
LITERATURE SURVEY

3.1 Opinion mining and sentiment analysis

AUTHORS: B. Pang and L. Lee

An important part of our information-gathering behavior has always been to find out what other people think. With the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. The sudden eruption of activity in the area of opinion mining and sentiment analysis, which deals with the computational treatment of opinion, sentiment, and subjectivity in text, has thus occurred at least in part as a direct response to the surge of interest in new systems that deal directly with opinions as a first-class object. This survey covers techniques and approaches that promise to directly enable opinion-oriented information seeking systems. Our focus is on methods that seek to address the new challenges raised by sentiment aware applications, as compared to those that are already present in more traditional fact-based analysis. We include material on summarization of evaluative text and on broader issues regarding privacy, manipulation, and economic impact that the development of opinion-oriented information-access services gives rise to. To facilitate future work, a discussion of available resources, benchmark datasets, and evaluation campaigns is also provided.

3.2 Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena

AUTHORS: J. Bollen, H. Mao, and A. Pepe

Microblogging is a form of online communication by which users broadcast brief text updates, also known as tweets, to the public or a selected circle of contacts. A variegated mosaic of microblogging uses has emerged since the launch of Twitter in 2006: daily chatter, conversation, information sharing, and news commentary, among others. Regardless of their content and intended use, tweets often convey pertinent information about their author's mood status. As such, tweets can be regarded as temporally-authentic

microscopic instantiations of public mood state. In this article, we perform a sentiment analysis of all public tweets broadcasted by Twitter users between August 1 and December 20, 2008. For every day in the timeline, we extract six dimensions of mood (tension, depression, anger, vigor, fatigue, confusion) using an extended version of the Profile of Mood States (POMS), a well-established psychometric instrument. We compare our results to fluctuations recorded by stock market and crude oil price indices and major events in media and popular culture, such as the U.S. Presidential Election of November 4, 2008 and Thanksgiving Day. We find that events in the social, political, cultural and economic sphere do have a significant, immediate and highly specific effect on the various dimensions of public mood. We speculate that large scale analyses of mood can provide a solid platform to model collective emotive trends in terms of their predictive value with regards to existing social as well as economic indicators

3.3 From tweets to polls: Linking text sentiment to public opinion time series

AUTHORS: B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith

If we want to know, say, the extent to which the U.S. population likes or dislikes Barack Obama, an obvious thing to do is to ask a random sample of people (i.e., poll). Survey and polling methodology, extensively developed through the 20th century (Krosnick, Judd, and Wittenbrink 2005), gives numerous tools and techniques to accomplish representative public opinion measurement. With the dramatic rise of text-based social media, millions of people broadcast their thoughts and opinions on a great variety of topics. Can we analyze publicly available data to infer population attitudes in the same manner that public opinion pollsters query a population? If so, then mining public opinion from freely available text content could be a faster and less expensive alternative to traditional polls. (A standard telephone poll of one thousand respondents easily costs tens of thousands of dollars to run.) Such analysis would also permit us to consider a greater variety of polling questions, limited only by the scope of topics and opinions people broadcast. Extracting the public opinion from social media text provides a challenging and rich context to explore computational models of natural language, motivating new research in computational linguistics. In this paper, we connect measures of public opinion derived from polls with sentiment measured from analysis of text from

the popular microblogging site Twitter. We explicitly link measurement of textual sentiment in microblog messages through time, comparing to contemporaneous polling data.

3.4 Mining and summarizing customer reviews

AUTHORS: M. Hu and B. Liu

Merchants selling products on the Web often ask their customers to review the products that they have purchased and the associated services. As e-commerce is becoming more and more popular, the number of customer reviews that a product receives grows rapidly. For a popular product, the number of reviews can be in hundreds or even thousands. This makes it difficult for a potential customer to read them to make an informed decision on whether to purchase the product. It also makes it difficult for the manufacturer of the product to keep track and to manage customer opinions. For the manufacturer, there are additional difficulties because many merchant sites may sell the same product and the manufacturer normally produces many kinds of products. In this research, we aim to mine and to summarize all the customer reviews of a product. This summarization task is different from traditional text summarization because we only mine the features of the product on which the customers have expressed their opinions and whether the opinions are positive or negative. We do not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as in the classic text summarization. Our task is performed in three steps: (1) mining product features that have been commented on by customers; (2) identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative; (3) summarizing the results. This paper proposes several novel techniques to perform these tasks. Our experimental results using reviews of a number of products sold online demonstrate the effectiveness of the techniques.

3.5 Learning user and product distributed representations using a sequence model for sentiment analysis

AUTHORS: T. Chen, R. Xu, Y. He, Y. Xia, and X. Wang

In product reviews, it is observed that the distribution of polarity ratings over reviews written by different users or evaluated based on different products are often skewed in the real world. As such, incorporating user and product information would be helpful for the task of sentiment classification of reviews. However, existing approaches ignored the temporal nature of reviews posted by the same user or evaluated on the same product. We argue that the temporal relations of reviews might be potentially useful for learning user and product embedding and thus propose employing a sequence model to embed these temporal relations into user and product representations so as to improve the performance of document-level sentiment analysis. Specifically, we first learn a distributed representation of each review by a one-dimensional convolutional neural network. Then, taking these representations as pretrained vectors, we use a recurrent neural network with gated recurrent units to learn distributed representations of users and products. Finally, we feed the user, product and review representations into a machine learning classifier for sentiment classification. Our approach has been evaluated on three large-scale review datasets from the IMDB and Yelp. Experimental results show that: (1) sequence modeling for the purposes of distributed user and product representation learning can improve the performance of document-level sentiment classification; (2) the proposed approach achieves state-of-the-art results on these benchmark datasets.