Different Aggregation Strategies for Generically Contextualized Sentiment Lexicons

Stefan Gindl

Department of New Media Technology, MODUL University Vienna, Austria

Abstract. Sentiment detection has gained relevance in the last years due to the vast amount of publicly available opinion in the form of Web forums or blogs. Yet, it still suffers from the ambiguity of language, lowering the efficacy and accuracy of sentiment detection systems. Thus, it is important to also invoke context information to refine the initial values of sentiment terms. Moreover, domain-independence is desirable to avoid using a topic determination beforehand. This work investigates strategies for extracting non-generic features to be integrated into a so-called contextualized sentiment lexicon, capable of getting the context correctly and assigning sentiment terms the proper sentiment value. The proposed approach will be applied in an online-media aggregation and visualization portal, covering a vast number of news media sources.

1 Introduction

Sentiment detection handles affect expressed in written text, more exactly it tries to classify documents into positively, negatively or neutrally opinionated. The classification can either be coarse-grained (i.e. positive, negative, neutral) or fine-grained (i.e. strong-positive, weak-positive, etc.). The research area experienced a leap in relevance with the upcoming availability of online opinions in reviews, forums or blogs. Applications range from the political area (tracking a political campaign online) over the economic area (acceptance studies for new products or services) to the purely scientific application, helping to understand human language. Thus, sentiment detection can play a major role in Web mining systems. It also adds value to Social Web applications. Trend analyses on fast moving platforms such as www.twitter.com become possible; websites hosting images or videos (such as www.flickr.com or www.youtube.com) can be exploited to measure the affect of the community towards celebrities or popular technical devices.

Many approaches rely on so-called sentiment lexicons, containing terms assumed to express sentiment. Sentiment lexicons suffer from term ambiguity - one and the same term can have different meanings under different circumstances. Table 1 shows three sentence, where one and the same sentiment term can be used in positive and negative context. The intuitively negative term “repair” can be used positively, when a person is satisfied with his/her repaired car. “Unpredictable” applied to the movie genre refers to an exciting movie; on the other

hand, if the breaks of a car are unpredictable, this is normally something undesirable. Finally, the term "peace" will be express a positive fact in the most cases. Yet, it can also refer to a negative state, such as in the sentence "This peace is a lie".

Positive	Negative
The **repair** of my car was satisfying.	I had many complaints after my camera's **repair**.
This movie's plot is **unpredictable**.	The breaks of this car are **unpredictable**.
The long **peace** brought wealth and safety to the people.	This **peace** is a lie.

Table 1. Examples for sentiment terms occurring in positive and negative contexts.

This work examines possible refinement strategies of the already existing context-sensitive sentiment detection system described in [7]. It takes into account the context of a sentiment term, and, based on the context, refines the sentiment value of the term. Naïve Bayes as a simple, fast and yet powerful technique serves as the method to train the model. To overcome the effects of domain-specificity the approach also merges features of the trained models and creates a domain-independent model. In the presented paper refinement strategies for creating a domain-independent lexicon are discussed, together with a preliminary evaluation of the planned strategies.

Temporal Sentiment Analysis Applied to Online Media

The proposed system will be used for temporal sentiment detection in the so-called "Media Watch on Climate Change". This portal aggregates climate change related issues and provides efficient visualization means, such as a semantic map for related keywords with strong media coverage and an ontology map for relations among significant phrases.

The sentiment map in the upper left corner of Figure 1 allows for tracing the sentiment towards relevant topics. For example, the phrases "oil spill" and "gulf oil" receive clearly negative media attention, whereas the term "Hayward" received positive attention until May 10, which turns into negative afterwards. Such a tool, i.e. accurate sentiment detection combined with efficient visualization techniques, strongly supports research on relevant topics and offers a specialized view on the online world.

