Potential and Limitations of Commercial Sentiment Detection Tools

Mark Cieliebak, Oliver Dürr, and Fatih Uzdilli*

Zurich University of Applied Sciences
Winterthur, Switzerland
{ciel, dueo, uzdi}@zhaw.ch
*Author names in alphabetic order

Abstract. In this paper, we analyze the quality of several commercial tools for sentiment detection. All tools are tested on nearly 30,000 short texts from various sources, such as tweets, news, reviews etc. In addition to the quality analysis (measured by various metrics), we also investigate the effect of increasing text length on the performance. Finally, we show that combining all tools using machine learning techniques increases the overall performance significantly.

Keywords: Sentiment Detection, Opinion Mining, Machine Learning, Classification, Corpus Analytics

1 Introduction

How good is the state-of-the-art in sentiment detection? If you look at scientific literature, there exist numerous approaches to the topic and many of them have been proven in experiments to perform very well, both in precision and recall. For instance, basic text-based sentiment detection seems to be "solved", in the sense that precision and recall of current algorithms are typically above 80% [14, 22]. On the other hand, if one looks at real-world applications that use or include sentiment detection, the picture changes dramatically. In fact, there exist various blog posts on the web that state something like this: "More often than not, a positive comment will be classified as negative or vice-versa" [16]. Is there really such a large gap between research and real-life systems?

In this paper, we will tackle this question by evaluating the performance of several commercial sentiment detection tools. More precisely, we will explore how good existing tools perform on different sentence-based test corpora. This will allow us to identify the potential for improvements, and to indicate relevant directions for future research on sentiment detection. We then combine all tools using machine learning techniques (Random Forest) to unleash a hidden portion of the commercial landscape's potential.

2 Related Work

2.1 Sentiment Detection in General

For the purpose of this paper, "sentiment detection" means to find the polarity (positive, negative, or neutral) of a given text. The texts are single sentences or very short texts from a single source ("sentence-based"). This includes the special case of Twitter documents.

There exist several other types and tasks in the realm of sentiment detection, e.g. emotion detection (is a text emotional or not?), document-based sentiment detection, target-specific sentiment detection (e.g. for a product), or rating prediction, where the number of stars for product reviews is predicted from the text. For a good overview of sentiment detection and its variants in general, see e.g. [12], [22], or [15].

2.2 Comparison of Tools and Algorithms

We are not aware of any scientific study on commercial sentiment detection tools that tackles questions as presented in this paper. However, there exist several comparison studies on sentiment detection algorithms, which have a somewhat different focus. In the following, we briefly summarize some of these studies. On the one hand, there exist scientific survey papers that explore the abilities of different algorithmic approaches to sentiment detection. Padmaja et al. list the results of 19 sentiment analysis papers and categorize each approach to a machine learning algorithm. Typical accuracy of the approaches is about 80% [14]. Cui et al. analyze the performance of different machine learning algorithms on a large test set of product reviews for predicting the number of "stars". Precision, recall and F1 score are above 85% for most algorithms they tested, reaching up to 90% [6]. Annett et al. compare basic sentiment analysis techniques on movie blog entries. They show that lexical methods are 50-60% accurate, while machine learning approaches are between 66 and 77 percent [1]. On the other hand, there are several comparisons of sentiment detection tools that focus on business needs. These studies are mostly done by companies or agencies, targeted for the non-scientific reader, and aim at guiding users to select appropriate tools. For instance, Bitext.com compares 10 sentiment APIs, using a negative sentence, a comparative sentence and a conditional sentence. They conclude that most of the APIs have problems with polarity modifiers or intensifiers and conditional sentences. Also they argue that most APIs do not show multiple opinions found in some sentences [4]. Hawskey analyzes the performance of two sentiment APIs using only tweets. The precision for polar text is around 20% [9].

Sentiment detection is an integral part of social media monitoring tools. For this reason, comparisons of social media monitoring tools typically also explore their sentiment detection abilities. Freshnetworks.com's comparison of 7 social media monitoring tools show that on average they coded positive and negative sentiment correctly for about 30% of the texts [8]. Toptenreviews.com provides a ranking of social media monitoring tools by different aspects, including sentiment analysis [21]. Sponder compares social media monitoring tools on sentiment analysis features [19].

Finally, Kmetz describes how to evaluate sentiment analysis, and presents advice for choosing a sentiment analysis tool for analyzing social media content [11].

