Usefulness of Sentiment Analysis

Jussi Karlgren, Magnus Sahlgren, Fredrik Olsson, Fredrik Espinoza, and Ola Hamfors

Gavagai, Stockholm, Sweden

Abstract. What can text sentiment analysis technology be used for, and does a more usage-informed view on sentiment analysis pose new requirements on technology development?

1 Human emotion, attitude, mood, affect, sentiment, opinion, and appeal

Analysis of sentiment in text is a new and rapidly growing field of study and application. This paper outlines some application areas for sentiment analysis technology, and discusses what requirements a technology for sentiment analysis of text should be able to answer to. The human sensations of emotion, attitude, mood, affect, sentiment, opinion, and appeal all contribute to the basic categories of sentiment analysis of text, but they have been studied in their own right for a long time. Traditionally, this has been done in the behavioural sciences; [9] but today also by information technologists, especially with respect to interaction design . "Emotion" , "attitude" , "mood" "affect", "sentiment" , and "appeal" are everyday words. No consensus beyond the general vernacular usage of the most common terms can currently be assumed, but mostly the usage tends to hold that affect or affective state is the more general term, emotion a momentary, mostly conscious sensation, and mood an affective frame over a longer time span, not necessarily consciously acknowledged by its holder.

These affective aspects of human behaviour and information processing are studied in various ways with variously differing perspectives, but the assumptions of most researchers is that people are in continuously changing affective states of some sort; and that activities people engage in have emotional impact and that their decision making, behaviour, and performance are informed by the affective state of the user. This appears to be true even for very mundane tasks such as workplace tasks or accessing information items, but most importantly for the purposes of this paper, in *producing* and *understanding information items*, and, it is assumed, even to the extent that mood, with respect to some topic or facet of life, will colour and influence the understanding, generation, or processing of information on another quite different topic.

Sentiment analysis of text typically assumes that lexical items found in the text carry attitudinal loading. Previous work on the loading of individual features and the affective reaction of human subjects to linguistic items on the level of words and terms [16] or still images [15] quite often take "emotion labels" to be

The final publication is available at Springer
via http://dx.doi.org/10.1007/978-3-642-28997-2

given, accepted, and comprehensible to test subjects as a basis for the study of correlation between emotions of various kinds. [13] This is a fairly far-reaching simplified model of human emotion, but human emotion in sentiment analysis of text is typically simplified further to be measurable somewhere on a scale from positive to negative. This paper argues that this simplified knowledge model in fact makes the task of informed sentiment analysis of text more complex than it should be.

2 Sentiment Analysis of Text as an Applied Technology

2.1 Consumer Attitude

It is widely understood that word of mouth phenomena play an important role for informing and affecting consumer decisions, and in building and destroying brand reputation. User-generated Internet content such as forums, blogs, and BBSes facilitate this process, and are undermining the authoritative status historically carried by traditional media, especially in markets where the authority and independence of traditional media is low for political or commercial reasons. Applications to address the analysis of consumer mood are drivers for much of the research done in the area of textual sentiment analysis. [10]

The application of sentiment analysis to consumer attitude can be viewed from two perspectives:

1. tools for market analysts to refine the offerings produces wish to make available and known to consumers or
2. tools for consumers to mine experiences of peers in face of a challenging purchase decision.

In the latter case, consumers might wish to find reviews or comments made about some product or service by others, especially critical ones, and rank them by authoritativeness, reliability and throughness or some other quality criteria. In the former case, producers might be most interested in aggregating the general mood of consumers visavi their product or service. In both cases, a broad recall of opinion is a first requirement for useful exploitation of the technology; a further analysis will benefit from a fine-grained break-down by facet.

In the latter case, the hypothesis is that a broad spectrum of peer opinions is a useful supplementary or overriding source of knowledge for making informed purchase decisions. In the former case, the hypothesis is that a broad range of consumer opinions can be mined by running a high-recall sieve over vaguely expressed opinionated text, capturing signal which otherwise might not have been detected.

2.2 Investment Trends

Herding behaviour is an attractive model for explaining and modelling certain price movements on the stock market. The recent and emerging research direction to model actors' tendencies and biases to move in concert or in averting

and seeking risk and rewards asymmetrically in accordance with some underlying latent behavioural variables takes as a point of departure that most of the movements that can be observed in trading data have unknown, unknowable or even random causes. [31, 4] The quote: *"... we consider a set of N investors, each of whom has either bullish or bearish opinion ... At every time step each of N investors can change [opinion]. ... The probability of [changing opinion] ... depends only on the bullish sentiment described as the number of bullish investors among the total of N investors. The number of bullish investors then forms a Markov chain ... "* [25] is typical in that the model described may be computationally sophisticated, but that the information sources which cause the processes it describes are viewed to be beyond the scope of the model itself.