During the U.S. elections 2008 another portal website using a former version of the proposed appraoch traced media attention towards the presidential candidates. Figure 2 shows the main window of the portal, with the presidential candidates in the upper part, a list of used media sources in the middle and the

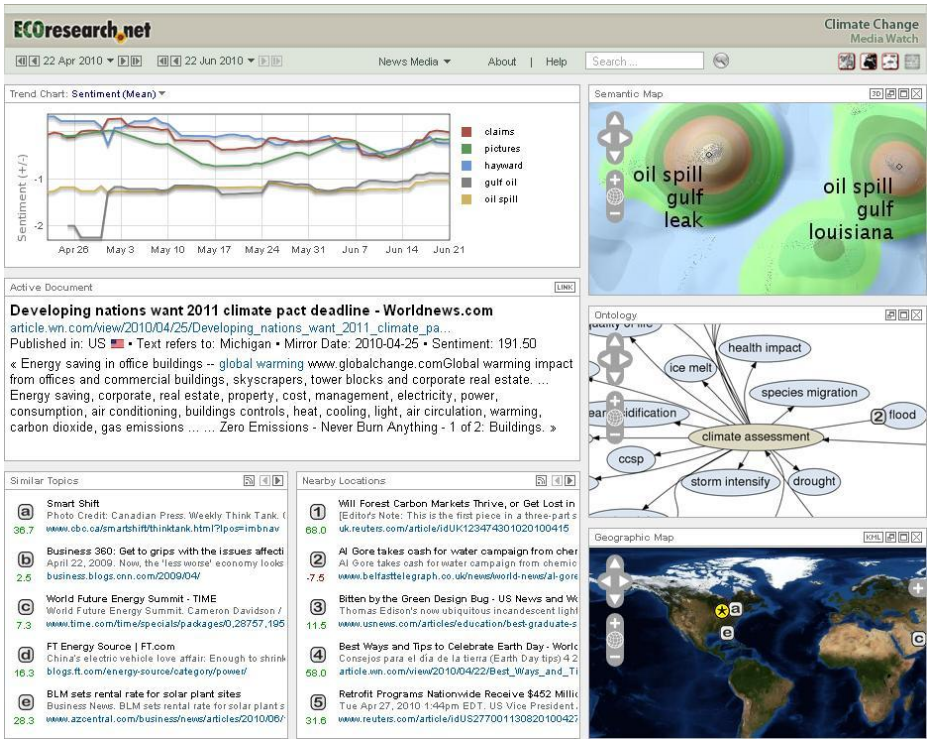

Fig. 1. The Media Watch on Climate Change, www.ecoresearch.net/climate/; see the sentiment map

sentiment map at the bottom. Such tools can complement or even replace traditional opinion surveys, and are a permanent source of feedback during a political campaign. Adapted to different application fields they can support enterpises to trace their reputation (e.g. in connection with the current oil spill in the Gulf of Mexico) or to measure the acceptance of a previously launched new product in the online community.

The paper is structured as follows: Section 2 summarizes existing work, Section 3 outlines the already existing approach and the refinement strategies. The evaluation follows in Section 4. Section 5 concludes the paper and contains an outlook on further work regarding the discussed refinement strategies.

2 Related Work

Sentiment detection as a research area dates back to the 1990s with the work of Wiebe [20] and Hatzivassiloglou and McKeown [9]. In [20] Wiebe started to identify subjective sentences, whereas Hatzivassiloglou and McKeown exploited syntactical relations to identify sentimental adjectives [9]. Turney and Littman

apply two different association measurements to identify new sentimental terms in [17]. In [13] Pang and Lee present a fine-grained approach to detect the exact sentiment (i.e. the star rating) of reviews using Support Vector Machines. Subrahmanian and Reforgiato base sentiment detection on a syntactical level by using adjective-verb-adjective combinations [16].

