Meta-Classifiers Easily Improve Commercial Sentiment Detection Tools

Mark Cieliebak, Oliver Dürr, Fatih Uzdilli*

Zurich University of Applied Sciences
Winterthur, Switzerland
Email: {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. The best commercial tools have average accuracy of 60%. We then apply machine learning techniques (Random Forests) to combine all tools, and show that this results in a meta-classifier that improves the overall performance significantly.

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

1. Introduction

How good is state-of-the-art 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% (Padmaja & Sameen, 2013; Vinodhini & Chandrasekaran, 2012). But: If you look 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 basically state something like this: "More often than not, a positive comment will be classified as negative or vice-versa" (Rhodes, 2010). Is there really such a large gap between research and real-life systems?

In this paper, we tackle this question by evaluating the performance of several commercial sentiment detection tools. More precisely, we explore how well existing tools perform on different sentence-based test corpora. It turns out that even the best tools only achieve an average accuracy of 60%. We then combine all tools using machine learning techniques (Random Forests), and show that this meta-classifier is better than any single tool. Based on these results, we conclude that there is still large potential to improve existing commercial sentiment detection tools.

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. emotionality 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. Liu, 2012; Vinodhini & Chandrasekaran, 2012; or Pang & Lee, 2008.

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: On one hand, there exist scientific survey papers that explore the abilities of different algorithmic approaches to sentiment detection: Padmaja and Sameen list the results of 19 sentiment analysis papers; typical accuracy of the approaches is about 80% (2013). Cui et al. analyze 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% (2006). Annett and Kondrak 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 (2008).

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: Bitext.com compares 10 sentiment APIs, using one negative sentence, one comparative sentence and one conditional sentence. They conclude that most of the APIs have problems with polarity modifiers or intensifiers and conditional sentences (bitext, 2013). Hawksey analyzes the performance of two sentiment APIs using only tweets: The precision for polar text is around 20% (2011).

Finally, comparisons of social media monitoring tools typically also explore their sentiment detection abilities, but mostly restrict themselves to qualitative aspects (see e.g. freshnetworks, 2011; Sponder, 2010; or Kmetz, 2010).

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 2 in the Appendix.

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, as shown in Table 1.

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

Table 1: Tools

4. Results

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

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.

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 approximately 60% or even lower.

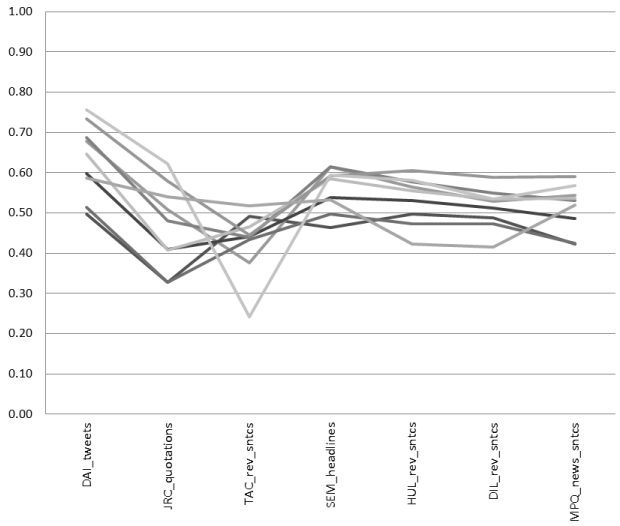

Figure 1: Accuracy of All Tools on Test Corpora.

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 (continuous 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.

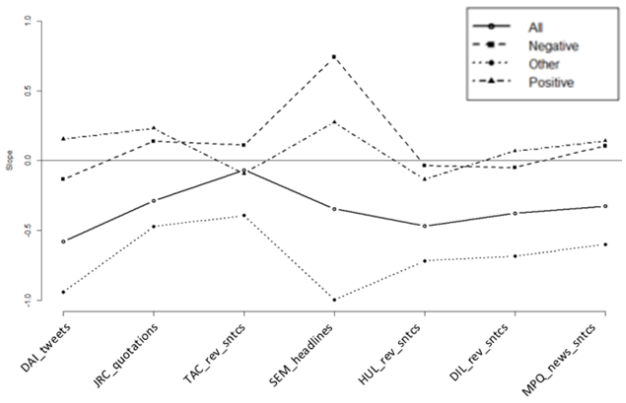

Figure 2: Impact of Increasing Text Length on Analysis Performance. Shown is slope of a linear model fitted into a performance vs. text length mapping (for details see main text). Negative values indicate a decrease of performance for longer texts, positive values the opposite.

