Information Processing and Management 54 (2018) 958–968

Contents lists available at ScienceDirect

Information Processing and Management

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

Using language models to improve opinion detection

Faiza Belbachir*,[a], Mohand Boughanem[b]

[a] Institut Polytechnique des Sciences Avancées (IPSA), France
[b] Institut de Recherche en Informatique de Toulouse (IRIT), France

ARTICLE INFO

Keywords:
Information retrieval
Opinion detection
Blog
Language model

ABSTRACT

Opinion mining is one of the most important research tasks in the information retrieval research community. With the huge volume of opinionated data available on the Web, approaches must be developed to differentiate opinion from fact. In this paper, we present a lexicon-based approach for opinion retrieval. Generally, opinion retrieval consists of two stages: relevance to the query and opinion detection. In our work, we focus on the second state which itself focusses on detecting opinionated documents . We compare the document to be analyzed with opinionated sources that contain subjective information. We hypothesize that a document with a strong similarity to opinionated sources is more likely to be opinionated itself. Typical lexicon-based approaches treat and choose their opinion sources according to their test collection, then calculate the opinion score based on the frequency of subjective terms in the document. In our work, we use different open opinion collections without any specific treatment and consider them as a reference collection. We then use language models to determine opinion scores. The analysis document and reference collection are represented by different language models (i.e., Dirichlet, Jelinek-Mercer and two-stage models). These language models are generally used in information retrieval to represent the relationship between documents and queries. However, in our study, we modify these language models to represent opinionated documents. We carry out several experiments using Text REtrieval Conference (TREC) Blogs 06 as our analysis collection and Internet Movie Data Bases (IMDB), Multi-Perspective Question Answering (MPQA) and CHESLY as our reference collection. To improve opinion detection, we study the impact of using different language models to represent the document and reference collection alongside different combinations of opinion and retrieval scores. We then use this data to deduce the best opinion detection models. Using the best models, our approach improves on the best baseline of TREC Blog (baseline4) by 30%.

1. Introduction

The large volume of opinionated data on the Web has caused a recent increase in a number of online phenomena, such as online shopping and online elections. These opinionated data need to be manipulated in order to analyze, deduce or predict users choices in a variety of domains. Unlike traditional topic-based retrieval, the documents returned by opinion mining should not only be relevant to the topic but contain opinions about it.

While blogs are a rich source of opinions, they makes opinion detection more difficult because bloggers have a specific language that incorporates emoticons and does not respect grammatical rules. In 2006, TREC (Voorhees, 2006) debuted a special track with the

* Corresponding author.
E-mail addresses: phdups@gmail.com (F. Belbachir), Mohand.Boughanem@irit.fr (M. Boughanem).

https://doi.org/10.1016/j.ipm.2018.07.001
Received 27 February 2017; Received in revised form 21 April 2018; Accepted 2 July 2018

main task of gathering opinions on various topics. Over the years, several research groups at TREC have developed different approaches for opinion retrieval. Their aim was to retrieve a set of opinionated documents for a given set of topics. These approaches ranked the retrieved list of opinionated documents in three different ways: using machine learning-based classifiers (Liu, 2015; Mullen & Collier, 2004; Pang & Lee, 2004; Riloff & Wiebe, 2003), lexical sentiment dictionaries (Hannah, Macdonald, Peng, He, & Ounis, 2007; Lafferty & Zhai, 2001; Song, Qin, Shi, Lin, & Yang, 2007) or probabilistic models (Huang & Croft, 2009; Mei, Ling, Wondra, Su, & Zhai, 2007a; Zhang & Ye, 2008). Most opinion retrieval approaches are designed to work in two phases: topic-relevance retrieval and opinion retrieval. During the topic-relevance retrieval phase, a list of relevant documents is retrieved and ranked according to documents relevance to the given topic. In the opinion retrieval phase, this list is then ranked again by combining each documents relevance and opinion scores. Different works combine both scores (relevant and opinion) and have reported some improvements in opinion detection. TREC organizers have published baselines to encourage participants to experiment further with TREC approaches. In this study, we focus on the opinion retrieval phase. We therefore use the strongest provided TREC baseline (baseline4) Ounis, Macdonald, and Soboroff (2009) for determining topic relevance, while relying on language modeling techniques for opinion retrieval. To determine if a document is opinionated or not, we match its language model with the language models of documents containing subjective information (i.e., our reference collection). If a document is determined to be similar to our collection, we conclude that it is opinionated. The novelty and effectiveness of this approach rely on the following key features:

- We use various open and available subjective resources;
- We use different methods to calculate opinion scores;
- We adapt the language models used for opinion detection;
- We compare different language models for representing the analysis document and reference collection in order to find the models that most improve opinion detection as compared to the best baseline provided by TREC (baseline4).

