# Machine Learning assisted Sentiment Analysis

Akshi Kumar
Department of Computer Engineering
Delhi Technological University
Delhi, India
akshi.kumar@gmail.com

Teeja Mary Sebastian
Department of Computer Engineering
Delhi Technological University
Delhi, India
teeja.sebastian@gmail.com

*Abstract*—***Enormous importance has been attributed to public opinion and sentiment as what people think has always influenced new business decisions, political mood, governance policies and personal choices. Also, the availability of opinionated user generated data has increased as people have started freely expressing their views on various cyber platforms like blogs, forums, review sites and social networks. This has spurred the unabated growth of opinion mining and sentiment analysis as important research areas. Opinion mining and sentiment analysis are computational techniques that seek to understand opinion and sentiment, subjectivity by analyzing unstructured opinion text. This paper illustrates the convergence of two prominent research areas, namely, Sentiment Analysis and Machine Learning where the latter has proven its merit as a technique for automated Sentiment Analysis.***

*Keywords*-***Sentiment Analysis; Opinion Mining; Machine learning***

## I. INTRODUCTION

Recent years manifest the beginning and growth of the social web, in which individuals freely express, articulate and respond to opinion on a whole variety of topics. Simultaneously, today's information society challenges companies and individuals to create and employ mechanisms to search and retrieve relevant data from the huge quantity of information available and mining for opinions thereafter. Consequently, Sentiment Analysis [1] which automatically extracts and analyses the subjectivities and sentiments (or polarities) in written text has emerged as an active area of research. The enthusiasm shown by researchers in this field has been because they realize the importance attributed to public opinion and sentiment in businesses, governance and decision making processes. The goal of sentiment analysis is to create market/business/governance intelligence, to detect opportunities and issues, understand the public sentiment conveyed in different forms of textual communications. Sentiment analysis has found applications in gauging the success of particular campaigns, understanding potential consumers who are not favorably responding to products, understanding the competitions standing and in picking up on promising trends. It can also be used as an augmentation to present recommendation systems.

Most researchers have defined the Sentiment Analysis problem as essentially a text classification problem and machine learning techniques have proved their dexterity in resolving the sentiment analysis tasks. Although we can generalize Sentiment Analysis as a text classification problem, it comes with its own set unique challenges and issues which have to be addressed by the machine learning techniques.

To find the opinion bearing portions in a document we have to first understand that sentences in a document maybe objective or subjective. It is generally considered that subjective sentences are the opinion carriers in a document as subjective sentences express some personal belief or view and objective sentence expresses some factual information. But an objective sentence may state a fact but they might also be conveying some sentiment, for e.g. "Women are getting more opportunities in offices". Models cannot infer opinions from facts as they are trained on opinion words.

One may further think that sentiment orientation can be easily identified by a set of keywords. But coming up with the right set of keywords is not a trivial task as shown by Pang et al.[2]. This is because sentiment can also be expressed in subtle ways or sarcastically which makes its identification problematic when considered separately at the sentence level. Another problem is that some words have both strong positive and negative sentiment. Classifying them without knowing the context is a difficult task. Order dependence also manifests itself at more fine-grained levels of analysis: "A is better than B" conveys the exact opposite opinion from "B is better than A" [3]. Also, the increased use of informal English, abbreviations and bad spellings on Web platforms makes the sentiment conveyed difficult to comprehend and classify.

Addressing the issues mentioned above that range from tackling the vague definition of sentiment and the complexity of its manifestation in text, brings up new questions providing ample opportunities for both quantitative and qualitative work. To automate sentiment analysis, different approaches have been applied to predict the sentiments of words, expressions or documents. These include Natural Language Processing (NLP) and Machine

Learning algorithms like supervised learning[2,4,5,6,7], unsupervised learning [8,9,10,11] and semi-supervised learning[12,13,14].

This paper aims to probe the role machine learning as a prominent assisting technology that has ascertained substantial gains in automated sentiment analysis research and practice by developing standards and improving effectiveness.

