Expert Systems With Applications 62 (2016) 243–249

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Short text opinion detection using ensemble of classifiers and semantic indexing

Johannes V. Lochter[a,*], Rafael F. Zanetti[b], Dominik Reller[a], Tiago A. Almeida[a]

[a] Department of Computer Science, Federal University of São Carlos – UFSCar, Sorocaba, 18052-780, Brazil
[b] Department of Computer Sciences, University of Wisconsin-Madison – Madison, WI 53703 USA

ARTICLE INFO

Article history:
Received 15 March 2016
Revised 2 June 2016
Accepted 12 June 2016
Available online 16 June 2016

Keywords:
Sentiment analysis
Text normalization
Semantic indexing
Classification
Machine learning

ABSTRACT

The popularity of social networks has attracted attention of companies. The growing amount of connected users and messages posted per day make these environments fruitful to detect needs, tendencies, opinions, and other interesting information that can feed marketing and sales departments. However, the most social networks impose size limit to messages, which lead users to compact them by using abbreviations, slangs, and symbols. As a consequence, these problems impact the sample representation and degrade the classification performance. In this way, we have proposed an ensemble system to find the best way to combine the state-of-the-art text processing approaches, as text normalization and semantic indexing techniques, with traditional classification methods to automatically detect opinion in short text messages. Our experiments were diligently designed to ensure statistically sound results, which indicate that the proposed system has achieved a performance higher than the individual established classifiers.

1. Introduction

Digital inclusion has allowed an increasing number of Internet users, which recently has been responsible for the most success of social networks. In such applications, users are able to share and read information, and perform many activities. Among shared information, users often post opinions and rate products. According to a press release of ComScore[1], online reviews have a significant impact on purchasing behavior. Consequently, companies noticed how important it is to be able to analyze a huge amount of messages in a fast way to discover tendencies and opinion of users.

The employment of classification methods in opinion detection were presented in some works (Denecke, 2008; Luo, Zeng, & Duan, 2016; Pang, Lee, & Vaithyanathan, 2002). However, in most cases, it is still very difficult to identify the polarity of text samples extracted from social networks because, besides being very short, they are often rife with idioms, slang, symbols, emoticons and abbreviations which make even tokenization a challenge task (Denecke, 2008).

Noise in text messages can appear in different ways. The following phrase offers an example: *"dz ne1 knw h2 ripair dis terrible LPT? :("*. There are misspelled words *"dz,ne1,knw,h2,dis"*, abbreviation *"LPT"* and symbol *":("*. In order to transcribe such phrase to a proper English grammar, a *Lingo* dictionary[2] would be needed along with a standard dictionary, which associates each slang, symbol or abbreviation to a correct term. After a step of text normalization, the input phrase would be translated to *"Does anyone know how to repair this terrible printer? :("* and the symbol at the end would mean the author has a sad or dissatisfied sentiment about the product.

In addition to noisy messages, there are other well-known problems described in literature such as sarcasm, ambiguous words in context (polysemy) and different words with the same meaning (synonymy). When such cases are properly handled, better results can be achieved (Mostafa, 2013; Pang & Lee, 2008).

Both synonymy and polysemy problems can have their effect minimized by semantic indexing for word sense disambiguation (Navigli & Ponzetto, 2012; Taieb, Aouicha, & Hamadou, 2013). Such dictionaries associate meanings to words by finding similar terms given the context of message. In general, the effectiveness of applying such dictionaries relies in the quality of terms extracted from samples. However, common tools for natural language pro-

* Corresponding author.
E-mail addresses: jlochter@acm.org, alemaoyo@gmail.com (J.V. Lochter), ferrazzanett@wisc.edu (R.F. Zanetti), dreller@acm.org (D. Reller), talmeida@ufscar.br (T.A. Almeida).

[1] ComScore press release. Available at http://goo.gl/PRlHmS, accessed in March 30, 2015.

[2] Lingo is an abbreviated language commonly used on Internet applications, such as chats, emails, blogs and social networks.

http://dx.doi.org/10.1016/j.eswa.2016.06.025
0957-4174/

cessing can not be suitable to deal with short texts, demanding proper tools for working in this context (Bontcheva et al., 2013; Maynard, Bontcheva, & Rout, 2012).

Even after dealing with problems of polysemy and synonymy, resulting terms may not be enough to detect opinion because the original messages are usually very short. Some recent works recommend to employ ontology models to analyze each term and find associated new terms (with the same meaning) to enrich original sample with more features (Kontopoulos, Berberidis, Dergiades, & Bassiliades, 2013).

Terms achieved by ontology models and semantic indexing (called expansion process) are more representative for classification methods if they can be related to an individual polarity. This way, recent works also demonstrate that lexical dictionaries can enhance classification performances (Mostafa, 2013; Nastase & Strube, 2013).

