Knowledge-Based Systems 89 (2015) 14–46

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A survey on opinion mining and sentiment analysis: Tasks, approaches and applications

Kumar Ravi [a,b], Vadlamani Ravi [a,*]

[a] Center of Excellence in CRM and Analytics, Institute for Development and Research in Banking Technology, Castle Hills Road No. 1, Masab Tank, Hyderabad 500057, AP, India
[b] School of Computer & Information Sciences, University of Hyderabad, Hyderabad 500046, AP, India

ARTICLE INFO

Article history:
Received 28 January 2015
Received in revised form 13 May 2015
Accepted 23 June 2015
Available online 29 June 2015

Keywords:
Opinion mining
Sentiment analysis
Social media
Micro blog
Lexica creation
Machine learning
Ontology

ABSTRACT

With the advent of Web 2.0, people became more eager to express and share their opinions on web regarding day-to-day activities and global issues as well. Evolution of social media has also contributed immensely to these activities, thereby providing us a transparent platform to share views across the world. These electronic Word of Mouth (eWOM) statements expressed on the web are much prevalent in business and service industry to enable customer to share his/her point of view. In the last one and half decades, research communities, academia, public and service industries are working rigorously on sentiment analysis, also known as, opinion mining, to extract and analyze public mood and views. In this regard, this paper presents a rigorous survey on sentiment analysis, which portrays views presented by over one hundred articles published in the last decade regarding necessary tasks, approaches, and applications of sentiment analysis. Several sub-tasks need to be performed for sentiment analysis which in turn can be accomplished using various approaches and techniques. This survey covering published literature during 2002–2015, is organized on the basis of sub-tasks to be performed, machine learning and natural language processing techniques used and applications of sentiment analysis. The paper also presents open issues and along with a summary table of a hundred and sixty-one articles.

1. Introduction

Ever increasing use of Internet and online activities (like chatting, conferencing, surveillances, ticket booking, online transactions, e-commerce, social media communications, blogging and micro-blogging, clicks streams, etc.) leads us to extract, transform, load, and analyze very huge amount of structured and unstructured data, at a fast pace, referred to as Big Data. Such data can be analyzed using a combination of Data Mining, Web Mining and Text Mining techniques in various real life applications. Huge amount of information related to customer opinions/reviews is quite cumbersome to analyze and needs extant approaches to get a generalized opinion summary. Numerous forums, blogs, social networks, e-commerce web sites, news reports and additional web resources serve as platforms to express opinions, which can be utilized for understanding the opinions of the general public and consumers on social events, political movements, company strategies, marketing campaigns, product preferences, and monitoring reputations [26]. To accomplish these tasks, research communities and academicians are working rigorously on sentiment analysis for last one and half decade. Sentiment analysis (SA) is a computational study of opinions, sentiments, emotions, and attitude expressed in texts towards an entity [138]. Sentiment analysis (also called opinion mining, review mining or appraisal extraction, attitude analysis) is the task of detecting, extracting and classifying opinions, sentiments and attitudes concerning different topics, as expressed in textual input [2,84]. SA helps in achieving various goals like observing public mood regarding political movement, market intelligence [90], the measurement of customer satisfaction [158], movie sales prediction [131] and many more.

Sentiments, evaluations, and reviews are becoming very much evident due to growing interest in e-commerce, which is also a prominent source of expressing and analyzing opinions. Nowadays, customers on e-commerce site mostly rely on reviews posted by existing customers, and producers and service providers, in turn, analyze customers' opinions to improve the quality and standards of their products and services. For example opinions given on e-commerce sites like Amazon, IMDb, epinions.com, etc. can influence the customers' decision in buying products and subscribing services [18].

* Corresponding author. Tel.: +91 40 23294042; fax: +91 40 23535157.
E-mail addresses: ankitaravi.ravi00@gmail.com (K. Ravi), rav_padma@yahoo.com (V. Ravi).

http://dx.doi.org/10.1016/j.knosys.2015.06.015
0950-7051/

In developing countries, online and social media is taking the place of offline media swiftly, which encourages common people to involve in political discussions and enable them to put across unilateral thoughts on Global issues interactively. Online media provides the platform for wide sharing of ideas and encouraging public for group discussions with open views. Online media provides better means to get quick response and feedback on different Global issues and entities in the form of textual posts, news, images, and videos. Thus, it can be utilized to analyze peoples' opinions for learning the behaviors of consumer, patterns market, and trends of society [206]. Twitter has 255 million monthly active users and it oversees 500 million tweets every day.[1] Thus, it serves as a good resource to extract heterogeneous opinions published by people from diverse societies for different purposes like improvement of quality of products and services, prediction of consumers' demand and taste, etc.

Online media and social networking sites (SNS) are used to express and share public experiences in the form of product reviews, blogs, and discussion forums. Collectively, these media contain highly unstructured data combining text, images, animations and videos that are useful in making public aware of various issues.

