Building sentiment Lexicons applying graph theory on information from three Norwegian thesauruses

Hugo Hammer[1], Aleksander Bai, Anis Yazidi and Paal Engelstad
Oslo and Akershus University College of Applied Sciences
Institute of Information Technology

Abstract

Sentiment lexicons are the most used tool to automatically predict sentiment in text. To the best of our knowledge, there exist no openly available sentiment lexicons for the Norwegian language. Thus in this paper we applied two different strategies to automatically generate sentiment lexicons for the Norwegian language. The first strategy used machine translation to translate an English sentiment lexicon to Norwegian and the other strategy used information from three different thesauruses to build several sentiment lexicons. The lexicons based on thesauruses were built using the Label propagation algorithm from graph theory. The lexicons were evaluated by classifying product and movie reviews. The results show satisfying classification performances. Different sentiment lexicons perform well on product and on movie reviews. Overall the lexicon based on machine translation performed the best, showing that linguistic resources in English can be translated to Norwegian without losing significant value.

1 Introduction

With the increasing amount of unstructured textual information available in internet, sentiment analysis and opinion mining have recently gained a groundswell of interest from the research community as well as among practitioners. In general terms, sentiment analysis attempts to automate the classification of text materials as either expressing positive sentiment or negative sentiment. Such classification is particularity interesting for making sense of huge amount of text information and extracting the "word of mouth" from product reviews, and political discussions etc.

Most of research in the field of sentiment analysis has been centered on the English language while little work has been reported for smaller languages. In this paper, we tackle the problem of building sentiment lexicon for the Norwegian language. The quality of a sentiment lexicon was assessed using annotated product reviews. In this perspective, we shall state that the main motivation behind our work is to build reliable sentiment lexicons for the Norwegian language. To the best of our knowledge, there is no publicly available sentiment lexicon for the Norwegian language. We hope that our work can facilitate the sentiment analysis in Norwegian and pave the work towards further research in the field. Another latent motivation in this article is to investigate the potential of generating lexicon in an automatic manner without any human intervention or refinement. We try to achieve this by increasing the sources of information, namely three different thesauruses, instead of solely relying on a single thesauruses as commonly done in the literature. In fact, we suggest that by increasing the number of thesauruses we can increase the quality of the generated lexicon.

[1]hugo.hammer@hioa.no

Possessing beforehand a sentiment lexicon is a key element in the task of applying sentimental analysis on a phrase or document level. Sentiment lexicon is merely composed of sentiment words and sentiment phrases (idioms) characterized by sentiment polarity, positive or negative, and by sentimental strength. For example, the word excellent has positive polarity and high strength whereas the word good is a positive having lower strength. Once a lexicon is built and in place, a range of different approaches can be deployed to classify the sentiment in a text as positive or negative. These approaches range from simply computing the difference between the sum of the scores of the positive lexicon and sum of the scores of the negative lexicon, and subsequently classifying the sentiment in the text according to the sign of the difference. More sophisticated approaches exist. The review of classification approaches are beyond the scope of this paper.

In order to generate a sentiment lexicon, the most obvious and naive approach involves manual generation. Nevertheless, the manual generation is tedious and time consuming rendering it an impractical task.

Due to the difficulty of manual generation, a significant amount of research has been dedicated to presenting approaches for automatically building sentiment lexicon. To alleviate the task of lexicon generation, the research community has suggested a myriad of semi-automatic schemes that falls mainly under two families: dictionary-based family and corpus-based family. Both families are semi-automatic because the underlying idea is to bootstrap the generation from a short list of words with polarity manually chosen (seed words), however they differ in the methodology for iteratively building the lexicon.

In this paper, we resort to a dictionary-based scheme for generating a sentiment lexicon for the Norwegian language. The corpus-based approach has also been followed in an other paper [2]. For a representative work on corpus based approaches, the informed reader can refer to [19, 20] for representative work. In order to put our work in the right perspective we shall review the most prominent dictionary-based approaches.

