An opinion about opinions about opinions: subjectivity and the aggregate reader

Asad Sayeed
Computational Linguistics and Phonetics / M^{2}CI Cluster of Excellence
Saarland University
66123 Saarbrücken, Germany
asayeed@coli.uni-saarland.de

Abstract

This opinion piece proposes that recent advances in opinion detection are limited in the extent to which they can detect important categories of opinion because they are not designed to capture some of the pragmatic aspects of opinion. A component of these is the perspective of the user of an opinion-mining system as to what an opinion really is, which is in itself a matter of opinion (metasubjectivity). We propose a way to define this component of opinion and describe the challenges it poses for corpus development and sentence-level detection technologies. Finally, we suggest that investment in techniques to handle metasubjectivity will likely bear costs but bring benefits in the longer term.

1 Introduction

Opinion mining, also known as sentiment analysis (Pang and Lee, 2008), is a relatively recent area of research in natural language processing. It has grown very quickly as a research area, developing around a small number of basic approaches. However, these approaches are based on particular definitions of opinion, assumptions about opinion expressions, and evaluation practices that we believe need to be expanded in order for sentiment analysis to reach new domains and applications.

We are not the first to express concern over the direction of sentiment analysis as a field. This paper seeks to further expand upon the views expressed in Alm (2011) that prevailing evaluation concepts in sentiment analysis limit the kinds of models we can build, particularly through the encouragement of a focus on "high-performing" systems.

The central thread that connects our view of the field is the idea that the basis of standard techniques and evaluation in information retrieval and extraction that underlie existing approaches needs to be rethought for applications that are inherently subjective and that the field needs to return to more theoretical groundwork. This will entail sacrificing some of the performance gains made in recent times, as well as potentially reducing the capacity for easily comparable research that has been gained by the rapid adoption of corpora that are very easily produced, shared, and used.

This problem is particularly relevant in the expansion of sentiment analysis techniques to areas such as market prediction (Bollen et al., 2010) and social science. In these areas, it is not enough to detect opinions in predefined areas of text or even to mine for the locations of opinions in large corpora, but it is necessary to be able to connect opinions across documents and to reconstruct the social networks that underlie social trends. Furthermore, it must be possible to do this in text that can have an arbitrary number of opinions intertwined in ways that go beyond the base case of product review text. This requires both additional consideration of the perspective of the user and attention to the finer-grained details of sentiment expression.

Do existing resources and techniques really reflect the ultimate goals and end-uses of fine-grained opinion-mining, particularly focusing on the sentential and sub-sentential levels? Consider an "ideal case" of a marketing director or a political campaign manager requesting a forecast of how a product or concept will unfold in the media and market. How do the present conceptions of opinion mining relate to this among other real-world problems of affect?

Proceedings of NAACL-HLT 2013, pages 691–696,
Atlanta, Georgia, 9–14 June 2013.

In the remainder of this position paper, we briefly describe three closely related issues in sentiment analysis that pertain to expanding beyond the current limits of the field.

2 Challenges

2.1 Metasubjectivity and pragmatic opinion

Recent efforts in opinion mining (Ruppenhofer et al., 2008) technology have often tended to take the position that opinion is an internal characteristic of the speaker, a "private state", and that the overall aim of the opinion mining field is to discover techniques that allow us to infer the that latent state from the evidence presented in text. But this may not always be appropriate to all circumstances.

A very simple boundary example comes from Somasundaran and Wiebe (2009): *The blackberry is something like $150 and the iPhone is $500.* This comes from a corpus of opinions on cell phone preference, and this sentence is intended to be a negative opinion about the iPhone. According to Somasundaran and Wiebe, this kind of opinion-expression requires a model of world-knowledge that is either not practical under current technologies, or it requires the development of techniques that can recruit a larger context in the text in order to make the correct inference. They refer to this phenomenon as "pragmatic opinion".