These - from the standpoint of information science - rather weak starting points allow for the injection of new information into the predictive models. Sentiment analysis of text has been tested and found to carry some signal in several recent reports [20, 24, 21, 30, 3, 2]. The hypothesis in these cases is one or a combination of the following:

1. the people who trade indicate their preferences and reveal their deliberations in advance of action by writing about them in public fora, or
2. the public sentiment visavi some tradable asset can be found through judicious analysis of public expressions of opinion with respect to this asset and that this sentiment is a fair estimate of the sentiment of traders, or
3. there is such a large volume of high-quality informed analyses available in published media texts that even with a competent topical search engine, no human reader can make sense of it all without an automatic opinion aggregator, or
4. the public mood in general, not necessarily bound to expressions of sentiment visavi some tradable asset, will act as a filter which informs trading decisions traders make with respect to tradable assets of all or alternatively some specific kinds.

2.3 Security Concerns

Similarly to market and financial analysts, intelligence and security analysts want to identify and keep track of certain user-initiated discussions and postings on forums, blogs, newsgroups, and other user generated web content. This domain has at least two distinct usage scenarios:

1. tracking public mood to detect and predict security threats or disruptive public behaviour, the hypothesis being that e.g. the tendency for public protest can be monitored and public action may be ignited or catalysed by public communicative behaviour and reflected in the sentiment expressed in public text; and
2. identifying and monitoring certain individuals or certain documents as being threatening, risky, abusive or help-seeking, the hypothesis being that individuals who express threatening or abusive sentiments in public texts can

be reliably identified and that bluster can be usefully distinguished from imminent bite through text analysis.

The first task is clearly related to the one of tracking public mood in the previous two domains of consumer sentiment and investment mood: a broad-recall tracker to find risk of public sentiment (or the sentiment of some targeted group) boiling over. The second task is a very different task, that of identifying specific sentiments and specific indication of future action in fine-grained analysis of specific texts or text streams. This is not unlike the general task of *author profiling*, and *authorship attribution* where the technical hypothesis is that a computational analysis of language and observable linguistic items may be more exact and more revealing than a human reader would be able to achieve.

3 Sentiment Analysis as an Engineering Question

Given the above descriptions of information need and potentially lucrative and fruitful application domains for large-scale sentiment analysis of text the application of known text analysis tools from information retrieval would seem to be straightforward. Many recent approaches to sentiment analysis have been patterned on search technology, under the reasonable assumption that most of the attitudinal signal in text is lexical.[17] Under this assumption, the procedure for computing sentiment loading for text is straightforward: occurrences of lexical items can be counted, those occurrence counts aggregated and tabulated, the items weighted according to previously observed occurrences in attitudinally loaded texts and the resulting statistics processed by categorisation algorithms originally developed for lexically based topical categorisation.

And the sound quality - my God!
Raymond left no room for error on his recordings and it shows.
Definitely one of the better tracks on the album.
Wow, could have been a expansion pack.

Table 1. Some benchmark example sentences (From [27])

Benchmarks, which drive research in this area are consumer reviews with a text-level annotation of author attitude and some collections of sentence- or clause-level attitudinal items. [29, 18, 1] Different types of benchmark give rise to different optimisation strategies for the algorithms employed. If the task is understood to be driven by a text level analysis the weighting of lexical items and the ideal categorisation of texts given those items will be different from the task of identifying mentions, clauses, or sentences. The first task — that of classifying texts such as review texts as being positive or negative typically achieves an accuracy of about 70 to 90 per cent, depending on topical area. Agreement for human annotators is quite high. Results for the second task — that of classifying

sentences as being positive or negative typically are distinctly lower, about 60 to 80 per cent, as is the agreement among human language annotators. This is not surprising given the open-ended nature of human language — the examples in Table 1 are difficult to assess as being positive or negative without a broader discourse context and knowledge of the expressive habits of the author.

4 Challenges

Some of the application scenarios given above are based on a fairly closely patterned high-precision analysis of attitude visavi some target of interest, akin to information extraction; some, by contrast are based on a broad high-recall analysis of public mood. This distinction is obviously addressable through an informed service design — but what sorts of analyses, processing models, and knowledge representations would best contribute to testing some of the above underlying hypotheses and best provide a basis for acting upon them if they are found fruitful? What are the immediate challenges for bringing the technology to productive use?

