Knowledge-Based Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Meta-level sentiment models for big social data analysis

Q1 Felipe Bravo-Marquez [a,b,*], Marcelo Mendoza [c], Barbara Poblete [d]

[a] Department of Computer Science, The University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand
[b] Yahoo! Labs Santiago, Av. Blanco Encalada 2120, 4h floor, Santiago, Chile
[c] Universidad Técnica Federico Santa María, Av. Vicuña Mackenna 3939, Santiago, Chile
[d] Department of Computer Science, University of Chile, Av. Blanco Encalada 2120, Santiago, Chile

ARTICLE INFO

Article history:
Received 22 November 2013
Received in revised form 6 May 2014
Accepted 15 May 2014
Available online xxxx

Keywords:
Q3 Sentiment classification
Twitter
Meta-level features

ABSTRACT

People react to events, topics and entities by expressing their personal opinions and emotions. These reactions can correspond to a wide range of intensities, from very mild to strong. An adequate processing and understanding of these expressions has been the subject of research in several fields, such as business and politics. In this context, Twitter sentiment analysis, which is the task of automatically identifying and extracting subjective information from tweets, has received increasing attention from the Web mining community. Twitter provides an extremely valuable insight into human opinions, as well as new challenging *Big Data* problems. These problems include the processing of massive volumes of streaming data, as well as the automatic identification of human expressiveness within short text messages. In that area, several methods and lexical resources have been proposed in order to extract sentiment indicators from natural language texts at both syntactic and semantic levels. These approaches address different dimensions of opinions, such as subjectivity, polarity, intensity and emotion. This article is the first study of how these resources, which are focused on different sentiment scopes, complement each other. With this purpose we identify scenarios in which some of these resources are more useful than others. Furthermore, we propose a novel approach for sentiment classification based on meta-level features. This supervised approach boosts existing sentiment classification of subjectivity and polarity detection on Twitter. Our results show that the combination of meta-level features provides significant improvements in performance. However, we observe that there are important differences that rely on the type of lexical resource, the dataset used to build the model, and the learning strategy. Experimental results indicate that manually generated lexicons are focused on emotional words, being very useful for polarity prediction. On the other hand, lexicons generated with automatic methods include neutral words, introducing noise in the detection of subjectivity. Our findings indicate that polarity and subjectivity prediction are different dimensions of the same problem, but they need to be addressed using different subspace features. Lexicon-based approaches are recommendable for polarity, and stylistic part-of-speech based approaches are meaningful for subjectivity. With this research we offer a more global insight of the resource components for the complex task of classifying human emotion and opinion.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Q4 Humans naturally communicate by expressing feelings, opinions and preferences about the environment that surrounds them. Moreover, the emotional load of a message, written or verbal, is extremely important when it comes to understanding its true meaning. Therefore, opinion and sentiment comprehension are a key aspect of human interaction. For many years, emotions have been studied individually and also collectively, in order to understand human behavior. The collective or social analysis of opinions and sentiment responds to the need to measure the impact or polarization that a certain event or entity has on a group of individuals. Social sentiment has been studied in politics to understand and forecast election outcomes, and also in marketing, to predict the success of a certain product and to recommend others.

Before the rise of online social media, gathering data on opinions was expensive and usually achieved at very small scale. When users on the Web started communicating massively through this channel, social networks became overloaded with opinionated

* Corresponding author at: Department of Computer Science, The University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand. Tel.: +64 7 838 Q2 4466x8766; fax: +64 7 838 4155.
Q1 *E-mail addresses:* fjb11@students.waikato.ac.nz (F. Bravo-Marquez), mmendoza@inf.utfsm.cl (M. Mendoza), bpoblete@dcc.uchile.cl (B. Poblete).

http://dx.doi.org/10.1016/j.knosys.2014.05.016
0950-7051/© 2014 Elsevier B.V. All rights reserved.

UNCORRECTED PROOF

data. In that aspect, social media has opened new possibilities for human interaction. Microblogging platforms, in particular, allow real-time sharing of comments and opinions. Twitter,[1] an extremely popular microblogging platform, has millions of users that share millions of personal posts on a daily basis. This rich and enormous volume of user generated data offers endless opportunities for the study of human behavior.

