Sentiment Analysis in Web text: An overview

JOŽE BUČAR and JANEZ POVH
Laboratory of Data Technologies
Faculty of Information Studies, University of Novo mesto
SLOVENIA
joze.bucar@fis.unm.si janez.povh@fis.unm.si http://www.fis.unm.si

Abstract: The growing interest in the efficient analysis of informal, subjective, opinionated web content has caused exceptional development in a field of sentiment analysis. In recent years there have been added blog and web tracks, tasks dealings with people and organization search, their branding and reputation detection, opinion analysis, news clustering, chats analysis, SMS-based FAQs, passage retrieval and retrieval from forums. It is a challenging natural language processing and text mining problem, which brings together scientists from different fields like computational linguistics, data mining, computer science, machine learning, graph theory, neural networks, sociology and psychology. More and more business, sale, finance and other companies are realizing the importance of people's opinion. Knowing and understanding people's feelings and satisfaction about their products and services has a great impact on structural, organizational, business solutions and future of the company.

This paper presents a short overview of sentiment analysis, defines its basic concepts, results, open problems, and challenges of working with emotionally-colored web text.

Key-Words: sentiment, sentiment analysis, opinion mining, text mining, natural language processing

1 Introduction

Textual information can be categorized into facts and opinions. Facts are objective expressions about entities, events and their properties, items of information, or state of affairs existing, observed, or known to have happened, and which is confirmed or validated to such an extent that is considered reality [10]. Sometimes, when developing breaking research or just purchasing decision about certain product, we often look for people that had experiences in a field of our research or product we are interested in. It is completely natural that we are looking for other's opinions.

The World Wide Web experienced a great success all around the world with tremendous media support. It became new media and source for advertising and providing information. The growing interest has emerged rush to integrate new processes, features that can contribute to more efficient work. It has made it possible to extract the opinions of people regardless of whether we are looking for opinions on the candidates related to upcoming elections or opinions about the movie that we would like to see. Increasing number of blogs, web sites, newsgroups, forums, chat rooms, etc. generated a new area in text analysis, where we can detect people's feelings about specific event, topic, popularity of political candidates, etc.

Sentiment analysis is very topical issue; both industry and academia understand advantages of sentiment extraction from web text. Especially business companies and institutions quickly realized the importance of caring quality control as well as marketing research for selling their products and services.

The rest of the paper is organized as follows: Section 2 defines basic concepts and introduces sentiment analysis. Section 3 gives a brief overview of studies so far, summarizes related work and performance. Section 4 describes in detail problems and challenges in sentiment analysis. Finally, the paper ends with conclusion in Section 5.

2 Introducing basic concepts

Sentiment analysis is also known under other names, such as opinion mining, subjectivity analysis and appraisal extraction, which generally deals with subjective elements as sentiment units (words, phrases, sentences, or whole document).

We often use terms opinion and sentiment. On one hand, there are some differences among authors in interpretation regarding sentiment. In fact, in practice many researchers avoid tightly definition of the "sentiment". Boiy and co-authors [3] for example, compare sentiments with emotions, judgments and ideas, which are prompted or colored by emotions. On the other hand, different authors define opinions in a similar way. Liu [10], for example, treats opinions as expressions with subjective annotation that describe people's sentiments, appraisals or feelings toward entities, events and their properties.

However, most of researchers like to look only the polarity and the target of the sentiment, which is either positive, (neutral), and negative ([14], [16], [25]). Semantic orientation, which usually captures positive or negative evaluative factor, is a measure of subjectivity and opinion in text.

In everyday discussions and debates, we use very often subjective reviews. Content of subjective opinions is highly context-sensitive, and its expressions often differ from person to person. Hence, we have to distinguish between subjective statements and false statements since subjective does not mean not true. Let me give an example with sentence *"Andy loves candy!"* Regardless of whether given sentence is true or not, in any case reflects Andy's feelings towards candy, which is that he likes to sweeten and enjoys candies.

Semantics is actually related to syntax. In most languages the syntax is how you say or write something, where semantics is the meaning behind what you said or wrote. Let me use previous example *"Andy loves candy!"* The syntax is represented with all the letters, words and punctuation in sentence, where semantics is actually the true meaning behind these words. At this point let's change upper sentence with *"Andy ♥ candy!"* As you can see we have changed the syntax, however, notice that semantics of the sentence stays the same.