One crucial piece of world-knowledge that provides an opinion its polarity is that of the perspective of the reader or listener to the opinion; we can minimally represent this as the "application" to which the opinion will be put. We refer to variation in the application-specific interpretation of the concept of opinion as "metasubjectivity." Metasubjectivity is a serious problem in extending sentiment analysis work to other domains, particularly for reasons that we describe in the next section.

Metasubjectivity is closely related to the underlying relative nature of veridicality assessment. The veridicality of an utterance is the level to which the listener may judge it as a factual statement about the world. de Marneffe et al. (2012) note that this requires, in some cases, extensive pragmatic knowledge. They present this sentence as an example: *FBI agents alleged in court documents today that Zazi had admitted receiving weapons and explosives training from al-Qaeda operatives in Pakistan last year.* There is an interplay between the trustworthiness of the source of the sentence, the mentioned entities, and the veridicality of words *alleged* and *admitted*, all of which are mediated by the perspective of the reader. For example, if the reader is strongly inclined to trust the FBI, then there may be a high level of veridicality in "alleged" than otherwise. But it could also be the case that the reader believes that Zazi is misleading the FBI.

These distinctions operate directly in the context of determining polarity in opinion mining. Consider the following example sentence from a major information technology (IT) business journal: *Lloyd Hession, chief security officer at BT Radianz in New York, said that virtualization also opens up a slew of potential network access control issues.*

This sentence can be taken to represent an opinion or merely a factual statement. A casual reader without experience in the domain of IT might be convinced that this sentence is simply a neutral statement of fact. But from the perspective of an interested reader such as an investor, this may actually represent a mildly negative statement about virtualization, or it may represent a negative statement about network access control. From the perspective of the manager of an IT support department, it may well be very negative. But from the perspective of Lloyd Hession, we have no idea outside of the pragmatic context. Mr. Hession could be a developer of IT solutions, in which case he would view this as a positive development for the market in new network access control technologies, or, for that matter, he may be invested in a set of technological approaches that compete with virtualization.

This extends to the vocabulary used to express opinions. The use of the word "slew", in this case, has negative connotations, but only if the whole statement is construed by the perspective of the reader to represent an opinion. However, if Lloyd Hession is a provider of new network access control solutions, then the use of "open" may convert this negative context into a positive context.

This is not merely a matter of the perspectives of individual users and participants. It is a matter of how providers of sentiment analysis applications choose to represent these choices to the user, which is in turn reflected in the way in which they create

resources, models, and algorithms. If, for example, our goal is to provide sentiment analysis for domain-specific market prediction or social science, then we need to model the reactions not of the private state of Mr. Hession or of the writer of the article, but of an "aggregate reader" with a presumed interest in the text. Here is a definition of this external state aggregate reader model that might apply to the IT business domain:

> Opinion source A expresses opinion about opinion target B if an interested third party C's actions towards B may be affected by A's textually recorded actions, in a context where actions have positive or negative weight.

This accounts for the cases in which the opinion of interest in the IT example happens to be held by an investor or a IT support manager or other interested readers, and it can be generalized to apply to other domains in which the world's opinion matters.

It is once again within the area of veridicality assessment that we suggest that a possible form of solution exists. de Marneffe et al. (2012) present a model in which the uncertainty in veridicality is represented as a distribution rather than a discrete labelling problem.

In the case of veridicality, there is generally an ultimate ground truth in verifiable facts about the world, apart from the relative veridical nature of a statement. For sentiment, however, there is no such foundation: opinion presence and opinion polarity exist entirely relative to the perspective of the aggregate reader. This requires a different process of annotation, the challenges of which we describe in the next section.

2.2 Corpus development and evaluation

Considering the prevalence of machine learning techniques in opinion mining research, addressing the issue of metasubjectivity must mean addressing the matter of the corpus development.

Existing evaluation techniques depend on a notion of "gold standard data" that are produced by expert judges or crowdsourced annotators (Wilson, 2007; Kessler et al., 2010; Hsueh et al., 2009). There are NLP areas in which popular notions of objectivity may partly apply, such as query relevance; due, among other things, to metasubjectivity, opinion mining is not entirely one of these. However, gold standard data for opinion mining is typically produced using procedures that are standard for information retrieval research, and the quality measures that are generally used happen to assume the presence of an underlying objective truth.

