# Mining Opinion from Text Documents: A Survey

Khairullah Khan, Baharum B.Baharudin, Aurangzeb Khan, Fazal-e-Malik

Department of Computer and Information Sciences , Universiti Teknologi PETRONAS, Malaysia
email: khairullah_k@yahoo.com baharbh@petronas.com.my aurangzebb_khan@yahoo.com
fazal_i_malik@yoo.com

*Abstract*—**Opinion Mining is a process, used for automatic extraction of knowledge from the opinion of others about some particular topic or problem. With the growing availability of online resources on web and popularity of fast and rich resources of opinion sharing such as online review sites and personal blogs, Opinion Mining has become an interesting area of research. World Wide Web is a fastest medium for opinion collection from users. Human perception and user opinion has greater potential for knowledge discovery and decision support. In this paper we have presented a survey which covers techniques and methods that promise to enable us to get opinion oriented information from text. This research effort deals with techniques and challenges related to sentiment analysis and Opinion Mining. We have followed systematic literature review process to conduct this survey. Our focus was mainly on machine learning techniques on the basis of their usage and importance for opinion mining. We have tried to identify most commonly used classification techniques for opinionated documents to assist future research in this area.**

***Index Terms*— Opinion Mining, knowledge discovery, sentiment, text classification.**

## I. INTRODUCTION

Opinion Mining (OM) is a new and emerging area of research which deals with information retrieval and knowledge discovery from text using Data Mining (DM) and Natural Language Processing (NLP) techniques. The goal of OM is to make computer able to recognize and express emotions. A thought, view, or attitude based on emotion instead of reason is called sentiment. Thus OM is also referred as sentiment analysis. Business organizations are spending a lot of money through consultants and surveys to find consumer sentiments and opinions about their products. Similarly individuals are interested in others opinion about products, services, issues and event for finding best choices This type of survey is now become easy to collect through web forums, blogs, discussion groups and comment boxes. Opinion can be collected from any person in the world about any thing through review sites, blogs and discussion groups etc. Extraction of information and knowledge discovery is an important area of research. The problem for knowledge extraction from World Wide Web is even more challenging because the data stored in the web is very dynamic in nature. The data is rapidly changing due to continuous updating and addition of latest information every time. Websites can be used for a variety of applications. One of an important application of web data is to collect user opinion and extract meaningful patterns from it. During decision making process most of us get help from others. It is a natural phenomenon that good decision can be taken on the basis of opinion of others. Before the World Wide Web, opinion was to share verbally or through letters, we had to ask our friends to suggest which item is the best among the rest or to explain what features of an item are good and what are bad. Due to the World Wide Web, it has become possible to share knowledge and to get advantage from each other experience. Over 75,000 new blogs are created daily along with 1.2 million new posts each day and 40% of people in modern world rely on opinion, reviews, and recommendations collected from blogs, forums and other related sites [1]. This shows the growing importance and need of OM.

Automatic detection of emotions in texts is becoming increasingly important from an applicative point of view. Survey, blogs and review site are used to collect customer opinion about products to get knowledge about the reputation of the company in the market. Companies are interested to know about the people demand. These surveys are to be then summarized to produce a report about the good and bad aspects of particular products. The summary reports are then to be used for decision making equally by manufacturer, customer and merchant. For business intelligence, it is useful to classify each opinion according to the aspect of the business or transaction e.g. product quality, ordering or credibility [2]. This summarization task is different from traditional text summarization. On the other hand, OM is based on the features of the product on which the customers have expressed their opinions which helps to decide whether the opinions are positive or negative [3]. OM can be used for recommendation system, government intelligence, citation analysis, human-computer interaction and its computer assisted creativity [4]. Similarly information extraction from formally written scientific literature is as measurable by precision and recall process that is used to find levels of correctness and exhaustiveness [5].

The rest of the paper is organized as follows. In Section 2 we have discussed methodology, Section 3 gives an overview of OM problems and linguistics approaches, Section 4 presents document classification models, in section 5 challenges and issues are discussed, while section 6 concludes the paper.

