What are Sentiment, Affect, and Emotion? Applying the Methodology of Michael Zock to Sentiment Analysis

Eduard H. Hovy

Abstract In Natural Language Processing or Computational Linguistics (NLP or CL), researchers assume almost universally that speakers hold some affective value or sentiment with regard to (some aspects of) a topic such as a film or camera, that this sentiment has a fixed value (typically, something like *good* or *bad*), and that the sentiment is expressed in text through a word or small combination of words. However, one finds in the NLP literature essentially no discussion about what 'sentiment' or 'opinion' really is, how it is expressed in actual language usage, how the expressing words are organized and found in the lexicon, and how in fact one can empirically verify cognitive claims, if any, implied in or assumed by an NLP implementation. Even the Wikipedia definition, which is a little more careful than most of the NLP literature, uses words like "polarity", "affective state", and "emotional effect" without definition. In this situation we can usefully try to duplicate Michael's mindset and approach. What do people actually do? How does what they do illustrate the complexities of the problem and disclose unusual and interesting aspects that computer scientists are simply blind to? In this paper I first provide interesting examples of real-world usage, then explore some definitions of sentiment, affect, opinion, and emotion, and conclude with a few suggestions for how computational studies might address the problem in a more informed way. I hope in the paper to follow the spirit of Michael's research, in recognizing that there is much more to language usage than simply making some computer system mimic some annotated corpus, and that one can learn valuable lessons for NLP by looking at what people do when they make language.

Keywords Sentiment analysis · Opinion mining · Computational sentiment determination · Cognitive view on opinion

Dedicated to Michael Zock.

E.H. Hovy (✉)
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: hovy@cmu.edu

1 Introduction

Throughout his career, Michael has been interested in one of the core issues of psycholinguistics: how we access and use the lexicon. Time and again he has pointed out the mismatch between computational and psycholinguistic approaches to the lexicon: in Computational Linguistics research the focus is always the content and use of lexical items, while finding the word(s) we need is never a problem; however real human language behavior shows clearly that lexical access is complicated and error-prone, and is just as interesting a problem as lexical content.

In this paper I address a similar mismatch between computational studies and real-world uses, on a different topic. Over the past decade, computational linguistics has developed a new area commonly called *sentiment analysis* or *opinion mining*. As Wikipedia defines it.

> **Sentiment analysis** or **opinion mining** refers to the application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in source materials. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation (see appraisal theory), affective state (that is to say, the emotional state of the author when writing), or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader).

In Natural Language Processing or Computational Linguistics (NLP or CL), researchers assume almost universally that speakers hold some affective value or sentiment with regard to (some aspects of) a topic such as a film or camera, that this sentiment has a fixed value (typically, something like *good* or *bad*), and that the sentiment is expressed in text through a word or small combination of words. However, one finds in the NLP literature essentially no discussion about what 'sentiment' or 'opinion' really is, how it is expressed in actual language usage, how the expressing words are organized and found in the lexicon, and how in fact one can empirically verify cognitive claims, if any, implied in or assumed by an NLP implementation. Even the Wikipedia definition, which is a little more careful than most of the NLP literature, uses words like "polarity", "affective state", and "emotional effect" without definition.

In this situation we can usefully try to duplicate Michael's mindset and approach. What do people actually do? How does what they do illustrate the complexities of the problem and disclose unusual and interesting aspects that computer scientists are simply blind to?

In this paper I first provide interesting examples of real-world usage, then explore some definitions of sentiment, affect, opinion, and emotion, and conclude with a few suggestions for how computational studies might address the problem in a more informed way. I hope in the paper to follow the spirit of Michael's research, in recognizing that there is much more to language usage than simply

making some computer system mimic some annotated corpus, and that one can learn valuable lessons for NLP by looking at what people do when they make language.

2 Current Tasks and Approaches

There has been a rapidly growing desire for automated sentiment/opinion detection systems. Companies are eager to read popular critiques of their products and learn the effects of their advertising campaigns; politicians are eager to assess their image among the electorate, and normal people overwhelmed by the variety of text on the web about almost any topic, with too many different voices and too little trustworthiness, are eager for some automated assistance.

But the lack of standardization—the absence even of clear definitions of the major topics under discussion—hampers any serious work. For example, restaurant reviews in Yelp, book reviews in Amazon, and similar crowdsourcing sites all include some sort of star rating system, the more stars meaning the more positive. But even a cursory glance shows that different raters apply very different standards, and that almost no review discusses just a single aspect, making a single star rating system a crude average of little specific value.