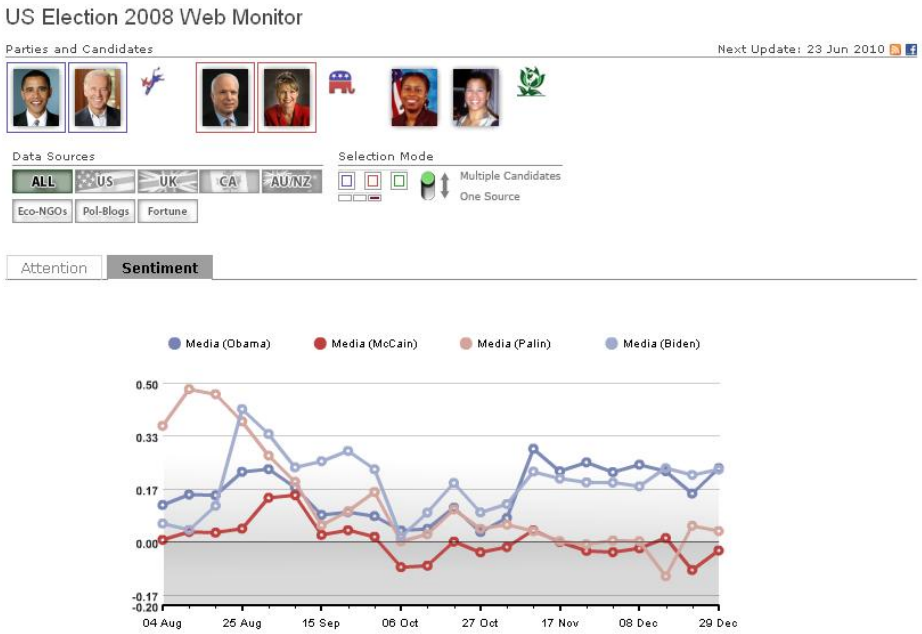

Fig. 2. The US Election 2008 Web Monitor, www.ecoresearch.net/election2008/; see the sentiment map

Some works also use context information to refine sentiment indicators. According to Nasukawa and Yi [12] sentiment detection is a three step process, where the identification of sentiment expressions is followed by the determination of their polarity and strength. The last step of the procedure identifies the subject the sentiment terms are related to. They model such relationships for verbs, which either directly transfer their own sentiment or another term's sentiment to the subject. With this model they are capable of treating expressions such as t_i *prevents trouble* [12]. The verb *prevents* passes the opposite sentiment of the term *trouble* to the target t_i. Sentence particles different from verbs directly transfer their sentiment to the subject. Kim and Hove [10] specify subjects with a Named-Entity-Recognition and assign them the overall sentiment value of the sentence. A list of 44 verbs and 34 adjectives expanded by WordNet [6] synonyms and antonyms serves as sentiment lexicon. To handle complex sentence structures such as "the California Supreme Court *disagreed* that the state's

new term-limit law was *unconstitutional*" [10] they developed a strategy, where several negative sentiment terms in one and the same sentence eliminate each other. Polanyi and Zaenen present a number of "contextual valence shifters" in their eponymous work [14]. Agarwal et al. propose syntactical capturing of context in [1]. Wilson et al. evaluate a large number of textual features, including context, in [21] on different machine learning algorithms; they use a two-stage process, firstly filtering neutral expressions from polar ones and afterwards disambiguating the sentiment of the polar expressions. In [22] they present a similar procedure with an expanded set of machine learners.

Turney and Littman [17] use Pointwise Mutual Information (PMI) and Latent Semantic Analysis (LSA) to identify sentiment terms in a large Web corpus. Terms with sufficient co-occurrence frequency with one of 14 paradigm terms (i.e. a gold standard list of seven positive and negative terms) are assigned the same sentiment value as the respective paradigm term. Evaluated on the General Inquirer [15] PMI shows results comparable with the algorithm of Hatzivassiloglou and McKeown [9]. Using three different extraction corpora and the sentiment lexicon of [9] Turney and Littman show that PMI does not outperform Hatzivassiloglou's and McKeown's algorithm but is more scalable [19]. LSA also provided better results, but was not as scalable as PMI too. In [18] Turney uses the same techniques to identify new sentiment terms from a paradigm list of only two terms (*excellent* and *poor*). This procedure performed well on the review corpus. Beineke et al. re-interpret the previously discussed mutual association as a Naïve Bayes approach [2]; they also expand this perspective (which is an unsupervised approach) and create a supervised approach using labeled data.