3 Experimental Setup

Our basic question in this experiment is simple: How good are commercial sentiment detection tools? To answer this question, we evaluated the quality and performance of nine commercial sentiment detection tools on a test set of annotated texts. The texts were from different media sources (news, reviews, twitter etc.); however, no context information about the texts was provided to the tools during the evaluation. We implemented a uniform evaluation framework to submit all documents to the tools' API and evaluate the responses automatically.

3.1 Test Data

For the evaluation, we searched for publicly available test corpora that contained annotated short texts from different media sources. We found 7 appropriate corpora, which contained in total 28653 texts. Most of these corpora have already been used in other research and experiments. Each text is either a complete short document, or a single sentence. We used the annotations provided by the corpora to classify each text as "positive", "negative", or "other" (e.g. for neutral or mixed sentiment). For more details on test corpora, see Table 1.

Corpus Name	Text Type	# of Texts	Polar Text Ratio			Average Word Count	Reference
			pos	neg	oth		
DAI_tweets	Tweets	4093	19%	13%	67%	14	[13]
JRC_quotations	Speech Quotations	1290	15%	18%	67%	30	[2]
TAC_reviews	Product Review Sentences	2689	34%	49%	17%	20	[20]
SEM_headlines	News Headlines	1250	14%	25%	61%	6	[17]
HUL_reviews	Product Review Sentences	3945	27%	16%	57%	18	[10]
DIL_reviews	Product Review Sentences	4275	31%	18%	51%	16	[7]
MPQ_news	News Sentences	11111	14%	30%	55%	23	[23]

Table 1. Test Corpora

Technical Remarks: Sizes of corpora might differ slightly from their original sizes, since we skipped some texts in our evaluation, where no proper sentiment annotation was available. As DAI_tweets and JRC_quotations provided several annotations per text we used only those texts where all annotations were identical. For TAC_reviews,

categories MIX (for "mixed sentiment") and NEU (for "neutral sentiment") were merged and texts with category NR (for "not relevant") were not used. SEM_headlines uses numeric annotations. In accordance with its documentation, we used positive sentiment for texts with value >= 50, other for values from -49 to 49, and negative for values <= -50. HUL_reviews, DIL_reviews and MPQ_news annotate features and chunks within a text; we aggregated these annotations as follows: if there were only positive annotations in a text, the entire text was labeled positive; analogously, texts with only negative annotations were labeled negative; all other texts were labeled other.

3.2 Tools

For the evaluation, we used commercial state-of-the-art tools for automatic sentiment detection. There exist literally hundreds of such tools. In order to obtain comparable results, the tools had to fulfill the following criteria: stand-alone sentiment detection tool (i.e., not part of a larger system, such as social media monitoring systems); ability to analyze arbitrary texts (i.e., not specialized on single text types like tweets); API access; free-of-charge access for the purpose of this evaluation. Based on these criteria, we selected nine tools[1], as shown in Table 2.

Tool	Short Name	URL
AlchemyAPI	alc	www.alchemyapi.com
Lymbix	lym	www.lymbix.com
ML Analyzer	mla	www.mashape.com/mlanalyzer/ml-analyzer
Repustate	rep	www.repustate.com
Semantria	sma	www.semantria.com
Sentigem	sen	www.sentigem.com
Skyttle	sky	www.skyttle.com
Textalytics	tex	core.textalytics.com
Text-processing	txp	www.text-processing.com

Table 2. Tools

Technical Remarks: Repustate returns values between -1 and 1, indicating negative to positive sentiment. We asked the tool provider for appropriate threshold values and used thresholds -0.05 and 0.05 to separate negative, other, and positive sentiment, respectively. Skyttle returns categories POS and NEG for chunks within the text. We aggregated these data to entire texts as follows: if there were only positive chunks in the text, result was "positive"; if it was only negative chunks, result was "negative"; in all other cases, result was "other" (similar to adaption of corpus annotations).

[1] We also had access to webknox.com, which we had to remove from our test because it only provides positive and negative classes, and this did not fit our experimental setup.

4 Results

Table 3 summarizes the results per corpus. This table and all raw data are also available at www.zhaw.ch/~ciel/sentiment.