This paper is organized as follows: In Section 2, we describe related work on the opinion retrieval phase. In Sections 3 and 4, we describe our proposed approach in detail. In Section 5, we discuss the experiments and the obtained results. Finally, we conclude the paper and summarize our findings.

2. Related work

There are several extant studies in the field of opinion mining. Some of these use approaches with a classifier that takes data as its input and produces output against testing data. With these approaches, difficulties include determining the features that represent opinionated documents, identifying the best classifiers for a better result, and choosing the training collection that represents the most subjective words. The features used to represent opinionated documents include, for example, the number of adjectives, verbs and adverbs (Bifet & Frank, 2010; Li, Mukherjee, Si, & Liu, 2015; Pang, Lee, & Vaithyanathan, 2002; Wang, Sun, Mukhtar & Rohini, 2008; Yang, Callan, & Si, 2006; Zhang, 2006). Some studies, meanwhile, examine the grammatical relations between different terms in a document. These studies suggest that subjectivity can only be measured in context, because some words, such as “like”, are considered subjective, when they are used in other sentences they may no longer express an opinion, such as in the sentence “it looks like a cat” (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011; Liu, Liu, Zhang, Kim, & Gao, 2016; Saif, He, Fernandez, & Alani, 2014; Saif, He, Fernández, & Alani, 2016; Seki & Uehara, 2009; Wang, Chen, & Liu, 2016). Some studies, such as the work of Aidan, Kushmerick, and Smyth (2002), classify features of opinionated documents into two categories: those that depend on the query and incorporate relevance and opinion into the learning phase (Saif et al., 2014; Seki, Kino, Sato, & Uehara, 2007), and those that use characteristics independent of the topic and do not incorporate relevance into the learning phase. Furthermore, while some studies use a single classifier like support vector machine (SVM), naive bayes or logistic regression to return opinionated documents, others use multiple different classifiers to compare their impacts on opinion detection (Balahur, 2016; Balahur & Jacquet, 2015; Bauman, Liu, & Tuzhilin, 2016; Fu, Abbasi, Zeng, & Chen, 2012; Lu, Mamoulis, Pitoura, & Tsaparas, 2016; Mullen & Collier, 2004; Pang & Lee, 2004; Riloff & Wiebe, 2003; Seki et al., 2007; Tu, Cheung, Mamoulis, Yang, & Lu, 2016). Finally, some pre-existing approaches use internal collections built directly from the collection to be analyzed for collections training, while others use external collections built from independent collections of the analyzed collection (Aidan et al., 2002; Baccianella, Esuli, & Sebastiani, 2010; Bifet & Frank, 2010; Pak & Paroubek, 2010; Seki et al., 2007). Due to their reliance on machine learning, these approaches are dependent on learning the data, features and choices of the classifier being used. Other studies use dictionaries of subjective words to identify opinionated documents, considering documents that contain many subjective words to be opinionated. Sometimes, these dictionaries are directly prepared from the test data collection. Other times, ready-made lexical dictionaries like General Inquiry (Stone, Dunphy, Smith, & Ogilvie, 1966) or SentiWordNet (SWN) (Esuli & Sebastiani, 2006) are used, although some studies report that these are less effective than lexicons extracted from the test data collection itself (Andreevskaia & Bergler, 2006; Lafferty & Zhai, 2001). While many approaches use individual words for processing semantic information, some approaches use natural language processing techniques that consider not only the word but the entire sentence (Agarwal et al., 2011; Castro-Espinoza, Gelbukh, & González-Mendoza, 2013; Liu et al., 2016; Tang, Tan, & Cheng, 2009; Thelwall, Buckley, & Paltoglou, 2012; Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010). These approaches are typically based on the presence or absence of a subjective term in a document, and do not incorporate the frequency of the subjective term in question. However, a term that reoccurs several times in a document suggests a stronger opinion than one that appears only once. This issue helps to explain why relatively few studies employ language models to identify opinionated documents. Language modeling has proved its worth in the field of information retrieval (Song & Croft, 1999) for ad-hoc information retrieval tasks based on probability to represent a model of document and a model of query.