2.1 Sentiment analysis and objectives

It is a field that is closely related to computational linguistics, natural language processing, and text mining, where we try to identify and extract subjective information in source material.

Objectives of sentiment analysis are usually presented in three stages: sentiment detection and identification, polarity classification, and discovery of the opinion's holder and target [10], [13], [18].

The purpose of sentiment analysis is to determine sentiment of source material through the expression and contextual polarity of the source. Opinion can be reflected through judgment or evaluation, emotional state of the subject (source), or state of emotional communication, by which they would like to impact on people's opinion or decision. Thus, it is a way, where people try to determine person's state of mind on specific subject they are talking about. Information as such can be extracted on-line from news articles, blogs, chats, forums, tweets, reviews, comments, and other web texts.

There are several qualities that we may want to detect in text. Often we want to categorize text by specific topic, which may involve dealing with whole taxonomies of topics. Another issue is detecting and identification of subjectivity and objectivity. Our attempt is to determine whether a sentence or document is either ([5], [10]):

- objective or subjective (subjectivity classification),
- and further if subjective text includes positive or negative sentiment (sentence/document-level sentiment classification).

An objective sentence expresses some factual information about the world, while a subjective sentence expresses some personal feelings of beliefs. *"Cockta tastes better than Coca-Cola"* is a subjective statement that gives a regular or comparative opinion. Objective opinion implies a regular or comparative opinion, which usually expresses a desirable or undesirable fact, e.g., *"The more expensive Dell UltraSharp U3014 30" has better screen resolution than Dell UltraSharp U2412M 24"."*

3 Studies and performance

A brief overview of studies is shown in Appendix A. It represents only selection of work done in a field of sentiment analysis.

First studies in this field were made around 2000. Two early methods for polarity detection at document level focused on detection of movie reviews [16] and product reviews [27]. Turney considered sentiment extraction from the whole document as basic information, which is commonly known as the document-level sentiment classification. He treated sentiment classification as detection of positive or negative sentiment. Others like to expand the basic task to classify a document's polarity or strength of opinion on a multi-way scale, for example: reviews measured in scaled range from -5 to 5, where -5 refers to extremely negative, 0 to neutral, and 5 to extremely positive; movie reviews rated on either a 3 or a 4 star scale [18], or restaurant reviews that rated various aspects of the given restaurant on a 5 star scale [22]. We score every concept by its relation to sentiment words and their associated score. It is a way, which allows movement towards understanding more complex text. Advanced sentiment classification seeks even more complex emotional states such as excited, happy, sad, angry, scared, or tender.

ISBN: 978-960-474-304-9

It is difficult to compare performance of different studies between each other, because variety of resources and collections of documents were used for training and testing. As seen in selection of studies in Appendix A, there are cases, when achieved accuracy is over 90% (see [7]). Godbole et al. worked on sources like newspapers and blog posts at the level of words, and achieved accuracy 82.7-95.7%, while others, who worked on documents, such as blog posts and full web pages, have in general accuracy of around 65-85%. Despite the fact that we are able to build comprehensive lexicons of sentiment-annotated words, there is still an issue how to locate it in text correctly. Some studies have been done in domain of long documents like product reviews, and in some tricky domains like political commentaries. Although relatively high accuracy in document polarity labeling has been achieved, it is still a challenge to extract sentiment orientation, complete with emotion's intensity, its holder, and target.

4 Problems and opened questions

Sentiment analysis challenges many difficulties. One of the main reasons for the lack of study on opinions is the fact that before the emergence of the web there was not much opinionated text available. Some big concerns represent defining opinions and subjectivity, detection of negotiations, sarcasm, humor, opinion citations, quotations, speculations, and problems related with emotion and content perception. People commenting on the Internet often pay no attention to grammar. They change, duplicate, and omit letters, overreact, use figurative meaning, slang phrases, superlative words, abbreviations, uppercase letters, exclamation marks, etc. Liu, for example, introduced the problem of sentiment analysis by following review segment on iPhone [10]:

"(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) Although the battery life was not long, that is ok for me. (6) However, my mother was mad with me as I did not tell her before I bought it. (7) She also thought the phone was too expensive, and wanted me to return it to the shop. …"

A quick overview reveals that customer made this review after purchasing iPhone, as seen in sentence (1), which is actually neutral statement. We can detect positive sentiment or opinion in sentences (2, 3, and 4), while sentences (5, 6, and 7) include negative connotation. Furthermore, we can notice that there are different targets on which customer expresses the opinions. Sentence (2), for example, considers iPhone as a whole, while the following sentences (3,4, and 5) reflect user's opinion on *"touch screen"*, *"voice quality"*, and *"battery life"*. The opinion in sentence (6) actually refers to user and not iPhone. The last sentence (7) applies to iPhone again, more precisely on its price. It is also not so difficult to distinguish between two sources or holders of opinions in this text. Sentences (2, 3, 4, and 5) are related to the author of this iPhone review, while next two sentences (6 and 7) represent author's mother opinion.