	DAI	JRC	TAC	SEM	HUL	DIL	MPQ
Number of Texts	4093	1290	2689	1250	3945	4275	11111
Text Type	tweet	quotation	sentence	headline	sentence	sentence	sentence
Ratio of Positive Text	19%	15%	34%	14%	27%	31%	14%
Ratio of Negative Text	13%	18%	49%	25%	16%	18%	30%
Ratio of Other Text	67%	67%	17%	61%	57%	51%	55%
Average Accuracy	0.63	0.47	0.43	0.56	0.53	0.51	0.51
Maximum Accuracy	0.76	0.62	0.52	0.61	0.60	0.59	0.59
Average F1 Score	0.57	0.39	0.39	0.46	0.49	0.47	0.44
Average Precision: Pos	0.44	0.24	0.52	0.33	0.48	0.51	0.30
Average Precision: Neg	0.51	0.30	0.69	0.43	0.35	0.36	0.51
Average Precision: Oth	0.82	0.75	0.14	0.67	0.70	0.62	0.66
Average Recall: Pos	0.65	0.52	0.55	0.40	0.67	0.59	0.46
Average Recall: Neg	0.53	0.35	0.37	0.47	0.40	0.38	0.43
Average Recall: Oth	0.65	0.48	0.34	0.63	0.51	0.51	0.57
Average F1 Score: Pos	0.51	0.31	0.52	0.34	0.54	0.53	0.33
Average F1 Score: Neg	0.50	0.31	0.47	0.42	0.35	0.35	0.43
Average F1 Score: Oth	0.71	0.55	0.19	0.63	0.57	0.54	0.57

Table 3. Summary of Main Results

Remarks: Some tools skipped some of the sentences, due to too long requests (mla, sma, lym), wrong language (alc), or other errors (lym). tex says on 2% of all texts that they have no polarity (handled as skips).

5 Key Findings

5.1 Tools are Wrong for Almost 50% of All Documents

We found that average accuracy of all tools on all documents is 54%. This means that if you pick a random tool and submit any of the documents, you have to expect a wrong result for almost every second document.

Of course, there are tools that have better average accuracy. But even the tool with maximum accuracy over all documents, sky, achieves only an accuracy of 60%. Hence, even with this tool, 4 out of 10 documents will be classified wrong.

It is very likely that commercial classifiers have not been trained with the test corpora we used. If they were, the accuracy figures could potentially be much different and even match the accuracies reported in scientific literature.

5.2 Tweets are Easier than All Other Text Types

Figure 1 shows that commercial tools can achieve maximum accuracy for tweets (corpus DAI_tweets). Here, the best tools achieve an accuracy of 76%. For all other text types, best accuracy is approx. 60% or even lower.

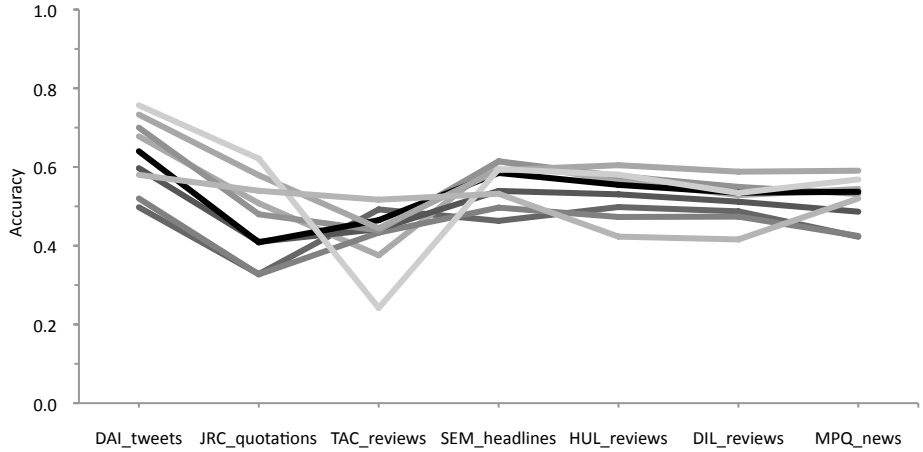

Fig. 1. Accuracy of All Tools on Test Corpora. Lines corresponding to tools from top to bottom for corpus DAI_tweets: tex, sky, sma, lym, sen, rep, txp, mla, alc

5.3 Longer Texts are Hard to Classify

How is sentiment detection performance affected by text-length? To answer that question we first have to define what we understand by "performance". Since the focus of this study is more on general trends than on the individual performance of the tools, we measure performance p as number of tools (0-9) classifying a given text correctly. We found that p can be modeled by linear regression using $p = a*x + b$, with x being the square-root of the text length (data not shown). In Figure 2 we display the slope a for all corpora. A positive value of a indicates that performance increases with increasing text length.

