From newspaper to microblogging: What does it take to find opinions?

Wladimir Sidorenko and **Jonathan Sonntag** and **Manfred Stede**
Applied Computational Linguistics
University of Potsdam/Germany
sidarenk|sonntag|stede@uni-potsdam.de

Nina Krüger and **Stefan Stieglitz**
Dept. of Information Systems
University of Münster/Germany
nina.krueger|stefan.stieglitz@uni-muenster.de

Abstract

We compare the performance of two lexicon-based sentiment systems – SentiStrength (Thelwall et al., 2012) and SO-CAL (Taboada et al., 2011) – on the two genres of newspaper text and tweets. While SentiStrength has been geared specifically toward short social-media text, SO-CAL was built for general, longer text. After the initial comparison, we successively enrich the SO-CAL-based analysis with tweet-specific mechanisms and observe that in some cases, this improves the performance. A qualitative error analysis then identifies classes of typical problems the two systems have with tweets.

1 Introduction: Twitter, SentiStrength and SO-CAL

In recent years, microblogging has been an attractive new target for sentiment analysis. The question studied in this paper is how the methods used for “standard” newspaper text can be transferred to microblogs. We focused on the Twitter network because of its widespread use, and because Twitter communication, in response to emerging issues, is fast and especially ad hoc, making it an effective platform for the sharing and discussion of crisis-related information (Bruns/Burgess, 2011). Furthermore, Twitter is characterized by a high topicality of content (Milstein al., 2008).

Specifically, we present experiments involving two sentiment analysis systems that both employ a combination of polarity lexicon and sentiment composition rules: (i) SentiStrength (Thelwall et al., 2012), a system that is geared toward short social-media text, and (ii) SO-CAL (Taboada et al., 2011), ‘Semantic Orientation Calculator’, a general-purpose system that was designed primarily to work on the level of complete texts. While both are lexicon-based approaches, there are certain differences in the roles of the various submodules. For our purposes here, it is important that SentiStrength was designed to cope specifically with “user-generated content”. Among the features of the system, as stated by Thelwall et al., the following four are especially important for tweets: (i) a simple spelling correction algorithm deletes repeated letters when the word is not found in the dictionary; (ii) repeated letters lead to a boost in sentiment value; (iii) an emoticon list supplements the polarity lexicon; (iv) positive sentences ending in an exclamation mark receive an additional boost, and multiple exclamation marks further strengthen the polarity.

SO-CAL, on the other hand, does not include social-media-specific measures. In contrast, it was designed for determining semantic orientation on the text level; in our experiments here, we are thus using it for the non-intended purpose of sentence-level sentiment, on tweet “sentences”.

Next, we review related work on twitter sentiment analysis (Section 2), and describe the data sets for our experiments in Section 3. Then we investigate the relative performance of SentiStrength and SO-CAL on newspaper text and on tweets (Section 4), including experiments with preprocessing steps. In Section 5, we present observations from a qualitative evaluation, and we interpret the results and conclude in Section 6.

Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 81–86,
Atlanta, Georgia, 14 June 2013. ©2013 Association for Computational Linguistics

2 Related work

Following the work on “standard” text, sentiment classification on tweets is often treated as a two-step task, e.g., (Barbosa/Feng, 2010): subjectivity classification followed by polarity classification. For subjectivity classification, (Pak/Paroubek, 2010) found that the distribution of POS tags is a useful feature, due to, for example, the presence of modal verbs in subjective tweets.

For polarity assignment, one approach is to automatically build large sets of training data and then train classifiers on token n-grams; in this vein, (Pak/Paroubek, 2010) found that in their approach, bigrams outperform unigrams and trigrams, and they report f-measures around 0.6 for the three-way pos/neg/neutral classification. The other, non-learning, approach is to rely on a polarity wordlist (or a collection of several, as in (Joshi et al., 2011; Mukherjee et al., 2012)). Mukherjee et al. report an accuracy of 66.69% for pos/neg, and 56.17% for pos/neg/neut classification.

