

Chapter 44

A Machine Learning– Based Lexicon Approach for Sentiment Analysis

Tirath Prasad Sahu

NIT Raipur, Raipur, India

Sarang Khandekar

NIT Raipur, Raipur, India

ABSTRACT

Sentiment analysis can be a very useful aspect for the extraction of useful information from text documents. The main idea for sentiment analysis is how people think for a particular online review, i.e. product reviews, movie reviews, etc. Sentiment analysis is the process where these reviews are classified as positive or negative. The web is enriched with huge amount of reviews which can be analyzed to make it meaningful. This article presents the use of lexicon resources for sentiment analysis of different publicly available reviews. First, the polarity shift of reviews is handled by negations. Intensifiers, punctuation and acronyms are also taken into consideration during the processing phase. Second, words are extracted which have some opinion; these words are then used for computing score. Third, machine learning algorithms are applied and the experimental results show that the proposed model is effective in identifying the sentiments of reviews and opinions.

1. INTRODUCTION

Sentiment analysis is becoming a trending and popular amongst researchers. Sentiment analysis is a special case of text classification which aims to categorize opinions based on polarities for e.g. positive or negative. People can express their opinions on the web in the form of reviews. The reviews can be extracted from different fields such as product review, movie review, tweets from Twitter, etc. Basically, sentiment analysis is done to express what a person thinks for a particular product, movie, twitter etc.

DOI: 10.4018/978-1-6684-6303-1.ch044

A Machine Learning-Based Lexicon Approach for Sentiment Analysis

In recent years lots of people are expressing their opinions on the web. Everyday millions of people express what they think in the form of reviews on platform such as blogs, forums, social networking sites like twitter etc. These platforms help people to connect to other people whom they don't even know and get the opinion from them. Basically, these platforms serve as intermediate between the end user and the service provider. From the end users perspective, these platforms are useful to get an idea about the product. From the service provider point of view, it helps to improve the standard of their product and services.

The data on the web is increasing everyday as many people are using it as a platform to express their reviews and opinions. Therefore, it becomes a very hectic task for people to take instant decision about the reviews and opinions. People are confused that on which review they can trust and on which review they cannot. Due to this problem, an automatic system must be designed to make the process of analysis, summarization and classification easy. The general approach for sentiment analysis is the Bag of Word (BOW) Approach (Dave, Lawrence, & Pennock, 2003). In this approach, a document is divided into bag of words to make feature vector which will be used in the classification process. However, BOW approach is failed to generate the desired results as it doesn't capture word sequence and semantic relation. The works have been carried out in the field of sentiment analysis to improve BOW in combination with linguistic knowledge (Dave, et al., 2003; Gamon, 2004; Kennedy & Inkpen, 2006; Na, Sui, Khoo, Chan, & Zhou, 2004; Ng, Dasgupta, & Arifin, 2006; Pang, Lee, & Vaithyanathan, 2002; Whitelaw, Garg, & Argamon, 2005; Xia, Zong, & Li, 2011). However, they failed to improve the classification accuracy.

The polarity shift is an important issue in sentiment analysis. Many approaches have been suggested in the literature to overcome the polarity shift problem (Council, McDonald, & Velikovich, 2010; Das & Chen, 2001; Ikeda, Takamura, Ratinov, & Okumura, 2008; Li & Huang, 2009; Li, Lee, Chen, Huang, & Zhou, 2010; Wilson, Wiebe, & Hoffmann, 2009). However, most of them required either complex linguistic knowledge or extra human annotations. Such high-level dependency on external resources makes the system difficult to be widely used in practice.

Sentiment Analysis can be carried out at different level such as aspect level, sentence level and document level. The proposed method uses sentiment analysis at sentence level using the lexical resources SentiWordNet3.0 (Baccianella, Esuli, & Sebastiani, 2010) and Affin111 (Zol & Mulay, 2015) to decide the overall polarity of the reviews based on feature vector and machine learning algorithms.

2. RELATED WORK

A lot of research has been carried out on sentiment analysis in recent years. The center research areas are sentiment classification and feature-based sentiment analysis. Sentiment classification explores ways to characterize reviews as positive or negative. The efforts on classification at the document level sentiment analysis have been presented in (Das & Chen, 2001; Dave, et al., 2003; Gamon, Aue, Corston-Oliver, & Ringger, 2005; Hearst, 1992; Pang & Lee, 2005; Pang, et al., 2002; Riloff & Wiebe, 2003; Turney, 2002). Most of the methods are based on sentence and document level classification which involves reorganization of opinion words. Corpus based method and dictionary-based method is the two types of method that are basically used. In corpus based, method co-occurrence patterns of words are found out to determine the sentiments of words. The research efforts (Hatzivassiloglou & Wiebe, 2000; Turney,

14 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/a-machine-learning-based-lexicon-approach-for-sentiment-analysis/308522

Related Content

A Hybrid Method for High-Utility Itemsets Mining in Large High-Dimensional Data

Guangzhu Yu, Shihuang Shao, Bin Luo and Xianhui Zeng (2009). *International Journal of Data Warehousing and Mining* (pp. 57-73).

www.irma-international.org/article/hybrid-method-high-utility-itemsets/1823

Distributed Privacy Preserving Clustering via Homomorphic Secret Sharing and Its Application to (Vertically) Partitioned Spatio-Temporal Data

Can Brochmann Yildizli, Thomas Pedersen, Yucel Saygin, Erkey Savas and Albert Levi (2011). *International Journal of Data Warehousing and Mining* (pp. 46-66).

www.irma-international.org/article/distributed-privacy-preserving-clustering-via/49640

On-Demand ELT Architecture for Right-Time BI: Extending the Vision

Florian Waas, Robert Wrembel, Tobias Freudenreich, Maik Thiele, Christian Koncilia and Pedro Furtado (2013). *International Journal of Data Warehousing and Mining* (pp. 21-38).

www.irma-international.org/article/demand-elt-architecture-right-time/78285

A Decision Support System for Privacy Compliance

Siani Pearson and Tomas Sander (2013). *Data Mining: Concepts, Methodologies, Tools, and Applications* (pp. 1496-1518).

www.irma-international.org/chapter/decision-support-system-privacy-compliance/73508

Decision Rule Extraction for Regularized Multiple Criteria Linear Programming Model

DongHong Sun, Li Liu, Peng Zhang, Xingquan Zhu and Yong Shi (2011). *International Journal of Data Warehousing and Mining* (pp. 88-101).

www.irma-international.org/article/decision-rule-extraction-regularized-multiple/55080