## II. METHODOLOGY

In this survey we have used systematic literature review process. We have followed standard steps for searching, screening, data-extraction, and reporting.

### A. Searching

First of all we tried to search for relevant papers, presentations, research reports and policy documents that were broadly concerned with opinion mining from text. We identified appropriate electronic databases and websites. Potentially relevant papers were identified using the following electronic databases and websites.

- IEEE Explore
- Springer Linker
- Science Direct
- ACM Portal
- Googol Search Engine

### B. Development of a search strategy

For best and consistent search, a systematic search strategy was adopted. Proper keywords, queries, and phrases were derived from the desired research question. These keywords were arranged into categories and related keyword words were arranged. Some facilities of digital libraries like sort by year etc were also used. The search key words were refined to include only those key words which have produced successful results. We used Boolean logic for efficient searching for example (Sentiment OR Opinion OR review OR recommendations). We also tried combination of words like Opinion Mining, Sentiment Analysis, Subjectivity Analysis, Opinion Spam etc.

### C. Screening

Each search results were checked and assessed on screen to find relevance for inclusion and exclusion with the given criteria as below. We made two categories of papers i.e. in or before 2002 and after 2002.

The following studies were included

- The result statement is written in English;
- The research is conducted after 1980.
- Published and/or unpublished research
- Focus on Opinion mining
- Focus on Machine Learning and
- Focus on Natural Language Processing

The following studies were excluded:

- Non English writing;
- Study before 1980
- Study with no experimental approach;
- Based on single person opinion;
- not focused on Opinion Mining

### D. Data Extraction

To find evidence and check the quality of papers we carried out an in-depth study of the results provided in the studies. In our future work we will try to make this step more strong and effective.

### E. Synthesis

A little attention was given to this step and we will make sure to add this in our future work. We will try to development a framework for data analysis and identification of key themes.

### F. Reporting

We have tried to get some reports drawn using tables and graphs on the basis of existing studies.

## III. OM PROBLEMS

The term OM appears in a paper by Dave et al. [6], according to Dave the idea of opinion-mining tool is to "process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinion". But with the passage of time more interesting application and developments came in existence in this area and now its main goal is to make computer able to recognize and generate emotion like human. In this section we will discuss OM research problems. The OM problem is to extract human perception from user generated text. OM problem is a mixed problem of Information Retrieval (IR) and Natural language processing (NLP) [45]. Its main concern is to automate extraction of opinionated, sentimental, and emotional expression from typically unstructured text.

### A. Analysis of linguistic resources for OM

For opinion extraction it is required to know the linguistic terms and get the idea from the text. Classification of contents of document into positive and negative, and subjective and objective terms is the basic problem of opinion mining. The terms are identified by syntactic features. According to Livia Polanyi and Annie Zaenen, "The most salient clues about attitude are provided by the lexical choice of the writer, but the organization of the text also contributes information relevant to assessing attitude" [7]. Another main focus is on subjectivity detection. Subjectivity is used to express private states in the context of a text or conversation. Private state is a general term for opinions, evaluation, beliefs, perception, emotions, peculation and etc [8]. Objective statement conveys information in accordance with the intention of the author. If a user feedback has no judgment or opinion on the source content then it is called objective. Jaehui Park et al. in their work [9] categorized objective statements into summary and additional information. Where the summary explain the idea of the source contents and additional information are those facts which do not appear in the source contents. Ahmed Abbas et al. [10] have presented a very good taxonomy about OM linguistic aspects. They have categorized the OM linguistic job as classification, features, techniques and domains. Changli Zhang et al. [11] in their work have used bag-of-word(BOW) and appraisal phrase and get 79.0% result through BOW and 80.26 with the combination of BOW and appraisal phrase. In [12] Xiaowen Ding and Bing Liu, by experiments they have shown that context rules are helpful to improve the recall without much loss in precision. In [13] Minqing Hu and Bing Liu have used NL Processor linguistic parser to parse each review to split text into sentences and to produce part of speech tags for each word like noun, verb, adjective etc [14,15]. Some authors have taken term senses into account

and assume that a single term can be used in different sense and can present different opinion. They use WordNet Synsets for different senses of the same term [16].