Typical preprocessing steps employed by the approaches discussed are the correction of misspellings, the replacement of URLs and hashtags, the translation of emoticons and of slang words. Sometimes, stop word removal and stemming is used; sometimes deliberately not. Few authors evaluate the influence of the various measures; one exception is (Mukherjee et al., 2012).

A recent branch of research deals with fine-grained target-specific analysis (as proposed recently by (Jiang et al., 2011)). In our work, however, we tackle the more coarse-grained problem of assigning a single sentiment value to a complete tweet. However, we will return to the issue of target-specificity in our conclusions.

An interesting result from analysing the state of the art is that apparently no consensus has been reached yet on the question of “extra difficulty” of tweet sentiment analysis. While everybody agrees that tweets are noisy and can pose considerable difficulty to any standard linguistically-inspired analysis tool, it is not clear to what extent this is a problem for sentiment analysis. Some authors argue that the noise renders the task more difficult than the analysis of longer text, while others maintain that the brevity of tweets is in fact an advantage, because – as (Bermingham/Smeaton, 2010) put it, “the short document length introduces a succinctness to the content”, and thus “the focused nature of the text and higher density of sentiment-bearing terms may benefit automated sentiment analysis techniques.” In their evaluation, the classification of microblogs indeed yields better results than that of blogs.

In correspondence with this open question, there are only few investigations so far on the performance differences for existing sentiment tools operating on newspaper versus social media text. To shed more light on the issue, we chose to run a set of comparative experiments with the two aforementioned lexicon/rule-based systems, on both newspaper and twitter corpora.

3 Data sets

MPQA The well-known MPQA corpus[1] (Wiebe et al., 2005) of newspaper text has fine-grained annotations of ‘private states’ at phrase level. For our purposes these need to be reduced to a more coarse-grained labelling of sentence-level sentiment. To avoid ambiguity, we ignored those sentences that include both positive and negative sentiment annotations. From the remaining sentences, we selected 100 positive and negative sentences each, where the former target-specific sentiment is now taken to represent sentence-level sentiment. The data set is a difficult one, given that we are dealing with isolated sentences from newspaper reports.

Qantas To track Twitter data we used a self-developed prototype (see (Stieglitz/Kaufhold, 2011)). We concentrate our analysis on Qantas, an Australian leading carrier for long-haul air travel, for which we assume substantial interest in public communication. We furthermore expect that – caused by some management crises in 2011 – online communication around Qantas-related topics is characterized by a strong emotional investment of stakeholders.

The tracking tool captures all those tweets that contain the keyword ‘Qantas’ in their content, in the username of the sender, or in a URL. After spam removal, we had a dataset of some 27,000 tweets, collected between mid-May and mid-November 2011.

[1] `http://mpqa.cs.pitt.edu/`

Topic	#pos	#neut	#neg	#irrelevant
Apple	219	581	377	164
Google	218	604	61	498
Microsoft	93	671	138	513
Twitter	68	647	78	611

Table 1: Distribution of tweets and labels across subcorpora

For evaluation purposes, 300 Tweets have been manually annotated by two annotators in parallel, using a polarity scale ranging from -2 to 2. 190 Tweets of those (63%) received identical labels, and we used only this set in our experiments described below. That means we also discarded cases of “minor” disagreement such as a -1/-2 annotation.

Sanders The Sanders corpus[2] is a corpus consisting of 5513 tweets of various languages which have been annotated for sentiment. The tweets have been sampled by the search terms „@apple“, „#google“, „#microsoft“ and „#twitter“. Each tweet is accompanied by a date-time stamp and the target of its polarity. Possible polarity values are *positive*, *negative*, *neutral* (simple factual statements / questions without strong emotions / neither positive nor negative / both positive and negative), and *irrelevant* (spam / non-English). The positive and negative tweets thus contain judgements on the companies or their products/services. Along with the corpus comes an annotation scheme and statistics about the corpus. Some numbers of the size and distribution within the corpus are given in Table 1.

