orientation is usually expressed in consecutive sentences.

Generally, it is very hard to gather a general opinion lexicon for all domains, because different opinion words may be used in different domains. For this reason, there is a number of articles, which are concerned with the topic of domain-specific opinion lexicons. In paper [12], the authors investigate the domain of product features. They propose a propagation approach that exploits the relations between sentiment words and topics that the sentiment words change to extract new opinion words. This method is called double propagation, because it propagates information through both opinion words and product features. The extraction rules in this paper are designed based on relations described in dependency trees. In [13], Goeuriot et al create an opinion lexicon for the medical sciences. They first generate a lexicon, which contains opinion words from the general domain. After that, the existing
lexicon is extended using a corpus of medical reviews. The article shows that there are several words with different polarity in the general and the medical dependent lexicon.

1.2 Roadmap

The paper is organized as follows. Section 2 gives the concise description of the use of subjectivity clues lexicons to detect the primary set of opinions words. Section 3 is the main part of the paper and discusses the way, how the given seed of opinion words was extended with specific words in relation to political campaigns expressed on Twitter. In this part, the semi-supervised clustering technique has been described and two different algorithms have been applied. The work shows that the use of EM algorithm is superior in relation to the K-Means clustering one. Evaluation of the results is done in Section 4. Section 5 draws conclusions and examines possibilities of future work.

2 Building General Lexicon

Most unsupervised sentiment classification approaches try to generate a general or domain dependent opinion lexicon for words or opinion phrases. In our work, the subjectivity clues lexicon, which was presented in [14] was used to detect the seed words with their semantic orientation. The resulting subjective clues were later annotated with their prior polarity using different manually developed sources and consist of 2296 positive, 4138 negative and 444 neutral distinct opinion words. The lexicon offers additional information regarding stemming, degree of sentiment (strong or weak) and is not limited to adjectives. In this paper, we are only concerned with the seed list of adjectives and their prior polarity in the subjective clues lexicon to be able to infer unknown domain specific adjectives.

When a specific problem domain is known, general opinion lexicon may not be able to detect the sentiment of domain specific vocabulary. In that case, domain specific lexicon extensions, like one described in the next section, are necessary.

3 Opinion Lexicon Expansion

Most established general opinion lexicons were created considering no particular domain and suffer from limited coverage and inaccuracies when applied to the highly informal domains like tweets or social networks communication. Therefore, in our work we extract domain specific adjectives from the Twitter corpus and expand the general lexicon based on [10].

The following subsections describe semi-supervised clustering concerning domain specific lexicon expansion.

3.1 Word Graph

Sentiment consistency is the idea of linked adjectives having the same semantic orientation. Based on this idea, adjective pairs connected by conjunctions were extracted from all tweets published in relation to Republican presidential candidates in 2012. Those pairs include the comparative and superlative forms of adjectives, too. Furthermore, extraction takes account of second adjectives of conjunctions being preceded by an adverb like “more” or “so”, such as:

“He seems to be nice and so funny”.

In the sentence above, “nice” and “funny” are connected by the conjunction “and”, although the adverb “so” follows directly after the conjunction. (Adverbs of this type are called modifiers of an adjective phrase and can typically be removed without affecting the grammatical structure of the sentence.) This results in 178 unique adjectives where the prior polarity of 106 of them can be inferred from the general lexicon. (It is assumed that the 72 unknown words are domain specific, since they do not appear in the general lexicon.) Furthermore, 8 neutral adjectives might have not neutral polarity in this new target domain. Table 1 shows the number in different groups of extracted adjectives.
<table>
<thead>
<tr>
<th>Known Positives</th>
<th>Known Negatives</th>
<th>Known Neutrals</th>
<th>Unknown</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>46</td>
<td>8</td>
<td>72</td>
<td>178</td>
</tr>
</tbody>
</table>

