Applied Soft Computing 67 (2018) 822–833

Contents lists available at ScienceDirect

Applied Soft Computing

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

An approach to improve the accuracy of probabilistic classifiers for decision support systems in sentiment analysis

Vicente García-Díaz [a], Jordán Pascual Espada [a], Rubén González Crespo [b,*], B. Cristina Pelayo G-Bustelo [a], Juan Manuel Cueva Lovelle [a]

[a] University of Oviedo, Department of Computer Science, Sciences Building, Oviedo, Asturias, Spain
[b] Universidad Internacional de La Rioja (UNIR), Av. de la Paz, 137, 26006 Logroño, La Rioja, Spain

ARTICLE INFO

Article history:
Received 25 January 2017
Received in revised form 16 May 2017
Accepted 18 May 2017
Available online 25 May 2017

Keywords:
Sentiment analysis
Classifier
Bayesian network
Soft computing
Domain-specific language
Machine learning

ABSTRACT

Social networks link people and machines, providing a huge amount of information that grows very fast without the possibility to be handled manually. Moreover, opinion mining is the process of using natural language processing, text analytics and computational linguistics to identify and extract subjective information in different sources such as social networks. To that, classification methods are used but due to the limitless number of topics and the breadth and ambiguity of natural language, with its peculiarities in social networks, the results can be greatly improved. In this work, we present DSociaL, a platform to automate the processing of information obtained from social networks, focusing on improving the accuracy of decision support systems for sentiment analysis. We focus on machine learning-based simple probabilistic classifiers, evaluating a naive Bayes classifier, the basis of one of the most used soft computing techniques. Thus, we show a use case in which the proposal, with definitions and refinements made by experts, helps to improve the prediction of users' feelings towards a movie compared to what would happen with a conventional approach.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Networks are groups of interrelated or interconnected elements, especially over large areas. Thus, networks of people exist from the origin of the human race. They are very important because those groups of people bound together may undertake an objective with greater ease. There are different types of groups such as informal networks, project teams, formal work groups or communities of practice, all of them composed under different circumstances and useful in complementary ways [1].

In addition, when computers link people and machines, they become social networks. Social Network Sites (SNS) such as Facebook, Twitter, Google+ or LinkedIn have attracted millions of users that are using those sites in their daily practices [2]. More than a billion of people in the world is connected together to create, collaborate and contribute their knowledge and wisdom, having social related factors the most significant impact on the intention of use [3]. In this regard, there are many works that use the current capacity of social networks, where the geo-localization plays an important role, to perform different types of research [4–7].

Microblogging is a form of social communication in which users describe their feelings in short messages distributed through the Internet by using Web, desktop or mobile applications. People use such services mainly to talk about their daily activities and to look for or share any type of information [8]. For example, Twitter is a microblogging service that basically allows users to write about any topic within the 140-character limit and follow others to receive their messages (called *tweets*). It has some features such as non-power-law follower distribution, a short effective diameter and low reciprocity [9], which determines a differentiation from other existing non-virtual networks [10].

With the evolution of social networks, there is available a great amount of digital content and shared knowledge resources that produce environments with a mixture between the architecture of the underlying information and the social structure of the groups of people that are part of the communities [11]. This has lead to the appearance of a huge amount of data that is impossible to be manually treated by people as it grows in millions of bytes every day. In fact, it has been estimated that the amount of data stored in databases in the world doubles every 20 months [12].

* Corresponding author.
E-mail addresses: garciavicente@uniovi.es (V. García-Díaz), pascualjordan@uniovi.es (J.P. Espada), ruben.gonzalez@unir.net (R.G. Crespo), crispelayo@uniovi.es (B.C. Pelayo G-Bustelo), cueva@uniovi.es (J.M. Cueva Lovelle).

https://doi.org/10.1016/j.asoc.2017.05.038
1568-4946/© 2017 Elsevier B.V. All rights reserved.

The Knowledge Discovery in Databases (KDD) process deals with the huge amount of available data with the aim of discovering new knowledge from the information that is stored in different systems. It has usually five steps [13]: 1 – selection; 2 – pre-processing; 3 – transformation; 4 – data mining; and 5 – interpretation. Data mining is an interdisciplinary field with lots of applications that refers to the process of discovering patterns in large data sets by the use of other fields such artificial intelligence, statistics or programming algorithms [14]. Thus, text mining or text data mining [15] is the process of deriving high-quality information from written texts. Moreover, sentiment analysis (a.k.a. as opinion mining) [16] is one of the most interesting applications of text mining. We will focus on opinion mining in this work. It refers to the use of Natural Language Processing (NLP) techniques, together with text analytics and computational linguistics to identify and extract subjective information from texts.