It often occurs that in one situation a word contributes to positive sentiment, while in other situation may be considered as negative. For example, let us take under consideration the word *"long"*. It makes a huge difference if a customer applies word *"long"* in conjunction with, for example, smart-phone's battery life expectancy, which would be a positive opinion; or if a customer relates it with smart-phone's start-up time, which, however, would give negative opinion. Understanding this problem and distinction between them means awareness that system trained to gather sentiment may not perform very well on another. Most work has been done on product and movie reviews (see [16], [25]), where it is easy to identify the topic of the text. It is useful to pay attention to which characteristics of this product or service the writer is talking about: is it perhaps the smart-phone battery life or its start-up time that concerns consumers the most?

Another issue is negation. Sentence *"the lunch at your restaurant was tasty"*, however, has completely different meaning from *"the lunch at your restaurant wasn't tasty."*

One of the challenges in sentiment detection is how to detect and identify holder or target. An opinion without its target being identified is of limited use. Understanding the significance of opinion holder and target contributes to the improvement of the sentiment analysis algorithms. For example, although the sentence *"although the service is not that great, I still love this restaurant"* clearly has a positive notation, however, we cannot say that this sentence produces only positive sentiment. Upper sentence has positive sentiment about the restaurant, but express negative about its service.

Another problem is also quotation. Unlike usual topical analysis, sentiment statement authorship represents an integral part in the series of problems. If we follow daily reports or the debates in parliament, we can discern that both news and political arguments are full of quotations and opinion citations. Some structures of sentences can be so complex that there can be some difficulties in sentiment detection. An article that includes daily news with political discussions, for example, would include not only quotations from the debaters, but also the pundits commenting on the debate, and perhaps even the author's stance on the issues.

Different people express their opinions differently. Even more, some people are contradictory in their statements, because they vary while expressing their own opinions. It is completely normal that people express both positive and negative opinions in their reviews or discussions. In most cases this does not pose computer major obstacles, because it is quite successful by analyzing sentences one at a time. Many traditional text processing algorithms are not efficient enough, when small differences between two pieces of text occur. However, the more informal the media, the more likely people are to combine different opinions in the same sentence [10]. Sentence, for example, *"Service was terrible but the food was excellent"* does not represent any problem to human perception; however it is hard nut to crack for a computer. People often encounter the problem, how to express opinions, thoughts, or update their status in social network media like Twitter with limited number of signs. The consequence is that even other people sometimes find it difficult to understand what someone thought based on a short commentary or status line because there is simply lack of context. Meaning of commentary *"Food was as good as the last time"*, for example, is not clear enough because customer expressed the opinion based on previous experience, which is probably unknown to the vast majority of other readers.

One day, while checking forums related with textual sentiment analysis and natural language processing, I encountered commentary, which I find amusing and entertaining at the same time [12]:

A linguistics professor was lecturing to her class one day. "In English," she said, "A double negative forms a positive. In some languages, though, such as Russian, a double negative is still a negative. However, there is no language wherein a double positive can form a negative." A voice from the back of the room piped up, "Yeah . . .right."

5 Conclusion

Today's problem is that there's tons of information (customer feedback, competitor information, client emails, tweets, press releases, legal filings, product & engineering documents, etc.), which rapidly grow. In addition to saturation of information, humankind is still hungry of knowledge derived from retrieved information. For many reasons, it is absolutely impossible to employ readers that could extract important knowledge about your customers, competitors, or your company operations, organization, marketing, sales, engineering, and product & service quality. Summarizing consumer reviews is an important problem. Opinion mining can be useful in several ways.