We observe a slope $a < 0$ for All Texts (dotted line), thus, longer texts are in general harder to classify. However, this effect is governed by texts with "other" sentiment: For all corpora, performance to detect "other" sentiment is negatively affected by the text-length. For texts with positive or negative sentiment, we find both slightly increasing and decreasing performances for longer texts. Only exception is corpus SEM_headlines, where we find a strong increase of performance for longer texts. The later might be due to the fact that headlines are very short texts (typically between 4-8 words), and longer texts give better indications on its sentiment.

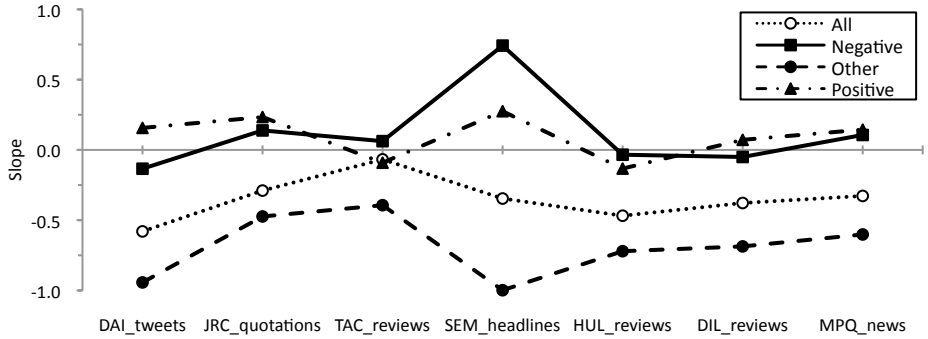

Fig. 2. Impact of Increasing Text Length on Analysis Performance. Shown is the slope of a linear model fitted into a performance vs. text length mapping (for details see main text). Negative values indicate a decrease of performance for longer texts, positive values indicate an increase of performance.

5.4 Corpus Annotations Might be Erroneous

In NLP research, one usually uses annotations of test corpora as "gold standard", in the sense that they provide a ground truth about the texts. Whenever a tool differs from this annotation, it is wrong. But our results imply that a non-negligible fraction of annotations might be wrong: for 9.2% of all texts, at least 7 of the tools agree on its tonality, but the corpus annotation is different (see Table 4). That is, 7 or more out of nine tools think a text is, say, positive, but the annotation is negative or other. For one corpus, this value reaches up to 15%.

	Disagree by >= 4 Tools	Disagree by >= 5 Tools	Disagree by >= 6 Tools	Disagree by >= 7 Tools	Disagree by >= 8 Tools	Disagree by >= 9 Tools
DAI_tweets	0.35	0.21	0.12	0.06	0.02	0.005
JRC_quotations	0.56	0.35	0.22	0.10	0.04	0.007
TAC_reviews	0.57	0.40	0.26	0.15	0.07	0.022
SEM_headlines	0.48	0.33	0.21	0.12	0.06	0.023
HUL_reviews	0.45	0.32	0.21	0.13	0.07	0.018
DIL_reviews	0.47	0.34	0.22	0.13	0.06	0.011
MPQ_news	0.50	0.29	0.14	0.06	0.02	0.003
ALL Texts	0.48	0.31	0.18	0.09	0.04	0.010

Table 4. Uniform Disagreement of Tools with Corpus Annotations.
Each column "Disagree by >= k Tools" shows proportion of texts in a corpus, for which at least *k* tools output the same sentiment classification for a text, and this classification differs from corpus annotation for this text.

Of course, it would be possible that all these tools are wrong; but manual inspection of sample texts showed that we - the authors - would often agree with the tools. Hence, there is a good chance that the annotations in the test corpora are erroneous.

One explanation might be that good corpus annotations are not easy to obtain: It is a well-known fact that human agreement on sentiment is far from perfect [24, 3]. Moreover, not all human annotators are equally qualified: Snow et al. have shown that it takes on average four non-expert annotators to achieve equivalent accuracy to one expert annotator [18].

It is out of scope of this paper to further investigate the reasons and implications of this issue in detail, nevertheless this will be an interesting and important research question.

For the purpose of this paper, we use the corpus annotations "as-is", since their impact on our findings is only marginal, some measurements might need to be adapted slightly due to errors in the corpora; however, our main results on quality of commercial sentiment analysis tools will remain unchanged.