4.1 Multilinguality and Cross-Domain Portability

A first and very obvious challenge — well discussed in the field — is the issue of moving research developed on a few resource-rich languages to further *low-density languages*. Focussing on the largest languages is not a sustainable strategy in an increasingly multilingual world; it is arguably not sufficient today, and will most definitely not be sufficient in the (near) future, not if high recall is a crucial quality criterion for the services under consideration. The understanding that attitude is a cross-linguistic pan-human communicative basic category is well understood in the field and even several benchmarks are cross lingual. [23] A closely related challenge is that of *domain specificity* or overtraining. Several research efforts focus on transferring models found reliable in one topical domain to another (e.g. [1]). While these two challenges share several central research questions, the contact surfaces between methodologies to address them appear to be limited.

4.2 Scalability

A second and as obvious challenge is that of *scalability*. Since much of the application potential is motivated by high recall and coverage any approach for automatic analysis must cope with very large data streams and have readiness to accomodate even more (and continuously increasing amounts of) data. It would not be convincing to deploy methods which must fall back on sampling in face of constantly growing and changing data streams if the hypothesis is that each expression of attitude contributes to the understanding. It would also not be realistic to deploy methods which require periodic non-incremental recompilation or re-organization to keep up with the evolving nature of the data:

the data are in constant flux, and the methods should accomodate that steady change continously.

4.3 New text

A third obvious challenge is that of *robustness*. Real human-produced language data is not lean, clean and neat. New text, non-edited, is different from the language of traditional linguistic grammars. [11] Every processing model which presumes stability, order and consistency will break down when exposed to actual language use; a model intended to accommodate new text should accept that every sentence in our language *is in order as it is*, without pre-processing, re-editing, or normalisation, and with mechanisms ready to accept new conventions, misspellings, non-standard usage, and code switching. Models which rely on non-trivial knowledge-intensive preprocessing (such as part-of-speech tagging, syntactic chunking, named entity recognition, language identification, etc) or external resources (such as thesauri or ontologies) will always be brittle in face of real-world data. Examples in Table 2 are examples which any fixed lexical resource will have trouble accommodating.

good		bad	
great 0.91	√	weird 0.86	√
prefect 0.83	√	sucky 0.86	√
perfect 0.83	√	scary 0.86	√
pristine 0.81	√	cool 0.85	≠
stable 0.80	√	nasty 0.84	√
grat 0.80	√	dumb 0.84	√
fantastic 0.80	√	sad 0.84	√
flawless 0.79	√	lame 0.84	√
mint 0.79	√	creepy 0.84	√
immaculate 0.79	√	stupid 0.84	√
geat 0.78	√	dog 0.84	?
excellent 0.78	√	shitty 0.83	√
working 0.77	√	quiet 0.83	?
decent 0.77	√	romantic 0.83	?
excelent 0.77	√	wierd 0.83	√
ggod 0.77	√	blind 0.83	?
nice 0.77	√	prayer 0.83	×

Table 2. The nearest neighbors in a distributionally aggregated semantic word space to "good" and "bad." √ indicates viable (near) synonyms, × indicates errors, **?** indicates uncertain cases, and ≠ indicates antonyms. The numbers are closeness scores, 1.0 denoting a perfect synonym. (From [19]).

4.4 Texts, Utterances, Clauses, and Fragments

A fourth — somewhat less obvious challenge — is that of *granularity of analysis*. The question of what entity is the best carrier of mood or attitude is as yet undetermined. Documents would appear to be the wrong kind of entity to operate on if the goal is to track attitude visavi some target of interest, since propositions or predications are likely to be expressed in utterances rather than entire arbitrarily long texts, the topical coherence of which can be called into question, especially in the case of spontaneously authored text by inexperienced writers. The general mood of an author may be something that can be detected on text level, or possibly on some other aggregation of lower level utterances.[12] Some recent approaches explicitly work with the distinction and interplay between text-level author attitude with respect to some topic and clause- or predication-level opinion visavi some facet of that topic.[28]

5 Requirements for an Appropriate Knowledge Representation: Beyond Positive and Negative

Given the above observations we argue that the benchmark and the approaches of research hitherto only partially address the needs given by the example applications. Optimising the task of weighting lexical items with respect to the two macro-level sentiments POSITIVE and NEGATIVE and using those resulting lexical resources for categorisation is obviously feasible: numerous experimental publications testify to this. For the purposes of fruitful application to task, however, the question is what the application potential of that generated knowledge is. While most sentiments, once identified, can be mapped to a positive-negative polarity dimension with some accuracy, the value of doing so is doubtful since the primary sentiments are more likely to be of value than the projection onto the polarity axis, and the task of coercing lexical items into categorising text into positive and negative may even be more challenging than working towards the primary sentiments.