Benefits can be seen in evaluation of sentiment and monitoring changes over time, identification of review sources to define potential targets, improving customer experience and competitive position with time, promoting research and development activities with a closed-loop and integrated analysis environment. No matter, if you are owner of a company, consumer that seeks for advise on products or employed for instance in sales or marketing, it can help you define product or service popularity that you provide or seek, the success of promotion campaign or new product launch, even identify where products, services, and features are desirable and where not. For example, a review about certain restaurant might be broadly positive, but be specifically negative about the service. Identifying this kind of information in a systematic way gives the provider a more realistic insight of public opinion than surveys or focus groups because the data is created by its customers.

Acknowledgment:
"Work supported by Creative Core FISNM-3330-13-500033 'Simulations' project funded by the European Union, The European Regional Development Fund. The operation is carried out within the framework of the Operational Programme for Strengthening Regional Development Potentials for the period 2007-2013, Development Priority 1: Competitiveness and research excellence, Priority Guideline 1.1: Improving the competitive skills and research excellence."

References:

[1] M. Annett and G. Kondrak, A comparison of sentiment analysis techniques: Polarizing movie blogs, *Advances in Artificial Intelligence. Springer Berlin Heidelberg*, 2008, pp. 25-35.

[2] T. Berners-Lee, J. Hendler, and O. Lassila, The semantic web, *Scientific American*, 2001, pp. 28-37.

[3] E. Boiy, et al., Automatic sentiment analysis in on-line text, *Proceedings of the 11th*

International Conference on Electronic Publishing, 2007, pp. 349-360.

[4] P. Carmichael and M. Tscholl, Cases, simulacra, and semantic web technologies, *Journal of Computer Assisted Learning*, Vol.29, No.1, 2013, pp. 31-42.

[5] H. Chen, *Dark web: Exploring and data mining the dark side of the web*, Springer-Verlag New York Incorporated, 2012, pp. 171-201.

[6] P. Ferguson, et al., Exploring the use of paragraph-level annotations for sentiment analysis in financial blogs, *1 st Workshop on Opinion Mining and Sentiment Analysis*, 2009.

[7] N. Godbole, M. Srinivasaiah and S. Skiena, Large-scale sentiment analysis for news and blogs. *Proceedings of the International Conference on Weblogs and Social Media*, Vol.2, 2007.

[8] V. Hatzivassiloglou and K. R. McKeown, Predicting the semantic orientation of adjectives, Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics, 1997, pp. 174-181.

[9] D. Kushal, S. Lawrence and D. M. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, *Proceedings of the World Wide Web Conference*, 2003, pp. 519-528.

[10] B. Liu, Sentiment analysis and subjectivity, *Handbook of natural language processing 2*, 2010, pp. 1-38.

[11] B. Liu, Sentiment analysis and opinion mining, *Synthesis Lectures on Human Language Technologies*, Vol.5, No.1, 2012, pp. 19-21.

[12] MARKUSQ, Sentiment analysis for Twitter in Python, 3.3.2009. Available at: http://stackoverflow.com/questions/573768/sentiment-analysis-for-twitter-in-python (3.4.2013)

[13] Y. Mejova, *Sentiment analysis: An overview.* Comprehensive paper, 2009, pp. 1-34.

[14] P. Melville, W. Gryc, and R. D. Lawrence, Sentiment analysis of blogs by combining lexical knowledge with text classification, *Proceedings of the Conference on Knowledge Discovery and Data Mining*, 2009.

[15] V. Nahar, et al., Sentiment analysis for effective detection of cyber bullying, *Web Technologies and Applications. Springer Berlin Heidelberg*, 2012, pp. 767-774.

[16] B. Pang, L. Lee and S. Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques, *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 79-86.

[17] B. Pang and L. Lee, A sentiment education: Sentiment analysis using subjectivity summarization based on minimum cuts, Proceedings of the 42nd annual meeting for Association of Computational Linguistics, 2004, pp. 271-278.

[18] B. Pang and L. Lee, Opinion mining and sentiment analysis, Foundation and Trends in Information Retrieval, Vol.2, 2008.

[19] A. Popescu and O. Etzioni, Extracting product features and opinions from reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2007, pp. 9-28.

[20] J. Ratkiewitcz, et. al., Detecting and tracking political abuse in social media, *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[21] P. Rosso, M. Errecalde and D. Pinto, Analysis of short texts on the web: Introduction to special issue. *Springer Science+Business Media Dordrecht*, 2013.

[22] B. Snyder and R. Barzilay, Multiple aspect ranking using the good grief algorithm, *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference*, 2007, pp. 300-307.

[23] M. Steyvers and J. B. Tenenbaum, The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, Vol.29, No.1, 2005, pp. 41-78.