### B. Text features Identification and Orientation

The text features identification has three different levels words, sentences and documents. Existing research work presents different techniques and ideas for extraction of sentimental terms from text. According to linguist rules words and phrases are categorized as noun, verbs, adjectives and adverbs, most of the work use part of speech (POS), stop words removal, fuzzy pattern matching, stemming, phrase patterns, punctuation, polarity tags, appraisal groups, semantic orientation, link-based patterns, document citations, and stylistic measures for extract of sentiments [17,18,19,50].

### C. Adjectives, Noun, Verbs, and Adverbs

while in comparative sentences authors of text compare different aspects of the object or topic under discussion. Existing research of polarity classification mainly focus on adjectives and adverbs to identify subjectivity [20,21,22]. From experiments they have shown that opinion extraction using adjective has precision of 64.2% and a recall of 69.3%. Most commonly used tool for adjective identification is WordNet [23]. WordNet used is by OM researchers for adjective words identification and semantic orientation [24,25,26,27]. Farah Benamara et al. have proposed that adjective and adverbs are better than adjective alone [49]. In most of the exiting work, sentiment expressions mainly depend on some words, which can express subjective sentiment orientation. For example, good for positive and bad is used for negative sentiment orientation. Such subjective words are actually called adjective in linguistic terms. Verb identification plays an important role in finding relationship between subjective and objective terms. For purposes of natural language processing, several researchers have looked into the acquisition of verb meaning, and sub categorizations of verb frames in particular. Claire Nedellec [28] have presented an interactive machine learning system called ASIUM, which is able to acquire taxonomic relations and sub categorization frames of verbs based on syntactic input. According to Turney, Adjectives, Nouns, Verbs and Adverbs are grammatical categories which have the capacity to express emotion or subjectivity [30].

### D. Semantic Orientation of Text

Classification of sentimental expression according to their meaning and background knowledge is called semantic orientation. Although syntactic analysis plays a key role in document classification but it is not sufficient to extract the concept from the text only through syntax. L. Cai and T. Hofmann [29] combined information-theoretic measures and semantic knowledge of a hierarchy using WordNet to extract concept from text automatically. Their model is based on the distribution of predicates and their arguments. Breaking multi-word expression, mapping of synonymous words into different components, and words with multiple meaning as one single component are the issues which can be resolved through semantic analysis. Turney [30] and Pu Wang et [31] have used bag of word (BOW) and semantic concept to enrich the representation of text classification and to extract concept from text.

### E. Ontology Based Learning

Ontology based learning is a growing area of research for extracting opinion from text. Ontology integrates the domain knowledge of individual words into the terms for learning and capture concept from text. The relationship between terms in text is helpful in understanding the background knowledge Ontology can be defined as a formal knowledge representation system (KRS) which has three main components: classes (or concepts or topics), instances (which are individuals which belongs to a class) and properties (which link classes and instances allowing to insert information regarding the world represented into the ontology) [48]. Ontologies based clustering combines lexical and concept hierarchies to improve results in both supervised and unsupervised clustering [33,34,35]. Wen Zhang et al [33] have worked on text classification based on multi-word using ontology. Hotho, A. et al. [34] have proposed to integrate core ontologies as background knowledge into the process of clustering text documents. Takahiro [35] combines conventional natural language processing like syntactic parsing for the unstructured part together with semantic information such as metadata and ontology retrieved by the structured part.

## IV. Machine Learning and Text Classification

There are two main approaches for sentiment classification. The knowledge based and supervised machine learning [36]. In the knowledge-based approach predefined affect dictionaries of opinion words are used to search the input words and find its effects. While in supervised machine learning a trained statistical classifier is used for sentiment classification. The trained classifier predicts the sentiment orientation of input documents. Both of these approaches rely on affective vocabulary although its use is different [37,38,39,40].