According to the annotation guidelines, positive and negative labels were only assigned to clear cases of sentiment. Ambigious tweets have been annotated as neutral.

4 Experiments and results

4.1 Performance on MPQA sentences

In order to establish a basis for the comparison, we first ran a small comparative evaluation on “standard” text, i.e., on the sentences from the MPQA newspaper corpus. The results, given in Table 2, show that both systems perform considerably better on negative than on positive sentences, and overall there is a slight advantage for SO-CAL.

	SentiStrength	SO-CAL
acc pos	0.2727	0.4717
acc neg	0.7071	0.6542
weighted avg	0.4899	0.5634

Table 2: Accuracy on MPQA sentences

	Senti-Strength	SO-CAL	SO-CAL preproc.
Qantas			
acc	0.3754	0.3953	0.3887
acc pos	0.3091	0.2545	0.2545
acc neg	0.2857	0.2857	0.2857
acc neut	0.6164	0.6781	0.6644
avg sentiment	1.1075	1.2756	1.3316
Sanders total			
acc	0.5945	0.5899	0.5790
acc pos	0.6171	0.5694	0.6032
acc neg	0.4572	0.5301	0.5519
acc neut	0.6230	0.6092	0.5802
avg sentiment	0.8517	1.3761	1.5233
Sanders twitter			
acc	0.4985	0.5804	0.5387
acc pos	0.4286	0.3750	0.4821
acc neg	0.4590	0.4754	0.5246
acc neut	0.5099	0.6121	0.5245
avg sentiment	0.8393	1.4054	1.6978

Table 3: Accuracy on tweet corpora

4.2 Performance on Qantas and Sanders tweets

In Table 3, we show the system performance on the Twitter corpora: Qantas, the complete Sanders corpus, and the Sanders subcorpus with target “Twitter”. We ran evaluations on all four separate subcorpora, but only “Twitter” showed interesting differences from the results for the total corpus, and that is why they are included in the table. The “acc” row gives the overall weighted accuracy. “Avg sentiment” is the absolute value of the sentiment strength determined by SentiStrength and SO-CAL; notice that these should not be compared between the two systems, as they do not operate on the same scale. (We will return to the role of sentiment strength in Section 6.)

[2] `http://www.sananalytics.com/lab/twitter-sentiment/`

4.3 Preprocessing steps

Since SO-CAL was not intended for analyzing Twitter data, we implemented three preprocessing steps to study whether noise effects of this text genre can be reduced. Similarly to the steps suggested by (Mukherjee et al., 2012), we first unified all URLs, e-mail addresses and user names by replacing them with unique tokens. Additionally, in step 1 all hash marks were stripped from words, and emoticons were mapped to special tokens representing their emotion categories. These special tokens were then added to the polarity lexicons used by SO-CAL.

In step 2, social media specific slang expressions and abbreviations like *"2 b"* (for *"to be"*) or *"imsry"* (for *"I am sorry"*) were translated to their appropriate standard language forms. For this, we used a dictionary of 5,424 expressions that we gathered from publicly available resources.[3]

In the last step, we tackled two typical spelling phenomena: the omission of final *g* in gerund forms (*goin*), and elongations of characters (*suuuper*). For the former, we appended the character *g* to words ending with *-in* if these words are unknown to vocabulary,[4] while the corresponding 'g'-forms are in-vocabulary words (IVW). For the latter problem, we first tried to subsequently remove each repeating character until we hit an IVW. For cases resisting this treatment, we adopted the method suggested by (Brody/Diakopoulos, 2011) and generated a squeezed form of the prolongated word, subsequently looking it up in a probability table that has previously been gathered from a training corpus.

Altogether, SO-CAL does not benefit from preprocessing in the Qantas corpus, but it does help for the pos/neg tweets from the Sanders corpus, especially for the Twitter subcorpus. The observation that the accuracy on neutral tweets decreases while the average sentiment increases will be discussed in Section 6. We also measured the effects of the three individual steps in isolation, and the only noteworthy result is that SentiStrength, when subjected to our "extra" preprocessing, benefits slightly from slang normalization for the Qantas corpus, and from noise cleaning for some parts of the Sanders corpus.