## II. TERMS AND DEFINITIONS

Formally stating, Sentiment Analysis is the computational study of opinions, sentiments and emotions expressed in text[1]. That is, sentiment analysis aims at detecting subjective information in a document and then determining the orientation of the sentiment expressed in the document. Researchers use the terms sentiment analysis, opinion mining, subjectivity analysis, review mining and appraisal extraction interchangeably.

The terms frequently used to define the Sentiment Analysis problem has been given below:

- Opinionated Document: A review, forum post, blog or tweet that contain sentences which expresses any kind of opinion, sentiment or emotion.

- Target Entity or Object: An individual, organization, product, service, issue or event that is being discussed in the opinionated document.

- Object Feature: The target entity can have features or components associated with it. For e.g. When the target entity is a camera, the shutter speed is a feature. Opinions can be expressed specifically about a feature instead of generally about the object e.g., the *battery life* of the phone is good.

- Opinion Holder: A person or an organization which expresses an opinion is called an opinion holder or opinion source. Authors of the blog post or a review are the opinion holders. In case of news articles the opinion holder is explicitly mentioned.

- Sentiment Orientation or Polarity: The orientation or polarity of the opinion is whether the opinion on a feature is positive, negative or neutral. Opinions vary in intensity from very strong to weak. For example a positive sentiment can range from content to happy to ecstatic. Thus, strength of opinion can be scaled and depending on the application the number of levels can be decided.

The following example in Figure 1 illustrates the basic terminology of sentiment analysis:

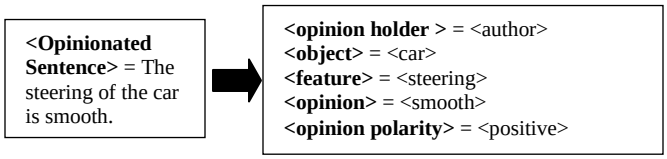


**Figure 1:** Example corresponding to Terminology of Sentiment Analysis

The objective in a sentiment analysis task is to find four parameters (the opinion, the opinion holder, the object, the feature) corresponding to each other. This is a challenging problem, as finding each parameter in itself is difficult enough.

Since 1940's, many knowledge-based systems have been built that acquire knowledge manually from human experts, which is very time-consuming and labor-intensive. To address this problem, Machine Learning algorithms have been developed to acquire knowledge automatically from examples or source data [15]. It is defined as "any process by which a system improves its performance" [16].

Machine Learning is programming computers to optimize a performance criterion using example data or past experience[17]. A learner (a computer program) processes data D representing past experiences and tries to either develop an appropriate response to future data, or describe in some meaningful way the data seen [18].

Machine Learning Algorithms can be broadly divided into Supervised, Unsupervised and Semi-Supervised learning algorithms. In supervised learning, the aim is to learn a mapping from the input to an output whose correct values are provided by a supervisor[17] Supervised Algorithms use labeled training data to create a function which would predict the correct output. In unsupervised learning, there is no supervisor and there is only input data. The aim is to find the regularities in the input. There is a structure to the input space such that certain patterns occur more often than others and these patterns help in prediction [17]. Semi-Supervised Algorithms use both labeled and unlabelled data for training purposes.

The following sections introduce and discuss the sentiment analysis as a text-classification problem and the multiplicity of machine learning techniques that have been employed by various researchers over the years for sentiment analysis problem.

## III. SENTIMENT ANALYSIS AS A CLASSIFICATION TASK

Sentiment analysis is a challenging interdisciplinary task which includes natural language processing, web mining and machine learning. It is a complex task and encompasses several separate tasks, viz:

- Subjectivity Classification
- Sentiment Classification
- Object Feature Identification
- Opinion Holder Identification