Reflecting this fact, the linguistic expression of opinion is often quite complex. For example, the opinion expressed in the following excerpt from the web

> I have been working as a stylist at great clips for about 7 months now, and I like it and I hate it at the same time I only have to work a 6 h shift and I get tips from customers but I hate it cuz I don't really like being close to strangers and I hate getting hair all over my shoes…

is nuanced in a way that makes a simple star rating for the job impossible.

In response, many NLP researchers have taken the approach of trying to identify 'aspects' or 'facets' of the topic and then discovering the author's opinion for each of them; for example, positive on hours and tips but negative on proximity to strangers and messy shoes. This resembles the periodical *Consumer Reports*, whose well-known faceted rating system of nicely composed tables rate facets of hundreds of products. But while this approach may work well for products, real life is often much harder to compartmentalize into facets. Unlike the price, weight, and lens quality of a camera, 'getting hair over my shoes' is not a common and easily quantifiable aspect of life.

The problem goes beyond nuancing and facets. People's attitudes change, and change again. For example, also from the web

> the movement…. er.… sometimes I like it and sometimes I hate it. It would've been sooooo much better if they had used head tracking rather than rotating the shoulders

Even assuming one can identify which facet is being discussed, the author holds about it two radically opposite opinions. What star rating system can express this?

Finally, there is the assumption that when one has identified the author's opinion then one has done one's job. But just assigning a sentiment value is not always enough. In

Why I won't buy this game even though I like it.

the author is probably saying something of prime importance to the makers of the game, far more serious than just some reasons for satisfaction of a happy customer.

3 Current Computational Sentiment Determination

There is a great deal of recent research on automated sentiment detection. Politicians, advertisers, product manufacturers, and service providers are all driven by the exciting promise of being able to determine easily and quickly the general public's response to them and/or their ideas or products. Computational systems and services like Media Tenor (http://us.mediatenor.com/en/) provide beautifully crafted results that show in dramatic color the level of popularity of, for example, the US President George W. Bush in 2005, shown in Fig. 1.

Despite its apparent sophistication, automated sentiment determination is almost universally very simple. The very simplest systems match words in the immediate context of the target item's name against lists of positive-affect and negative-affect words, and compute some sort of possibly weighted average. That is, the sentences

I hate George Bush
I have trouble with Bush's policies

are given the label *bad* for Bush because of the presence of "hate" and "trouble"; often also a strength score is provided when "hate" is scored as 'more-negative' than "trouble". The fact that one person's 'having trouble with' may be worse than another's 'hate' is simply ignored.

Recent research papers extend this simple scheme in one of three ways: (1) more elaborate signals of affect not just longer lists of words, but also other features such as part of speech tags, negation as expressed by words like "not", "don't", "never…", etc. (Pang et al. 2002; Turney 2002); (2) additional facets or components of the problem including facets of the topic, the holder of the affect, etc. (Kim and Hovy 2006; Snyder and Barzilay 2007); and (3) more elaborate methods to compose individual signals, in order to handle mixed-affect sentences such as