Background work

The underlying idea of a dictionary-based approach is to build the lexicon based on a dictionary containing synonyms and antonyms, and possibly hyponyms. A set of seed words are used to iteratively extend the list by resorting to the synonym and antonym structure. The intuitive idea behind the approach is that polarity of a sentiment word is preserved by a synonym relationship and inverted by an antonym relationship. Dictionary-based approaches were first introduced by Hu and Liu in their seminal work [11]. They stop the generation when no more words can be added to the list. Mohammad, Dunne and Dorr [15] used a rather subtle and elegant enhancement of Hu and Liu's work [10, 11] by exploiting the antonym-generating prefixes and suffixes in order to include more words in the lexicon. In [12], the authors constructed an undirected graph based on adjectives in WordNet [14] and define distance between two words as the shortest path in WordNet. The polarity of an adjective is then defined as the sign of the difference between its distance from the word "bad" and its distance from the word "good". While the strength of the sentiment depends on the later quantity as well as the distance between words "bad" and "good".

Blair and his colleagues [4] employs a novel bootstrapping idea in order to counter the effect of neutral words in lexicon generation and thus improve the quality of the lexicon. The idea is to bootstrap the generation with neutral seed in addition to a positive and a negative seed. The neutral seeds are used to avoid positive and negative sentiment propagation through neutral words. The later work uses a modified version of the label propagation algorithm proposed in [21]. The label propagation algorithms have an initial phase: where a score $+1$ is assigned to positive seed words and a score -1 to negative seed words, and 0 to the rest of words obtained through bootstrapping, then the transfer of score is performed from the seed to the rest of words. Note that the scores are updated in an iterative manner.

Rao and Ravichandran [18] used semi-supervised techniques based on two variants of the Mincut algorithm [5] in order to separate the positive words and negative words in the graph generated by means of bootstrapping. In simple words, Mincuts algorithms [5] are used in graph theory in order to partition a graph into two partitions minimizing the number of nodes possessing strong similarity being placed in different partitions. Rao and Ravichandran [18] employed only the synonym relationship as a similarity metric between two nodes in the graph. The results are encouraging and show some advantages of using Mincut over label propagation.

In [9], Hassan and Radev use elements from the theory of random walks of lexicon generation. The distance between two words in the graph is defined based on the notion of hitting time. The hitting time $h(i|S)$is the average number hops it takes for a node i to reach a node in the set S. A word w is classified as positive if $h(w|S_+) > h(w|S_-)$ and vice versa, where S_+ denotes the set of positive seed words, and S_- refers to the set of negative seed words.

Kim and Hovy [13] resorts to the Bayesian theory for assigning the most probable label (here polarity) of a bootstrapped word.

In [1], a quite sophisticated approach was devised. The main idea is to iteratively bootstrap using different sub-sets of the seed words. Then to count how many times a word was found using a positive sub-set seed and how many times the same word was found using a negative sub-set seed. The sentiment score is then normalized within the interval $[0,1]$ using fuzzy set theory.

In [17], a hybrid approach was proposed that combines both elements from corpus based approaches and dictionary based approaches. In fact, the bootstrapping is done based on a dictionary while the score assigning is based on corpus.

It is worth mentioning that another research direction for building lexicon for foreign language is based on exploiting the well developed English sentiment lexicon. A representative work of such approaches is reported in [8]. In [8], Hassan and his co-authors employ the hitting time based bootstrapping ideas introduced in [9] in order to devise a general approach for generating a lexicon for a foreign language based on the English lexicon.

2 Material and methods

In this section we describe how we created a set of sentiment lexicons from thesauruses (synonyms and antonyms) and how we evaluated them.

Linguistic resources

Sentiment lexicons are developed based on the information from three different Norwegian thesauruses

- Norske synonymer (Norwegian synonyms)
- Norsk ordbok (Norwegian dictionary)

- Din ordbok (Your dictionary)

The two first thesauruses were avaible from `ordnett.no` and the last from `dinordbok.com`. All thesauruses contained synonyms and Norsk ordbok also contained antonyms.

SCARRIE is a dictionary which, among other things, contains all forms of different words in Norwegian [6]. E.g. other forms of the word good are better and best. We used the dictionary to automatically expand developed sentiment lexicons to full-form.

We also generated a sentiment lexicon translating the well-known English sentiment lexicon AFINN [16] to Norwegian using machine translation (Google translate). We denote this sentiment lexicon AFINN in the rest of the paper. We generated a second lexicon bycorrecting several different errors from the machine translation using manual check. We denote the corrected sentiment lexicon AFINN_M in the rest of the paper and is considered our gold standard sentiment lexicon.

We tested the quality of the created sentiment lexicons using 13296 product reviews from the Norwegian online shopping site `komplett.no` and 4149 movie reviews from `filmweb.no`. Each product review contained a rating from one to five, five being the best and the movie reviews a rating from one to six, six being the best. We assume the ability a sentiment lexicon has to predict the ratings of reviews is a measure of the quality of the sentiment lexicon. The reviews were collected using web crawling.