Original samples can be processed by different text processing techniques and resulting text samples become inputs to classification methods. Since there are several techniques to perform feature processing and different established classification methods, an ensemble system that naturally integrates these approaches could overcome individual drawbacks, achieve better hypothesis and consequently enhance the overall prediction performance. Ensemble strategies are commonly applied in literature to combine outputs of several classifiers in an integrated final output (Dietterich, 2000; Wang, Sun, Ma, Xu, & Gu, 2014; Xia, Zong, & Li, 2011).

In this scenario, we have designed and evaluated an ensemble system to perform opinion detection in short text messages extracted from social networks. Our model combines text normalization methods along with state-of-the-art natural language processing techniques to improve quality of extracted features which are then used by established machine learning approaches. The results demonstrate that our proposal clearly outperforms established methods available in literature.

This paper is organized as follows: Section 2 presents the most relevant related work. Text normalization and semantic indexing techniques are described in Section 3. Section 4 presents the proposed ensemble system. Experimental methodology is described in Section 5. Section 6 presents the achieved results and main conclusions are provided in Section 7.

2. Related work

Opinion detection is the task of analyzing huge amounts of information from thousands (or millions) of users to detect the majority opinion about anything in discussion. The understanding and fast reaction about such opinions allows companies to guide their marketing and to aid in decision making (Mostafa, 2013; Pang et al., 2002). According to results available in literature, this task is far from being properly solved due to many reasons, such as difficulties to deal with sarcasm, irony, and sentences with multiple polarities. In addition, another important well-known problem is related to the amount and quality of features extracted from messages. Often, text messages extracted from social networks are short and usually rife with noise (slangs, symbols, abbreviations, and so on), causing bad vector representation that decreases the classifiers performances (Go, Bhayani, & Huang, 2009; Navigli & Lapata, 2010).

In text categorization, a challenge that remains in dealing with short text is the lack of information about its content. The limit size usually imposed by the channel (e.g., Twitter), not rarely, leads classification methods to face problems like polysemy and synonymy. Polysemy is the capacity for a single term has multiple meanings represented by only one attribute in a feature vector. Synonymy is related to the capacity for multiple terms have same meaning represented by more than one attribute in a feature vector. In this scenario, there are recent works that successfully applied semantic indexing and lexical normalization to avoid these problems in order to improve the quality of features (Nastase & Strube, 2013; Navigli & Ponzetto, 2012).

Lexical normalization or text normalization is the task of replacing lexical variants of standard words and expressions normally obfuscated in noisy texts to their canonical forms, in order to allow further processing of text processing tasks. It is closely related to spell checking, and in fact, many approaches in literature share techniques from this task (Cook & Stevenson, 2009; Xue, Yin, Davison, & Davison, 2011).

Semantic indexing or Query Expansion is the task of replacing words in texts by their synonyms according to the concept the target word belongs to (Hidalgo, Rodríguez, & Pérez, 2005). As an example, the semantic network WordNet represents synonyms sets as following: {car, auto, automobile, machine, motorcar} (a motor vehicle with four wheels) or {car, railcar, railway car, railroad car} (a wheeled vehicle adapted to the rails of railroad) for the word "car".

Output samples produced by semantic indexing add complexity in the task of identifying the most appropriate concepts for each word in the message given its context. This problem can be handled using *Word Sense Disambiguation* (WSD) which is a popular technique used in deep natural language processing (Agirre & Edmonds, 2006). In this work, we have used the BabelNet semantic network along with WSD unsupervised algorithm (Navigli & Lapata, 2010), following the Semantic Expansion method described in Gómez Hidalgo et al. (2005).

After lexical normalization and semantic indexing, the original noisy samples are processed and expanded by adding new concepts related to the context of terms in sample. Therefore, besides the sample being normalized, it is also enriched with more information in order to aid classification methods to improve their prediction capacities (Kontopoulos et al., 2013; Nastase & Strube, 2013).

The abundance of text processing techniques and classification methods to handle short text messages demands some way to combine them in order to acquire a generic and good hypothesis. In this scenario, an ensemble system is highly recommended to find out a good classification model in an automatic way (Dietterich, 2000).

Ensemble of classifiers is a technique developed to achieve generic hypotheses by combining different classifiers. The ensemble works like a committee in which each classifier is a voting member and the committee produces a final prediction based in their votes. This technique is commonly applied to minimize specific drawbacks, such as overfitting and the curse of dimensionality (Dietterich, 2000). Although ensemble systems can adopt different strategies, they usually achieve better results than individual classifiers (Wang et al., 2014).

Widely adopted, weighted ensemble systems are often found in literature. The effectiveness of these techniques relies on assigning an appropriated weight for each vote. Thus, a less-accurate classifier should not have the same or more a significant vote than a more-accurate one (Kim, Kim, Moon, & Ahn, 2011; Xia et al., 2011).

As short text samples can be processed by different text processing techniques, and moreover, there are several established classification methods recommended to opinion detection, a sophisticated ensemble system that combines these approaches can lead to generic hypotheses and consequently achieve good performance.

3. Text processing techniques

In scenarios where messages are short and rife with idioms, symbols and abbreviations, just employing a simple bag of words