Sentiment Analysis and the Use of Extrinsic Datasets in Evaluation

Ann Devitt, Khurshid Ahmad

School of Computer Science & Statistics,
Trinity College Dublin Ireland
Ann.Devitt@cs.tcd.ie, Khurshid.Ahmad@cs.tcd.ie

Abstract

The field of automated sentiment analysis has emerged in recent years as an exciting challenge to the computational linguistics community. Research in the field investigates how emotion, bias, mood or affect is expressed in language and how this can be recognised and represented automatically. To date, the most successful applications have been in the classification of product reviews and editorials. This paper aims to open a discussion about alternative evaluation methodologies for sentiment analysis systems that broadens the scope of this new field to encompass existing work in other domains such as psychology and to exploit existing resources in diverse domains such as finance or medicine. We outline some interesting avenues for research which investigate the impact of affective text content on the human psyche and on external factors such as stock markets.

1. Introduction

Proponents of sentiment and polarity analysis have undertaken the challenge of automatically identifying levels of meaning in language which may not be explicitly represented in surface linguistic structure. The expression of sentiment is one of the key factors in making natural language based communication ambiguous, imprecise and uncertain. Given the twin challenges of identifying levels of implicit meaning and its inherent ambiguity, it is critical not only to have programs for extracting sentiment automatically from text but also to have transparent and objective batteries of tests for evaluating such programs. This is the burden of argument in this paper. We ask what effect the articulation of (implicit) sentiment has on readers/listeners, in particular in terms of their mood, beliefs and evaluation perspectives.

The articulation of sentiment in language involves linguistic devices related to affect and metaphor. The symbiotic interplay between sentiment and language has been the subject of studies in areas as diverse as cognitive linguistics and investor psychology, and in tasks from deciphering the plans of terrorists to understanding the emotional charge of emergency phone calls. Given the broad swathe of coverage, it is critical that in evaluating programs designed to identify implicit levels of meaning, the evaluators investigate two key issues:

1. the use of existing extrinsic datasets in conjunction with other more traditional measures of emotion;
2. the effects of emotion on cognitive processes and behaviour.

It is these two issues which are under investigation within a multi-disciplinary sentiment analysis project underway at Trinity College Dublin. Analysis and development of evaluation methodologies for the outputs of sentiment analysis programs are hence of primary importance, as in the EAGLES project (King, 1999), and the topic of discussion here. Our proposals can be considered as complementary to existing approaches, a means of validating results and opening new avenues for research rather than replacing them.

The current state of the art for evaluating sentiment analysis systems in computational linguistics is set out in section 2. Some limitations to existing approaches and the motivations for devising alternatives are discussed in section 3. Possible evaluation methodologies and proposed studies to examine their possible merits and demerits are discussed in section 4. We conclude with a synopsis of the challenges presented here and our proposals to tackle them in section 5.

2. Current Evaluation Methodologies

In computational linguistics and information extraction, performance evaluation in general is reported as precision and recall of the system on a gold standard of human judgments of "correct" output in the domain, be it parse trees or disambiguated word senses. In sentiment analysis, the gold standard consists of human judgments of sentiment in text where the type and level of human annotation depends on the granularity and purpose of the sentiment analysis system. As regards level, the gold standard tags may be at word, phrase, sentence or text level. For example, in sentence 1 below the terms "disastrous" and "strength" could be tagged as conveying negative and positive affect, respectively, while the sentence as a whole is a positive evaluation of the current company status. The sentiment conveyed by the full text from which this excerpt is drawn could be strongly negative or neutral or weakly positive, or whatever the writer intended or the reader interprets.

(1) After a disastrous start, the company appears to be going from strength to strength.

As for the type of annotation, different sentiment analysis tasks, such as subjectivity or polarity identification or attribution, require different kinds of annotation. For example, for subjectivity identification, binary subjective/objective tags are sufficient for testing. For example 1, this tag would be subjective as the text conveys a subjective evaluation of a company. While for most polarity identification systems, binary positive/negative tags or a graded scale of positive-negative ratings are required. In example 1 the annotator will assign a positive tag, degree of this tag could range

from weak to strong depending on the annotators perspective and the specification of the evaluation task. Given the range of potential annotation criteria for sentiment analysis, it is imperative that researchers define appropriate resources for evaluation of specific sentiment analysis tasks. The following sections outline the main sources for evaluation resources currently available.

2.1. Manually Annotated Corpora

One of the most valuable and comprehensive resources available in the domain of sentiment analysis is the Univeristy of Pittsburgh MPQA Database which consists of 10,000 sentences of world news tagged at levels of granularity up to phrase–level and annotated for private states as well as private state source and target (see (Wiebe et al., 2005) for details). With 40 hours of training, the annotation project took between 3-6 months (part-time) for each of 3 annotators and achieved inter-annotator agreement of approximately $\kappa > 0.8$. This tagged corpus is publicly available and a first of its kind. However, there are two key limitations. Firstly, with only 10,000 sentences, it is quite small to be used as both training and test data for the myriad of sentiment analysis systems. Secondly, it was not intended to provide text–level annotations nor to give a compositional account of subjectivity from phrase up to text–level, therefore it cannot be used for text–level analysis systems.