Building a sentiment word graph from thesauruses

We built a large undirected graph of synonym and antonym relations between words from the three thesauruses. The words were nodes in the graph and synonym and antonym relations were edges.

We started with a set of set of 109 seed words (51 positive and 57 negative). The words were manually selected based on the following criterias

- The words should be used frequently in the Norwegian language
- The words should have the same sentiment independent of context. E.g. the word 'long' is not a suitable seed word since the sentiment can be negative ("waiting for a long time") or positive ("the battery life is long")
- The words should span different dimensions of positivity (happy, clever, intelligent, love etc) and negativity (lazy, aggressive, hopeless, chaotic etc)

We used web crawling to crawl each thesaurus to depth three for each seed word (synonyms of seed words, synonyms of synonyms and synonyms of synonyms of synonyms). All the information was included in one graph using the MultiGraph tool in the NetworkX package in Python. The full graph consisted of a total of 6036 nodes (words) and 16475 edges.

Generating sentiment lexicons from word graph

We generated sentiment lexicons from the word graph using the Label Propagation algorithm [13]. The Label Propagation algorithm initial phase consists of giving each positive and negative seed a word score 1 and -1, respectively. All other nodes in the graph are given score 0. The algorithm propagates through each non-seed words updating the score using a weighted average of the scores of all neighbouring nodes (connected with an edge). When computing the weighted average, synonym and antonym edges are

Rating	1	2	3	4	5
Average sentiment score	-0.23	-0.06	0.04	0.13	0.24

Table 1: Average computed sentiment score for reviews with different ratings.

given weights 1 and -1, respectively. The algorithm is iterated to changes in score is below some threshold for all nodes. The resulting score for each node is our sentiment lexicon.

We generated a sentiment lexicon for the full graph and for 14 subgraphs. We denode the full graph ALL_D3, since all thesauruses to depth 3 were used. We also used all thesauruses to depth 1 and 2, denoded ALL_D1 and ALL_D2, respectively. Further we generate sentiment lexicons using only one thesaurus at the time. We denote the thesaurus Norske synonymer with SYS, Din ordbok with DO and synonyms and antonyms from Norsk ordbok with NOS and NOA. E.g. a sentiment lexicon generated from Din ordbok crawled to depth two is denoted DO_D2. We wanted to evaluate if depth is related to the quality of the sentiment lexicons. If we crawl deep, we get many words which potentially are good since many of the words in the text is taken in to consideration when sentiment is computed. On the other hand, we also expect that crawling deep introduces noise in the sence that words far from the seed words is less likely to be reliable sentiment words. Building a good sentiment lexicon therefore is a trade-off between the number of words and the quality of the words.

We also use SCARRIE to generate full-form sentiment lexicons. This is computed giving each form of the word the same sentiment score. This is not necessarily an optimal strategy. For example best (superlative of good) normally have a stronger positive sentiment than better which is used for comparisons. A full-form version of a lexicon is denoted using an _S, e.g the full-form of DO_D2 is denoted DO_D2_S.

Finally we also consider the seed words as a sentiment lexicon, denoted SEED.

Evaluation of sentiment lexicons

We evaluated the quality of the sentiment lexicons measuring the classification performance of the `komplett.no` and `filmweb.no` reviews.

We computed the sentiment score of a review by simply adding the score of each sentiment word in a sentiment lexicon together, which is the most common way to to it [3]. If the sentiment shifter 'ikke' (not) was one or two words in front of a sentiment word, sentiment score was switched. E.g. 'glad' (happy) is given sentient score 0.8, while 'ikke glad' (not happy) is given score -0.8. Finally the sum is divided by the number of words in the review, giving us the final sentiment score for the review. We also considered other sentiment shifter, e.g. 'aldri' (never), and other distances between sentiment word and shifter, but our approach seem to be the best for such lexicon approaches in Norwegian [7].