Sentiment models in the study of human emotion are of two major types. Categorial models list emotions are in a palette of salient and recognisable basic emotions such as in the most well established "Big 6" (later amended to the "Big 18") list of emotions, based on work by Paul Ekman [8] and originally inspired by Charles Darwin [6]. Dimensional representations assess emotions along dimensions such as "Pleasure", "Arousal", and "Dominance", based on work by Albert Mehrabian. [14] Categorial models are typically used in studies where the objective is to recognise one of a set of emotions for some purpose, or to test the efficiency of some analysis algorithm. However, most projects or studies which apply models of human emotion or affective state to some task or interaction in general use variants of dimensional models, and sometimes define a model specifically suited for application to information access [7, 5]. Mood in this context can be thought to be an underlying moderator of human action and a representation of background information used as a basis for assessment and

judgment even without conscious attribution of an emotion towards a target notion. [22] These various more complex models could provide a basis for a deeper sentiment analysis of text, and a more applicable one.

Going back to the application scenarios given in Section 2 above, rather few of them are immediately related to a positive-negative axis. For the consumer attitude application scenario, "recommend", "endorse", "surprise", "disappoint", "satisfy" or "delight" are more salient examples of attitudes than mere positive or negative. For the investment trends application scenario, "optimism", "pessimism", "worry", and "uncertain" are crucial sentiments to mine (as shown in some of the published experiments made on micro-blog data with respect to trading indices). For the security concern application scenario, "anger", "frustration", "violence" are examples of the most central sentiments. A useful knowledge representation should not coerce the analysis into one single dimension, especially if that projection requires more tuning work than a multi-polar model would. This has been recognised as a challenging modelling task in content analysis experimentation e.g. in the General Inquirer project [26] which uses dozens and scores of special purpose word lists which are then convoluted by a editorial decision into a task-specific processing model.

The challenges given in Section 4 above add to the requirements. Any model should be adaptible to new signal streams such as new topics or new languages with a minimum of manual intervention, it should ideally scale incrementally without compilation or compression steps, it should be robust and adaptible, and it should aggregate fine grained assertions by the author without assuming coherence in text.

These requirements are only partially met by research following along the lines given in Section 3.

Most of the recently published research efforts accept the positive-negative axis of analysis as given, and work hard at improving experimental scores for a benchmark test set of sentences annotated along it. That effort is misguided if the baseline given by human performance at the task is unimpressive — the task may be simply be impossible to do well and is likely to be of little application potential if optimised beyond its reasonable level of accuracy. Research results, even if they show improvement over previous effort, will be of little sustainable impact if they do not address a knowledge model which the application in mind might actually find useful.

Much of the recently published research efforts make use of pre-compiled knowledge resources of various types, and of processing models which presuppose pre- or post-processing and repeated recompilation of some central processing component. Research results which do not scale or which rely on non-transferable knowledge are likely to be of little use for a commercial stakeholder unless they expressly address the question of transfer, maintenance, and incremental adaptation.

In conclusion, a knowledge representation for sentiment analysis of text for real world applications must be multi-polar, not restricted to positive and negative sentiment. It must be adaptible and transferable from language to language,

domain to domain, and sentiment model to sentiment model without too much training and too much knowledge engineering. It must be robust, dynamic, and scalable. It cannot be built to go off-line for training and recompilation but must accommodate on-line updating of its knowledge models. And it does not necessarily need to do well on current research benchmarks. The bottleneck for building real and useful applications is not a question of parameter tuning but of a realistic knowledge and processing model.

