

Sentiment Analysis: An Overview

Dr. Sachin A. Kadam¹, Mrs. Shweta T. Joglekar²

^{1,2}Bharati Vidyapeeth Deemed University
Institute of Management and Entrepreneurship Development
Pune, India – 411038

Abstract:

With the explosive growth of the social media content on the Internet in the past few years, people now express their views on almost anything in discussion. Finding the opinion sites and monitoring them on the web is difficult task. Thus there is a need for automatic opinion discovery and summarization systems. Sentiment Analysis or Opinion Mining is the computational study of opinions, sentiments and emotions expressed in text. This paper describes the field of Sentiment Analysis and its latest developments.

Keywords: Sentiment Analysis, Opinion Mining, Machine Learning, Natural Language Processing.

1. Introduction:

“What other people think” has always been an important piece of information during the decision making process. In the past, when an individual needed to make a decision he typically asked for opinions from friends and family. When an organization wanted to find opinions of the general public about its products and services, it conducted surveys.

With the explosive growth of the social media content on the internet in the past few years, the world has been transformed. E Commerce sites, online communities, forums, discussion groups, web logs, product rating sites, chat rooms are some of the sources on which people can now express their views on almost anything in discussion. The fast growth of such content has not been fully harnessed yet [1]. Finding the opinion sites and monitoring them on the web is a difficult task. This is because there are large numbers of diverse sites and each of these sites has huge volume of text that expresses opinions. In addition to this information posted by the user is unstructured and disorganized and is hidden in long forum posts and blogs. Hence it becomes difficult for a person to find relevant sites, extract related sentences with opinions, read them, summarize them and organize them into usable forms [2]. Thus there is a need for automated opinion discovery and summarization systems.

In the past few years, it attracted a great deal of attentions from both academia and industry due to many challenging research problems and a wide range of applications. This research paper opts for the same.

2. Overview of Domain:

2.1.Sentiments:

In general sentiments are either emotions, or they are judgments or ideas prompted or colored by emotions [3]. An emotion is usually caused by a person consciously or unconsciously evaluating an event. The study of emotions in text can be conducted from two points of view. Firstly, one can investigate how emotions influence a writer of a text in choosing certain words and/or other linguistic elements. Secondly, one can investigate how a reader interprets the emotion in a text, and what linguistic clues are used to infer the emotion of the writer [4].

2.2.Sentiment Analysis:

“Sentiment Analysis is the task of identifying positive and negative opinions, emotions, and evaluations”. Sentiment Analysis has many names. It’s often referred to as subjectivity analysis, Opinion mining, and appraisal extraction, with some connections to affective computing (computer recognition and expression of emotion) [5]. It is a technology for extracting opinions from unstructured human-authored documents. In simple words it is used to track the mood of the public. It is an evolving field having roots in Natural Language Processing, Computational Linguistics and Text Mining.

2.3. Terminologies:

2.3.1. Object and Features:

Opinions are expressed on target entity. It can be a product, a service, an individual and so on. An object has a set of components and a set of attributes called features. E.g. A phone is an object. Battery and screen are its components and size and voice quality are its features. Opinions are expressed on an object itself or on any of the feature of an object.

The attributes on which opinions are expressed can be explicit attributes or Implicit attributes. Explicit attributes directly appear in the opinion. E.g. the battery life of this phone is too short. On the other hand implicit attributes do not appear in the opinion. E.g. This phone is too large. This opinion is expressed on feature “size”.

2.3.2. Opinion Holder:

Opinion holder is the person or organization that expresses the opinion.

2.3.3. Opinion and Orientation:

An opinion is a positive or negative view on Object/Feature from an opinion holder. Opinion can be either direct opinion or comparative opinion.

2.3.4. Opinion strength:

It is the scale or intensity of opinion indication how strong it is.

2.4. Goals of Sentiment Analysis:[2]

2.4.1. Object Identification:

Object identification refers to identifying an object on which an opinion is expressed. This problem is important because without knowing the object on which an opinion has been expressed, the opinion is of little use.

2.4.2. Feature Extraction and Synonym grouping:

Feature extraction refers to identification of features/attributes on which an opinion is expressed. In addition it is also required to group synonym features as people often use different words or phrases to describe the same feature (e.g., “voice” and “sound” refer to the same feature).