### A. Commonly Used Machine Learning Models

Different supervised categorization algorithms have been used so far for polarity classification tasks. Most commonly used methods are Support Vector Machine (SVM), Naïve Bayesian Classifier. While other methods like Maximum Entropy, Decision Tree, Neural Network, Latent Dirichlet Allocation (LDA), and Probability Latent Semantic Analysis (PLSA) are also used for this purpose. We randomly selected 336 related papers categorized the papers according to the use of Machine Learning Algorithms as shown in table 1. We found from this survey that the using graph of SVM is going up while Naïve Bayesian is also consistently used for this purpose.

| Year | SVM | NBA | Other | Total |
|---|---|---|---|---|
| 2008 | 45 | 19 | 11 | 74 |
| 2007 | 32 | 23 | 17 | 72 |
| 2006 | 25 | 28 | 13 | 60 |
| 2005 | 11 | 9 | 35 | 55 |
| 2004 | 5 | 11 | 19 | 35 |
| 2003 | 6 | 9 | 8 | 23 |
| 2002 | 2 | 8 | 6 | 17 |

Table-1. Usage of Machine Learning Techniques

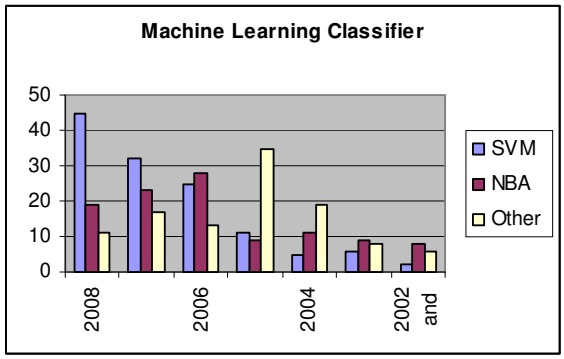


Figure 1. Machine Learning Classifiers

## V. General Challenges and Issues

OM suffers from several different challenges, such as determining which segment of text is opinionated, identifying the opinion holder, determining the positive or negative strength of pinion. OM is concerned with the human reviews, emotions and sentimental discussion. Every one has their own perception and concern about a particular problem, issue, or topic. Opinionated text may be fake, irrelevant and or ambiguous information. Opinions are far harder than facts to describe. Opinion sources are typically informally written and highly diverse. The following general challenges are pointed out so far by the different authors [41,42,43,44].

- Authority: An accepted source for the information or advice, either an expert on the subject or a persuasive force [41].
- Credibility: A quality of opinion being believable, trustworthy [41].
- Spam: Analysis of spam opinion. Nitin Jindal and Bing Liu in their paper [42] have identified spam as shown in the table-2.
- Non Expert opinion: Open forums and blogs are often suffered from non expertise. They can not provide review text in a proper manner.
- Domain Dependent: Normally opinions are on specific issue, problem, or topic. Therefore the techniques are normally domain dependent. But it leads to the problem non-generalization [44].
- Language differences: different use different language context in their opinion, even in English forum they write in roman words of their own language, which makes the OM task difficult. e.g someone can add the text about the book roman English of urdu language as "ye book mujey passand hey, koink ye poori course ko cover karatha hey ".
- Effects of syntax on semantics: Breaking multi-word expression, mapping of synonymous words into different components, and words with multiple meaning as one single component (polysemous) Sentence document Complexity, Contextual Sentiments, Heterogeneous documents, Reference Resolution, Modal operators: might, could, and should are still remain challenging problems in this area [5,45,46,47].
- Effect of sense on terms, finding subjective terms, and multi-word document analysis [48,49].

| Spam Type | Number of Reviews |
|---|---|
| Different userid on the same product | 3067(104) |
| Same Userid on different products | 50869(4270) |
| Different userid on different products | 1383(114) |
| Total | 55319(4488) |

Table-2. Spam Analysis by Nitin Jindal and Bing Liu

## VI. Conclusion

Since OM is an emerging and rapidly growing field of interest so in this paper we have mainly focused on the existing research work to explore the field in order to find a clear direction for future work. The year 2001 and 2002 seem to mark the beginning of the awareness regarding this problem but proper name "Opinion Mining" was actually given by Dave in 2003. Now with rapid interest in machine learning and improvement in NLP sentiment analysis becomes a challenge for researcher. Hundreds of papers have been published on the subject.

