Everything Before "But"

Jussi Karlgren

***Abstract*—This paper reports on baseline experiments to improve large-scale real-time sentiment analysis. 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. There are many ways of improving a naïve model and this paper describes a practical and low footprint method based on linguistic principles, without relying on costly, cumbersome, and brittle syntactic analysis machinery. The method has practical utility and is being introduced in a commercial system at present.**

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 some target of interest — a brand, a corporate entity, a political organisation, politician, or political issue, e.g. — are scored for sentiment and tabulated to accumulate a sentiment score for the target in question over time. These resulting timelines are used to inform e.g. market strategies or communication strategies and may have considerable commercial value.

In real-time analysis of internet-scale text streams, heavy natural language processing machinery is rendered impracticable for 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. This paper takes as its point of departure simple but effective lexical sentiment analysis approach and investigates how it can be improved from first principles without costly intellectual superstructures at processing time.

II. Argumentation markers

Certain lexical items break the neat progression of topicality along the utterance. The topic-comment structure does not work linearly, if argumentative terms such as *but*, *however*, *yet*, *instead* or other explicit markers of argumentation are used.

(1) a. It uses an old-time formula, it's not terribly original and it's rather messy – *but* you just have to love the big, dumb, happy movie.

b. It is ridiculous, of course *but* it is also refreshing, disarming, and just outright enjoyable *despite* its ridiculousness.

[1]

III. Hypothesis: Subordinating conjunctions neutralise attitude before them

Certain items, such as "*but*", in an utterance indicate that what comes after is of greater argumentative import than what came before. A scoring scheme enhanced by a suitable hiccup at points where it encounters argumentative markers will provide better precision than one which does not.

IV. Starting points

Take a number of human-assessed utterances and establish whether a method which notes the presence of subordinating conjuncations does better than one which does not.

A. Target notion

We have in practice found that more fine-grained sentiment analyses than positive and negative would seem to be necessary. (Karlgren et al., 2012) There are no benchmarking resources available outside our laboratory for experimentation on a broader palette of attitude or sentiment, and thus we will in these experiments focus on those macro sentiments on which other sentiments are typically and simplistically projected. We report precision for positive and negative sentiment separately and recall for each.[2]

B. Data

The below experiments make use of 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

[1] Interestingly, while we might assume that these basic level argumentative operators would be invariant across human languages, von Klopp (1994), among others, finds that there are differences, even between closely related languages. Thus, the English *but* is equivalent to two different argumentational operators in e.g. Spanish, German, or apparently, even its most closely related Scandinavian languages.

(i) a. English: I cut my elbow but I didn't cry ↔ Spanish: ... *pero* ...
b. English: It is not difficult but impossible ↔ Spanish: ... *sino* ...

[2] We do not report F-scores. We have yet to find a use case for which F-score would be a valid quality measure, an algorithm development effort where it would be a useful target metric, or a research direction for which it would provide insights.

up to 140 characters, and may contain several sentences within that limit. Details are given in 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 in 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 in Socher et al. (2013).

In these experiments, for each data-set, only sentences which were tagged positive or negative were considered.

Set	Size	Positive	Negative	Other
RepLab	62 886	36 548	8 618	17 599
T & McD	3 836	923	1 320	1 593
Stanford	11 855	4 963	4 650	2 242

TABLE I
THE DATA SETS.

C. The sentiment lexicon

The sentiment analysis scheme in these experiments is entirely lexical and consists of a large number of unweighted multi-word terms taken from the sentiment lexicon published by Liu et al Liu et al. (2005). In our commercial practice, the lexicon we use is built semiautomatically using data from recent text streams (Sahlgren et al., 2016); for the purposes of these present experiments and for replicability reasons, the lexicon is used in the form as it is publicly available.

A two-category case of two polar categories such as the present one is arguably an oversimplification of human expression of sentiment, attitude, appeal, and emotion, as argued in earlier case studies (Karlgren, 2009) Here, this experiment will be used as a template, but I stress the fact that in a real applied commercial use case, other more domain-specific categories will be of greater practical utility for customers. The argument and results here can be assumed to be extensible to any lexically encoded attitude and any practical task where a fixed lexicon is used as a knowledge source.

V. 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.

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.

A *conservative* approach is to assess as positive only texts with positive terms and no negative terms, and equivalently, to require 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, to assume that any occurrence of a polar term indicates a polar text irrespective of other terms in the text.

The solution space under consideration in optimising lexical polarity analysis lies between these two bounds, which are given in Table III, as conditions "greedy" and "conservative", respectively.