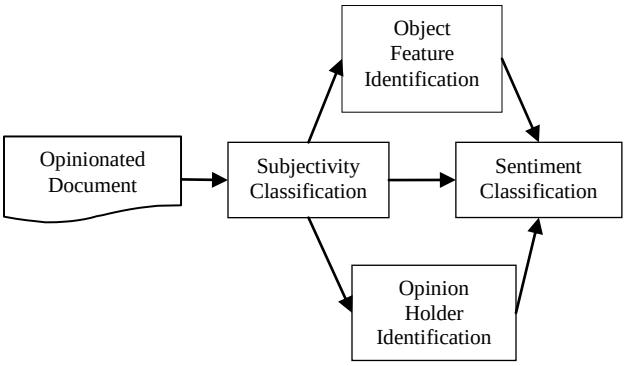


**Figure 2**: Tasks in Sentiment Analysis

### *A. Subjectivity Classification*

As mentioned earlier, a document may contain both subjective and objective sentences. For sentiment analysis, it is beneficial to be able to differentiate between opinionated and non-opinionated sentences. Subjectivity classification is thus the task involving the classification of sentences/ phrases/ words as opinionated or not.

### *B. Sentiment Classification*

After establishing that a sentence is opinionated, it is required to know the orientation of the opinion too. Sentiment classification can be a binary classification (positive or negative) [2], multi-class classification (extremely negative, negative, neutral, positive or extremely positive), regression or ranking[4].

### *C. Object and Feature Identification*

Blogs and social media sites do not have a set target or predefined topic and tend to discuss diverse topics. Therefore in these scenarios it becomes essential to know the target entity [11, 19]. Also in case of review sites, the reviewer could talk about certain features of the target object and may like a few and dislike others. It thus, becomes necessary to differentiate between the different features of the object. Feature based opinion sentiment analysis which involves feature extraction and the corresponding opinion is also a goal pursued by researchers

### *D. Opinion Holder Identification*

Detection of opinion holder [20] is to identify direct or indirect sources of opinion or emotion. They are important in genres like news articles and other formal documents. In such documents, the holder of the opinion maybe explicitly mentioned. In blogs and review sites the opinion holder is usually the author who can be identified by the login id.

All text processing approaches require converting text into a feature vector or engineer a suitable set of features. These representations may make the significant features available for machine learning approaches. Few of the features used in practice are given below [3]:

- Words and their frequencies
  Unigrams, bigrams and n-grams along with their frequency counts are considered as features. There has been contention on using word presence rather than frequencies to better describe this feature. Pang et al.[2] so showed better results by using presence instead of frequencies.
- Parts of Speech Tags
  Parts of speech like adjectives, adverbs and some groups of verbs and nouns are good indicators of subjectivity and sentiment.
- Syntax
  Syntactic patterns like collocations, are used as features to learn subjectivity patterns by researchers. The syntactic dependency patterns can be generated by parsing or dependency trees.
- Opinion Words and Phrases
  Apart from specific words, some phrases and idioms which convey sentiments can be used as features, e.g. ”cost someone an arm and leg”[1].
- Position of Terms
  The position of a term within a text can effect on how much the term affects overall sentiment of the text
- Negation
  Negation is an important but tricky feature to incorporate. The presence of a negation usually changes the polarity of the opinion but all appearances do it. For e.g., “no doubt it is the best in the market” .

Sentiment Analysis is formulated as a text-classification problem [3] and therefore traditional machine learning techniques are used for the subjectivity/sentiment classification task. High accuracy classification has been achieved by using a variety of techniques, most of which are heavily reliant on machine learning. Like most machine learning applications, the main task of sentiment classification is to engineer a suitable set of features.

## IV. MACHINE LEARNING TECHNIQUES

We now discuss the various machine learning techniques that have been employed by various researchers over the years for sentiment analysis problem. This section focuses on the unique aspects of the machine learning techniques in sentiment analysis mainly because of the different features involved in case of supervised and semi-supervised techniques. Unsupervised techniques use sentiment driven pattern to obtain labels for words and phrases.

### A. *Supervised Learning*

Supervised learning generally functions as follows: in the initial training phase, an inductive process learns the characteristics of a class based on a feature set of pre classified documents (reference corpus) and it then applies the acquired knowledge to categorize unseen documents, during testing. Several classical classifiers like Naïve Bayes,, Maximum Entropy and Support Vector Machines are most commonly used supervised methods used.