To deal with opinion mining related problems, soft computing techniques (a.k.a. computational intelligence) are usually used. It consists in the use of inexact solutions to computationally difficult tasks, for which there is no known algorithm that gives correct solutions in a constraint amount of time. So, soft computing is tolerant to imprecision or uncertainty. In addition, machine learning plays an important role in soft computing. It is a subfield of computer science that gives computers the ability to learn without being explicitly programmed [17]. Machine learning is widely used to categorize anything for which there is no exact precision on how to deal with it. A clear example is sentiment classification.

In order to classify the sentiment of a text, classifiers are used. In machine learning and statistics, classification focuses on identifying to which category a new observation belongs on the basis of a training set of data that has observations, whose category is known beforehand. Classification is considered part of supervised learning.

The most well-known type of classification is probabilistic classification. To that, statistical inference is needed to find the best class for a given observation, based on the probability for each of them. Naive Bayes classifier is the most common probabilistic classifier and refers to a family of simple classifiers based on applying Bayes theorem with strong independence assumptions among the different variables or features. Being able to know the characteristics of each of the classifiers, and how to configure them for all the possible cases is not a simple task and requires the knowledge of expert developers, among other factors [18].

Working with all the concepts, parameters, algorithms and tools available for sentiment analysis is not a trivial task and requires the use of expert developers. Thus, Domain-Specific Languages (DSLs) can play an important role. According to Walton [19], a DSL is a small, usually declarative, language expressive over the distinguishing characteristics of a set of programs in a particular problem domain. Those little languages, tailored towards the specific needs of a particular domain can significantly ease building software systems for such a domain [20]. Even, domain-experts can directly use the DSL to make required routine modifications [21] or even program applications for their domain of knowledge (e.g., Garcia-Diaz et al. [22] for building food traceability applications under a Model-Driven Engineering approach [23]).

Getting a good sentiment analysis is key to recommendation or feedback systems, in which the users sentiment towards a product or service can be decisive in the future actions of the provider or to other potential customers. Thus, the main goal of this research work is to provide a novel approach to improve the accuracy of probabilistic classifiers, which are applied in the field of sentiment analysis in messages related to current issues. For example, messages related to a new movie, social events, political decisions, etc. As the characteristics and expressions used in social networks have their particular peculiarities (e.g., colloquial expressions, abbreviations, emoticons, dynamism, etc.), we will focus on the fourth step in the KDD process for improving opinion mining accuracy on social networks.

Our approach allows experts in specific domains of knowledge to identify positive and negative features that current classifiers are not able to automatically identify as key elements, to correctly classify the polarity of a given text. By using a DSL, even people without a background in traditional programming languages can use it to define rules and metarules that can improve the performance of a given classifier.

This approach is based on a post-processing task performed using the cosine similarity between the analyzed messages and a set of text snippets and logical expressions. This set could be, for example, few phrases or expressions related to a current event. The output modifies directly the probability for a message to be positive, negative o neutral. So, this solution can be used after the machine learning algorithm for detecting the message polarization.

The remainder of this work is structured as follows: In Section 2 we present a background with advances of the state of the art related to sentiment analysis, focusing on social networks. In Section 3 we propose our alternative to improve access to data in social networks focusing on getting the sentiment of a message. Section 4 shows an evaluation of the proposal. Finally, Section 5 deals with the conclusions and future work to be done.

2. Background

A wide range of research work is focused on the sentiment classification [24–26]. There are lot of different approaches, being very common the division into two well differentiated types of methods [27]:

- Lexicon based methods. There are a lot of methods used in some research works such as the pointed out by Taboada et al. [28]. These techniques are mainly based on dictionaries of words annotated with their semantic polarity.
- Machine learning based methods. They are divided also in groups: supervised and unsupervised techniques [29]. In addition, some authors mention a hybrid between these both: semi-supervised learning [30,31]. These techniques are typically more complex and, in the most usual case, require to create a model by training a classifier with labeled examples.

The different works rely on NLP. For example, Alonso et al. [32], propose a linguistic consensus model for Web communities, with some feedback mechanisms to improve speed and convergence. Godbole et al. [33] work on the sentiment analysis of online news and blogs texts, assigning scores to each distinct entity in the text corpus. Other works such as Gonzalez et al. [34], also explore the use of Twitter. They propose a novel approach to obtain a fine grain sentiment analysis with semantics.

Many polarity analysis use “feature selection” to classify the message polarity of the text or some snippets of it [35]. These approaches extract key parts of the message like nouns, verbs and adjectives to create n-grams. These keywords are used as features in machine learning algorithms like Naive Bayes, which are a family of algorithms commonly used in messages polarity classification [36].

Many authors have used techniques based on analyze sets of n-grams, like bigrams and trigrams to consider consecutive words, because these sets could have a different meaning than every single word [37,28]. Other approaches have used dependency trees to model the correlation between sets of words [38]. But there are other approaches which have achieved an improvement in the

Download English Version:

https://daneshyari.com/en/article/6903888

Download Persian Version:

https://daneshyari.com/article/6903888

Daneshyari.com