International Conference on Sustainable Design, Engineering and Construction

Text analytics for supporting stakeholder opinion mining for large-scale highway projects

Xuan Lv[a], Nora El-Gohary[b]*

[a]*Graduate Student, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Ave., Urbana, IL 61801, United States*
[b]*Assistant Professor, Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 North Mathews Ave., Urbana, IL 61801, United States*

Abstract

For large-scale highway projects, late identification of stakeholder concerns often leads to design changes and duplication of effort, which may cause major project delays. This paper proposes a stakeholder opinion mining approach for helping transportation practitioners better identify the types of concerns in the early project stage. The proposed approach includes two major components: (1) stakeholder concern extraction, and (2) stakeholder concern classification. This paper focuses on presenting the proposed methodology and experimental results for stakeholder concern extraction, which extracts the words and phrases that describe stakeholder concerns from stakeholder comments on large-scale highway projects. In developing the proposed stakeholder concern extraction methodology, several supervised machine learning (ML) algorithms were tested and evaluated, and the effect of using a predefined name list as feature was also investigated. All the algorithms were tested on a testing data set of 200 comment sentences, which were selected from a comment collection including 1,849 stakeholder comments on five large-scale highway projects.

Keywords: Data analytics; Opinion mining; Information extraction; Natural language processing; Machine learning.

---

* Corresponding author. Tel.: +1-217-333-6620; fax: +1-217-265-8039.
  E-mail address: gohary@illinois.edu

1. Introduction

Because large-scale highway projects tend to have significant impact on the surrounding natural environment, the everyday life of the affected public, and the regional economic development, they are required to undergo a comprehensive environmental review, during which the opinions of a wide spectrum of different stakeholders ranging from government agencies to the general public are solicited. Transportation agencies spend a large amount of time and money on the environmental review process. For example, the median time to complete the environmental review process for large-scale highway projects was over 7 years in 2013, and the cost can be several million dollars [1]. Despite that, late identification of stakeholder concerns has been identified as one of the major causes for the lengthy and costly project development process [2]. There is, thus, a need for identifying stakeholder concerns in the early project stage to improve the efficiency of transportation decision making. To address this need, this paper proposes a stakeholder opinion mining approach, which consists of two major components: stakeholder concern extraction and stakeholder concern classification. Stakeholder concern extraction aims at extracting words and phrases that describe stakeholder concerns from stakeholder comments on large-scale highway projects. Stakeholder opinion classification aims at classifying the extracted concern words and phrases into different concern types. This paper focuses on describing the proposed methodology and experimental results for stakeholder concern extraction.

2. Background and knowledge gaps

A stakeholder concern is an issue that is affected, positively or negatively, by the project, such as property value, farmland, fuel tax, population growth, and nearby environmental resources. The concern extraction problem can be defined as an aspect extraction problem. Aspect extraction problems have long been studied in the field of aspect-based opinion mining. There are three main aspect extraction approaches that have been proposed in recent years: (1) language rule-based approach, (2) supervised machine learning (ML)-based approach, and (3) topic model-based approach.

The language rule-based approach extracts aspects using predefined rules, which capture the contextual patterns and/or grammatical relations between the terms in the text [3]. Hu and Liu [4] proposed an extraction method based on association rules, which (1) find frequent aspects through frequent nouns and noun phrases, and (2) identify infrequent aspects using relations between aspects and opinion words. The relationships of opinion words and aspects can be generalized as dependency relations, which are widely used rules for aspect extraction. For example, Qiu et al. [5] developed the double propagation methods to extract aspects and opinions simultaneously based on direct dependency relations. Poria et al. [6] exploited common-sense knowledge and sentence-dependency trees to detect both explicit and implicit aspects from product reviews. One limitation of the rule-based approach is the adaptability of language rules, because the performance of rules depends largely on the type of collection; rules that work well on one collection may not work well on another collection.

The supervised ML-based approach learns to extract aspects from manually labeled data. Some methods utilized sequence models, which treat aspect extraction as a sequence-labeling task. For example, Jin et al. [7] utilized lexicalized hidden Markov model (HMM), which incorporated linguistic features such as part-of-speech and lexical patterns to extract aspects from product reviews. Jakob and Gurevych [8] evaluated the performance of a conditional random field (CRF)-based method for aspect extraction in a single and cross-domain environment. Shariaty and Moghaddam [9] also employed CRF for identifying product aspects and proposed a technique for defining and filtering features to enhance the performance. Other researchers used other supervised learning models that treat aspect extraction as a binary or multi-class classification task. For example, Ghani et al. [10] used both supervised and semi-supervised algorithms to extract attribute and value pairs from product descriptions. Yu et al. [11] trained a one-class support vector machine (SVM) to identify aspects in the candidate noun phrases extracted from Pros and Cons consumer reviews.

The topic model-based approach assumes that the comments are generated though mixtures of topic models, and each topic model is a unigram language model that represents a type of aspect. Mukherjee and Liu [12] developed two joint aspect-opinion models for extracting and categorizing aspects at the same time given user-provided seed words. Chen et al. [13] proposed an aspect extraction framework to extract more coherent aspects by exploiting the