1.1. Earlier reviews

Pang and Lee [1] performed an extensive survey of more than three hundred papers by covering applications, common challenges for sentiment analysis, major tasks of opinion mining viz., opinion extraction, sentiment classification, polarity determination, and summarization. Then, Tang et al. [16] discussed four problems related to opinion mining, i.e., subjectivity classification, word sentiment classification, document sentiment classification and opinion extraction. For subjectivity classification, they highlighted some approaches like similarity dependent, NB classifier, Multiple NB classifier, and cut-based classifier.

O'Leary [89] presented a survey on blog mining, which includes introduction on blog search and mining, type of blogs to be analyzed, unit and type of opinions to be extracted from blogs, and their applications. Montoyo et al. [84] listed some open issues along with achievements obtained thus far in the area of subjectivity analysis and sentiment analysis. Tsytsarau and Palpanas [137] presented a survey on SA by focusing on opinion mining, opinion aggregation including spam detection and contradiction analysis. They compared opinion mining methods, which were employed on some common dataset.

Liu [181] presented different tasks possible and works published in SA and opinion mining. Major tasks listed are subjectivity and sentiment classification, aspect-based SA, sentiment lexicon generation, opinion summarization, analysis of comparative opinions, opinion search and retrieval, opinion spam detection and quality of reviews. Cambria et al. [15] pointed out complexities involved in SA with respect to current demand along with possible future research directions. Recently, Feldman [14] focused on five specific problems in the field of SA: Document-level SA, sentence-level SA, aspect-based SA, comparative SA and, sentiment lexicon acquisition. They also listed some open issues like SA of composition statement, automatic entity recognition, discussion on multi-entity in same review, sarcasm detection and subjectivity classification at finer level.

Most recently, Medhat et al. [138] presented a survey on feature selection and sentiment classification methods. A very brief description is presented about feature selection methods (mainly pointwise mutual information and Chi-square) and a detailed discussion is presented on sentiment classification methods and related papers. They summarized fifty-four articles listing out task accomplished, domain, algorithm utilized, polarity, data scope, data source, and type of language. The authors' major concern is to discuss the techniques applied in surveyed papers.

Along with these surveyed papers, a considerable amount of work has been reported in this area and a number of lexica have been created by research community to evaluate new devised sentiment analysis algorithm. Especially in the last four years, major concerns of researchers are micro-blogs, which have been successfully applied for market prediction [18], social advertising [43], and box-office prediction [51]. Tsytsarau and Palpanas [137] presented very limited discussion on this major domain and its applications. In addition to that, several other issues are reported in recently published papers. Therefore, there is an urgent need to focus on several other issues raised in currently published papers, which were not the part of the extant surveys.

This survey work differs from existing literature surveys in various ways (i) we classified existing studies on the basis of opinion mining tasks, approaches and applications as presented in Fig. 1, (ii) this paper presents articles related to tasks and major issues pointed out by existing articles like subjectivity classification, sentiment classification from coarse-grained to fine-grained level, review usefulness measurement, opinion spam detection, lexicon creation, and opinion word and product aspect extraction as presented throughout the paper, (iii) we summarized each of surveyed articles in four aspects viz. problem addressed, exploited dataset details, feature representation and selection method (if applied), techniques applied, obtained results, and indicated future directions along with our views, (iv) we included some recently proposed feature selection techniques for SA, (v) we provided a detailed list of online available datasets, (vi) classification of articles on the basis of SA performed at various granular levels as presented in Table 1, (vii) the exploited lexica are listed in Table 10, and (viii) summary of one hundred and sixty-one articles is presented in Table 10 before concluding the paper.

Therefore, works addressing these issues are considered for this survey. This rest of the paper is organized as follows: Section 2 presents background information related to the survey. Section 3 presents state-of-the art discussion on SA covering common issues listed in previous paragraph. Detailed discussion on the existing work, open issues, and possible applications of sentiment analysis is presented in Section 4. The paper is concluded in Section 5.

2. Preliminary steps of sentiment analysis

Sentiment analysis, opinion mining and subjectivity analysis are interrelated areas of research which use various techniques taken from Natural Language Processing (NLP), Information Retrieval (IR), structured and unstructured Data Mining (DM). Major part of data available worldwide, being unstructured (such as text, speech, audio, and video), poses important research challenges. To deal with such unstructured text data, traditional methods of NLP i.e. information retrieval and information extraction came into existence [84]. In order to get a sense of the extracted text, numerous research efforts have been witnessed in recent years leading to automated SA, an extended NLP area of research [23].

Sentiment analysis is not a single problem; instead it is a multi-faceted problem [134]. Various steps are needed to perform opinion mining from given texts, since texts for opinion mining is coming from several resources in diverse format. Data acquisition and data preprocessing are most common sub-tasks required for text mining and SA, which are discussed in this section.

[1] https://about.twitter.com/company.