Opinion mining from noisy text data

Lipika Dey · Sk. Mirajul Haque

Received: 7 January 2009 / Revised: 2 July 2009 / Accepted: 9 July 2009 / Published online: 21 August 2009
© Springer-Verlag 2009

Abstract The proliferation of Internet has not only led to the generation of huge volumes of unstructured information in the form of web documents, but a large amount of text is also generated in the form of emails, blogs, and feedbacks, etc. The data generated from online communication acts as potential gold mines for discovering knowledge, particularly for market researchers. Text analytics has matured and is being successfully employed to mine important information from unstructured text documents. The chief bottleneck for designing text mining systems for handling blogs arise from the fact that online communication text data are often noisy. These texts are informally written. They suffer from spelling mistakes, grammatical errors, improper punctuation and irrational capitalization. This paper focuses on opinion extraction from noisy text data. It is aimed at extracting and consolidating opinions of customers from blogs and feedbacks, at multiple levels of granularity. We have proposed a framework in which these texts are first cleaned using domain knowledge and then subjected to mining. Ours is a semi-automated approach, in which the system aids in the process of knowledge assimilation for knowledge-base building and also performs the analytics. Domain experts ratify the knowledge base and also provide training samples for the system to automatically gather more instances for ratification. The system identifies opinion expressions as phrases containing opinion words, opinionated features and also opinion modifiers. These expressions are categorized as positive or negative with membership values varying from zero to one. Opinion expressions are identified and categorized using localized linguistic techniques. Opinions can be aggregated at any desired level of specificity i.e. feature level or product level, user level or site level, etc. We have developed a system based on this approach, which provides the user with a platform to analyze opinion expressions crawled from a set of pre-defined blogs.

Keywords Noisy text · Context-dependent cleaning · Opinion mining · WordNet · Text analytics for market knowledge discovery

1 Introduction

The increased use of Web as a medium of communication has led to the generation of a huge quantity of unstructured data in the form of blogs, chats, emails, reviews, feedbacks, news, etc. Since Internet is a crucial driving force in today’s world, these texts are rich pointers to the collective opinion of the global population on almost every topic. Thus these texts are considered as rich sources for raw inputs to market research and knowledge discovery. Given the volume and growth rate of these sources, efficient mechanisms are required to aggregate, assimilate and interpret all the information with minimal human intervention. Opinion mining refers to the task of opinion extraction and sentiment analysis from unstructured text documents. The key components of opinion mining are opinion extraction and structurization that can help in aggregation and analysis of opinions about pre-defined subjects. Opinion extraction involves identification of opinion holder, the subject being reviewed, the part or the feature of the subject that is being evaluated, and finally classifying the opinion as either positive or negative or some other pre-defined opinion category. Structurization involves transformation of
the extracted opinion expressions into structures suitable for assimilation and analysis.

In this paper, we present the design and analysis of a system that has been designed to mine opinions mainly from blogs, though the principles are generic and work for other documents also. The quality of texts encountered in blogs and chats is extremely noisy. Noisy text data typically comprises spelling errors, ad-hoc abbreviations and improper casing, incorrect punctuation and malformed sentences. Though an array of systems has been developed in the recent past to analyze sentiments and mine opinions from text, most of them assume that the underlying texts are linguistically correct. Compared to the volume of work dedicated to sentiment analysis which identifies whether a sentence expresses positive or negative emotion, fewer attempts have been made toward structuring and mining the content extracted from noisy data sources. In Sect. 2 we provide a detailed review of the related work in this field.

This paper presents the framework for a generic web-based opinion finder system. It can be customized to get a summarized view of all relevant opinions expressed about chosen products. It is a semi-automated system which extracts and organizes information according to analysts’ needs, and also acts as an aid to analysts to validate new knowledge to update the knowledge base. Of particular interest to market analysts are dedicated blogs where users post their opinions and exchange views on product-related issues. This system is therefore designed to learn and extract all product-related features and concepts that are being discussed and opinionated about. The system provides facilities for mining the data and viewing the results at desired levels of specificity.

The opinion mining framework comprises two core modules:

Data acquisition and pre-processing to reduce noise in the text—Data acquisition is done using site-specific crawlers. The acquired data are stored along with as much associated information as is available. Associated data often contains useful information about blogger, time, product name, etc. Since the system is primarily designed to analyze opinions from blog data, one of the core tasks is to reduce as much noise as possible. Pre-processing itself comprises various sub-tasks like identifying meaningful segments of text, spelling correction, etc. It may be noted that the focus of the system is not to correct all the errors in text. Rather these methods have been designed to focus on minimizing error in opinion mining and are customizable to adapt to different domains. Though a number of systems have been proposed earlier for extracting opinions from product reviews, most of these have assumed linguistically correct sentences as inputs.

Opinion extraction and mining—Opinion extraction identifies fragments within a sentence that express opinions about a relevant subject and stores them in pre-defined templates. These templates are subjected to different analytics for generating collective opinions.

The distinguishing features of the proposed approach are:

(i) The system employs a linguistic approach which exploits surface dependency rules to determine opinion expressions within noisy text data. It can extract and summarize opinions about features or parts of a product or product as a whole from user generated data. It can extract multiple opinion expressions from a single sentence and assign orientation to each expression individually, as also to the whole sentence. Rather than working with a fixed set of opinion words, the system is initialized with a seed set, and then expanded to learn domain-specific opinion words, exploiting WordNet. The process of knowledge base creation is semi-automated.

(ii) A new approach is proposed to deal with adverbial modifiers like very or mildly, etc. which can modify the semantic orientation of the descriptive words.

(iii) The extracted opinions are stored as structured templates and can be aggregated using domain ontology to generate statistics at multiple levels of resolution.

(iv) Based on the proposed framework, opinion mining systems have been built for diverse domains like vehicles, movies, etc. These systems are equipped with Graphical User Interface to allow users to interact with them. Opinion orientations can be viewed at individual feature or product levels, which can also be used to provide comparisons among competing products. On the other hand, the system also provides an insight into viewing sentence-level, document-level or site-level statistics.

2 Related research

As the Web gains power as a social media, the task of extracting and analyzing meaningful content from these noisy sources is attracting a fair deal of attention from the researchers. Initiated by Hatzivassiloglou and McKeown [11], opinion mining gained momentum only recently. While a section of research is dedicated toward identifying opinionated expressions and their orientation, there has been a lot of interest in mining the web for extraction of user generated product reviews and their classification. Another class of systems is dedicated to track opinions about people or events from news sources. We have provided an in-depth review of the first class of systems since they are functionally more similar to ours than the others.

Traditional sources of noisy text included (a) text obtained from automatic transcription of speech data and (b) text obtained by using OCRs on text images. More recent sources