We have collected 450 papers directly relevant to the area. From the graph shown in figure 2, signifies the growing emergence of this area. We also tried to explore challenges and issues in this area and compare the most commonly used techniques. The key issues in this area are authority, credibility, spam detection, language difference, non-expert opinion, domain dependency, effect of syntax on semantic.

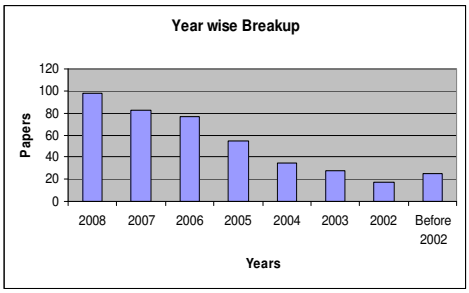


Figure 2. Year wise published papers.

## VII. REFERENCES

[1] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008) 1–135 2008.

[2] Hu and Bing Liu, "Mining Opinion Features in Customer Reviews, Mining", American Association for Artificial Intelligence 2004.

[3] Sushmita Mitra et al, "Data Mining", Wiley, 2003.

[4] Bing Liu, "Web Data Mining – Exploring hyperlinks, Contents, and Usage Data", 2006, Springer, December, 2006.

[5] Seth Grimes, "Sentiment Analysis: Opportunities and Challenges",http://www.b-eye-network.com/view/6744, 2008.

[6] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion, extraction and semantic classification of product reviews," in Proceedings of WWW, pp. 519–528, 2003.

[7] Livia Polanyi and Annie Zaenen, "Contextual valence shifter", Computing Attitude and Affect in Text: Theory and Applications chapter 1, pages 1–10. Springer, 2006.

[8] Nitin Jindal and Bing Liu, "Mining Comparative Sentences and Relations", American Association for Artificial Intelligence, www.aaai.org., 2006.

[9] Jaehui Park et al., "web content summarization using social bookmarks", WIDM08, 2008.

[10] Ahmed Abbas, etl. "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums" ACM Transactions on Information Systems, Vol. 26, No. 3, Article 12, 2008.

[11] Changli Zhang et al., "Sentiment Classification for Chinese Reviews Using Machine Learning Methods based on String Kernel", International on Convergence and Hybrid Information Technology, 2008.

[12] Xiaowen Ding and Bing Liu, "The Utility of Linguistic Rules in Opinion Mining", SIGIR'07, Amsterdam, Netherlands, 2007.

[13] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews", KDD'04, August 22–25, 2004, Seattle, Washington, USA, 2004.

[14] Yejin Choi et al., "Joint Extraction of Entities and Relations for Opinion Recognition", Proceedings of Empirical Methods in Natural Language Processing (EMNLP), 2006.

[15] Veselin Stoyanov et al., "Partially Supervised Coreference Resolution for Opinion Summarization through Structured Rule Learning", Proceedings of Empirical Methods in Natural Language Processing (EMNLP), 2006.

[16] Alina Andreevskaia and Sabine Bergler, "Mining Word-Net for fuzzy sentiment: Sentiment tag extraction from WordNet glosses", In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06), pages 209–216, Trento, IT. 2006.

[17] Janyce Wiebe et al., "Recognizing and Organizing Opinions Expressed in the World Press", AAAI Spring Symposium on New Directions in Question Answering, 2003.

[18] Janyce Wiebe et al., "NRRC Summer Workshop on Multiple-Perspective Question Answering: Final Report", Theresa Wilson. 2002.

[19] David Pierce et al, "User-Oriented Machine Learning Strategies for Information Extraction: Putting the Human Back in the Loop", Working Notes of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining, 2001.

[20] Wiebe, J., "Learning subjective adjectives from corpora", Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000), Austin, Texas, 2000.

[21] P. Chesley, B. Vincent, L. Xu, and R. Srihari, "Using verbs and adjectives automatically classify blogs sentiment," in AAAI Symposium on Computational Approaches to Analyzing Weblogs (AAAI-CAAW), pp. 27–29, 2006.