Table 1: Extracted adjectives

After this analysis, an undirected weighted graph is formed out of the adjective pairs. Nodes represent adjectives and edges represent the connections between the adjectives. The weights of the edges depend on the conjunction type: “and” and “or” conjunctions are presented as weight 1, whereas “but” conjunctions are presented as -1. Thus, weight 1 implies similarity and weight -1 dissimilarity. (The weight 0 means that adjectives have no relation to each other.) Additionally, every node is connected to itself with weight 1 which initializes the adjacency matrix of the graph as identity matrix. Once insertion of all edges is finished, the weighted adjacency matrix is used as input of the clustering algorithm. The following short example shows this process:

For instance, three adjectives, “nice”, “liberal” and “tragic” can be represented as the following mathematical graph:

\[ G = (V, E) \]
\[ V = \{ \text{nice, liberal, tragic} \} \]
\[ E = \{ \{ \text{nice, liberal, weight}=1 \}, \{ \text{nice, tragic, weight}=-1 \}, \{ \text{nice, nice, weight}=1 \}, \{ \text{liberal, liberal, weight}=1 \}, \{ \text{tragic, tragic, weight}=1 \}, \{ \text{tragic, liberal, weight}=0 \}, \ldots \} \]

The corresponding adjacency matrix \( M \) is as follows:

\[
M = \begin{pmatrix}
1 & 1 & -1 \\
1 & 1 & 0 \\
-1 & 0 & 1 \\
\end{pmatrix}
\]

with vectors:

\[
\vec{v}_{\text{nice}} = (1,1,-1), \quad \vec{v}_{\text{liberal}} = (1,1,0), \quad \vec{v}_{\text{tragic}} = (-1,0,1)
\]

The word “but” in English means “contrary”. A sentence containing “but” is assumed to bear the most information and is generally handled in the following way: the opinion orientation before and after “but” are opposite to each other. Therefore, “but”-conjunctions are propagated to the next related nodes having a positively weighted connection.

### 3.1.1 Clustering

The clustering task is to separate all nodes in the input graph into two polarity classes: positive and negative. We used two techniques to split all adjectives into two groups: K-Means clustering and Expectation-Maximization (EM) algorithm.

The unsupervised K-Means clustering algorithm did not suffice in splitting up the adjectives into two groups. In this case, the parameters of K-Means are \( k=2 \) and cosine distance as distance measure function. Cluster 1 encompasses 156 out of 178 adjectives which leaves only 22 adjectives for cluster 2. Since both clusters provide a greater number of positive adjectives, it is not trivial to decide which cluster represents the positive class and which one the negative. If we declare the bigger cluster to be positive, the accuracy is 51%, meaning that the clustering process provides almost the same performance as random guessing.

Using the EM algorithm we achieved better results in partitioning all nodes than K-Means clustering when two calculated vectors are given as initial means. One initial mean represents a vector connected to all known positive nodes derived from the general opinion lexicon and one represents a negative vector connected to all negative nodes. Therefore, these initial means can be seen as a fictional total negative and total positive connected word. As this approach makes use of both labelled and unlabelled data, it is considered as semi-supervised.
Figure 1: Graph of Adjectives

According to the example of the previous section, assuming prior polarity of “nice” is known to be positive and prior polarity of “tragic” is known to be negative, the initial means are:

\[
\mu_{\text{pos}} = (1,0,0), \quad \mu_{\text{neg}} = (0,0,1)
\]

The EM algorithm of the Natural Language Toolkit (NLTK) uses a Gaussian Mixture Model [15]. In the E-step we calculated the membership probabilities for each vector and for each cluster. Then the M-step updates the parameters of the probability distribution in using the maximum likelihood estimate of the membership probabilities. This process continues until no more significant increase in the likelihood of the data is achieved. In general, the EM algorithm may only converge to a local maximum and not to an optimal solution [16]. Table 2 shows the confusion matrix for the output data.