Although I hate Bush's policies on immigrants, I really love his fiscal policy

Modern techniques propagate affect up the sentence parse tree and perform various kinds of affect combination at nodes where values meet. The most sophisticated sentence sentiment computation engine at the time of writing is that
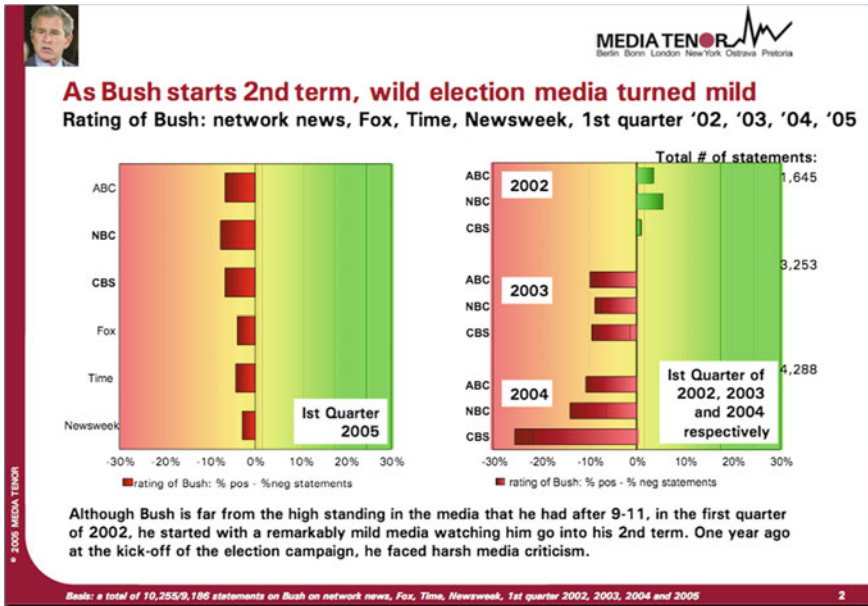
Fig. 1 Media Tenor's analysis of public popularity of then US President George W. Bush in 2005
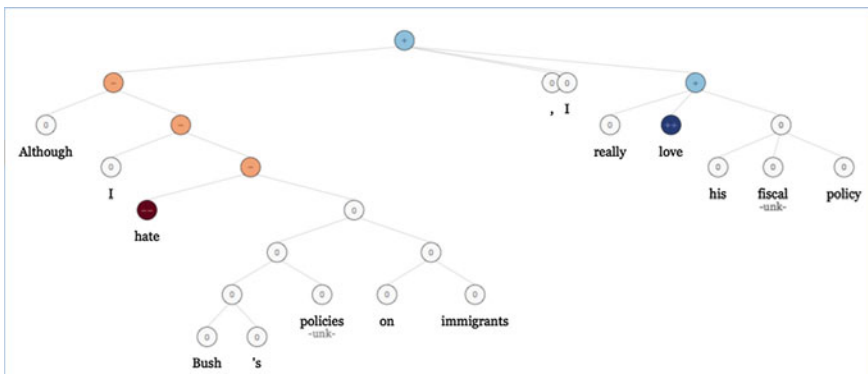

Fig. 2 Analysis of "Although I hate Bush's policies on immigration, I really love his fiscal policy"

of Richard Socher, whose online demo system at http://nlp.stanford.edu:8080/sentiment/rntnDemo.html produces for the above sentence the analysis in Fig. 2. Here brown nodes (in the left half of the sentence) reflect negative and blue nodes (right half) positive sentiment, and the intensity of the node's color expresses the strength of affect.

This system embodies a complex model in which words are represented as vectors of values and are combined using a recursive neural network that is trained for exactly this task, described in Socher et al. (2013).

Recently, attention has focused on determining the sentiment/affect of tweets. Tweets can be viewed as a socially diversified, rapidly changing, and wide-ranging 'sensor network' that allows politicians and product manufacturers to gauge popular opinion. Annotated collections of tweets have been made available for evaluation; see Saif et al. (2013).

Despite all the research, there has never been a serious attempt to define the concepts of sentiment/opinion or to establish generally accepted criteria for judgment in the computational community. While corpora of user judgments of the affective values are used to train systems and evaluate output, they justify their trustworthiness on high enough annotation agreement, for example even across languages (Steinberger et al. 2011). However, people on average tend to agree about sentiment only at the level of 79 % (Ogneva 2012): one in five decisions is in contention, even under the very crude granularity of *good–neutral–bad*. Therefore researchers continue to face a serious definitional problem before sentiment analysis can be considered mature and trustworthy enough to be truly scientific.

4 What Would Michael Do?

The intense computational effort (in some cases associated with considerable private sector funding) attracts many people. The relative ease of building wordlists and training word or feature matching algorithms (even with original variations) generates a plethora of sentiment analysis systems. But I can't help wondering what Michael would do if he were to address this problem. It would not be "let's make a better feature list and break up the sentence a little more and build another classifier". His approach —thoughtful and deeply engaged, interested in both the computational and the psychological/cognitive— would, I venture to imagine, proceed along the following lines:

- first, he would analyze the phenomena,
- then he would define his terms,
- and finally he would propose an algorithm, and look for people to implement it.

In other words, one has to address the following open questions if one wants to know what one is talking about:

1. Definitions: What are *Sentiment*? *Opinion*? *Affect*?
2. Theory: What is the structure of these concepts?
3. Practice: Is sentiment recognition all just a matter of identifying the appropriate keyword(s) (perhaps in combinations)? What do people do when they assign (generate) and understand (interpret) sentiment?
4. Evaluation: How do people assign values? Do they agree?