[24] I. Subašić and B. Berendt, Peddling or Creating Investigating the Role of Twitter in News Reporting. *Advances in Information Retrieval. Springer Berlin Heidelberg*, 2011, pp. 207-213.

[25] M. Taboada, et. al., Lexicon-based methods for sentiment analysis, *Computational linguistics*, Vol.37, No.2, 2011 pp. 267-307.

[26] M. Thomas, B. Pang and L. Lee, Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2006, pp. 327–335.

[27] P. D. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the Association for Computational Linguistics*, 2002 pp. 417-424.

[28] J. Wiebe, et al., Learning subjective language. *Computational linguistics*, Vol.30, No.3, 2004, pp. 277-308.

[29] T. Wilson, J. Wiebe and P. Hoffman, Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of HLT-EMNLP*, 2005, pp. 347-354.

[30] T. Wilson, J. Wiebe and P. Hoffman, Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, Vol.35, No.3, 2009, pp. 399–433.

[31] H. Xia, et al., Exploiting social relations for sentiment analysis in microblogging, *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013, pp. 537-546.

Appendix A: Studies and performance

Studies	Polarity mining techniques used	Text granularity	Features	Data sources/Domains	Performance (accuracy)
Hatzivassiloglou and McKeown (1997)	Log-linear regression model	document	conjunctions, part-of-speech	Wall Street Journal corpus	Adjectives: precision > 90%
Pang and Lee (2002)	Nave Bayes, maximum entopy classification support vector machines	document	unigram, bi-gram, contextual effect of negation, feature presence of frequency, position	IMDb (Movie review)	82.90%
Turney (2002)	pointwise mutual information	document	bi-grams	Known positive terms such as exellent and negative terms auch as poor movies, cars, banks	66-84%
Kushal et al. (2003)	Support Vector Machines (SVM)	document	semantic features based on substitutions and proximity	Amazon Cnn.Net	88.90%
Popescu and Etzioni (2005)	relaxation labeling clustering	phrase	syntactic dependency templates, conjunctions and disjunctions, WordNet	Amazon Cnn.Net	Opinion phrase polarity: precision: 86% recall: 97% relationships
Wilson et al. (2005)	Adaboost	phrase	subjectivity lexicon	multiperspective Question, Anwsering Opinion Corpus	Contextual polarity: 65.7%
Thomas and B.Pang (2006)	SVM	speech segment	reference classification	2005 U.S. floor debate in the House of Representatives	With same-speaker links and agreement links: 71.16%
Godbole et al. (2007)	lexical (WordNet)	word	graph distance measurements between words based on relationships of synonymity and anonymity, commonality of a words	newspapers, blog posts	82.7-95.7%
Annet and Kondrak (2008)	lexical (WorldNet) and SVM	document	number of positive/negative adjectives/adverbs, presence, absence or frequency of words, minimum distance from pivot words in WorldNet	movie reviews, blog posts	65.4-77.5%
Ferguson el al. (2009)	Multinomial Naive Bayes (MNB)	phrase	binary word feature vectors	financial blog articles	75.25%
Wilson et al. (2009)	boosting, memory-based learning, rule learning, and support vector learning	phrase	words, negation, polarity modification features	MPQA Corpus	83.60%
Melville et al. (2009)	Bayesian classification with lexicons and training documents	document	words	Blog posts reviewing software, political blogs, movie reviews	Blogs: 91.21% Political:63.61% Movies: 81.42%
Taboada et al. (2011)	Semantic Orientation CALtulator (SO-CAL),	words	all words, nouns, verbs, adjectives, adverbs, negation, irrealis blocking, modals	Epinions 1, Epinions 2, Movie (IMDb), Camera	Epinions: 65.5-81.5%, Epinions: 2 65.25-80%, Movie: 68.05-76.37%, Camera: 64.70-81.16 %
Nahar et al. (2012)	LibSVM, Probabilistic Latent Semantic Analysis (PLSA)	document	cases, bully, non-bully	XML files of MySpace Dataset (MSD) and SlashDot Dataset (SDD)	MSD & SDD: > 96.61%
Xia et al. (2013)	least squares, Lasso, MinCuts, LexRatio, Sociological Approach to handling Noisy and short Texts (SANT)	document	bag of words, n-gram	Stanford Twitter Sentiment (STS) and Obama-McCain Debate (OMD)	STS: 67-79.6% OMD: 61.5-76.3%