References

1. Blitzer, J., Dredze, M., Pereira, F.: Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In: Annual Conference of the Association of Computational Linguistics (ACL) (2007)
2. Bollen, J., Mao, H., Zeng, X.J.: Twitter mood predicts the stock market. Journal of Computational Science 2, 1–8 (March 2010)
3. Bollen, J., Pepe, A., Mao, H.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: Proceedings of the WWW conference (2010)
4. Brody, D.C., Hughston, L.P., Macrina, A.: Credit risk, market sentiment and randomly-timed default. In: Crisan, D. (ed.) Stochastic Analysis. Springer Verlag (2010)
5. Chanel, G., Rebetez, C., Bétrancourt, M., Pun, T.: Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In: MindTrek 08. ACM, New York (2008)
6. Darwin, C.: The Expression of the Emotions in Man and Animals. John Murray, London (1872)
7. Dunker, P., Nowak, S., Begau, A., Lanz, C.: Content-based mood classification for photos and music: a generic multi-modal classification framework and evaluation approach. In: MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval. ACM, New York (2008)
8. Ekman, P.: An argument for basic emotions. Cognition and Emotion pp. 169–200 (1992)
9. James, W.: What is an emotion? Mind pp. 188–205 (1884)
10. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. JASIST 60, 2169–2188 (2009)
11. Karlgren, J. (ed.): New Text. Proceedings from the workshop on New Text: Wikis and blogs and other dynamic text sources, held in conjunction with EACL. ACM, Trento, Italy (2006)
12. Karlgren, J.: The relation between author mood and affect to sentiment in text and text genre. In: ESAIR'11, Fourth Workshop on Exploiting Semantic Annotation in Information Retrieval. Glasgow, Scotland (October 2011)
13. Kuppens, P., van Mechelen, I., Smits, D.J.M., de Boeck, P.: Associations between emotions: Correspondence across different types of data and componential basis. European Journal of Personality 18, 159–176 (2004)
14. Mehrabian, A., Russell, J.A.: An approach to environmental psychology. M.I.T. Press, Cambridge, Massachusetts (1974)
15. Mikels, J., Fredrickson, B., Larkin, G., Lindberg, C., Maglio, S.: Emotional category data on images from the International Affective Picture System. Behavior Research Methods pp. 626–630 (2005)

16. Morgan, R.L., Heise, D.: Structure of Emotions. Social Psychology Quarterly 51(1), 19–31 (1988)
17. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundation and Trends in Information Retrieval 2(1-2), 1–135 (2008)
18. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of EMNLP 2002 (2002)
19. Sahlgren, M., Karlgren, J.: Terminology mining in social media. In: The 18th ACM Conference on Information and Knowledge Management (CIKM 2009). Hong Kong (Nov 2009)
20. Schumaker, R.P., Chen, H.: Evaluating a news-aware quantitative trader: The effects of momentum and contrarian stock selection strategies. Journal of the American Society for Information Science and Technology 59(2), 247–255 (2008)
21. Schumaker, R.P., Chen, H.: A discrete stock price prediction engine based on financial news. Computer 43(1), 51–56 (2010)
22. Schwarz, N.: Feelings as Information: Implications for Affective Influences on Information Processing. In: Martin, L., Clore, G. (eds.) Theories of Mood and Cognition. Lawrence Erlbaum, Mahwah (2001)
23. Seki, Y., Evans, D.K., Ku, L.W., Sun, L., Chen, H.H., Kando, N.: Overview of multilingual opinion analysis task at NTCIR-7. In: Proceedings of the 7th NTCIR Meeting. NII, Tokyo (2008)
24. Shih, C.C., Peng, T.C.: Building topic/trend detection system based on slow intelligence. In: DMS2010 (2010)
25. Shmatov, K., Smirnov, M.: On some processes and distributions in a collective model of investors' behavior. SSRN (2005), http://ssrn.com/abstract=739504
26. Stone, P.J., Dunphy, D.C., Smith, M.S.: The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Oxford, England (1966)
27. Täckström, O., McDonald, R.: Discovering fine-grained sentiment with latent variable structured prediction models. In: Proceedings of European Conference on Information Retrieval. Dublin (2011)
28. Täckström, O., McDonald, R.: Semi-Supervised Fine-Grained Sentiment Analysis with Latent Variable Structured Conditional Models. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland (2011)
29. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation 39, 165–210 (2005)
30. Zhang, X., Fuehres, H., Gloor, P.A.: Predicting stock market indicators through twitter "I hope it is not as bad as I fear". Social and Behavioral Sciences (2010)
31. Zhou, W.X., Sornette, D.: Renormalization group analysis of the 2000-2002 anti-bubble in the US S & P 500 index: Explanation of the hierarchy of 5 crashes and prediction. Physica A: Statistical Mechanics and its Applications 330, 584–604 (December 2003)