5 Qualitative evaluation

Having computed the success rates, we then performed a small qualitative evaluation: What are the main reasons for the misclassifications on tweets? In addition, we wanted to know why the Qantas corpus yielded much worse results than the Sanders corpus, and thus we looked into its results.

5.1 Problems for SO-CAL

We chose SO-CAL's judgements as the basis for this evaluation and randomly selected 120 tweets from the Sanders corpus that were not correctly classified. The distribution across the manual annotations pos/neg/neut was 40/40/40.

In Table 4, we provide a classification of the reasons for problems. The first group are cases where we would not agree with the annotation and thus cannot blame SO-CAL. The second group includes problems that are beyond the scope of the system and hence, strictly speaking, not its fault. Among the typos, there are cases of misspelled opinion words, but also a few where the typo leads to problems with SO-CALs linguistic analysis and in consequence to a misclassification. The slang words include items like "wow!" but also shorthands such as "thx". Most important are "domain formulae": expressions that require inferences in order to identify the sentiment. An example is "I now use X instead of TARGET". We encounter these most often in negative tweets, where complaints are expressed, as in "My phone can send but not receive texts."

In the third group, we find problems that are or could be in the scope of SO-CAL. Occasionally, negation or irrealis rules misfire. Gaps in the lexicon are noticeable especially on the positive side (examples: "loving", "better", "thanks to"). 'Lexical ambiguity' refers to words that may or may not carry polarity; by far the most frequent example here is "new", which SO-CAL labels positive, but in technology-related tweets often is neutral. Also in neutral tweets, we often find high complexity, i.e., cases where both positive and negative judgements are mixed. And finally, a fair number of problems stems from sentiment expressed on the wrong target of the tweet.

[3] http://www.noslang.com/dictionary/, http://onlineslangdictionary.com/, http://www.urbandictionary.com/

[4] For vocabulary check, we used the open Hunspell dictionary (http://hunspell.sourceforge.net/).

Problem	Pos	Neg	Neut
Annotation ambig.	15%	0%	2%
Typo	3%	5%	10%
Slang words	12%	10%	0%
Sarcasm	0%	2%	0%
Domain formula	23%	60%	5%
Wrong rule	3%	5%	3%
Lexicon gap	30%	12%	0%
Lexical ambiguity	5%	5%	50%
Complexity	0%	0%	18%
Wrong target	8%	0%	12%

Table 4: SO-CAL error types on 120 Sanders tweets

Problem	Pos	Neg	Neut
Annotation ambig.	45%	25%	12%
Typo	18%	0%	0%
Slang words	0%	0%	0%
Sarcasm	0%	16%	0%
Domain formula	9%	42%	4%
Wrong rule	9%	0%	10%
Lexicon gap	9%	16%	0%
Lexical ambiguity	0%	0%	16%
Complexity	9%	0%	16%
Spam / news	0%	0%	41%

Table 5: Error types on 75 Qantas tweets

5.2 Observations on the Qantas corpus

The analysis of 75 Qantas tweets that have been misclassified by both SentiStength and SO-CAL yielded the results in Table 5: Again, many annotation cases are ambiguous, and domain formulae are the major problem with negative tweets. Sarcasm is much more frequent than in the Sanders corpus. The central problem for neutral tweets stems from the fact that spam and tweets containing headlines and URLs of news messages have been annotated as neutral, but these may very well contain polarity-bearing words, which are then detected by the systems.

6 Interpretation and Conlusions

News versus tweets. Since the Sanders corpus is much larger than Qantas, we regard it as the tweet representative for the comparison to MPQA (a difficult data set, as argued above). For positive text, both SentiStrength and SO-CAL yield better results on tweets, while for negative texts, the results on tweets are much lower than on news sentences. Within the news genre, however, both systems perform much better on negative than on positive text. So we conclude a “polarity flip” in the performance of both systems when going from news to tweets.