A. Incidence of But

The presence of "but" is a significant indication that attitude may change over the course of an utterance, and this and similar items have been used in previous research. The point of tracking "but" in terms of improving sentiment analysis results is to find items where observing a "but" would invalidate the analysis so far in the utterance. The sentences in Example (1) conform to this processing template. To identify the power of this approach we identify the number of utterances in the test corpora where this approach has a potential of improving results. Table II shows a breakdown of items which contain "but". The items under consideration are those which contain both positive and negative polar terms, i.e. the first column of the table.

	Incidence of "But"			
RepLab	Pos & Neg	Pos only	Neg only	No polar terms
All	74	98	65	91
Positive	34	61	22	46
Negative	18	13	19	9
T & McD	Pos & Neg	Pos only	Neg only	No polar terms
All	194	168	83	93
Positive	36	50	2	8
Negative	55	39	39	39
Stanford	Pos & Neg	Pos only	Neg only	No polar terms
All	588	398	219	138
Positive	220	206	40	46
Negative	209	94	120	52

TABLE II
INCIDENCE OF "BUT" RELATED TO THE GOLD STANDARD AND TO POLAR ITEMS

VI. Baseline Lexical Sentiment Analysis

Using the simplest conceivable approach, scoring polar items and assigning polarity by largest score we get the results given in Table III.

VII. Experiment: But

This experiment treats utterances with "*but*" differently from others: in the linear sequence, when "*but*" is encountered, the scores are reset, under the assumption that a "*but*" signals that what follows it is of more import than what came before. This, as shown in Table III, increases precision for both positive items and negative items with a fairly low cost to recall.

Set	Micro		Macro		Positive		Negative	
	p	r	p	r	p	r	p	r
	BOUNDED RESULTS							
	Conservative							
RepLab	61.3	36.2	49.3	36.3	68.8	36.1	29.7	36.4
T & McD	51.2	37.1	50.1	39.7	43.9	53.8	56.3	25.5
Stanford	64.2	31.3	64.2	30.9	64.7	45.9	63.6	15.8
	Greedy							
RepLab	57.4	44.3	45.8	45.1	64.6	43.9	27.0	46.2
T & McD	39.0	64.2	38.3	66.1	33.9	76.6	42.6	55.6
Stanford	50.8	75.0	50.8	74.8	53.2	80.8	48.3	68.8
	BASELINE LEXICAL BAG OF WORD RESULTS							
Set	Micro		Macro		Positive		Negative	
	p	r	p	r	p	r	p	r
RepLab	52.4	38.3	48.2	40.1	68.1	37.2	28.4	42.9
T & McD	44.2	55.7	44.3	56.9	40.4	63.7	48.3	50.1
Stanford	58.4	60.9	58.7	60.9	63.5	60.4	53.9	61.5
	SUBJUNCTION CLEARED WEIGHTING							
Set	Micro		Macro		Positive		Negative	
	p	r	p	r	p	r	p	r
RepLab	**52.7**	37.9	**48.4**	39.6	**68.3**	36.9	**28.5**	42.3
T & McD	**44.8**	54.5	**44.9**	55.8	**41.4**	62.7	**48.5**	48.8
Stanford	**59.4**	60.1	**59.8**	60.1	**65.2**	59.6	**54.4**	60.8

TABLE III
BOUNDS, BASELINES, AND RESULTS

VIII. CONCLUSIONS

We have here demonstrated how a practical low foot-print system can be enhanced with a simple, effective and efficient, incrementally improvable approach which is consistent with linguistic theory.

REFERENCES

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 Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *Lecture Notes in Computer Science*, pages 333–352. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-40801-4. doi: 10.1007/978-3-642-40802-1_31. URL http://dx.doi.org/10.1007/978-3-642-40802-1_31.

Jussi Karlgren. Affect, appeal, and sentiment as factors influencing interaction with multimedia information. In *Proceedings of Theseus/ImageCLEF workshop on visual information retrieval evaluation*, pages 8–11, 2009.

Jussi Karlgren, Magnus Sahlgren, Fredrik Olsson, Fredrik Espinoza, and Ola Hamfors. Usefulness of sentiment analysis. In *Advances in Information Retrieval*, pages 426–435. Springer Berlin Heidelberg, 2012.

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)*. 2005.

Magnus Sahlgren, Amaru Cuba Gyllensten, Fredrik Espinoza, Ola Hamfors, Anders Holst, Jussi Karlgren, Fredrik Olsson, Pelle Persson, and Akshay Viswanathan. The Gavagai Living Lexicon. In *Proceedings of LREC*. 2016.

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 *Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*. 2013.

Oscar Täckström and Ryan McDonald. Discovering fine-grained sentiment with latent variable structured prediction models. In *Advances in Information Retrieval*, pages 368–374. Springer, 2011.

Ana von Klopp. But and negation. *Nordic journal of linguistics*, 1: 1–33, 1994.