SENTIMENT ANALYSIS ON ONLINE PRODUCT REVIEWS USING MACHINE LEARNING TECHNIQUES

सिद्धिमूलं प्रबन्धनम्
भा. प्र. सं. इन्दौर
IIM INDORE

A Doctoral Dissertation Submitted in Partial Fulfillment of the Requirements for the Fellow Programme in Management

Indian Institute of Management Indore

By
Anuj Sharma

March 2013

Thesis Advisory Committee:

S. Dey
Prof. Shubhamoy Dey
(Chair)

Prof. Prabin K. Panigrahi
(Member)

Prof. Ashish Sadh
(Member)

Abstract

The growing popularity of Internet and Web 2.0 social media have led to the advent of many e-commerce websites, discussion forums and Weblogs. These in turn have facilitated the sharing of user provided feedback about post-purchase experience; specifically related to various products and services. This user-generated content comprises of opinions, appraisals, recommendations and evaluations associated with virtually anything people care about in products or services. The opinions and post purchase experiences shared by users through online product reviews constitute a major part of online word-of-mouth (WOM) communication. Online WOM is valuable to potential consumers for making product choices and purchase decisions. At the same time, by analyzing these reviews, business organizations can gain insights into what people are discussing about their products and services. Thus, the analysis of customer sentiments from 'freely available' online reviews can be a potentially cost effective and time efficient solution for eliciting consumer preferences. Also, deeper exploration of user opinions and feedback may lead to the discovery of interesting pattern of product usage (e.g., brand experience), product weaknesses and product-feature related opinion.

As more and more user-generated reviews are created and aggregated, a strong demand for automatic approaches capable of extracting overall as well as specific opinion from these unstructured texts has emerged. Sentiment analysis, often referred to as opinion mining, is a recent area of active research. It deals with the computational treatment of opinion and extraction of subjectivity knowledge from online user-generated content. Thus, sentiment analysis is the task of retrieving aggregated and fine-grained opinions related to an object or its attributes as expressed by users in the form of free text. However, there are many problems and challenges associated with extracting meaningful opinions articulated in unstructured user-generated texts, like product reviews.

Based on the broad objective of mining sentiments and opinions from online reviews, four conjoint studies were conducted. As the main step, a comprehensive study of automatic extraction of overall and fine-grained opinions from online reviews is presented. This study focuses on sentiment based product review classification to

discover product-sentiment. As an output, it identifies reviewed product(s) as recommended/not-recommended along with extraction of discussed features, feature-level opinion mining, and opinion summarization and visualization.

Sentiment based classification of text documents is a more challenging task compared to traditional topic based classification. Discrimination based on opinions, feelings, and attitudes is inherently more complex than classification based on topics due to the semantic relationships of the natural language involved. Further, extraction of the features or attributes about which opinion has been expressed is one of the major challenges of opinion mining. Feature-level opinion mining aims at identifying the relevant opinions associated with specific features or attributes of a product or service from a set of reviews. However, identifying and determining the relevance of features and the accuracy of the expressed opinion continue to pose challenges for this task.

This research work addresses some of the critical issues related to sentiment based classification of online reviews. Document-level sentiment analysis using supervised machine learning techniques faces many challenges like feature selection, dimensionality reduction, sentiment based visualization and domain dependency of sentiments. Text sentiment classification requires deep analysis and understanding of textual features and natural language semantics. Therefore, a part of this work has been devoted to the empirical comparative study of the applicability of feature selection techniques to sentiment analysis of text documents. This study also compares the performance of different machine learning classifiers on a benchmark dataset for document-level sentiment analysis and explores the synergy between feature selection techniques and various machine learning based classifiers.

Another contribution of this research to existing literature is the formulation of novel sentiment classification models using back-propagation artificial neural network (BPANN) and self-organizing maps (SOM). Domain independent sentiment classification models exploit sentiment lexicons in an attempt to classify online reviews from diverse domains. We have investigated the problems associated with domain dependency through sentiment analysis on documents from two different domains. By using large sentiment lexicons along with appropriate handling of negation, this study has shown that encouraging results are obtainable for domain independent sentiment analysis. Further, we have also demonstrated the efficacy of

supervised and unsupervised self-organizing map based approaches for sentiment based classification and visualization of opinion in text documents. Finally, the study establishes how the proposed sentiment analysis framework can be successfully employed for deriving marketing intelligence from online product reviews.

Sentiment analysis may be as simple as overall sentiment based categorization of text documents; but could as well be more complex and advanced procedures to extract opinion at different granularity levels. All the document-level and feature-level sentiment analysis approaches described in this study have been tested on a publicly available benchmark dataset and a real-life dataset created by us. The proposed methods have been found to yield significantly better accuracy in dealing with online subjective text compared to those previously reported. Thus, we have devised an effective way of domain-independent opinion summarization from online customer reviews using our unified framework for opinion retrieval, classification and summarization at various granularity levels.

Keywords: Machine Learning, Feature Selection, Sentiment Analysis, Classification, Opinion Mining, Performance, Experimentation.

Table of Contents

List of Tables

List of Figures