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 meta-classifier combining the individual results. A robust classifier, which can naturally handle categorical input by design, is the random forest classifier (Breiman, 2001). More precisely, we use the random forest implementation of the R-package "randomForest" with default settings. For each corpus, we train a separate classifier for each corpus using the three votes (negative, other, positive) as categorical input for the random forest. 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 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.

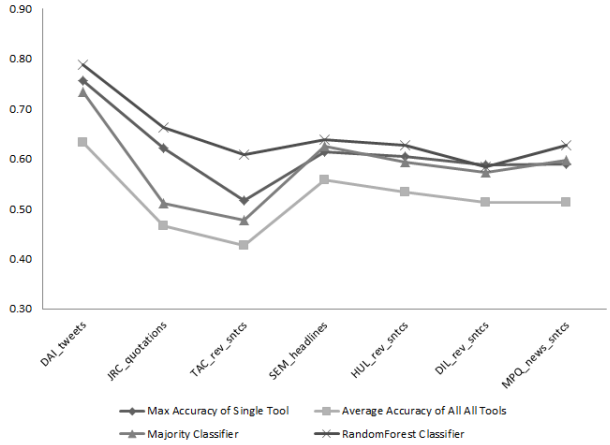

Figure 3: Accuracies for Tools and Meta-Classifiers, per Corpus

7. Summary and Future Challenges

In this work, we evaluated the quality of nine 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, accuracy decreases if texts get longer, which is due to the decline in the ability to detect "other" sentiments.

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

Our results show that accuracy for commercial tools is only mediocre in comparison to scientific papers, which 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.

On the other hand, we want to use smarter ensemble methods for combining the 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. It will be interesting to see what level of quality could be achieved at best.

8. 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. In addition, we would like to thank the LREC reviewers who gave valuable feedback.

9. References

Annett M.; Kondrak, G. (2008). A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs. In: *Proceedings of the Twenty-First Canadian Conference on Artificial Intelligence*.

Balahur, A.; Steinberger, R.; Kabadjov, M.; Zavarella, V.; Goot, E; Halkia, M.; Pouliquen, B. and Belyaeva, J. (2010): Sentiment Analysis in the News. In: *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010)*, pp. 2216-2220.

bitext (2013). Sentiment API Market comparison. URL http://www.bitext.com/2013/09.

Breiman, L. (2001). Random Forests. *Machine Learning* 45(1), p. 5-32.

Cui, H; Mittal, M. and Datar, M (2006). Comparative Experiments on Sentiment Classification for Online Product Reviews. In: *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI)*.

Ding, X.; Liu, B. and Yu, P.S. (2008). A Holistic Lexicon-Based Approach to Opinion Mining. In: *Proceedings of First ACM International Conference on Web Search and Data Mining (WSDM-2008)*, Stanford University, Stanford, California, USA.

Freshnetworks (2011). *Social media monitoring report - Turning conversations into insights*. URL http://www.freshnetworks.com/files/freshnetworks/FINAL%20FreshNetworks%20version_0.pdf

Hawksey, M. (2011). Sentiment Analysis of tweets: Comparison of ViralHeat and Text-Processing Sentiment APIs. URL http://mashe.hawksey.info/2011/11/ sentiment-analysis-of-tweets-comparison-of-viralheat-and-text-processing-sentiment-api/

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004)*, Seattle, Washington, USA.

Kmetz, J. (2010). *Measuring Social Sentiment: Assessing and Scoring Opinion in Social Media*. URL: http://www.visibletechnologies.com/resources/white-papers/measuring-sentiment/

Liu B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers (2012)

Narr, S; Hülfenhaus M. and Albayrak, S. (2012). Language- Independent Twitter Sentiment Analysis. In: *Knowledge Discovery and Machine Learning (KDML)*, LWA.