In addition to the MPQA database, there are other corpora available which have been manually labelled at different levels of granularity. For example, Devitt and Ahmad (2007) use a small corpus of news stories on an airline takeover bid annotated at text–level for sentiment polarity intensity with inter–rater agreement of $\kappa = 0.55$. Almas and Ahmad (2007) also use a manually annotated corpus of world news. The sentence–level annotations are binary polarity ratings with an option to tag sentences for absence or ambiguity of sentiment polarity. Such resources provide useful validated human responses for system testing but still are too small for machine learning training.

2.2. Reviews

A second set of resources are reviews for products and services posted on the internet (Pang et al., 2002; Turney, 2002). These provide a varied source of affective text usually with an explicit text–level polarity intensity ranking, such as recommendation stars. These ratings constitute a manual annotation of the associated texts. The disadvantage of these reviews is that they are entirely domain–specific and limited to consumer items. Furthermore, the relationship between the review and the rating is not transparent and may not reflect the text's affective content in the same way as the manual annotations set out above.

2.3. Taxonomies of emotion

For word–level sentiment analysis, the domains of psychology and content analysis offer several lexica of sentiment for evaluating system performance. These lexica, based on psycholinguistic experimentation, aim to validate models or taxonomies of emotion. For example, the lexica in Stone's General Inquirer or in Whissell (1989) are a dimensional representation of emotion whereas Ekman and Friesen (1971) set out a categorical model of basic universal emotions. These sources, alone or in combination, provide a gold standard for word–level semantic orientation analysis systems with the strong foundation of experimental psychology.

3. Limitations of Existing Resources

As in many emergent fields, the provision of resources can be the vital input to new research developments. However, what is currently available in this domain is limited, due to the novelty of the work and also the cost and difficulty of producing resources. This causes problems of data sparsity for analysis and training of computational approaches as well as an over-reliance on existing resources for evaluation, skewing performance results with respect to these resources. This phenomenon is not uncommon in computational linguistics where systems are tuned to optimum performance on an accepted gold-standard dataset but may under-perform on other datasets.

In addition, there are theoretical grounds for examining alternative means of evaluating the performance of systems which aim to estimate emotional content of text automatically. The primary issue regards the elicitation of human sentiment judgments for a gold standard which is by its nature a problematic task. Unlike parsing or word sense disambiguation, there may not be a single "correct" judgment for a text, given that sentiment is highly subjective. Aboulafia et al. (2001) note that in HCI experiments on affect, respondent responses may be pre-determined or at least strongly affected by their mood at a given time. Respondent mood is not a variable which has been examined in sentiment annotation tasks to date. It could be accounted for by respondents reporting a mood variable or perhaps by following experimental psychology and "inducing" a good/bad mood in respondents prior to the annotation task. The high inter-rater agreement of the MPQA annotation project would suggest that this subjectivity can be overcome. However, this project included a lengthy (and expensive) training period which suggests that estimating "emotion" in isolation is a difficult and costly process. It may be more useful to determine the *effect* of text on respondent behaviour, rather than relying on one-dimensional, potentially subjective, self-reported sentiment judgments. Existing work in the domains of cognitive psychology and behavioural finance, for example, which examine the influence of emotion on cognitive processes such as memory and decision–making can be leveraged here. The alternatives for evaluation set out in the following section go some way to addressing these issues.

4. Evaluation Methodologies

4.1. Extrinsic Data Sources

The use of existing non-linguistic data sources to evaluate a computational linguistics system can be an inexpensive way of testing systems on very large datasets. For example, econometrics has long used numeric variables as proxies for sentiment in the markets. These proxies could be used as an external measure of sentiment to correlate with

news data from the financial domain. The proxy indicators are often generated from financial variables such as stock returns or volatility measures but can also include explicit sentiment metrics such as the Yale investor confidence indices generated from surveys. The great advantage of using such a dataset is that very long time periods can be studied to give statistically significant results. Tetlock (2007) has adopted this approach to study the impact of news on financial markets using factor analysis to analyse the news content and Ghose et al. (2007) used sales data to evaluate opinion in on-line reviews. Other domains also offer similar extrinsic datasets which can be leveraged for evaluation, such as using tourist numbers or GDP to evaluate country-specific news, or medical diagnoses or treatment records to evaluate sentiment in doctor reports.

We have carried out preliminary analyses of financial market data with a human gold standard sentiment judgment which suggested a correlation between stock prices and evaluations (Devitt and Ahmad, 2007). This experiment will be extended to a longer time series and to other financial datasets (e.g. oil prices, stock market indices) to provide evidence for the hypothesis that financial data is correlated with human estimations of sentiment in relevant news and can thus be used as a proxy for human sentiment judgments in evaluation of a sentiment analysis system. This pilot study can be extended to domains outside finance.