Classification method

We divided the reviews in two parts, one part being training data and the other part for testing. We used the training data to estimate the average sentiment score of all reviews related to the different ratings. The computed scores could look like Table 1. We classified a review from the test set using the sentiment lexicon to compute a sentiment score for the test review and classify to the closest average sentiment score from the training set. E.g. if the computed sentiment score for the test review was -0.05 and estimated averages were

Rating	1	2	3	4	5
Average sentiment score	−0.23	−0.06	0.18	0.10	0.24

Table 2: Example were sentiment score were not monotonically increasing with rating.

as given in Table 1, the review was classified to rating 2. In some rare cases the estimated average sentiment score were not monotonically increasing with rating. Table 2 shows an example where the average for rating 3, is higher than for the rating 4. For such cases, the average of the two sentiment scores were computed, $(0.10+0.18)/2 = 0.14$, and classified to 3 or 4 if the computed sentiment score of the test review was below or above 0.14, respectively.

Classification performance

We evaluated the classification performance using average difference in absolute value between the true and predicted rating for each review in the test set

$$\text{Average abs. error} = \frac{1}{n}\sum_{i=1}^{n} |p_i - r_i| \tag{1}$$

where n is the number off reviews in the test set and p_i and r_i is the predicted and true rating of review i in the test set. Naturally a small average absolute error showing that the sentiment lexicon performed well.

Note that the focus in this article is not to do a best possible classification performance based on the training material. If that was our goal, other more advanced and sophisticated techniques would be used, such as machine learning based techniques. Our goal is rather to evaluate and compare the performance of sentiment lexicons and the framework described above is chosen with respect to that.

3 Results

This section presents the results of classification performance on product and movie reviews for different sentiment lexicons. The results are shown in Tables 3 and 4. Training and test sets were created by randomly adding an equal amount of reviews to both sets. All sentiment lexicons were trained and tested on the same training and test sets, making comparisons easier. This procedure were also repeated several times and every time the results were in practice identical to the results in Tables 3 and 4, documenting that the results are independent of which reviews that were added to the training and test sets.

For the product reviews (Table 3) we observe large variations in classification performance among the sentiment lexicons with average classification errors in absolute value ranging from below 1.0 rating points to above 2.5 point. We do not observe any clear relation between the number of sentiment words in the sentiment lexicons and the classification performance. An interesting observation is that the best three sentiment lexicons all contains 400 or less words showing that good classification performances are possible using only small amount of words. Statistically these three sentiment lexicons perform significantly better than the gold standard (AFINN_M). Recall that full-form sentiment lexicons were created using SCARRIE and that the names of the full-form sentiment lexicon ending with _S. The results do not show any systematic improvements in classification performance for the full–form sentiment lexicons compared to the original.

Sentiment lexicon	N	Mean (Stdev)	95% conf.int.
NOS_D1	105	0.91 (1.34)	(0.88, 0.95)
NOS_D2	264	0.98 (1.38)	(0.94, 1.01)
NOS_D1_S	400	0.99 (1.4)	(0.96, 1.02)
NOS_D2_S	951	1.11 (1.48)	(1.07, 1.14)
AFINN M	2161	1.35 (1.39)	(1.31, 1.38)
AFINN_M _S	7027	1.39 (1.41)	(1.36, 1.42)
AFINN	2260	1.39 (1.32)	(1.36, 1.43)
AFINN_S	7554	1.4 (1.37)	(1.36, 1.43)
SYN_D3	4483	1.54 (1.41)	(1.51, 1.58)
DO_D2_S	5227	1.58 (1.4)	(1.55, 1.61)
NOS_D3	562	1.59 (1.44)	(1.55, 1.62)
DO_D2	1470	1.61 (1.39)	(1.58, 1.64)
SYN_D2_S	8343	1.62 (1.48)	(1.59, 1.66)
ALL_D1_S	5023	1.63 (1.43)	(1.6, 1.67)
ALL_D2_S	11729	1.65 (1.48)	(1.62, 1.69)
NOS_D3_S	2108	1.67 (1.6)	(1.63, 1.7)
ALL_D2	3497	1.67 (1.49)	(1.63, 1.7)
SYN_D3_S	15818	1.69 (1.56)	(1.65, 1.72)
SYN_D1_S	2706	1.7 (1.46)	(1.66, 1.73)
SEEDS_S	410	1.7 (0.88)	(1.68, 1.72)
SYN_D2	2430	1.71 (1.51)	(1.67, 1.74)
ALL_D3	6036	1.73 (1.54)	(1.69, 1.76)
ALL_D3_S	20265	1.74 (1.54)	(1.7, 1.77)
DO_D1_S	2906	1.76 (1.39)	(1.72, 1.79)
SYN_D1	782	1.8 (1.54)	(1.76, 1.84)
NOA_D3_S	2361	1.81 (1.53)	(1.77, 1.84)
NOA_D3	576	1.82 (1.52)	(1.78, 1.85)
NOA_D2_S	1916	1.86 (1.53)	(1.83, 1.9)
ALL_D1	1466	1.87 (1.6)	(1.83, 1.91)
NOA_D2	467	1.87 (1.47)	(1.83, 1.9)
DO_D3_S	8329	1.96 (1.63)	(1.92, 2)
DO_D3	2402	1.97 (1.61)	(1.93, 2.01)
NOA_D1	262	2.06 (1.62)	(2.02, 2.1)
NOA_D1_S	1084	2.07 (1.62)	(2.03, 2.11)
DO_D1	808	2.12 (1.5)	(2.08, 2.15)
SEEDS	108	2.32 (1.22)	(2.29, 2.35)