Lau et al. [11] prove the importance of context by applying three different language models, whereof one is an inferential language model sensible for context. According to their evaluation the inferential language model outperforms the other two models, emphasizing the importance of context. Bikel and Sorensen apply a simple feature selection together with a perceptron classifier to reviews from Amazon.com [3]. They use all tokens with an occurrence frequency higher than four and achieve an accuracy of 89% in their experiments. Denecke [4] applies a machine learning approach to multi-lingual sentiment detection using movie reviews from six different languages. Google Translator (www.google.com/language_tools) translates foreign-language documents into English. The feature selection procedure extracts a total of 77 features out of four superclasses [4]: (1) the frequency of word classes (i.e. the number of verbs, nouns, etc.), (2) polarity scores for the 20 most frequent words and the averages scores for all verbs, nouns and adjectives are calculated using SentiWordNet [5]; other features are (3) the frequency of positive and negative words according to the General Inquirer and (4) textual features such as the number of question marks. Using all features the Simple Logistic classifier of the WEKA tool[8] reaches exorbitantly good results when applied to native English documents. When applied to non-native, translated documents the results are still higher than the baseline demonstrating the efficacy of using a lexical resource such as SentiWordNet.

Our contextualization method is different from the presented context-aware approaches. For example, we do not use linguistic relations such as synonymy as Esuli and Sebastiani in [5]. Furthermore, we also do not transfer sentiment from sentiment terms to subjects as done in [12], nor do we filter polar from neutral expressions as or use predefined syntactical features [21, 22]. Instead, the proposed method considers the term's context based on discriminators identified in the text and adjusts its sentiment value accordingly.

3 Methodology

The work is based on [7] and can be roughly divided into three steps (also see Figure 3). The first step comprises the enrichment of an initial sentiment lexicon with contextual information. The initial lexicon is a lexicon based on sentimental terms from the General Inquirer [15]. We applied "reverse lemmatization" on these terms, which adds inflected forms to the initial terms. The second step is the application of the created contextualized sentiment lexicon on unknown documents, using the Naïve Bayes technique to recalculate the original sentiment values in the sentiment lexicon. The last step comprises the identification of context features applicable across the domains of the training corpora. This step results in the creation of a generic contextualized lexicon. We compare the improvement achieved with this approach using a lexical algorithm as our baseline. This algorithm sums up the sentiment values of all sentiment terms occurring in a document:

$$Sent(doc) = \sum_{i=1}^{n} Sent(t_i)$$

$$Sent(t_i) = \begin{cases} 1, \text{ if } t_i \text{ is a positive term} \\ -1, \text{ if } t_i \text{ is a negative term} \\ 0, \text{ if the term is neutral} \end{cases}$$

In case of a negation trigger preceding a sentiment term its value is multiplied by -1. In the following, we describe each of these steps in more detail:

Generation of the contextualized lexicon The system identifies ambiguous terms in the initial sentiment lexicon by analyzing their usage in a labeled training set. The training set consists of documents with positive and negative labels. A sentiment term with equally high frequency in both parts is considered to be an ambiguous term. All ambiguous terms identified with that process undergo a so-called "contextualization". This means, that the system identifies terms frequently co-occurring with the ambiguous term in positive/negative reviews (i.e. context terms). The contextualization creates a contextualized lexicon. This lexicon stores the probability that a certain ambiguous term in combination with certain context terms is normally used in positive/negative reviews.