6 Combined Forces

Our results above show that many tools perform reasonably well on most of the corpora. But there is no tool that excels on all corpora. Even more important, maximum accuracy is only about 75% even for the best tools, which is far from perfect. But what if we combine the tools, to build a "meta-tool"? Will we get better results? We explore this idea next and analyze the potential of two different approaches.

6.1 Majority Classifier

Our first approach is a majority classifier: each input document is submitted to all nine tools for analysis. Each tool returns a vote for "positive", "negative", or "other". These votes are collected, and the sentiment that received most votes is chosen. If several sentiments with equal high number of votes exist, one of those sentiments is picked randomly.

6.2 Random Forest Classifier

A more advanced approach to predict the sentiment given the votes of the tools is to use a random forest classifier [5]. More precisely, we use the random forest implementation of the R-package randomForest with default settings. For each corpus, we train the classifier using the votes (negative, other, positive) as the numerical values (-1, 0, 1), respectively. In Figure 3, accuracy is reported as usual as one minus the out-of-bag error.

6.3 Result: Random Forest >> Best Single Tool ≈ Majority

Figure 3 shows the accuracy of both two meta-classifiers on all corpora. For comparison, we included average accuracy of all tools and the best classifier for each corpus in this figure.

The majority classifier outperforms the average of all tools. On the other hand, the best single tool for a corpus is always better than the majority classifier. Thus, if the type of a new document (tweet, review etc.) is known, the best single tool for this document type should be used; but if document type is unknown, the majority classifier could be used, which yields superior results in this case.

On the other hand, Figure 3 shows that the random forest classifier yields the best result of all tested classifiers. In fact, it is up to 9 percent better than even the best single tool for a corpus. This increase of the accuracy shows that there is still room for improvement of the existing tools.

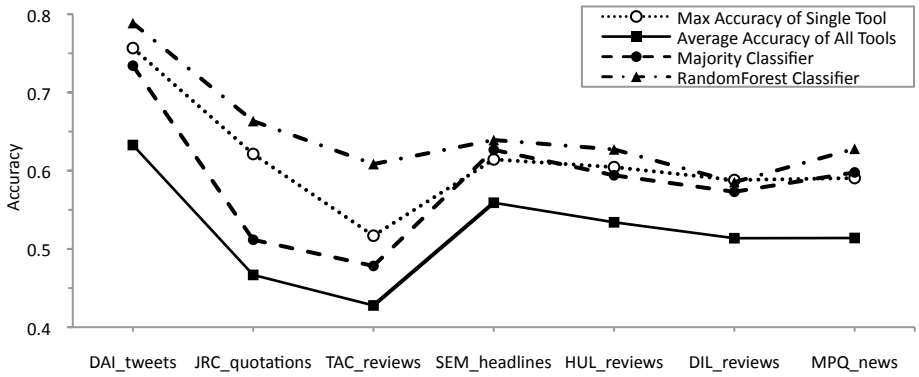

Fig. 3. Accuracies for Tools and Meta-Classifiers, per Corpus. Note that the vertical axis does not start with 0.0.

7 Summary and Future Challenges

In this work, we evaluated the quality of 9 state-of-the-art commercial sentiment detection tools for approx. 30,000 different short texts (tweets, news headlines, reviews etc.). The best tools have an accuracy of 75% for some document types (tweets), but the average accuracy over all documents is at best 60%. Surprisingly, the accuracy decreases if texts get longer, which is due to the decline in the ability to detect "other" sentiments. As an aside, we observed that existing sentiment corpora are prone to error, with error rates up to 15% per corpus.

Combining all tools with a meta-classifier can help to improve analysis results. In fact, using a random forest classifier can improve accuracy by up to 9 percent points, in comparison to the best single tools.

Our work gives rise to several interesting directions of future research. A first direction would be to explore the quality of existing sentiment corpora. How good are these corpora in reality? Our classification method could be used to find suspicious texts within a corpus which need further manual verification. This could, on one hand, lead to better "gold standard" data; on the other hand, we might have to re-analyze some of the results that are based on such corpora.

Our main motivation, as mentioned in the introduction, is to explore and understand the gap between commercial and scientific algorithms for sentiment detection. We saw that accuracy for commercial tools is only mediocre; on the other hand, scientific papers often claim excellent accuracy rates. Hence, our next step will be to apply up-to-date scientific algorithms and prototypes to all test corpora, and compare these results. From this, we expect interesting insights on how to further improve existing sentiment detection systems.