4.2. Emotional Response as Behaviour

An estimate of emotional responses to text in terms of behaviour rather than ratings promises another evaluation metric and has two advantages. Firstly, although it may be no less costly to derive a gold standard of human judgments than one of behavioural responses, behavioural experiments do not require a training period, relying on immediate reactions rather than considered, or potentially mediated, evaluations. Secondly, this approach looks at the broader research question of "emotion" in the context of other cognitive processes. The experimental set-up in this case should replicate and draw on existing work in cognitive psychology and related disciplines which examines the role of emotion and mood in processes such as risk assessment and problem-solving (Finucane et al., 2000; Kaufmann and K.Vosburg, 1997). A sample experiment to examine the effect of text on risk evaluation in the financial domain could use financial professionals, divided into a control and an experimental group, who report their willingness to trade before and after reading news extracts. The control group are shown the original extracts while the other group is shown texts modified to include/exclude sentiment-bearing terms, constructions or devices.[1] The experimental aim here is to determine whether these affective features do impact on behaviour and in what way. This experimental set-up could be adapted to other domains by varying the behavioural variable under investigation and of course the text sources. Results can be then be used in the design or analysis of a system to detect these features automatically.

[1] Our thanks to Professor Ravi Dhar at the Yale University Center for Customer Insights for his suggestions for conducting such an experiment.

5. Conclusions and Future Work

Given the expense in terms of money and time of generating human gold standard judgments and the tendency then to rely on a restricted set of gold standards for all evaluations, it is advisable in many fields of computational linguistics to look for alternative, cheaper means of evaluating system performance on a given computational linguistic task. This can be in the form of a data source which is non-linguistic or in fact external to the task at issue, positing a relationship between the task and the data source and extrapolating the results from one to the other. It may require broadening the terms of the evaluation. As regards extrinsic data sources, we are currently undertaking an analysis of potential long-term correlations between media reporting in the main Irish broadsheet, the Irish Times, with stock market data for the celtic tiger years in Ireland. In addition we are establishing the first Irish market sentiment survey which we intend to include in an investigation of current media sentiment and stock market movements. As regards evaluating the effects of sentiment in text, the focus must shift to sentiment as emotional response and evaluating this as is currently done in psychology. The proposed experiment in section 4 will be undertaken with finance students and professionals using an existing corpus of texts regarding an aggressive airline takeover bid with respondents reporting their willingness to buy shares in either company and their evaluation of the risk involved in this transaction. The control group responses constitute a baseline reliant on the facts of the takeover while the experiment group give their responses based on highly emotionally charged accounts of these same facts. In this way, we can draw on methodologies of other domains which may be more established in this area and focus on cognitive, rather than quantitative, aspects of the task in hand, giving us insights perhaps into the processes and effects rather than just the outputs of the task. While these proposals are being implemented, this abstract aims to open a debate on evaluation of sentiment analysis systems that may generate new questions and interesting avenues for future research in the field.

6. References

A. Aboulafia, L. Bannon, and M. Fernstrom. 2001. Shifting perspective from effect to affect: some framing questions. In *Proc. of Affective Human Factors Design*, pages 508–514.

Yousif Almas and Khurshid Ahmad. 2007. A note on extracting sentiments in financial news in english, arabic and urdu. In *Proc. of Workshop on Computational Approaches to Arabic Script-based Languages*, pages 1–12, Stanford.

Ann Devitt and Khurshid Ahmad. 2007. Cohesion-based sentiment polarity identification in financial news. In *Proceedings of ACL 2007*, Prague, Czech Republic.

Paul Ekman and W. V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17:124–129.

Melissa L. Finucane, Ali Alhakami, Paul Slovic, and Stephen M. Johnson. 2000. The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, 13(1):1–17.

A. Ghose, P. Ipeirotis, and A. Sundararajan. 2007. Opinion mining using econometrics: A case study on reputation systems. In *Proceedings of ACL 2007*, Prague, Czech Republic.

Geir Kaufmann and Suzanne K.Vosburg. 1997. Paradoxical mood effects on creative problem-solving. *Cognition & Emotion*, 11(2):151–170.

Maghi King. 1999. Evaluation design: the eagles framework. In *Evaluation of the Linguistic Performance of Machine Translation Systems, Proc. of the Konvens'98*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proc. of EMNLP'02*, pages 79–86.

Paul C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. of ACL'02*, pages 417–424, Pennsylvania. Association for Computational Linguistics.

Cynthia Whissell. 1989. The dictionary of affect in language. In Robert Plutchik and Henry Kellerman, editors, *Emotion: theory research and experience*, volume 4, The measurement of emotions. Academic Press, London.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.