This assumption can be coerced to fit particular cases. For example, a large proportion of opinion mining research is invested in predicting the ratings of product reviews and then aggregating results into a single ratings summary, sometimes based on a lower-level breakdown of product features (de Albornoz et al., 2011). Implicit in this type of work is the assumption of the existence of an ideal rater who uses language in a roughly predictable way to express his or her feelings about the text.

The users of these types of systems can be assumed, to some degree of safety, to share some of the expectations of the builders of these systems, particularly since groups of users as product raters are often the source of the information itself.

But in environments where the users of the system may have various different perspectives on the nature of sentiment, it does not make sense to assume that there would ever be significant agreement among annotators, particularly for market-relevant applications where prediction of reader reaction is central to the task. We attempted to annotate IT business press articles for sentence-level reader-perspective opinion occurrences and found that multiple trained annotators had very low inter-rater agreement by Cohen's κ. Multiple attempts at further annotator training and error analysis revealed that the annotators simply found it very difficult to agree on what the definition of an opinion was. Originally, we had two trained student annotators for this task, with repeated training and joint practice annotations in order to achieve consensus as to what counts as an opinion mention instance and what does not. Other groups of annotators and annotation designs had no better success.

However, we observed that this appears to be primarily a problem of conservativity where annotators differed in the quantity of sentences that they considered to be opinionated, and had a large amount of overlap in those that they did consider to be opinionated. Further discussion with the annotators found that some simply had a much lower threshold at which they would consider a sentence to contain an

opinion. In other words, this form of annotation is more affected by metasubjectivity than opinion annotation focused on opinion source perspective. It should be noted that this is a different task from finding opinion sources and labelling the textual evidence of their private states; we were attempting to model the "ideal case" we identified in section 1.

We suggest that the answer to this problem is to deploy the concept of the aggregate reader mentioned in the previous section and to pose the annotation question indirectly. The former requires the collection of data from a larger number of people and can be provided by existing crowdsourcing techniques (Snow et al., 2008). The latter, however, requires designing the annotation in such a way that it avoids letting the annotator consider the question: "What is an opinion?" This is most likely done by a user interface that simulates the behaviour of the intended aggregate reader (Sayeed et al., 2011).

2.3 Grammatical expression

There are a number of types of features with which one can construct and train supervised sentence-level sentiment detection models. Most recent techniques (Kim and Hovy, 2006; Choi et al., 2006; Jakob and Gurevych, 2010) take into account the syntactic context of the sentence but limit the amount of syntactic context thus used. These restrictions reduce the presence or absence of particular structures to binary features in the model. We argue that we need techniques that take into account more syntactic context, particularly without making use of predefined structures.

The latest techniques make use of larger syntactic contexts with potentially unlimited scope. One example is Nakagawa et al. (2010), who use factor graphs (McCallum et al., 2009) to learn a model that traces paths through the dependency trees of opinion-relevant sentences (de Marneffe and Manning, 2008). However, this is in the service of polarity classification, as it assumes that the appropriate sentences have already been identified; then it is a matter of correctly processing negations and other polarity-changing items. The challenge of metasubjectivity is a barrier to opinion sentence detection itself, well before polarity classification.

Another example is Qiu et al. (2011). They are more directly focused on detecting opinion-relevant language. However, they make use of a system of hard-coded heuristics to find opinion words in dependency parses. While these types of heuristics support longer-distance syntactic relations, they tend to focus on cases where some form of semantic compositionality holds. However, consider this sentence from the IT business press: *The contract is consistent with the desktop computing* ***Outsourcing*** *deals Citibank awarded EDS and Digital Equipment in 1996...* In this case, an interested aggregate reader might note that "awarded" is a word that puts "outsourcing" in a positive light. However, the syntactic relationship between these two words does not directly imply or permit any semantic compositionality, In order to find these relationships, we would need to invest in techniques that can learn from arbitrary non-compositional structure, thereby potentially capturing patterns in grammar that actually reflect some aspects of external pragmatic knowledge.

