Bounds of lexical sentiment analysis

Jussi Karlgren

***Abstract*—Sentiment analysis, the process where texts or sentences are categorised in positive, negative, or neutral, is in practical application most often based on purely lexical features – the presence or absence of attitudinally loaded terms. This paper, as part of an effort to improve low-footprint large-scale real-time sentiment analysis demonstrates the bounds of quantitative results given gold standards and lexical resources in use today. The paper shows that the gold standards differ considerably, and that potential gains are relatively small, given a certain lexical resource and a given gold standard.**

I. Practical low-footprint attitude analysis

The use case that motivates the experiments described here is that of sentence-level ("fine-grained") classification of utterances into positive, negative, or neutral, in many different languages, based on a sentiment lexicon of positively and negatively attitudinally loaded terms. The utterances in question are harvested from internet conversations: from editorial media, blogs, forums, micro blog posts, and potentially chat rooms and other informal sites, utterances which mention a concept of interest are scored for sentiment, scores tabulated, and the resulting timelines are used to inform e.g. market strategies or communication strategies.

This use case can be called into question in at least two different ways.

Firstly, that the method is arguably naive and could be improved by more sophisticated analyses. However, in real-time analysis of internet-scale text streams, heavy natural language processing machinery is rendered impracticable for portability, scalability, and maintenance reasons. The processing effort is prohibitive, the coverage of most analysis frameworks does not extend to new informal text types, the multi-lingual reality of commercial text analysis require cross-linguistic processing paradigms, and most importantly, the abstraction level and sophisticated knowledge representation inherent in dependency-based systems make systems which rely on syntactic models overly costly to maintain. Similarly, machine learning classifiers need sufficient amounts of training data, which is not always available for real-world commercial scenarios: annotating training data is costly and time-consuming. Such classifiers are also relatively tightly bound to the material they have been trained on, and transfer to other genres or data sets generally hurts performance considerably.

Secondly, that a two-category model is an oversimplification of human expression of sentiment, attitude, appeal, and emotion, as argued by us in earlier case studies (Karlgren et al, 2012). The extension from this overly simple model to an analysis using a more sophisticated palette of subjective and attitudinal language is straightforward, and the conclusions of this arguments are equally applicable for such cases.

To improve any models, we need experimental baselines and test sets. The range of variation in such test sets are bounds for the improvement we can expect to find using new methods.

II. Data

I use three human-annotated data sets. Quantitative data are given in Table I.

1) The Replab data set consists of several tens of thousands of microblog posts from Twitter, which are annotated for positive, negative, or neutral effect on the reputation of a commercial entity. These have been used extensively in shared tasks at the CLEF conferences. Items consist of up to 140 characters, and may contain several sentences within that limit. Details are given by Amigó et al (2013).
2) The data set used by Täckström and McDonald in experiments for inferring text-level sentiment from sentence level analyses consists of single sentences from consumer reviews on several topics, hand tagged for positive, negative, neutral, non-relevant, or mixed. Details are given by Täckström and McDonald (2011).
3) The Stanford sentiment treebank analysis data set consists of single sentences graded continuously constituent by constituent. In these experiments only the full sentence grading is used, with cutoffs for positive and negative scores set to be 0.4 and 0.6 respectively, as suggested in the data set documentation. Details are given by Socher et al (2013).

Set	Size	Positive	Negative	Other
RepLab	62 886	36 548	8 618	17 599
T & McD	3 564	923	1 320	1 321
Stanford	11 855	4 963	4 650	2 242

TABLE I
The data sets.

III. Bounds of lexical approaches

The recall bound for correct results in a gold standard, given a polarity lexicon, is bounded by the coverage of the lexical resource. A lexical model cannot transcend the coverage of the items in the lexicon and their occurrence in the target texts: documents with no terms in the lexicon cannot be reached by the analysis; documents with terms from both the positive and negative lexicon are inconclusive; documents may contain terms from a lexicon yet not be assessed as negative or positive in the gold standard. A coverage analysis of the material will give the bounds within which an experiment operates as shown in Table II.

As an extreme example, Table III shows how many items in the gold standard data sets contain the most prototypical

	Feature observed	Feature not observed
in category	Hit	Miss
not in category	Noise	Not seen

TABLE II
GENERAL CASE OF IMPROVING CATEGORISATION BASED ON FIXED FEATURES

polar terms *good* and *bad*, respectively. A sentiment classifier built on those two features alone will obviously yield low recall. Table IV shows more interestingly, what can coverage can be expected from an well-established sentiment lexicon published by Liu et al (2005), often and used as a basis for experimentation. (The argument posed here is obviously not crucially dependent on this specific resource.)