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1 (positive)</th>
<th>Cluster 2 (negative)</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known Positives</td>
<td>43 (TP)</td>
<td>9 (FN)</td>
<td>52</td>
</tr>
<tr>
<td>Known Negatives</td>
<td>17 (FP)</td>
<td>29 (TN)</td>
<td>46</td>
</tr>
<tr>
<td>Known Neutrals</td>
<td>8</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Unknown</td>
<td>34</td>
<td>38</td>
<td>72</td>
</tr>
<tr>
<td>Σ</td>
<td>102</td>
<td>76</td>
<td>178</td>
</tr>
</tbody>
</table>

Table 2: Confusion matrix of clustered adjectives
Since some positives and negatives are known, accuracy (see Equation 1) can be calculated by ignoring unknown and neutral words. (To avoid noisy adjectives, a frequency constraint could be used to discard low-frequency adjectives. Such a measure was not used in this work, because of data shortage.)

\[
\frac{TP + TN}{TP + FP + TN + FN} = \frac{72}{98} = 73\% \quad (\text{Eq. 1})
\]

Finally, the general opinion lexicon is augmented with the newly learned positive and negative adjectives. Neutral words were not replaced in the original lexicon. Table 3 shows the list of all newly learned adjectives.

<table>
<thead>
<tr>
<th>Learned Positives</th>
<th>Learned Negatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>young, serial-adulterer, cool, unhysterical, small, national, flush, conservative, dumber, unorganized, brightest, personable, many, foreign, accomplished, crankier, centrist, self-righteous, comfy, controlled, socialist, g-d, give, diverse, fourth, electable, laxative, short, faster, third, super-primary, corporate</td>
<td>religious, pro-moon, slogan-free, personal, local, cultural, mercurial, marry, common, smaller, past, economic, unmistakable, racial, current, fifth, sixth, anti-newt, military, other, political, final, representative, former, politician, harder, uh, more, topic, anti-semitic, comedian, libertarian, rancid, anti-american, physical, comprehensive, funnier, lower</td>
</tr>
</tbody>
</table>

Table 3: Learned adjectives

**4 Evaluation**

A simple classification method serves as a first baseline for comparison. The polarity class of each node is based on the number of adjacent adjectives for which we can infer polarity from the general opinion lexicon (seed list). Each node is assigned the most frequent polarity class of this neighbourhood (positive on a draw).

The second baseline is a slight modification of the first one so that “but” conjunctions switch polarity to the opposing class. Figure 2 displays an example of graph segment consisting of adjective \( a_1 \) and its adjacent neighbourhood. From the lexicon, polarity of \( a_3 \) and \( a_4 \) can be inferred, the others are unknown. The first baseline method would classify \( a_1 \) as positive, because of uniform class distribution in the neighbourhood. Assuming that only the link between \( a_1 \) and \( a_4 \) is based on a “but”-conjunction, the second baseline would classify \( a_1 \) as negative.

Table 4 presents classification performances of the two baselines compared to the EM-Clustering algorithm. The second baseline achieves higher accuracy than the first one. This verifies the
assumption that “but”-conjunctions often exhibit opposing polarity. EM-Clustering clearly outperforms the two baseline methods.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Baseline</td>
<td>62.24%</td>
</tr>
<tr>
<td>2nd Baseline</td>
<td>66.33%</td>
</tr>
<tr>
<td>EM-Clustering</td>
<td>73.47%</td>
</tr>
</tbody>
</table>

Table 4: Classification accuracy

5 Conclusions and Future Work

This paper is a part of the more comprehensive work in relation to aspect-based opinion summarization on Twitter data in the domain of politics, which falls into the application category of social media monitoring. Although Twitter data can easily be gathered, special considerations in retrieval and pre-processing are needed. Aspect extraction presented some noun phrases as results, which were difficult to interpret without more contextual information. (More accurate aspects are expected with the use of more data collected over a longer period of time to achieve vocabulary convergence.) Following the idea of sentiment consistency, the general lexicon was augmented with domain-specific adjectives which were learned during semi-supervised clustering. This approach can easily be transferred to other domains.

Possibilities for future work include the learning of other domain-specific opinion words like nouns and verbs. Such a classification task would probably need to involve syntactic dependencies as features. Both time and regional distinctions could show trends and allow a more detailed presentation of political topics and associated sentiment. This could reveal that a certain topic causes positive reactions in one state, while it gets mostly negative comments in another state. In terms of aspect-level sentiment, the simple distance-weighted score presented here can be improved when it is assured that particular opinion words are expressed in relation to the aspect or the opinion target.

References

[15] nltk.org