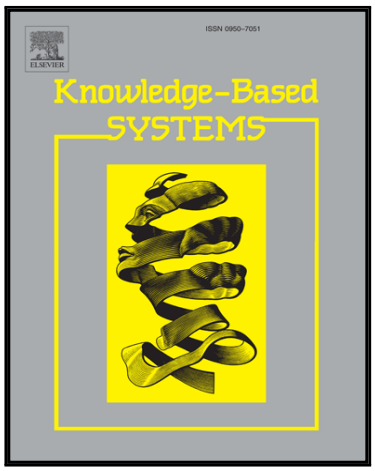

Imbalanced text sentiment classification using universal and domain-specific knowledge

Li Yijing , Guo Haixiang , Zhang Qingpeng , Gu Mingyun , Yang Jianying

PII: S0950-7051(18)30322-8
DOI: 10.1016/j.knosys.2018.06.019
Reference: KNOSYS 4392

To appear in: *Knowledge-Based Systems*

Received date: 7 March 2018
Revised date: 17 June 2018
Accepted date: 19 June 2018

Please cite this article as: Li Yijing , Guo Haixiang , Zhang Qingpeng , Gu Mingyun , Yang Jianying , Imbalanced text sentiment classification using universal and domain-specific knowledge, *Knowledge-Based Systems* (2018), doi: 10.1016/j.knosys.2018.06.019

Imbalanced text sentiment classification using universal and domain-specific knowledge

Yijing Li[1,2], Haixiang Guo[1,2,3,*], Qingpeng Zhang[4], Mingyun Gu[1,2], Jianying Yang[5]

1. *College of Economics and Management, China University of Geosciences, Wuhan 430074, China*
2. *Research Center for Digital Business Management, China University of Geosciences, Wuhan 430074, China*
3. *Mineral Resource Strategy and Policy Research Center of China University of Geosciences, Wuhan 430074, China*
4. *Department of Systems Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong SAR, China*
5. *Wuhan Center of China Geological Survey, Wuhan 430074, China*

Abstract

In this paper, a sentiment classification model is proposed to address two predominant issues in sentiment classification, namely domain-sensitive and data imbalance. Since words may embed distinct sentiment polarities in different contexts, sentiment classification is widely contended as a domain-sensitive task. Accordingly, this paper draws on label propagation to induce universal and domain-specific sentiment lexicons and builds a domain-adaptive sentiment classification model that incorporates universal and domain-specific knowledge into a unified learning framework. On the flip side, sentiment-related corpuses are usually formed with skewed polarity distribution because individuals tend to share similar assessment criteria on a given object and hence their sentiment polarities toward the same object are likely to be similar. We endeavor to address such imbalanced data problem by advancing a novel over-sampling technique. Unlike existing over-sampling approaches that generate minority-class samples from numerical feature space, the proposed sampling method directly creates synthetic texts from word spaces. Several experiments are conducted to verify the effectiveness of the proposed lexicon generation method, learning framework, and over-sampling method. Results show that the induced sentiment lexicons are interpretable and the proposed model is found to be effective for imbalanced and domain-specific text sentiment classification.

Keywords: Sentiment analysis, label propagation, imbalanced data, ensemble learning

1. Introduction

The Internet has given rise to electronic Word-of-mouth (eWOM) communication, particularly in the form of reviews, forums, and microblogs. As these eWOMs are usually embedded with rich opinions/sentiments towards various forms of objects such as products, movies, social issues, and political events, analyzing such opinioned texts could offer useful insights for practitioners and help them make better decisions. Existing research has shown that people tend to seek for others' opinions to reduce their risk of making bad decisions [1]. For example, customers may browse movie reviews

* Corresponding author
E-mail addresses: liyijing024@gmail.com (Y. Li), faterdumk0732@sina.com (H. Guo), qingpeng.zhang@cityu.edu.hk (Q. Zhang), gmingyun@gmail.com (M. Gu), 307198446@qq.com (J. Yang).