3 Conclusions

This paper has proposed a challenge for opinion mining, the challenge of metasubjectivity: where the answer to the question "What is an opinion?" is in itself an opinion and an intrinsic part of the task. We first established the context of metasubjectivity relative to existing characterizations of the opinion mining task, establishing the notion of an external aggregate reader as a way to extend from existing notions of sentiment as an internal state. Then we described how this affects the annotation process, given the as-yet-continuing dependence on supervised corpus-based detection techniques. Finally, we described how this affects sentence-level fine-grained opinion detection at the level of syntactic analysis.

One of the risks for the field in proceeding to investigations of how to deal with the question of metasubjectivity is one familiar in natural language processing as a whole: there is a strong risk that these techniques will—initially and for a non-trivial quantity of time—cause the incremental performance gains in existing research to be lost or damaged. It will also require the creation of new training corpora and related resources, temporarily threatening comparability. Nevertheless, we believe that these risks need to be accepted in order to make progress in sentiment analysis.

References

Alm, C. O. (2011). Subjective natural language problems: Motivations, applications, characterizations, and implications. In *ACL (Short Papers)*.

Bollen, J., Mao, H., and Zeng, X.-J. (2010). Twitter mood predicts the stock market. *CoRR*, abs/1010.3003.

Choi, Y., Breck, E., and Cardie, C. (2006). Joint extraction of entities and relations for opinion recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

de Albornoz, J., Plaza, L., Gervás, P., and Díaz, A. (2011). A joint model of feature mining and sentiment analysis for product review rating. In Clough, P., Foley, C., Gurrin, C., Jones, G., Kraaij, W., Lee, H., and Mudoch, V., editors, *Advances in information retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 55–66. Springer Berlin / Heidelberg.

de Marneffe, M.-C. and Manning, C. D. (2008). The stanford typed dependencies representation. In *CrossParser '08: Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, Morristown, NJ, USA. Association for Computational Linguistics.

de Marneffe, M.-C., Manning, C. D., and Potts, C. (2012). Did it happen? the pragmatic complexity of veridicality assessment. *Computational linguistics*, 35(1).

Hsueh, P.-Y., Melville, P., and Sindhwani, V. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, HLT '09, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jakob, N. and Gurevych, I. (2010). Extracting opinion targets in a single and cross-domain setting with conditional random fields. In *EMNLP*.

Kessler, J. S., Eckert, M., Clark, L., and Nicolov, N. (2010). The 2010 ICWSM JDPA sentment corpus for the automotive domain. In *4th Int'l AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*.

Kim, S.-M. and Hovy, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *SST '06: Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

McCallum, A., Schultz, K., and Singh, S. (2009). Factorie: Probabilistic programming via imperatively defined factor graphs. In *Neural Information Processing Systems (NIPS)*.

Nakagawa, T., Inui, K., and Kurohashi, S. (2010). Dependency tree-based sentiment classification using crfs with hidden variables. In *HLT-NAACL*.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2).

Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9—27.

Ruppenhofer, J., Somasundaran, S., and Wiebe, J. (2008). Finding the sources and targets of subjective expressions. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Sayeed, A. B., Rusk, B., Petrov, M., Nguyen, H. C., Meyer, T. J., and Weinberg, A. (2011). Crowdsourcing syntactic relatedness judgements for opinion mining in the study of information technology adoption. In *Proceedings of the Association for Computational Linguistics 2011 workshop on Language Technology for Cultural Heritage, Social Sciences, and the Humanities (LaTeCH)*. Association for Computational Linguistics.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP 2008*.

Somasundaran, S. and Wiebe, J. (2009). Recognizing stances in online debates. In *Proceedings of*

the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1, ACL '09.

Wilson, T. (2007). *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of private states*. PhD thesis, Intelligent Systems Program, University of Pittsburgh.