Padmaja, S. and Sameen, S.F. (2013). Opinion Mining and Sentiment Analysis - An Assessment of Peoples' Belief: A Survey. In: *International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC)* Vol.4, No.1.

Pang, B. and Lee L. (2008). *Opinion Mining and Sentiment Analysis*. Now Publishers Inc.

Rhodes, M. (2010). *The problem with automated sentiment analysis*. URL http://www.freshnetworks.com/blog/2010/05/the-problem-with-automated-sentiment-analysis/

Strapparava, C. and Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. In: *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 70-74.

Sponder, M. (2010). *Comparing Social Media Monitoring Platforms on Sentiment Analysis about Social Media Week NYC 10*. URL http://www.webmetricsguru.com/archives/2010/01/comparing-social-media-monitoring-platforms-on-sentiment-analysis-about-social-media-week-nyc-10/

Täckström, O. and McDonald, R. (2011). Discovering fine-grained sentiment with latent variable structured prediction models. In: *European Conference on Information Retrieval (ECIR 2011)*, Dublin, UK.

Vinodhini, G. and Chandrasekaran, R.M. (2012). Sentiment Analysis and Opinion Mining: A Survey. In: International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, pp. 282-292.

Wiebe, J.; Wilson, T and Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language. Int: *Language Resources and Evaluation*, Volume 39, Issue 2-3, p. 165-210.

Appendix: Tables

Corpus Name	Text Type	Number of Texts	Ratio of polar texts	Average Text Length	Reference
DAI_tweets	Tweets	4093 tweets	19% pos 13% neg 67% oth	14 words	Narr, Hülfenhaus & Sahin, 2012
JRC_quotations	Quotations from Online News	1290 quotations	15% pos 18% neg 67% other	30 words	Balahur et al., 2010
TAC_rev_sntcs	Product Review Sentences	2689 sentences	34% pos 49% neg 17% other	20 words	Täckström & McDonald, 2011
SEM_headlines	News Headlines	1250 headlines	14% pos 25% neg 61% other	6 words	Strapparava & Mihalcea, 2007
HUL_rev_sntcs	Product Review Sentences	3945 sentences	27% pos 16% neg 57% other	18 words	Hu & Liu, 2005
DIL_rev_sntcs	Product Review Sentences	4275 sentences	31% pos 18% neg 51% other	16 words	Ding et al., 2008
MPQ_news_sntcs	News Sentences	11111 sentences	14% pos 30% neg 55% other	23 words	Wiebe, Wilson & Cardie, 2005

Table 2: Test Corpora

Corpus	DAI	JRC	TAC	SEM	HUL	DIL	MPQ
Number of Texts in Corpus	4093 tweets	1290 quotations	2689 sentences	1250 headlines	3945 sentences	4275 sentences	11111 sentences
Proportion of Annotations	19% pos 13% neg 67% other	15% pos 18% neg 67% other	34% pos 49% neg 17% other	14% pos 25% neg 61% other	27% pos 16% neg 57% other	31% pos 18% neg 51% other	14% pos 30% neg 55% other
Average Accuracy	0.63	0.47	0.43	0.56	0.53	0.51	0.51
Max. 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: Positive	0.44	0.24	0.52	0.33	0.48	0.51	0.30
Average Precision: Negative	0.51	0.30	0.69	0.43	0.35	0.36	0.51
Average Precision: Other	0.82	0.75	0.14	0.67	0.70	0.62	0.66
Average Recall: Positive	0.65	0.52	0.55	0.40	0.67	0.59	0.46
Average Recall: Negative	0.53	0.35	0.37	0.47	0.40	0.38	0.43
Average Recall: Other	0.65	0.48	0.34	0.63	0.51	0.51	0.57
Average F1 Score: Positive	0.51	0.31	0.52	0.34	0.54	0.53	0.33
Average F1 Score: Negative	0.50	0.31	0.47	0.42	0.35	0.35	0.43
Average F1 Score: Other	0.71	0.55	0.19	0.63	0.57	0.54	0.57

Table 3: Summary of Main Results