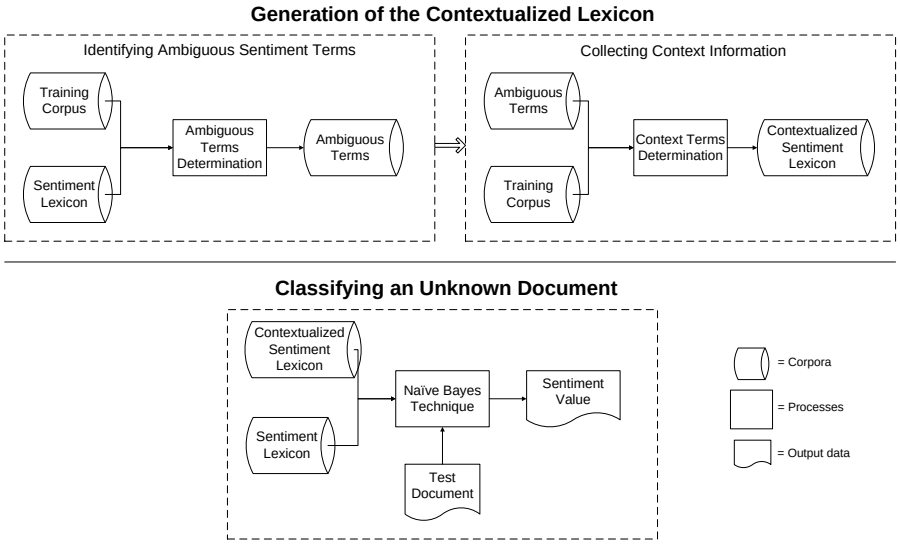

Fig. 3. Creation and application of a contextualized sentiment lexicon.

Application on unknown documents Each time a sentiment term occurs in a new document, the contextualized sentiment lexicon is consulted and decides, if the term is ambiguous. For non-ambiguous terms the lexicon returns the original sentiment value of the term. In case of an ambiguous term the system analyzes the context of the document. It uses the ten strongest context sentiment terms and calculates the probability of the ambiguous term being positive/negative given these ten context terms.

The system calculates an ambiguous term's sentiment given context $\mathbf{c}$ using the Naïve Bayes formula (c_i is a single context term):

$$p(Sent^+|\mathbf{c}) = \frac{p(Sent^+) \cdot \prod_{i=1}^{n} p(c_i|Sent^+)}{\prod_{i=1}^{n} p(c_i)}$$

The resulting value is the final sentiment value of the ambiguous term. Finally, the sentiment values of all sentiment terms (ambiguous and non-ambiguous) are summed up. The sum is the overall sentiment of the document.

Figure 4 shows an example of the context-sensitive sentiment detection. The system analyzes the document and finds the sentiment term "repair", which turns out to be ambiguous. So, it also analyzes the context, i.e. all other terms of the document. It identifies the three context terms "friendly", "quickly", and "reliable" as indicators for a positive meaning of "repair". Thus, the system assigns it a positive sentiment value and classifies the whole document as being positive. Note that the example is very simple - in reality a document usually contains more sentiment terms, both ambiguous and non-ambiguous.

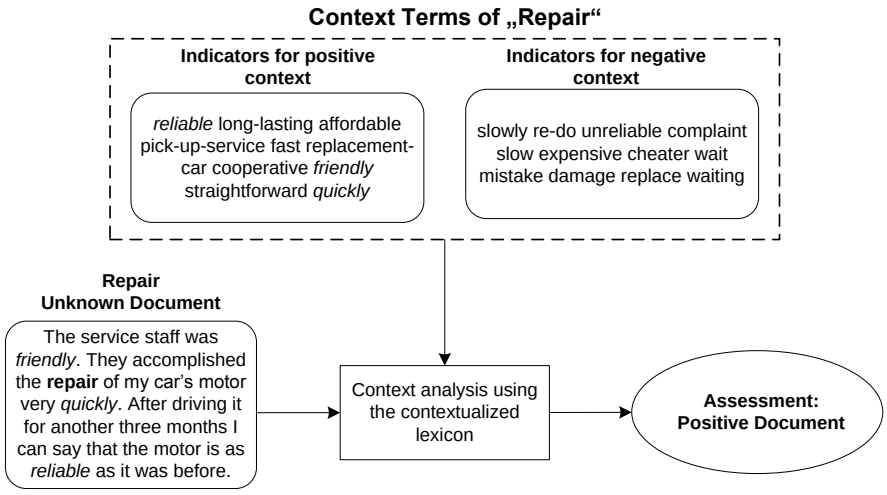

Fig. 4. Context invocation for the ambiguous term **repair** in an unknown document.