Differences among tweets. Based on the Sanders corpus, the SentiStrength and SO-CAL results are a little better than those reported by (Mukherjee et al., 2012), who achieved 56.17% accuracy for the three-way classification. As SO-CAL does not include tweet-specific analysis, we may conclude that the utility of such genre-specific measures is in fact limited. – An interesting question is why the “Twitter” subcorpus of Sanders behaves so different from the others: While overall accuracy is the same, the figures for the three categories differ widely. Also, SO-CAL here benefits heavily from preprocessing on the non-neutral tweets. One factor is the large proportion of neutral tweets (see Table 1); besides, we find that these tweets are not as target-related as those for Apple, Google, Microsoft; it seems that users often drop a ‘#twitter’ without actually talking *about* Twitter or its service.

Preprocessing. Of the four measures taken by SentiStrength to account for tweet problems (see Sct. 1), SO-CAL already implements the exclamation mark boost; the other three were added in our own preprocessing, but we did only minimal spellchecking. Overall, SO-CAL does not profit as much as we had expected, but we find a fair improvement (0.57–0.6) for the positive Sanders tweets. For neutral tweets, performance actually decreases.

The role of targets An interesting observation is that adding preprocessing to SO-CAL leads to detecting “more” sentiment: The average sentiment values increase for all the corpora in Table 3. At the same time, the accuracy on neutral tweets decreases, which indicates that “spurious” sentiment is being detected. The most likely reason is that SO-CAL indeed profits from tweet-preprocessing but then detects sentiment that is unrelated to the target and therefore not annotated in the gold data. An important direction for future work therefore is to pay more attention to target-specific sentiment identification, cf. (Jiang et al., 2011).

Acknowledgments

This work was funded by German Ministry for Education and Research (BMBF), grant 01UG1232D.

References

L. Barbosa and J. Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. Proc. of COLING (Posters), Beijing.

A. Bermingham and A. Smeaton. 2010. Classifying Sentiment in Microblogs: Is Brevity an Advantage? Proc. of the 20th ACM Conference on Information and Knowledge Management (CIKM), Toronto.

S. Brody and N. Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 562–570, Edinburgh.

A. Bruns and J.E. Burgess. 2011. The Use of Twitter Hashtags in the Formation of Ad Hoc Publics. 6th European Consortium for Political Research General Conference, Reykjavik, Iceland, pp. 25-27.

L. Jiang, M. Yu, M. Zhou, X. Liu and T. Zhao. 2011. Target-dependent twitter sentiment classification. Proc. of the 49th Annual Meeting of the ACL, pp. 151-160, Portland/OR.

A. Joshi, Balamurali A R, P. Bhattacharyya and R. Mohanty. 2011. C-Feel-It: a sentiment analyzer for micro-blogs. Proc. of the ACL-HLT 2011 System Demonstrations, pp. 127-132, Portland/OR.

S. Milstein, A. Chowdhury, G. Hochmuth, B. Lorica and R. Magoulas. 2008. Twitter and the Micro-Messaging Revolution: Communication, Connections, and Immediacy - 140 Characters at a Time.

S. Mukherjee, A. Malu, A.R. Balamurali and P. Bhattacharyya. 2012. TwiSent: a multistage system for analyzing sentiment in twitter. Proc. of the 21st ACM Conference on Information and Knowledge Management (CIKM).

A. Pak and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proc. of LREC, Valletta/Malta.

S. Stieglitz and C. Kaufhold. 2011. Automatic Full Text Analysis in Public Social Media – Adoption of a Software Prototype to Investigate Political Communication. Proc. of the 2nd International Conference on Ambient Systems, Networks and Technologies (ANT-2011) / The 8th International Conference on Mobile Web Information Systems (MobiWIS 2011), Procedia Computer Science 5, Elsevier, 776-781.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

M. Thelwall, K. Buckley, and G. Paltoglou. 2012. Sentiment strength detection for the social Web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):165–210.