Table 3: Classification performance for sentiment lexicons on `komplett.no` product reviews. The columns from left to right show the sentiment lexicon names, the number of words in the sentiment lexicons, mean absolute error with standard deviaton and 95% confidence intervals for mean absolute error.

Sentiment lexicon	N	Mean (Stdev)	95% conf.int.
DO_D3_S	8329	1.86 (1.07)	(1.81, 1.91)
SEEDS_S	410	1.86 (1.12)	(1.81, 1.91)
AFINN_M_S	7554	1.86 (1.09)	(1.81, 1.9)
ALL_D3_S	20265	1.86 (1.1)	(1.81, 1.91)
AFINN_M	2260	1.87 (1.13)	(1.82, 1.92)
ALL_D2_S	11729	1.9 (1.1)	(1.85, 1.95)
SEEDS	108	1.9 (1.08)	(1.86, 1.95)
ALL_D1_S	5023	1.93 (1.11)	(1.88, 1.97)
ALL_D1	1466	1.93 (1.14)	(1.88, 1.98)
NOA_D3_S	2361	1.96 (1.13)	(1.91, 2.01)
DO_D2_S	5227	1.97 (1.1)	(1.92, 2.01)
NOA_D2_S	1916	1.98 (1.11)	(1.93, 2.03)
AFINN	2161	1.98 (1.15)	(1.93, 2.03)
AFINN_S	7027	1.98 (1.12)	(1.94, 2.03)
ALL_D2	3497	1.98 (1.09)	(1.94, 2.03)
NOS_D3_S	2108	1.99 (1.14)	(1.94, 2.04)
NOA_D1_S	1084	2 (1.12)	(1.95, 2.05)
DO_D1_S	2906	2 (1.08)	(1.96, 2.05)
SYN_D1_S	2706	2 (1.15)	(1.95, 2.05)
NOS_D3	562	2.01 (1.13)	(1.96, 2.05)
ALL_D3	6036	2.05 (1.13)	(2, 2.1)
SYN_D1	782	2.05 (1.17)	(2, 2.1)
DO_D2	1470	2.05 (1.12)	(2, 2.1)
DO_D1	808	2.05 (1.13)	(2, 2.1)
SYN_D2_S	8343	2.05 (1.15)	(2, 2.1)
NOS_D1_S	400	2.06 (1.09)	(2.01, 2.11)
SYN_D3_S	15818	2.06 (1.15)	(2.01, 2.11)
SYN_D2	2430	2.07 (1.16)	(2.02, 2.12)
NOA_D2	467	2.11 (1.16)	(2.06, 2.16)
DO_D3	2402	2.14 (1.12)	(2.09, 2.18)
NOS_D2_S	951	2.14 (1.1)	(2.1, 2.19)
NOA_D3	576	2.16 (1.19)	(2.11, 2.21)
NOS_D1	105	2.17 (1.1)	(2.13, 2.22)
NOS_D2	264	2.2 (1.15)	(2.15, 2.25)
SYN_D3	4483	2.21 (1.17)	(2.16, 2.26)
NOA_D1	262	2.22 (1.1)	(2.17, 2.26)

Table 4: Classification performance for sentiment lexicons on `filmweb.no` movie reviews. The columns from left to right show the sentiment lexicon names, the number of words in the sentiment lexicons, mean absolute error with standard deviaton and 95% confidence intervals for mean absolute error.