5 Types and Definitions of Opinions

One can identify at least two kinds of Sentiment expressed in text:

- **Opinions**, such as *like/dislike/mixed/don't-know… believe/disbelieve/unsure… want/don't-want/sometimes-want…* . This is something the subject <u>decides</u>.
- **Feelings/emotions**, such as *happy/sad/angry… calm/energetic/patient/relaxed …* . This is something the subject <u>feels</u>.

What exactly these notions are is not simple to define. The concepts are connected; they cause and/or reinforce one another. Researchers in Emotion/Affect, mostly in Psychology, have written dozens of books on the topic (see *Affective Computing* in Wikipedia). We do not address Emotion in this chapter.

Turning to opinions, the Merriam-Webster dictionary defines an *opinion* as "a view, judgment, or appraisal formed in the mind about a particular matter", or "a belief stronger than an impression and less strong than positive knowledge". This indicates that there are at least two kinds of opinion:

Judgment opinions: *good, bad, desirable, disgusting…*: "The food is horrible"
Belief opinions: *true, false, possible, likely…*: "The world is flat"

Analysis of examples indicates that they both have the same internal structure, which can be defined at minimum as a quadruple: (Topic, Holder, Claim, Valence):

- **Topic** = theme/topic of consideration
- **Holder** = person or organization holding or making the opinion
- **Claim** = statement about the topic
- **Valence** (judgment opinions):

    - *Positive* or *Negative* or *Mixed* or
    - *Neutral*: "I don't care one way or the other about him" or
    - *Unstated*: "they had strong political feelings"

- **Valence** (belief opinions):

    - *Believed* or *Disbelieved* or *Unsure* or
    - *Neutral*: "I don't care one way or the other about him" or
    - *Unstated*: "perhaps he believed it, I don't know"

Armed with this knowledge, one can define opinions and its types as follows:

Definition An *opinion* is a decision made by someone (the Holder) about a topic (the Topic). This decision assigns the Topic to one of a small number of classes (the Valences) that affect the role that the topic will play in the Holder's future goals and planning decisions (discussed below).

Definition *Judgment opinions* express whether or not the Holder will follow goals to try to own/control/obtain the Topic.

Definition *Belief opinions* express whether or not the Holder will assume the Topic is true/certain/etc. in later communication and reasoning.

One can include additional components to extend the structure:

- **Strength** of opinion
  - This is very difficult to normalize across Holders
- **Facet(s) of topic**
  - It may be useful to differentiate subfacets of the Topic; not "the camera" but "the weight of the camera". This is simply a narrower Topic.
- **Conditions** on opinion
  - Adding conditions is possible, at the cost of complexity: "I like it only when X"/"If X then I like it".
- **Reasoning/warrant** for opinion
  - "The reason I like it is X". As argued below, this is important, even though it opens up the question of reasoning and argument structure.

6 Identification of Opinion

The question now arises, how are opinions of both kinds expressed in text? It is clear from the examples above that the current computational practice of using simple word lists is not adequate. Opinion are expressed by units of various sizes:

- **Word level**: individual words are opinion clues
  - Yes: "hate", "disgusting", "anger"
  - No opinion: "run", "announce", "tall"
- **Sentence level**: compositions of words
  - Yes: "Actions with negative consequences include the US attack on Iraq."
  - No opinion: "To receive a copy of our catalogue, send mail."
- **Text level** (implicature): opinions are obtained via rhetorical relations
  - "Not only did he eat the meat, he spoiled all the rest of the food as well"
  - "Sure he ate the meat. But he still didn't clean the kitchen!"

Computational linguistics research has devoted a lot of effort to creating lists of words and phrases to be used for opinion recognition, including, at the lexical level, wordlists, better word/feature combination functions, etc. (Yu and Hatzivassiloglou 2003; Riloff and Wiebe 2003; Kim and Hovy 2005; Agarwal and

Mittal 2013; and other recent work); at the structural level, sentence and discourse structure analysis (Socher et al. 2013; Wang et al. 2012a, b) and document-level sentiment (Pang et al. 2002; Turney 2002; Wiebe et al. 2005). Additional supporting information includes knowledge about user–user relationships in online social networks like Twitter (Balahur and Tanev 2012; Tan et al. 2011), or general ideology (Wang et al. 2012a, b; Gryc and Moilanen 2010; Kim and Hovy 2006).