Pang et al.[2] experimented with three classifiers(Naive Bayes, maximum entropy, and support vector machines) using features like unigrams, bigrams, term frequency, term presence and position, and Parts-of-speech to classify movie reviews as good or bad. They concluded that SVM classifier works best and that unigram presence information was most effective. Dave et al. [5] although claim that in some situations, bigrams and trigrams produce better product-review polarity classification.

Using supervised learning for predicting the rating scores has also been done (1-5 stars) in [4]. The problem is formulated as a regression problem since the rating scores are ordinal.

Supervised learning methods have been used for subjectivity classification too. Most works focus on adjectives and their effects on subjectivity of sentences [7]. Wiebe et al.[21] used the naive Bayes classifier to develop a gold standard data set for subjectivity classification.

Yu and Hatzivassiloglou [22] developed three approaches to classify opinions from facts at the sentence level. The first approach explored the hypothesis that "within a given topic, opinion sentences will be more similar to other opinion sentences than to factual sentences". The second method trained a Naive Bayes classifier , using sentences in opinion and fact documents as the examples of the two categories. The features included words, bigrams, and trigrams, as well as the parts of speech in each sentence. They also included in their features the counts of positive and negative words in the sentence , as well as counts of the polarities of sequences of semantically oriented words. Third approach involved training separate Naive Bayes classifier for each different subset of the features. The goal was to reduce the training set to the sentences that are most likely to be correctly labeled. They assumed as ground truth the information provided by the document labels and that all sentences inherit the status of their document as opinions or facts. Then they train the first classifier on the entire training set. Then they used the classifier to predict the labels of the training set. The sentences that were labeled incorrectly were removed. The second classifier then trained on the reduced training set and this went on until the training set could no longer be reduced.

Wilson et al. [23] also formulate sentiment detection as a supervised learning task. However, instead of using just text classification, they focus on the construction of linguistic features, and train classifiers using Boostexter [24]. Incorporating background knowledge, in terms of linguistic rules, in such classifiers is an interesting direction for future work.

There has been a growing interest in the use of background, prior or domain knowledge in supervised learning. Most of this work has focused on using such prior class-bias of features to generate labeled examples that are then used for standard supervised learning. Provided with some features associated with each class, Wu and Srihari [25] assigned labels to unlabeled documents, which were then used in conjunction with labeled examples to build a Weighted Margin Support Vector Machine.

Another paper that includes prior knowledge is [26]. They constructed a generative model based on a lexicon of sentiment-laden words, and a second model trained on labeled documents. The distributions from these two models were then adaptively pooled to create a composite multinomial Naïve Bayes classifier that captured both sources of information. By exploiting prior lexical knowledge they dramatically reduced the amount of training data required. In addition, by using some labeled documents they were able to refine the background knowledge, which is based on a generic lexicon, thus effectively adapting to new domains.

Adaption to different domains is crucial as the accuracy of sentiment classification can be influenced by the domain. Thus, classifiers trained in a certain domain give poor results in other domains. This is because phrases can be expressing different sentiments in different domains.

### B. *Unsupervised Machine Learning*

There has been shift from using supervised approaches to using unsupervised and semi supervised approaches as the manual effort to annotate a huge corpus is too much. Unsupervised learning approaches first build a sentiment lexicon in an unsupervised manner, and then resolve the strength of sentiment (or subjectivity) of a text using a function based on the orientation (or subjectivity) indicators.

Thus, an important task of applying this technique is the construction of the lexicon by means of unsupervised labeling of words or phrases with their sentiment orientation or subjectivity status.

To create a lexicon Turney [8] suggested comparing whether a phrase was more likely to co-occur with the word "poor" or "excellent". The basic idea was that a phrase has a positive semantic orientation when it has good associations and similarly negative semantic orientations when it has bad associations. The relationship between an unknown word and a set of manually-selected seeds defined by PMI (Point-wise mutual information), was used to place it into a positive or negative subjectivity class.

Kim and Hovy [9] manually created a small seed list of positive and negative words that contained verbs and adjectives. The synonyms and antonyms of the words were extracted from WordNet and then added to appropriate lists (synonyms would have same orientation and antonyms opposite). The seed lists were further developed by using the expanded list to extract another set of words. They then

calculate the sentiment strength of the unseen word by determining how it interacts with the sentiment seed list.