Finally, we want to use smarter ensemble methods for combining tools besides random forest. One could also use other ensemble approaches, such as bagging and boosting, to build new meta-classifiers on top of existing tools. Furthermore, other features such as text length or text type could be used to further improve analysis results. Since we have already shown that such approaches can improve analysis quality significantly, it will be interesting to see what level of quality could be achieved at best.

Acknowledgments

We would like to thank all tool providers for giving us the opportunity to test and evaluate their systems for free, and for their excellent support. Further we would like to thank Thilo Stadelmann for carefully reading the manuscript and Andreas Ruckstuhl for comments and suggestions on the statistical methods.

References

1. Michelle Annett and Grzegorz Kondrak: A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs. In: Proceedings of the Twenty-First Canadian Conference on Artificial Intelligence (2008)
2. Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva: Sentiment Analysis in the News. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), pp. 2216-2220. Valletta, Malta, 19-21 (May 2010)
3. Adam Bermingham and Alan F. Smeaton: A Study of Inter-Annotator Agreement for Opinion Retrieval. In: SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA (2009)
4. Sentiment API Market comparison, http://www.bitext.com/2013/08/comparing-apis-example.html (2013)
5. Leo Breiman: Random Forests. Machine Learning 45(1), 5-32 (2001)

6. Hang Cui, Vibhu Mittal, and Mayur Datar: Comparative Experiments on Sentiment Classification for Online Product Reviews. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-2006) (2006)
7. Xiaowen Ding, Bing Liu, and Philip S. Yu: A Holistic Lexicon-Based Appraoch to Opinion Mining. In: Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008), Stanford University, Stanford, California, USA (2008)
8. Social media monitoring report - Turning conversations into insights, http://www.freshnetworks.com/files/freshnetworks/FINAL%20FreshNetworks%20version_0.pdf (2011)
9. Martin Hawksey: Sentiment Analysis of tweets: Comparison of ViralHeat and Text-Processing Sentiment APIs, http://mashe.hawksey.info/2011/11/sentiment-analysis-of-tweets-comparison-of-viralheat-and-text-processing-sentiment-api/ (2011)
10. Minqing Hu and Bing Liu: Mining and summarizing customer reviews. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, Washington, USA (2004)
11. Jackie Kmetz: Measuring Social Sentiment: Assessing and Scoring Opinion in Social Media, http://www.visibletechnologies.com/resources/white-papers/measuring-sentiment/ (2010)
12. Bing Liu: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers (2012)
13. Sascha Narr, Michael Hülfenhaus, and Sahin Albayrak: Language-Independent Twitter Sentiment Analysis. In: Knowledge Discovery and Machine Learning (KDML), LWA (2012)
14. S. Padmaja and S. Sameen Fatima: Opinion Mining and Sentiment Analysis - An Assessment of Peoples' Belief: A Survey. International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol.4, No.1 (February 2013)
15. Bo Pang and Lillian Lee: Opinion Mining and Sentiment Analysis. Now Publishers Inc. (2008)
16. Matt Rhodes: The problem with automated sentiment analysis, http://www.freshnetworks.com/blog/2010/05/the-problem-with-automated-sentiment-analysis/ (2010)
17. SemEval Corpus. 4th Internation Workshop on Semantic Evaluations (2007)
18. Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng: Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 254–263, (2008)
19. Marshall Sponder: Comparing Social Media Monitoring Platforms on Sentiment Analysis about Social Media Week NYC 10, http://www.webmetricsguru.com/archives/2010/01/comparing-social-media-monitoring-platforms-on-sentiment-analysis-about-social-media-week-nyc-10/ (2010)

20. Oscar Täckström and Ryan McDonald: Discovering fine-grained sentiment with latent variable structured prediction models. European Conference on Information Retrieval (ECIR 2011), Dublin, UK. (2011)
21. Social Media Monitoring Review, http://social-media-monitoring-review.toptenreviews.com/ (2013)
22. G. Vinodhini and R.M. Chandrasekaran: Sentiment Analysis and Opinion Mining: A Survey. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, p. 282-292 (2012)
23. Janyce Wiebe, Theresa Wilson, and Claire Cardie: Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation 2005, Volume 39, Issue 2-3, pp 165-210 (2005)
24. Theresa Wilson, Janyce Wiebe, and Paul Hoffmann: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT 2005), p.347-354 (2005)