7 A Longer-Term Cognitive View on Opinion

As is so nicely described in the opening chapters of this book, Michael's primary research interest would focus on the cognitive aspects of the problem. It would be axiomatic to him that it is simply not interesting to assign labels in a simplistic word- or feature-matching manner (even though, for some corpora, this approach may work quite well). I can imagine him saying: let's look at *why* people say what they say. That is, sentiment reflects the deeper psychological state of the holder, enabling people to give reasons why they like or dislike something. Where does this lead one?

On analysis, one readily discovers two principal types of reason:

- Goals and plans regarding future actions
- Emotional attachments and preferences/attitudes toward objects and people

First, considering the goals and plans of the speaker: when the topic is something of utility to the holder, one can match up its characteristics to the holder's plans and goals. Characteristics that align with a goal or a plan pursued by the holder would receive a positive rating, and vice versa. For example, a sturdy and perhaps heavy camera would match up well with the goals and plans of a mountaineer (and hence be *positive*), but absolutely not with the goals and plans of a teenage girl going to the beach (and hence for her be *negative*). Simple keyword matching can never explain *why* someone makes his or her opinion judgment. But if you understand this first category of sentiment properly, you can explain why. (Ask anyone who does not study NLP why some sentiment decision is made, and they will readily and easily provide their view of the reasoning.)

Second, considering emotional attachments and preferences: when the topic is not of primarily utilitarian value, it does not relate to plans, but generally only to the holder's highest-level goal(s) to be interested/amused/happy. In this case, one can try to identify the general conceptual types of entity and event that he or she prefers or disprefers: Western or crime stories, Action or SciFi movies, conservative or informal clothes, classical or jazz or rock music. When people say "I prefer cotton over polyester even though it needs ironing because it is cooler", that reflects their goal to be comfortable at the cost of their goal of being groomed. But not everything relates back to a goal or plan. People will generally have a hard

time saying "I prefer jazz over rock because…". For the second category of sentiment, *de gustibus non est disputandem*. In this case, referring to the prototypes of the preferred/dispreferred actions and entities usually provides sufficient justification.

Following this line of thought, one can extend the structural definition of opinion as follows:

Opinion type 1:

- Topic: camera
- Claim: strength is good
- Holder: buyer
- Valence: +
- Reason:

    – Personal profile: mountaineer
    – Goals: climb mountains, protect camera
    – Plans: take photos while climbing

Opinion type 2:

- Topic: music
- Claim: jazz is good
- Holder: listener
- Valence: +
- Reason:

    – Features: free-form, complex harmonies and rhythms, etc.

To address the problem of reason determination computationally, one can perform automated goal and plan harvesting, using match patterns such as "*a * camera because *"* to link topic features relative to the relevant plans, as shown in Fig. 3.

While the causes for type 2 opinions will range widely over emotions, and hence probably not be tractable computationally, the causes for type 1 might be easily categorized into a small ontology of physical features relating to actions, including notions of movement, cognition, social acts, construction, money, and a few others.

In conclusion, the challenge for Opinion (Sentiment) Analysis is not just sentiment classification, but deeper explanation generation. (In fact, precisely this is what the companies and politicians really want!) discovering how to do so is an interesting and longer-term research challenge that will provide a rich dividend to the researcher. But it is not easy, and not likely to be popular with people interested in a lot of quick-win publications.
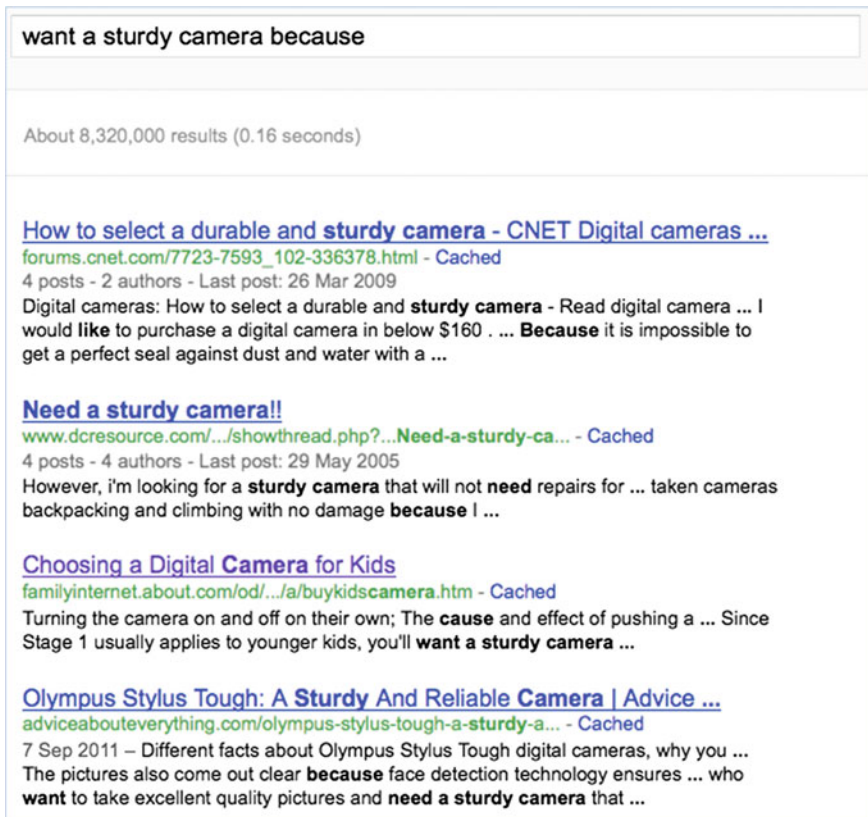
Fig. 3 Harvesting goal and plan information, and associated features, from the web