These coverage figures can be used to establish precision and recall bounds, which are given in Table V. A *conservative* approach is to assess as positive only texts with positive terms present and no negative terms present, and equivalently, to require a text to contain only negative polar items for texts assessed to be negative. This imposes a low ceiling on the recall. A higher recall for finding items is the *greedy* approach, which assumes that any occurrence of a polar term warrants categorising a text accordingly, irrespective of the presence of other terms in the text. This, of course, has consequences for precision. These two bounds are given in Table V, in columns ”Greedy” and ”Conservative”, respectively.

	Incidence of terms *good* and *bad*				
RepLab	Both	Pos only	Neg only	Neither	
All	18	1 046	370	61 452	62 886
Positive	5	693	179	35 671	36 548
Negative	3	107	97	8 411	8 618
T & McD	Both	Pos only	Neg only	Neither	
All	12	179	54	3 319	3 564
Positive	5	75	6	837	923
Negative	2	48	37	1 233	1 320
Stanford	Both	Pos only	Neg only	Neither	
All	24	336	193	11 302	11 855
Positive	4	164	17	4 778	4 963
Negative	15	105	159	4 371	4 650

TABLE III
COVERAGE OF *good* AND *bad* OVER THE GOLD STANDARD ITEMS

	Incidence of polar terms				
RepLab	Both	Pos only	Neg only	Neither	
All	5 355	18 939	8 976	29 616	62 886
Positive	2 715	13 033	3 662	17 413	36 548
Negative	1 213	1 477	2 682	3 246	8 618
T & McD	Both	Pos only	Neg only	Neither	
All	794	1 080	709	981	3 564
Positive	202	495	60	166	923
Negative	284	253	437	346	1 320
Stanford	Both	Pos only	Neg only	Neither	
All	3 839	3 530	2 630	1 856	11 855
Positive	1 659	2 288	493	523	4 963
Negative	1 474	713	1 675	788	4 650

TABLE IV
COVERAGE OF THE EXPERIMENT LEXICON OVER THE GOLD STANDARD ITEMS

IV. OBSERVATIONS

1) There are clear differences across the collections and the relative identifiability of the positive and negative items, which is an indication that the choice of test collection is a parameter of importance in designing system evaluation.
2) The difference between the *greedy* and *conservative* approaches are found in the texts in the ”Both” column of Table IV.
3) There are two ways to improve results of a lexical classifier. Firstly, and most obviously, the set of fixed features, the lexical resource, could be improved. This means adding items judiciously to a lexicon or removing items from a lexicon, increasing or decreasing recall at some cost or some benefit to precision, improving the lexicon with respect to coverage or purity. This would harvest items currently in the ”Neither” column of Table IV and move items from the ”Feature not observed column” to the ”Feature observed” column of Table II.
4) Secondly, the methods employed to make use of lexical features can be improved beyond scanning for their presence in a text, to find ways to pick up items from cell *Miss* or to disregard items from cell *Noise* without modifying the feature set.
5) For both approaches, the solution space under consideration lies between the bounds of the *greedy* and *conservative* approaches.

	Greedy		Conservative	
	Recall	Precision	Recall	Precision
	RepLab			
Positive	43.9	64.6	36.1	68.8
Negative	46.2	27.0	36.4	29.7
	T & McD			
Positive	76.6	33.9	53.8	43.9
Negative	55.6	42.6	25.5	56.3
	Stanford			
Positive	80.8	53.2	45.9	64.7
Negative	68.8	48.3	15.8	63.6

TABLE V
BOUNDED RESULTS

V. CONCLUSIONS

This quick analysis demonstrates how much experimental gain there is to find from a gold standard given an accepted lexical resource of attitudinal terms. If the goal is to e.g. find how much improvement an analysis of negation or logical connectives might yield, the results will be contingent on the choice of gold standard. This points at the necessity to perform a baseline analysis of the experimental materials (such as has been done here) to investigate the power of one’s results.

REFERENCES

1) Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, and Damiano Spina. Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. In *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. Proceedings of the Fourth International Conference of the CLEF Initiative (CLEF)*. Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, Springer, 2013.
2) Jussi Karlgren, Magnus Sahlgren, Fredrik Olsson, Fredrik Espinoza, and Ola Hamfors. Usefulness of sentiment analysis. In *Proceedings of the European Conference on Advances in Information Retrieval (ECIR)*. 2012.
3) Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of the 14th International World Wide Web conference (WWW)*. 2005.
4) Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2013.
5) Oscar Täckström and Ryan McDonald. Discovering fine-grained sentiment with latent variable structured prediction models. In *Proceedings of the European Conference on Advances in Information Retrieval (ECIR)*, 2011.