2.4.3. Opinion Orientation Classification:

The task of opinion orientation refers to determine whether a statement is objective or subjective. If subjective it determines whether the opinion expressed is positive, negative or neutral. Existing approaches for opinion orientation classification are based on supervised and unsupervised methods.

2.4.4. Integration:

It refers to the integrating the opinion, opinion holder on feature of an object at a particular time.

2.5. Levels of Sentiment Analysis:

Sentiment analysis can be done on:

2.5.1. Document Level: Identify if the document (e.g. product reviews, blogs, forum posts) expresses opinions and whether the opinions are positive, negative, or neutral. (Technique: Heuristics)

2.5.2. Sentence Level: Identify if a sentence is opinionated and whether the opinion is positive, negative, or neutral. (Technique: Part of Speech Tagging)

2.5.3. Attribute Level: Extract the object attributes (e.g. image quality, zoom size) that are the subject of an opinion and the opinion orientations. (Technique: n-gram classifiers, lexicons).

2.6.Challenges in Sentiment Analysis:

G.Vinodhini and RM.Chandrasekaran, in a paper titled “Sentiment Analysis and Opinion Mining: A Survey” [6], focused on the following challenges that makes the task of sentiment analysis complex.

1. Opinion words that are considered to be positive in one situation may be considered negative in another situation.
2. People don't always express opinions in a same way.
3. People can be contradictory in their statements.
4. Most reviews can have both positive and negative comments.
5. People combine different opinions in the same sentence which is easy for a human to understand, but more difficult for a computer to parse.
6. People have difficulty understanding what someone thought based on a short piece of text because it lacks context.

3. Present state of Sentiment Analysis :

A relatively new field, Sentiment Analysis, in its brief history has used natural language processing, data mining, and text retrieval tools to tackle the problem of extracting opinions from text. The research community has studied almost all main aspects of the problem. The research in the field started with sentiment and subjectivity classification, which treated the problem as a text classification problem. Sentiment classification classifies whether an opinionated document (e.g., product reviews) or sentence expresses a positive or negative opinion [6]. Subjectivity classification determines whether a sentence is subjective or objective [7].

The most well studied subproblem is opinion orientation classification where different supervised techniques as Support Vector Machines (SVM), Naive Bayes Multinomial (NBM), Maximum Entropy (Maxent), and Unsupervised and Weakly-Supervised Methods as using AltaVista, clustering are applied.

According to David Osimo and Francesco Mureddu in a paper titled, “Research challenge on Opinion Mining and Sentiment Analysis”[10], the current research is focusing on improving the accuracy of algorithm for opinion detection, reduction of human effort needed to analyze content ,Semantic analysis through lexicon/corpus of words with known sentiment for sentiment classification, Visual mapping of bipolar opinion.

According to Yelena Mejova [11], a lot of studies have been done on controlled collections of text like movie or product reviews, but algorithms that work for these collections fail miserably in a more complex setting like political commentary or news articles. He also stated that sentiment analysis is topic specific. The meaning of words changes from domain to domain and in some cases it gets reversed. Therefore a research challenge exists to adjust general lexicons and algorithms, and extend it to accommodate each topic and its peculiarities.

4. Existing Techniques for Sentiment Analysis:

There are two approaches to perform the task of sentiment analysis. They are Machine Learning (Supervised) or Natural language Processing (Unsupervised).

4.1 Machine Learning:

Machine learning approach uses training data to build a predictive model. Predictive models such as decision trees, logistic regressions or neural networks are used to make prediction on documents that are outside the training set.

This approach has an advantage as it is based on learning patterns that are useful for making automated and efficient predictions. Also the algorithms are capable of discovering unimagined and complicated patterns that would be beyond what a human could anticipate. However it has drawbacks as huge training data is required to build the model and validating it is time consuming and challenging. A rating is required to be provided for every document, and if there are attributes of documents it is needed to provide a rating for each of these as well. Another complication arises if two different reviewers assign two different sentiment ratings to the same document, then this can introduce unexpected errors in building and measuring the performance of model.