8 Conclusion

Over his long career, Michael has shown a remarkable talent to connect with people. The existence of this book demonstrates the respect and affection we hold for him. I think our regard has both emotional and intellectual grounds; we empathize with his humanity, personal humility, and genuine concern for people, but we equally much respect and value his tenacity, intellectual humility toward problems, and seriousness in addressing the questions that have occupied most of his research life. Michael's unwillingness just to follow an easy computational approach and accept as an answer "because it works", but instead to keep asking "but how, and why, do people do it?", despite having papers rejected and contributions ignored—this is the spirit that moves us who contribute to this book. These qualities have influenced me, and I believe, most of us.

I am grateful for having known Michael, and I hope that I have the privilege for a long time to come. Happy birthday!

References

Agarwal, A., & Mittal, N. (2013). Optimal feature selection methods for sentiment analysis. *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)* (Vol. 7817, pp. 13–24).

Balahur, A., & Tanev, H. (2012). Detecting entity-related events and sentiments from tweets using multilingual resources. *Proceedings of the CLEF Workshop Online Working Notes*.

Gryc, W., & Moilanenm, K. (2010). Leveraging textual sentiment analysis with social network modelling: sentiment analysis of political blogs in the 2008 U.S. Presidential Election. *Proceedings of the workshop 'From Text to Political Positions'* Amsterdam: Vrije Universiteit.

Kim, S.-M., & Hovy, E. H. (2005). Automatic detection of opinion bearing words and sentences. *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)* (pp. 61–66).

Kim, S.-M., & Hovy, E. H. (2006). Identifying and analyzing judgment opinions. *Proceedings of the Human Language Technology/North American Association of Computational Linguistics conference (HLT-NAACL 2006)*.

Ogneva, M. (2012). How companies can use sentiment analysis to improve their business. From http://mashable.com/2010/04/19/sentiment-analysis/.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 79–86).

Riloff, E., & Wiebe, J. (2003). Learning extraction patterns for subjective expressions. *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 105–112).

Saif, H., Fernandez, M., He, Y., & Alani, H. (2013). Evaluation datasets for twitter sentiment analysis. *Proceedings of the Workshop on Emotion and Sentiment Analysis (ESSEM)* at the AI*IA Conference.

Snyder, B., & Barzilay, R. (2007). Multiple Aspect Ranking using the good grief algorithm. *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)* (pp. 300–307).

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.

Steinberger, J., Lenkova, P., & Kabadjov, M., Steinberger, R., & van der Goot, E. (2011). Multilingual entity-centered sentiment analysis evaluated by parallel corpora. *Proceedings of the Conference on Recent Advancements in Natural Language Processing (RANLP)*.

Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., & Li, P. (2011). User-level sentiment analysis incorporating social networks. *Proceedings of the KDD conference*.

Turney P. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *Proceedings of the Association for Computational Linguistics* (pp.417–424). arXiv:cs.LG/0212032.

Wang, F., Wu, Y., & Qiu, L. (2012a). Exploiting discourse relations for sentiment analysis. *Proceedings of COLING conference posters* (pp. 1311–1320).

Wang, J., Yu, C. T., Yu, P. S., Liu, B., & Meng, W. (2012b). Diversionary comments under political blog posts. *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM)*.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, 39*(2–3), 165–210.

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 129–136).