Information Structure, Item Position in Clause, and Sentiment Analysis

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.**

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. Topic, Comment, and Information Structure

These experiments take instead as their starting point the notion of *information structure* of human language utterances. Linguistic theory suggests that human utterances can be understood in terms of *topic* and *comment*, where utterances are constructed to first establish the topic they are about and then to provide a comment on that topic. Halliday (1967-68); Hajičová et al. (1998) For the purposes of the present discussion, this would mean that the final portion of the utterance in the typical case is especially salient for understanding what is being said about the referents introduced in the former portion.

(1) a. What really surprises about X is its low-key quality and genuine tenderness.
b. X's "performance" is incredible!

Complicating this neat information theoretically sound convention, occasionally, for emphasis, informationally especially salient constituents in a sentence may be *topicalised*, i.e. moved to the beginning of the utterance from their default position later in the utterance.

(2) a. For its seriousness, high literary aspirations and stunning acting, the film can only be applauded.
b. It is the taste, not the nutritional value, which makes one want to go back.

III. The positional hypothesis

Position in an utterance matters: The location of the attitudinal word in the utterance is correlated with the strength it has as an indicator of author attitude. A scoring scheme which takes the position of the attitudinal item in an utterance will achieve better precision than one which does not. This can be operationalised to two more specific hypotheses:

(3) a. *Final position has higher weight than other positions*: by weighting up later mentions in a standard utterance, we improve precision for cases where both negative and positive terms are in an utterance.
b. *Topicalised position has higher weight than other positions*: by weighting up very early mentions in topicalised utterances, we improve precision for cases where both negative and positive terms are in an utterance.

The two hypotheses are partially contradictory, if no method beyond position is available to identify topicalisation.

IV. Starting points

Take a number of human-assessed utterances and establish whether a method which weights occurrences of attitudinal terminology differentially according to their position in the utterance.

A. Target notion

In practice more fine-grained sentiment analyses than positive and negative would seem to be necessary. (**?**) There are no benchmarking resources available outside our laboratory for experimentation on a broader palette of attitude or sentiment, and thus these experiments will focus on those macro sentiments on which other sentiments are typically and simplistically projected. Precision for positive and negative sentiment are reported separately and recall for each.[1]

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 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.. 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 IV, in sections ”greedy” and ”conservative”, respectively.

A. Distribution of polar items

First, each input utterance is divided into four sections, by length of the utterance, as shown in Figure II. Observed positions of positive and negative terms from the lexicon can then be tabulated per section. As shown in Table III, the attitudinal terms are unevenly distributed over the utterances. This lends some support to the positional hypothesis — apparently attitudinal terms tend to be utterance-initial or utterance-final rather than medial.

VI. BASELINE LEXICAL SENTIMENT ANALYSIS

The simplest conceivable approach, scoring polar items and assigning polarity by largest score yields the results given in Table IV.

[1]No F-scores are reported. I 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. (The only part way sensible argument for using the F-score given to me in repeated arguments with colleagues—even in face of lack of validity—is that it is easy to teach to undergraduate students.)

It has some special effects, but	you will not leave the theatre	with a good feeling, just a	wanting for your money back!
@twitter-id this little girl	in the Honda commercial	makes me think of	a young steph URL

TABLE II
EXAMPLE PARTITIONINGS

Set	q1	q2	q3	q4
	Positive			
RepLab	10 323	7 558	8 007	6 242
T & McD	1 008	798	775	876
Stanford	3 919	2 841	2 731	2 816
	Negative			
RepLab	5 168	4 228	4 458	4 099
T & McD	670	614	667	784
Stanford	2 994	2 452	2 377	2 679

TABLE III
DISTRIBUTION OF ATTITUDINAL TERMS OVER UTTERANCE LENGTH

A. Alternatives to lexical analyses

State-of-the-art performance is around 81% using standard machine learning classifiers, and best published results for binary classification on the Stanford data set appear to be around 87% accuracy. For a three-class setup (i.e. positive/negative/neutral), state-of-the-art accuracy is below 70%, and for fine-grained classification (five classes), state-of-the-art accuracy is below 50%. These figures are given by experiments machine learning classifiers trained over sufficient amounts of training data, which might not always be available for real-world commercial scenarios: annotating training data is costly and time-consuming. No such data are available for any of the commercial cases I have worked on in the past. Such classifiers are also relatively tightly bound to the material they have been trained on, and transfer to other genres or data sets hurts performance considerably.

VII. EXPERIMENT: POSITION

The observation motivates testing a modified attitudinal scoring scheme. A positional weighting scheme where the weight of every word in the utterance is increased by adding $0.1 \times position$ improves precision noticeably at some cost to recall for negative items, and the converse for positive items, as shown in Table IV.

Following the observation that attitudinal terms tend to occur in the beginning of the utterance, it is reasonable to experiment with the reverse strategy adding $(length\ of\ utterance - position) * 0.1$. This gives very similar results, again boosting precision at cost to recall for negative items and the converse for positive items, as given in Table IV.

The obvious third experiment is pivoting at the midpoint, weighting up both ends of the utterance visavi the middle stretch by adding $(length\ of\ utterance - position) * 0.1$ before the midpoint of the utterance and $0.1 \times position$ after, which gives the results given in Table IV. This combination gives on average similar results to the baseline, but with a clear recall improvement for positive items and a precision improvement for negative items.

VIII. CONCLUSIONS

These experiments show that taking position into account increases precision for negative items at a cost to recall and conversely increases recall for positive items at a cost to precision. This appears to be unconnected to the relative recall-precision vs negative-positive scores in the baseline condition, making it likely this is a constructional issue.

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.

Eva Hajičová, Barbara H. Partee, and Petr Sgall. *Topic–Focus Articulation, Tripartite Structures, and Semantic Content*. Studies in Linguistics and Philosophy. Kluwer, Dordrecht, 1998.

MAK Halliday. Notes on transitivity and theme in English (part 1–3). *Journal of Linguistics*, 3, 1967-68.

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.

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)*.

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.

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**
	INCREASING POSITIONAL WEIGHTING							
Set	Micro		Macro		Positive		Negative	
	p	r	p	r	p	r	p	r
RepLab	53.8	39.3	48.2	39.3	67.1	39.3	29.2	39.3
T & McD	42.7	53.8	43.5	55.6	37.8	66.0	49.3	45.2
Stanford	58.2	60.2	58.0	60.5	60.3	65.9	55.8	55.2
	DECREASING POSITIONAL WEIGHTING							
Set	Micro		Macro		Positive		Negative	
	p	r	p	r	p	r	p	r
RepLab	53.6	39.2	47.8	38.8	66.5	39.5	29.0	38.1
T & McD	43.2	54.4	44.3	56.5	38.2	68.4	**50.4**	44.7
Stanford	58.1	60.6	57.9	60.4	59.5	67.2	56.3	53.6
	MIDPOINT PIVOTED POSITIONAL WEIGHTING							
Set	Micro		Macro		Positive		Negative	
	p	r	p	r	p	r	p	r
RepLab	**53.9**	**39.4**	48.1	39.2	66.8	**39.5**	**29.4**	39.0
T & McD	43.1	54.3	44.0	56.2	38.0	**67.3**	50.0	45.2
Stanford	**59.1**	**61.6**	**58.9**	**61.4**	60.6	**67.4**	**57.1**	55.4

TABLE IV
BOUNDS, BASELINES, AND RESULTS