Decision Support Systems 55 (2013) 871–882

Contents lists available at ScienceDirect

Decision Support Systems

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

What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings

Alan S. Abrahams [a,*], Jian Jiao [b], Weiguo Fan [c,e], G. Alan Wang [a], Zhongju Zhang [d]

[a] Department of Business Information Technology, Pamplin College of Business, Virginia Tech, 1007 Pamplin Hall, Blacksburg, VA 24061, United States
[b] Department of Computer Science, Virginia Tech, 114 McBryde Hall, Blacksburg, VA 24061, United States
[c] Department of Accounting and Information Systems, Pamplin College of Business, Virginia Tech, 3007 Pamplin Hall, Blacksburg, VA 24061, United States
[d] Operations and Information Management Department, School of Business, University of Connecticut, 2100 Hillside Road, Unit 1041, Storrs, CT 06269, United States
[e] School of Information, Zhejiang University of Finance and Economics, Hang Zhou, 310018, P.R. China

ARTICLE INFO

Available online 30 December 2012

Keywords:
Social media analytics
Diagnostics
Text mining
User-generated content (UGC)

ABSTRACT

In the blizzard of social media postings, isolating what is important to a corporation is a huge challenge. In the consumer-related manufacturing industry, for instance, manufacturers and distributors are faced with an unrelenting, accumulating snow of millions of discussion forum postings. In this paper, we describe and evaluate text mining tools for categorizing this user-generated content and distilling valuable intelligence frozen in the mound of postings. Using the automotive industry as an example, we implement and tune the parameters of a text-mining model for component diagnostics from social media. Our model can automatically and accurately isolate the vehicle component that is the subject of a user discussion. The procedure described also rapidly identifies the most distinctive terms for each component category, which provides further marketing and competitive intelligence to manufacturers, distributors, service centers, and suppliers.

1. Introduction

Hundreds of millions of users are contributing social media content via blogs, product reviews, social networking sites, and content communities on the Internet [45]. Detecting "whispers of useful information in a howling hurricane of noise" is a huge challenge and filters are needed to extract meaning from the "blizzard of buzz", prompting automotive companies like Chrysler to employ "Twitter teams" to find and reply to complaint-laden tweets [94].

To gain a better understanding of the issues affecting their products, effective firms must gather product-relevant information both internally and externally [28,32,35]. It is widely accepted that consumer complaints are a valuable source of product intelligence [34,58,75,76]. The knowledge of outsiders or user communities is an important source for product-related business intelligence [5,23,25]. Historically, many companies have invested substantial effort in soliciting product usage stories from practitioners, for the purposes of diagnosing or understanding problems, or allocating issues to technicians that are able to solve them. Especially for firms selling a mechanical consumer product, these so-called 'communities of practice' are an important aspect of a firm's business intelligence repertoire, as they provide a repository of past usage experiences which can be drawn upon for operational issue resolution, product development, or other purposes [12,20,49,79,90,91].

Monitoring trending topics [11,52,99] and finding opinionated postings by outspoken customers [2,21,64,86,87] have been attempted in a variety of industries. However, it has recently been shown that, in the motor vehicle industry, customer anger (negative opinion) does not necessarily correlate with defect existence or severity, and special-purpose tools are required for diagnosing and prioritizing vehicle defects [4]. Pin-pointing the specific components referred to in these discussions, is a further issue of concern to automotive corporations, since diagnostic information must map to a sensible, industry-specific ontology (or 'decomposition') for describing vehicle components, in order for a proper hazard analysis (HA) or failure modes and effects analysis (FMEA) to be performed [43,89].

In this paper, we employ text mining to map automotive enthusiast discussion forum postings to such an ontology: effectively isolating the component under discussion. In the process, we uncover distinctive terminology that relates to each component category, which provides additional marketing and competitive insight to both vehicle and parts manufacturers and distributors.

This paper tackles three major research questions. *Firstly*, can text mining be employed to automatically isolate the vehicle component category under discussion in an online social media posting? *Secondly*, if component isolation is feasible, what text mining parameters (algorithm, feature selection method, and number of features/terms) produce optimal classification performance? *Finally*, what are the

[*] Corresponding author. Tel.: +1 540 231 5887; fax: +1 540 231 3752.
E-mail addresses: abra@vt.edu (A.S. Abrahams), jjiao@vt.edu (J. Jiao), wfan@vt.edu (W. Fan), alanwang@vt.edu (G.A. Wang), john.zhang@business.uconn.edu (Z. Zhang).

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2012.12.023

distinctive features that discriminate discussion threads that come from different component categories?

The primary contribution of this paper is the development and assessment of a text mining model for vehicle component isolation from discussion forum postings. For nine major component categories, our model is able to accurately pinpoint the component category that is the subject of the postings, with greater than 95% accuracy. In comprehensive evaluations, we determine the parameters (algorithm, feature selection, and number of terms) that appear to provide optimal classification performance. A further contribution of this paper is a consequence of the pursuit of the primary goal: a significant number of the top terms discovered by the feature selection methods are component-specific industry jargon terms, or brand names that are active in manufacturing or distributing replacement parts for the component. As we shall illustrate later, our approach can therefore be a helpful source of real-time marketing and competitive intelligence.