Comparing Tables 3 and 4, we see that the classification performance is poorer for movie reviews than for product reviews. It is known from the literature that sentiment analysis of movie reviews is normally harder than product reviews [3]. E.g. movie reviews typically contain a summary of the plot of the movie which could contain many negative sentiment words (sad movie), but still the movie can get an excellent rating. We see that some sentiment lexicons that performed well on product reviews perform poorly on product reviews and visa verca, e.g. NOS_D1 and NOS_D2. The lexicons with the best overall performance across the two review types seem to be the gold standard sentiment lexicons (AFINN_M and AFINN_M _S).

4 Closing remarks

In this paper we have created and evaluated a large set of sentiment lexicons in Norwegian. The lexicons were created by two approaches 1) running the Label propagation algorithm on information from three different Norwegian thesauruses (synonyms and antonyms) and 2) using machine translation to translate the well-known AFINN sentiment lexicon from English to Norwegian.

Overall the sentiment lexicons generated from Norwegian thesauruses using state-of-the-art graph methods do not outperform the automatically translated AFINN-list. This shows that machine translation of linguistic resources in English can be used successfully for the Norwegian language. Norwegian research can profit from the extensive amount of linguistic resources in English by transferring it to the Norwegian language without losing significant value.

That said, translation of English resources is not without challenges. The words with strongest sentiment are the most important in sentiment lexicons and translation of those words was especially difficult like for example motherfucker, bitch, fuck, son of a bitch etc. AFINN and AFINN_M suffer from this challenges.

Further research in developing linguistic resources in Norwegian is necessary to outperform the simple strategy of translating resources from English. We see our work as a first step towards this. It is known that automatic generation of sentiment lexicons using graph algorithms, as presented in this article, introduces noise in the sense that sentiment words are given wrong sentiment value or that the lexicons include words with no sentiment. Natural next steps in developing sentiment lexicons in Norwegian is to manually control and adjust the lexicons presented in this article, extract the best from different lexicons, test them on several different contexts etc.

The computed sentiment lexicons are openly available[2] for researchers and others. We hope this work and the available sentiment lexicons will initiate a growing activity among researchers to build reliable linguistic resources for the Norwegian language.

[2]The various sentiment lexicons can be found on GitHub under `https://github.com/aleksab/lexicon`

References

[1] Alina Andreevskaia and Sabine Bergler. Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *EACL*, volume 6, pages 209–215, 2006.

[2] Aleksander Bai, Hugo Lewi Hammer, Aniz Yazidi, and Paal Engelstad. Constructing sentiment lexicons in Norwegian from a large text corpus. In *The 17th IEEE International conference on Computational science and Engineering (submitted)*, 2014.

[3] Liu Bing. *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2011.

[4] Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, and Jeff Reynar. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*, page 14, 2008.

[5] Avrim Blum, John Lafferty, Mugizi Robert Rwebangira, and Rajashekar Reddy. Semi-supervised learning using randomized mincuts. In *Proceedings of the twenty-first international conference on Machine learning*, page 13. ACM, 2004.

[6] Koenraad de Smedt and Victoria Rosen. Automatic proofreading for Norwegian: The challenges of lexical and grammatical variation. In *NODALIDA '99, 12th Nordiske datalingvistikkdager*, pages 206–215, 1999.

[7] Hugo Lewi Hammer, Per Erik Solberg, and Lilja Øvrelid. Sentiment classification of online political discussions: a comparison of a word-based and dependency-based method. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 90–96. Association for Computational Linguistics, 2014.

[8] Ahmed Hassan, Amjad Abu-Jbara, Rahul Jha, and Dragomir Radev. Identifying the semantic orientation of foreign words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 592–597. Association for Computational Linguistics, 2011.

[9] Ahmed Hassan and Dragomir Radev. Identifying text polarity using random walks. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 395–403. Association for Computational Linguistics, 2010.

[10] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168–177, 2004.

[11] Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *Proceedings of AAAI*, pages 755–760, 2004.

[12] Jaap Kamps, MJ Marx, Robert J Mokken, and Maarten De Rijke. Using wordnet to measure semantic orientations of adjectives. 2004.

[13] Soo-Min Kim and Eduard Hovy. Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 483–490. Association for Computational Linguistics, 2006.

[14] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[15] Saif Mohammad, Cody Dunne, and Bonnie Dorr. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 599–608. Association for Computational Linguistics, 2009.

[16] Finn Årup Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903, 2011.

[17] Wei Peng and Dae Hoon Park. Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. *Urbana*, 51:61801, 2004.

[18] Delip Rao and Deepak Ravichandran. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682. Association for Computational Linguistics, 2009.

[19] Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 417–424, 2002.

[20] Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[21] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.