Identifying Generic Features Generic features are context terms which can be used across domains. Having obtained the contextualized lexicons from several training corpora the system distinguishes between three types of context term categories:

- **Helpful**: Using a helpful sentiment term improves the efficacy of sentiment detection.
- **Neutral**: These terms do not change the efficacy.
- **Harmful**: Harmful terms reduce the efficacy.

The categorization into helpful, neutral and harmful is accomplished as follows: if a review has been classified incorrectly by our baseline (i.e. the lexical algorithm explained at the beginning of this section), but correctly by the Naïve Bayes approach, the context terms of all ambiguous terms in this document are considered as helpful terms. If it has been correctly classified by the baseline but is incorrectly classified by Naïve Bayes all context terms are considered as harmful. Neutral context terms are those occurring in documents where Naïve Bayes and the baseline deliver the same classification. Using such a procedure means that a term helpful in document A can be neutral or even harmful in document B. A special exclusion strategy decides which of the harmful terms should be discared, and thus also their occurrences as helpful or neutral terms.

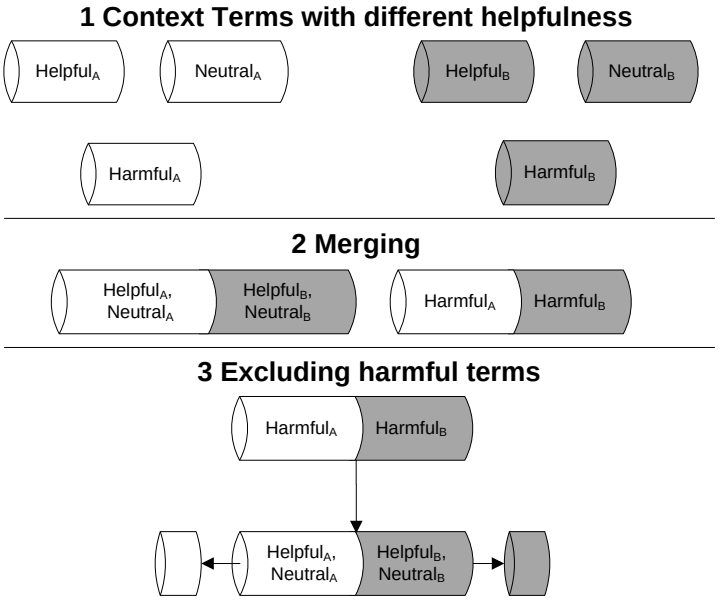

Fig. 5. Filtering harmful terms

4 Evaluation

We evaluated the contextualization refinements on the same corpora as in [7], which are a set of 2 500 products reviews from Amazon[1] and 1 800 holiday reviews from TripAdvisor[2] (which we call the "Amazon" and the "TripAdvisor" corpus later on). We accomplished a 10-fold cross-validation on both evaluation sets. A simple lexical approach serves as the baseline for the evaluation, summing up sentiment values of the sentiment terms occurring in the document to be classified. The sentiment values come from the initial lexicon described in Section 3.

We tested the following strategies for the exclusion of harmful terms:

- C_{all}: no harmful terms are excluded.
- $C \setminus H$: even terms with a single harmful occurrence are excluded.
- $C = \{c | \frac{F(c|\neg h)}{F(c|h)} > 5\}$: if a term has been helpful/neutral, but also has a harmful occurrence, its frequency in helpful/neutral cases must be five times higher than in harmful cases.
- $C = \{c | \frac{F(c|\neg h)}{F(c|h)} > 10\}$: if a term has been helpful/neutral, but also has a harmful occurrence, its frequency in helpful/neutral cases must be ten times higher than in harmful cases.

[1] amazon.com
[2] tripadvisor.com

– H: only terms with harmful occurrences are used.

In Table 2 we give the results (i.e. the F-measures) for all tested exclusion strategies. For each corpus we distinguish between positive and negative and list the F-measure for each type (indicated by $\oplus$ and $\ominus$). The evaluation shows that excluding harmful terms requires great care. Removing all terms with harmful occurrences ($C \setminus H$) gives worse results than leaving them untouched (C_{all}). Setting the ratio of non-harmful terms to harmful terms to high (i.e. > 10) gives the same results as keeping all harmful terms. Using only terms having harmful occurrences lowers the evaluation results strongly. Yet, the results are not low enough to judge them as completely useless. Finally, using a weaker ratio (i.e. > 5) delivers the best results.