4.2. Natural language Processing Approach:

Natural language processing (NLP) is a field of artificial intelligence that deals with automatically extracting meaning from natural language text. It uses entities and syntactic patterns in the text to understand its meaning. It also uses a combination of language dictionaries, linguistic constructs like parts of speech, and noun phrases along with a range of operators.

The major advantage of rule-based methods is that it provides freedom for the rule developers to use their domain knowledge to devise rules for analysis purpose. Rule-based methods are completely unsupervised and they do not require any training data. This is a big advantage in real-life applications where training data is scarce. Additionally it provides the facility to refine the rules over time based on the feedback from analysts or subject-matter experts to adjust the models. The major drawback of NLP approach is that they require a lot of human involvement in developing the rules and it completely rely on the domain knowledge of rule developers.

5. Tools for Sentiment Analysis:

There is a wide range of tools in market that performs automatic sentiment analysis on a given text. These tools utilize existing online textual content generated from sites such as Epinion, Amazon, Rotten Tomatoes, Twitter, Face book etc. Several sentiment search engines exist where users run typical queries on any topic of interest, and generate text results. Usually the results are coded and categorized into two or three polar categories. Some examples currently available are: Topsey, BackTweets, Twitterfall, TweetBeep, Reachli, Social Mention, Trackur , Twendz , Sentiment.ly, Sentiment140, Opinion Crawl, OpenAmplify , Amplified Analytics, Lithium, SAS

Sentiment Analysis Manager, Twittratr, IBM Social Sentiment Index, SAS Sentiment Analysis Studio, TweetSentiments etc.

6. Applications of Sentiment Analysis:

Sentiment analysis can be useful in Online Advertising, Hotspot detection in forums, Competitive intelligence, Online message sentiment filtering, Mail sentiment classification, Web blog author's attitude analysis, Summarization and analysis of customers reviews, detection of flames (overly heated or antagonistic language) etc.[6]

The capture and analysis of attitudes and opinions in an automated and structured fashion offers a powerful technology to a number of problem domains including Business Intelligence, Marketing, Security, Crime Prevention etc [7].

7. Open Research Directions:

The field of sentiment analysis is in its early stages of development. Even though a considerable amount of work is done in the field there are some issues which require attention from the research community. There is a need to develop new psychological models of sentiment, topic-dependent/independent sentiment identification, mass opinion estimation based on NLP and statistical models, Opinion, Sentiment, Emotion extraction, categorization and aggregation, sentiment corpora and annotation, Sentiment lexicon, Applications of sentiment analysis specially in Social Networking, Multimodal and multilingual Sentiment Analysis.

8. Concluding Remarks:

This paper indicated the usefulness of sentiment analysis, and has given an overview of the various methods used for this task. Sentiment detection has a wide variety of applications in information systems, including classifying reviews, summarizing review and other real time applications. Although the techniques and algorithms used for sentiment analysis are advancing fast, however, a lot of problems in this field of study remain unsolved.

References:

1. Adam Westerski, "Sentiment Analysis: Introduction and the State of the Art overview".
2. Bing Liu, "Sentiment Analysis: A Multi-Faceted Problem", IEEE Intelligent Systems: 2010.
3. Biber, D; Finegan, E. Styles of stance in english: Lexical and grammatical marketing of evidentiality and affect. Text 9, 1989, pp. 93-124.
4. Erik Boiy; Pieter Hens; Koen Deschacht; Marie-Francine Moens" Automatic Sentiment Analysis in On-line Text", Proceedings ELPUB2007 Conference on Electronic Publishing – Vienna, Austria – June 2007.

5. Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 2(1-2):1–135.
6. G.Vinodhini and RM.Chandrasekaran, “Sentiment Analysis and Opinion Mining: A Survey”, *International Journal of Advanced Research in Computer Science and Software Engineering* Volume 2, Issue 6, June 2012 ISSN: 2277 128X.
7. Pang, B. and Lee, L. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 10:79–86.
8. B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis.” *Foundations and Trends in Information Retrieval* 2(1-2), pp. 1–135, 2008.
9. J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, “Learning Subjective Language,” *Computational Linguistics*, vol. 30, pp. 277–308, September 2004.
10. David Osimo and Francesco Mureddu, “Research Challenge on Opinion Mining and Sentiment Analysis”.
11. Yelena Mejova, “Sentiment Analysis: An Overview”, November 16, 2009.