While we focus on a specific example from the vehicle industry, the method we propose can be generalized to various other industries and problem contexts. For instance, it could be well-suited to component isolation for a multitude of complex manufactured products, such as power tools [32], computers and electronics, fishing gear [33], and other items. Further, while our example specifically analyzes discussion threads, these are just one form of social media posting [45] and the techniques are applicable to other forms of social media posting such as user reviews, blog posts, micro-blogs, social media status updates, and feeds (such as RSS feeds [72], Atom feeds [41,42], or Twitter feeds).

The rest of this paper is structured as follows. First, we discuss and contrast related work. We describe our contributions and the research questions we aim to address. We lay out a workflow for automotive component diagnostics from online discussions. We test our text mining approach in a pilot study and then on a large sample data set. Finally, we discuss limitations, implications, and conclusions.

2. Background and related work

In this section, we set out our research motivation. We explore related work on text categorization and on social media analytics. We review the coverage and limitations of prior work, and the research questions raised.

2.1. Text categorization

Automated document categorization involves the automated assignment of documents to pre-defined categories, and has been well-studied over the past half-century [7,9,13,44,51,95]. Categorization may be guided by a training set of manually tagged documents [50,68,73,82], or may be entirely machine-directed [17,61] using classification-based techniques. In the information retrieval literature, search results (documents) can be partitioned by topic (class) to allow the user of the web search engine to rapidly locate a pertinent document in the user's intended subject area, thereby reducing information overload [18,38,96]. Text categorization problems characteristically have high dimensionality: hundreds of available input attributes, some of which may be highly correlated or violate the assumption of normality. Popular algorithmic approaches to text categorization include Bayesian approaches, decision trees, example-based classifiers, neural nets, and Support Vector Machines, with the latter (SVMs) showing particularly high performance for text characterization and discrimination tasks when used with a suitable selection of text features [1,82].

2.2. Social media analytics

The "social web" has recently received substantial attention [65]. 75% of Internet users have created or read some form of social media content: this user-generated content (UGC) includes blog postings, product reviews, social networking sites, and collaborative content [45]. We define social media as online services that provide for decentralized, user level content creation (including editing or tagging), social interaction, and open (public) membership. In our definition, public discussion forums, public listservs, public wikis, open online communities (social networks), public usenet groups, customer product reviews, public visitor comments, user-contributed news articles, micro-blogs, and folksonomies would fall within the gamut of social media. Online discussions contain a significant amount of information relating to companies and their product categories [32,33]. Navigation of this content is a significant research challenge [31,39,65] that may require filtering, semantic content grouping, tagging, information mining, or other techniques.

Social media analytics involves "developing and evaluating informatics tools and frameworks to collect, monitor, analyze, summarize, and visualize social media data, usually driven by specific requirements from a target application" [98]. Social media analytics revolves around

Table 1
Comparison of text analysis studies using traditional web and social media.

Study	Medium	Domain	Output variable
Coussement and vd. Poel [21]	Email	Customer complaints	Complaint classification
Spangler and Kreulen [83]	Email	Customer complaints	Issue categorization
Romano et al. [70]	Product reviews	Movie box office	Film popularity score
Cao, Duan, and Gan [14]	Product reviews	Software	Helpfulness votes
Wang, Liu, and Fan [88]	Online forums	Consumer electronics	Helpfulness classification
Duan, Gu, and Whinston [26]	Product reviews	Movie box office	Estimated revenues
Abbasi and Chen [1]	Email	Enron scandal communications	Topic, opinion, style, genre, interaction pattern
Schumaker and Chen [80]	News articles	Stock market	Stock price
Antweiler and Frank [6]	Online forums	Stock market	Market volatility and trading volume
Tetlock et al. [85]	News articles	Stock market	Company earnings and stock returns
Ma, Sheng, and Pant [56]	News articles	Stock market	Comparative revenue
Ma, Pant, and Sheng [57]	News articles	Stock market	Competitor relationships
Oh and Sheng [63]	Micro-blogs	Stock market	Directional movement
Loughran and McDonald [55]	10-K filings	Stock market	Tone (negativity); earnings; volatility; company internal control weakness
Vechtomova [87]	Blog postings	General	Topic, opinion
Kolari et al. [47]	Blog postings	General	Spam classification
Brooks and Montanez [11]	Blog postings	General	Topic
Santos et al. [77]	Blog postings	General	Relevance, trends
Li and Wu [52]	Online forums	Sports	Hot (i.e. trending) sports topics
Zhang et al. [99]	News articles	Health epidemics	Disease news classification
Wiebe et al. [69,92,93]	News articles	News/politics/business/travel/English literature	Subjectivity
Finch [32]	Newsgroup	Power tools	Message purpose, tone, product class[a]
Finch and Luebbe [33]	Public listserv	Fly fishing gear	Company name, product type[a]
Abbasi, Chen, and Salem [2]	Online forums	Movie reviews, political talk	Opinion classification
Decker and Trusov [24]	Online product reviews	Mobile phones	Sentiment towards each product attribute of each brand
Abrahams et al. [4]	Online forums	Vehicle defects	Defect existence and criticality
This study	Online forums	Vehicle defects	Component classification

[a] Manual analysis, rather than automated classification, was used in [32,33].