Computer Science Review 27 (2018) 16–32

Contents lists available at ScienceDirect

Computer Science Review

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

The evolution of sentiment analysis—A review of research topics, venues, and top cited papers

Mika V. Mäntylä [a,*], Daniel Graziotin [b], Miikka Kuutila [a]

[a] M3S, ITEE, University of Oulu, Finland
[b] Institute of Software Technology, University of Stuttgart, Germany

ARTICLE INFO

Article history:
Received 22 March 2017
Received in revised form 25 October 2017
Accepted 29 October 2017

Keywords:
Sentiment analysis
Opinion mining
Bibliometric study
Text mining
Literature review
Topic modeling
Latent Dirichlet Allocation
Qualitative analysis

ABSTRACT

Sentiment analysis is one of the fastest growing research areas in computer science, making it challenging to keep track of all the activities in the area. We present a computer-assisted literature review, where we utilize both text mining and qualitative coding, and analyze 6996 papers from Scopus. We find that the roots of sentiment analysis are in the studies on public opinion analysis at the beginning of 20th century and in the text subjectivity analysis performed by the computational linguistics community in 1990's. However, the outbreak of computer-based sentiment analysis only occurred with the availability of subjective texts on the Web. Consequently, 99% of the papers have been published after 2004. Sentiment analysis papers are scattered to multiple publication venues, and the combined number of papers in the top-15 venues only represent ca. 30% of the papers in total. We present the top-20 cited papers from Google Scholar and Scopus and a taxonomy of research topics. In recent years, sentiment analysis has shifted from analyzing online product reviews to social media texts from Twitter and Facebook. Many topics beyond product reviews like stock markets, elections, disasters, medicine, software engineering and cyberbullying extend the utilization of sentiment analysis.

Contents

* Correspondence to: P.O. Box 4500, FI-90014, Finland.
E-mail address: mika.mantyla@oulu.fi (M.V. Mäntylä).

https://doi.org/10.1016/j.cosrev.2017.10.002
1574-0137/

1. Introduction

"The pen is mightier than the sword" proposes that free communication (particularly written language) is a more effective tool than direct violence [1]. Sentiment analysis is a series of methods, techniques, and tools about detecting and extracting subjective information, such as opinion and attitudes, from language [2]. Traditionally, sentiment analysis has been about opinion polarity, i.e., whether someone has positive, neutral, or negative opinion towards something [3]. The object of sentiment analysis has typically been a product or a service whose review has been made public on the Internet. This might explain why sentiment analysis and opinion mining are often used as synonyms, although, we think it is more accurate to view sentiments as emotionally loaded opinions.

The interest on other's opinion is probably almost as old as verbal communication itself. Historically, leaders have been intrigued with the opinions of their subordinates to either prepare for opposition or to increase their popularity. Examples of trying to detect internal dissent can be found already at Ancient Greece's times [4]. Ancient works in East and West mingle with these subjects. "The Art of War" has a chapter on espionage that deals with spy recruiting and betrayal, while in the beginning of "Iliad" the leader of Greeks Agamemnon tries to gauge the fighting spirit of his men. Voting as a method to measure public opinion on policy has its roots in the city state of Athens in the 5th century BCE [5]. Efforts in capturing public opinion by quantifying and measuring it from questionnaires have appeared in the first decades of twentieth century [6], while a scientific journal on public opinion was established in 1937 [7].

We have seen a massive increase in the number of papers focusing on sentiment analysis and opinion mining during the recent years. According to our data, nearly 7000 papers of this topic have been published and, more interestingly, 99% of the papers have appeared after 2004 making sentiment analysis one of the fastest growing research areas. Although the present paper focuses on the research articles of sentiment analysis, we can see that the topic is getting attention in the general public, as well. Fig. 1 shows the increase in searches made with a search string "sentiment analysis" in Google search engine.

We observed that the first academic studies measuring public opinions are during and after WWII and their motivation is highly political in nature [8,9]. The outbreak of modern sentiment analysis happened only in mid-2000's, and it focused on the product reviews available on the Web, e.g., [3]. Since then, the use of sentiment analysis has reached numerous other areas such as the prediction of financial markets [10] and reactions to terrorist attacks [11]. Additionally, research overlapping sentiment analysis and natural language processing has addressed many problems that contribute to the applicability of sentiment analysis such as irony detection [12] and multi-lingual support [13]. Furthermore, with respect to emotions,[1] efforts are advancing from simple polarity detection to more complex nuances of emotions and differentiating negative emotions such as anger and grief [15].

The area of sentiment analysis has become so large that any individual researcher would face several issues when keeping track of all the activities in the area and the information overload. An academic literature review can only focus on one particular area of sentiment analysis as it typically includes between 10 and 100 studies, e.g., a recent systematic review of the prediction of financial markets with sentiment analysis reviewed 24 papers [10]. To overcome the challenges caused by the increasing number of articles about sentiment analysis, we present a computer-assisted literature review and a bibliometric study of sentiment analysis. Studies like the one we are presenting should be helpful when working in an area with large volumes of literature.

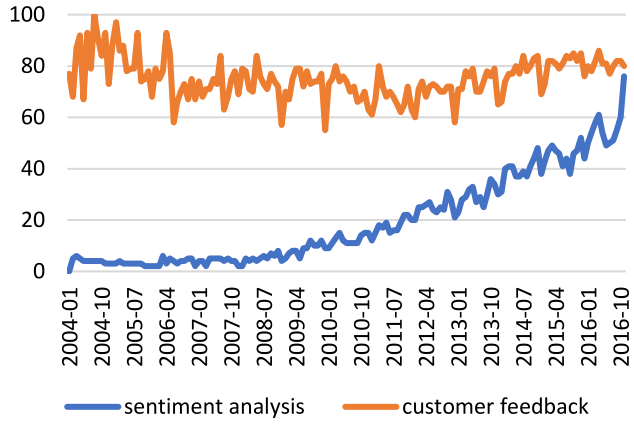

Fig. 1. Google Trends (www.google.com/trends) data showing the relative popularity of search strings "sentiment analysis" and "customer feedback".

We think that the present article can offer an overview of sentiment analysis to newcomers and it may provide valuable to more seasoned scholars for educational purposes. To provide such an overview, we characterize the field of sentiment analysis by answering the following research questions that are typical in bibliometric studies, e.g., [16–19]:

- *RQ1: What is the number of papers in sentiment analysis?* This question helps us in understanding the volume of work in sentiment analysis. We can also observe yearly trends that tell us about the history of the topic and can help in predicting the future.
- *RQ2: What is the number of citations in sentiment analysis?* This question addresses the impact of sentiment analysis. Similarly to RQ1, we can observe yearly trends about the history of the topic and can also help in predicting the future.
- *RQ3: What are the most popular publication venue for sentiment analysis?* This question shows the popular venues for publishing sentiment analysis studies. Understanding the different communities related to sentiment analysis helps understanding the entire field.
- *RQ3: What research topics have been investigated with sentiment analysis?* Given that the topic has rapidly grown very large, we use text clustering to get on overview of the different areas of sentiment analysis. Our text clustering approach originates from influential paper by Griffiths and Steyvers [20] titled "Finding scientific topics". We support the automated text clustering with manual qualitative analysis and they jointly provide a thematic overview of the research topics in this field.

[1] See [14] for the differences between feelings, opinions, sentiment, and emotions in the context of natural language processing.