[22] V. Hatzivassiloglou and K. McKeown, "Predicting the semantic orientation of adjectives," in Proceedings of the Joint ACL/EACL Conference, pp. 174–181, 1997.
[23] C. Fellbaum, ed., Wordnet: An Electronic Lexical Database. MIT Press, 1998.

[24] Eric Breck et al. "Identifying Expressions of Opinion in Context.", Twentieth International Joint Conference on Artificial Intelligence (IJCAI), 2007

[25] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in Proceedings of Language Resources and Evaluation (LREC), 2006.

[26] A. Esuli and F. Sebastiani, "PageRanking WordNet synsets: An application to opinion mining," in Proceedings

of the Association for Computational Linguistics (ACL), 2007.

[27] J. Kamps, M. Marx, R. J. Mokken, and M. de Rijke, "Using WordNet to measure semantic orientation of adjectives," in Proceedings of LREC, 2004.

[28] Claire Nédellec. Corpus-based learning of semantic relations by the ILP system, Asium. In James Cussens, editor, Proceedings of the 1st Workshop on Learning Language in Logic, pages 28--39, Bled, Slovenia, June 1999.

[29] L. Cai and T. Hofmann. Text Categorization by Boosting Automatically Extracted Concepts. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, 2003.

[30] Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL'02), 417-424, 2002.

[31] Turney P. and Littman, "Measuring praise and criticism: Inference of semantic orientation from association", ACM Transactions on Information Systems, 21(4), 315-346, 2003.

[32] Pu Wang and Carlotta D.," Building Semantic Kernels for Text Classification using Wikipedia", KDD'08, Las Vegas, Nevada, USA, 2008.

[33] Wen zhang , Taketoshida, and Xijin Tang, "Text classification based on multi-word with support vector machine" Elsevier knowledge-based systems, 2008.

[34] Hotho, A. Staab, S. Stumme, G, "Ontologies Improve Text Document Clustering", Data Mining, 2003. ICDM 2003. Third IEEE International Conference, 2003.

[35] Takahiro Kawamura Shinichi Nagano Yumiko Mizoguchi, "Ontology-based WOM Extraction Service from Weblogs", SAC'08, Fortaleza, Cear´a, Brazil, 2008.

[36] Bei Yu, Stefan Kaufmann,Daniel Diermeier, "Exploring the Characteristics of Opinion Expressions for Political Opinion Classification", Procedding of the 9th Annual International Digital Government Research Conference, 2007.

[37] Hatzivassiloglou, V. & McKeown, K, "Predicting the semantic orientation of adjectives", Proceedings of the 35th ACL Conference, 174-181, 1997.

[38] Dave, K., Lawrence, S., & Pennock, " Mining the peanut gallery: opinion extraction and semantic classification of product reviews", Proceedings of the 12th international conference on World Wide Web (WWW2003), 519-528,2003.

[39] Hu, M. & Liu, B., "Mining and summarizing customer reviews", Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD2004), 168-177, 2004.

[40] Pang, B., Lee, L., & Vaithyanathan, S, "Thumps up?, Sentiment classification using machine learning techniques", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2002), 79-86, 2002.

[41] Jack G. Cornrad et al. 'Professional Credibility: Authority on The web', WICOW08, 2008.

[42] Nitin Jindal and Bing Liu, "Opinion Spam and Analysis", WSDM08, 2008.

[43] Nitin Jindal and Bing Liu, "Review Spam Detection", WWW2007, ACM 978-1-59593-654-7/07/00005, 2007.

[44] G. Salton, "Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer", Addison-Wesley, 1989.

[45] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008) 1–135 2008.

[46] Carlo Strapparava, Rada Mihalcea, "Learning to Identify Emotions in Text", SAC'08 Fortaleza, Brazil 2008.

[47] Michael Gamon and Anthony Aue, "Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms", In Proceedings
Of the Workshop on Feature Engineering for Machine Learning in Natural Language Processing pages 57–64. ACL 2005.

[48] Antonio Lieto, "Manually vs semiautomatic domain specific ontology building", Thesis dissertation, Chapter 1, 2008.

[49] Farah Benamara, Carmine Cesarano, Diego Reforgiato, "Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone", ICWSM '2007 Boulder, CO USA.