Kamps[10] measured similarity of words by using distance between words based on WordNet lexical relation. They collected all words in WordNet, and related words that could be synonymous, i.e. were part of the same synset. A graph was created with edges connecting each pair of synonymous words. The distance between two words $w_i$ and wj was the length of a shortest path between wi and wj . The orientation of a term was determined by its relative distance from the two seed terms good and bad. The values ranged from [-1, 1] with the absolute value indicating the strength of the orientation.

Gamon et al. [11] used the unsupervised learning technique for identification of aspects or features. They presented an unsupervised aspect identification algorithm that employed clustering over sentences with each cluster representing an aspect. Sentence clusters were labeled with the most frequent non-stop word stem in the cluster.

### *C. Semi Supervised Machine Learning*

Semi Supervised Learning models learn from both tagged and untagged data. The untagged data provides no information about subjectivity or sentiment polarity but they contain information about the joint distribution of the classification features. Bootstrapping is usually the technique used in semi supervised learning Bootstrapping is fundamentally to use the output of an existing initial classifier to produce labeled data, to which a supervised learning algorithm is later applied. This method is also called self-training.

Riloff et al.[12] proposed a bootstrapping process to identify subjective patterns. A bootstrapping process is used that learns linguistically rich extraction patterns for subjective (opinionated) expressions. Two high-precision classifiers, Hp-Subj and Hp-Obj, label unannotated data to automatically create a large training set, which is then given to an extraction pattern learning algorithm. The learned patterns are then used to identify more subjective sentences. A set of syntactic templates was needed to represent the space of possible extraction patterns.

Co-training is another semi supervised method that has been applied. Jin et al. [13] created disjoint training sets for building two initial classifiers. The bootstrap document was then tagged using each of the trained HMM(Hidden Markov Model) based classifiers. The opinion sentences that were agreed upon by both classifiers were extracted and saved in the database if it was unique. The newly discovered data was randomly split and added to the training set of the two classifiers. This bootstrap process was continued until no new data could be discovered.

Graph based semi supervised technique has been used in the task of rating inference by Goldberg and Zhu[14].

Given below is a table which compares the accuracy of the different machine learning techniques implemented by researchers (This is not an exhaustive table):

TABLE I. COMPARATIVE RESULTS

| ***Literature and author*** | ***Machine Learning Method*** | ***Classifier/Training Set*** | ***Accuracy*** |
|---|---|---|---|
| Pang et al. (2002)[2] | Supervised | NB | .815 |
| | | ME | .810 |
| | | SVM | .829 |
| Dave et al. (2003)[5] | Supervised | SVM | |
| Turney and Littman (2002)[27] | Unsupervised | SO-PMI-IR using a two billion word corpus | .894 |
| Riloff et al. (2003)[12] | Semi Supervised | News document from FBIS | .733 |
| Pang and Lee (2004) | Supervised | SVM | .872 |
| Kim nd Hovy (2004)[9] | Unsupervised | WordNet | .81 |
| Kamps et al. (2004)[10] | Unsupervised | WordNet | .787 |
| Aue and Gamon (2005)[11] | Supervised | SVM | .905 |
| Jin et al. (2009)[13] | Semi Supervised | Online product review from Amazon | .771 |

a. .NB: Naïve Bayes
b. .ME: Maximum Entropy
c. SVM: Support Vector Machine.

While machine learning methods have established to generate good results, there are associated disadvantages. Machine learning classification relies on the training set used, the available literature reports detail classifiers with high accuracy, but they are often tested on only one kind of sentiment source, mostly movie review, thus limiting the performance indication in more general cases. Further, gathering the training set is also arduous; the noisy character of input texts and cross-domain classification add to the complexities and thus push the need for continued development in the area of sentiment analysis.