		C_{all}	$C \setminus H$	$C = \{c \mid \frac{F(c\mid\neg h)}{F(c\mid h)} > 5\}$	$C = \{c \mid \frac{F(c\mid\neg h}{F(c\mid h} \} > 10$	H
Amazon	$\oplus$	0.68	0.68	**0.69**	0.68	0.58
	$\ominus$	0.74	0.73	**0.75**	0.74	0.72
TripAdvisor	$\oplus$	**0.84**	**0.84**	**0.84**	**0.84**	0.81
	$\ominus$	0.78	0.78	**0.79**	0.78	0.78

Table 2. F-Measures achieved with different exclusion strategies

5 Conclusion & Further Work

The evaluation showed that particular aggregation strategies improve the overall result for sentiment detection using contextualized lexicons. Their sole impact is not too large, but they should be regarded as an integral component of a battery of refinement strategies for generically contextualized sentiment detection.

Future work comprises the investigation on further, more potential aggregation strategies. Moreover, an investigation of the semantic and syntactical sentence structure will be accomplished. The idea is that certain sentence types might mislead sentiment detection. For example, sentences which are too short or too long, or are in another way distorted might be counterproductive for sentiment detection. If used anyways those sentences worsen classification results. Sentiment detection would benefit from a-priori filtering of these. Machine-learning methods can accomplish this task.

References

1. Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 24–32, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

2. Philip Beineke, Trevor Hastie, and Shivakumar Vaithyanathan. The sentimental factor: Improving review classification via human-provided information. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 263, Morristown, NJ, USA, 2004. Association for Computational Linguistics.
3. Daniel M. Bikel and Jeffrey Sorensen. If we want your opinion. In *ICSC 2007. International Conference on Semantic Computing*, pages 493–500, Irvine, CA, September 2007.
4. Kerstin Denecke. How to assess customer opinions beyond language barriers? In *Third International Conference on Digital Information Management*, pages 430–435. IEEE, November 2008.
5. Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 417–422, 2006.
6. C. Fellbaum. WordNet - An electronic lexical database. *Computational Linguistics*, 25(2):292–296, 1998.
7. Stefan Gindl, Albert Weichselbraun, and Arno Scharl. Cross-domain contextualization of sentiment lexicons. In *ECAI 2010: Proceedings of the 19th European Conference on Artificial Intelligence*, in press.
8. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18, 2009.
9. Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181, Morristown, NJ, USA, 1997. Association for Computational Linguistics.
10. Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 1367, Morristown, NJ, USA, 2004. Association for Computational Linguistics.
11. R.Y.K. Lau, C.L. Lai, and Yuefeng Li. Leveraging the web context for context-sensitive opinion mining. In *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*, pages 467–471, Aug. 2009.
12. Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77, New York, NY, USA, 2003. ACM.
13. Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124, Morristown, NJ, USA, 2005. Association for Computational Linguistics.
14. Livia Polanyi and Annie Zaenen. Contextual valence shifters. In *Computing Attitude and Affect in Text: Theory and Applications*, The Information Retrieval Series, 2006.
15. Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. *The General Inquirer: A computer approach to content analysis*. M.I.T. Press, Cambridge, Massachusetts, 1966.
16. V.S. Subrahmanian and Diego Reforgiato. AVA: Adjective-Verb-Adverb combinations for sentiment analysis. *Intelligent Systems, IEEE*, 23(4):43–50, July-August 2008.

17. P.D. Turney and M.L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical report, National Research Council, Institute for Information Technology, 2002.
18. Peter D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424, Morristown, NJ, USA, 2002. Association for Computational Linguistics.
19. Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346, 2003.
20. Janyce M. Wiebe. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287, 1994.
21. Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Morristown, NJ, USA, 2005. Association for Computational Linguistics.
22. Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, 2009.