## V. CONCLUSION

Web is an ever expanding sea of information and sentiment analysis is one of the ways that can be used to analyse it and confer structure to it. Machine Learning is one of the foremost techniques used to achieve this end. This paper attempted at exploring the union of the two major research fields, Sentiment analysis and Machine Learning. We can conclude by saying that all sentiment analysis tasks are challenging and difficult. Our knowledge

and comprehension of the problems in this field is still developing. The many practical applications of sentiment analysis is urging researchers to make significant improvements to understand and work in the sentiment analysis domain.

## REFERENCES


[1] B. Liu . "Sentiment Analysis and Subjectivity". Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010

[2] B. Pang, L.Lee, and S. Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques". In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002

[3] B.Pang and L. Lee,"Opinion mining and sentiment analysis". Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008

[4] Bo Pang and Lillian Lee." Seeing stars: Exploiting class relationships for sentiment categorizationwith respect to rating scales". In Proceedings of the Association for Computational Linguistics (ACL),pages 115–124, 2005**.**

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

[6] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the Association for Computational Linguistics (ACL), pages 271–278, 2004.

[7] J. Wiebe, "Learning subjective adjectives from corpora," Proceedings of AAAI, 2000

[8] Peter Turney. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews". In Proceedings of the Association for Computational Linguistics (ACL), pages 417–424, 2002.

[9] Soo-Min Kim and Eduard Hovy."Determining the sentiment of opinions". In Proceedings of the International Conference on Computational Linguistics (COLING), 2004.

[10] Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. "Using WordNet to measure semantic orientation of adjectives". In LREC, 2004

[11] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. "Pulse: Mining customer opinions from free text". In Proceedings of the 6th International Symposium on Intelligent Data Analysis (IDA), 2005

[12] E. Riloff, J. Wiebe, and T. Wilson, "Learning subjective nouns using extraction pattern bootstrapping," Proceedings of the Conference on Natural Language Learning (CoNLL), pp. 25– 32, 2003

[13] Wei Jin, Hung Hay Ho, and Rohini K.Srihari." OpinionMiner: A novel machine learning system for web opinion mining". In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France,2009

[14] Andrew B. Goldberg and Jerry Zhu." Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization". In TextGraphs: HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing, 2006

[15] Bhatia, MPS. & Kumar, A., "Information Retrieval & Machine Learning: Supporting Technologies for Web Mining Research & Practice", Webology, Vol. 5, No.2 ,2008.

[16] Simon, H. A.. "Why Should Machine Learn?" In R. S. Michalski, J. Carbonell, & T. M. Mitchell (Eds.), Machine learning: An artificial intelligence approach (pp. 25-38). Palo Alto, CA Tioga Press.,1983

[17] Alpaydin E "Introduction to machine learning", vol 452. MIT Press, Cambridge,2004

[18] Vucetic,Slobodan,

http: //www.ist.temple.edu/~vucetic/cis526fall2003/lecture1.pdf

[19] Ana-Maria Popescu and Oren Etzioni."Extracting product features and opinions from reviews". In Proceedingsof the Human Language Technology Conference and the Conference on Empirical Methodsin Natural Language Processing (HLT/EMNLP), 2005

[20] Yejin Choi, Eric Breck, and Claire Cardie. "Joint extraction of entities and relations for opinion recognition".In Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), 2006

[21] J. Wiebe, R. F. Bruce, and T. P. O'Hara. "Development and use of a gold standard data set for subjectivity classifications." Proceedings of the Association for Computational Linguistics (ACL), pp. 246–253

[22] Hong Yu and Vasileios Hatzivassiloglou." Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences". In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003

[23] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis". In Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 347–354, 2005

[24] Schapire, R. E. and Singer, Y. 2000. "BoosTexter: a boosting-based system for text categorization". Machine Learning 39, 2/3, 135–168

[25] X. Wu and R. Srihari." Incorporating prior knowledge with weighted margin support vector machines". In KDD, 2004

[26] Melville, Prem, Wojciech Gryc, and Richard D. Lawrence. 2009. "Sentiment analysis of blogs by combining lexical knowledge with text classification". In KDD. ACM

[27] Turney, P. D., & Littman, M. L." Unsupervised learning of semantic orientation from a hundred-billion-word corpus". Technical Report ERB-1094. National Research Council Canada, Institute for Information Technology,2002.