Manual classification of millions of posts for opinion mining is an unfeasible task. Therefore, several methods have been proposed to automatically infer human opinions from natural language texts. Computational sentiment analysis methods attempt to measure different opinion dimensions. A number of methods for polarity estimation have been proposed in [3,12,13,23] discussed in depth in Section 2. Polarity estimation is reduced into a classification problem with three polarity classes – positive, negative and neutral – with supervised and unsupervised approaches being proposed for this task. In the case of unsupervised approaches, a number of lexicon resources with positive and negative scores for words exist. A related task is the detection of *subjectivity*, which is the specific task of separating factual from opinionated text. This problem has also been addressed with supervised approaches [33]. In addition, opinion intensities (strengths) have also become a matter of study, for example, SentiStrength [30] estimates positive and negative strength scores at sentence level. Finally, the emotion estimation problem has also been addressed with the creation of lexicons. The Plutchik wheel of emotions, proposed in [28], is composed of four pairs of opposite emotion states: *joy-trust*, *sadness-anger*, *surprise-fear*, and *anticipation-disgust*. Mohammad et al. [20] labeled a number of words according to Plutchik emotional categories, developing the NRC word-emotion association lexicon. All of the approaches described above perform sentiment analysis at a syntactic-level. On the other hand, there are approaches that use semantic knowledge bases to perform sentiment analysis at a semantic-level [5,24,25].

Regardless of the growing amount of work in this research area, sentiment analysis remains a widely open problem, due in part to the inherent subjectivity of the data, as well as language and communication subtleties. In particular, opinions are multidimensional semantic artifacts. When people are exposed to information regarding a topic or entity, they normally respond to this external stimuli by developing a personal point of view or orientation. This orientation reveals how the opinion holder is polarized by the entity. Additionally, people manifest emotions through opinions, which are the driving forces behind motivations and personal dispositions. This indicates that emotions and polarities are mutually influenced by each other, conditioning opinion intensities and emotional strengths.

In this article we analyze the existing literature in the field of sentiment analysis. Our literature overview shows that current sentiment analysis approaches mostly focus on a particular opinion dimension. Although these scopes are difficult to categorize independently, we propose the following taxonomy for existing work:

1. **Polarity**: These methods and resources aim to extract polarity information from a passage. Polarity-oriented methods normally return a categorical variable whose possible values are positive, negative and neutral. On the other hand, polarity-oriented lexical resources are composed of positive and negative words lists.
2. **Strength**: These methods and resources provide intensity levels according to a polarity sentiment dimension. Strength-oriented methods return numerical scores indicating the intensity or the strength of positive and negative sentiments expressed in a text passage. Strength-oriented lexical resources provide lists of opinion words together with intensity scores regarding positivity and negativity.
3. **Emotion**: These methods and resources are focused on extracting emotion or mood states from a text passage. An emotion-oriented method should classify the message to an emotional category such as sadness, joy, surprise, among others. Emotion-oriented lexical resources provide a list of words or expressions marked according to different emotion states.

We analyze how each approach can be used in a complementary way. In order to achieve this, we introduce a novel meta-feature classification approach for boosting the sentiment analysis task. This approach efficiently combines existing sentiment analysis methods and resources focused on the three different scopes presented above. The main goals are to improve two major sentiment analysis tasks: (1) Subjectivity classification, and (2) Polarity classification. We combine all of these aspects as meta-level input features for sentiment classification. To validate our approach we evaluate our classifiers on three existing datasets.

Our results show that the composition of these features achieves significant improvements over individual approaches. This indicates that strength, emotion and polarity-based resources are complementary, addressing different dimensions of the same problem.

To the best of our knowledge, this is the first study that combines polarity, emotion, and strength oriented sentiment analysis lexical resources, with existing opinion mining methods as meta-level features for boosting sentiment classification performance.[2] Furthermore, we perform lexicon analyses by comparing resources created manually to lexicons that were completely automatically created or partially automatically created. We explore the level of neutrality of each resource, and also their level of agreement. Our results indicate that manually generated lexicons are focused on emotional words, being very useful for polarity prediction. On the other hand, lexicons generated by automatic methods tend to include many neutral words, introducing noise in the detection of subjectivity. We observe also that polarity and subjectivity prediction are different dimensions of the same problem, but they need to be solved using different subsets of features. Lexicon-based approaches are recommendable for polarity, and stylistic part-of-speech based approaches are meaningful for subjectivity.

This article is organized as follows. In Section 2 we provide a review of existing lexical resources and discuss related work on Twitter sentiment analysis. In Section 3 we describe our meta-level feature space representation of Twitter messages. The experimental results are presented in Section 4. In Section 4.1 we explore the relationship between different opinion lexicons, and in Section 4.2 we present our classification results. Finally, we conclude in Section 5 with a brief discussion.

2. Related work

2.1. Twitter sentiment analysis

Twitter users tend to post opinions about products or services [26]. *Tweets* (user posts on Twitter) are short and usually straight to the point messages. Therefore, tweets are considered as a rich resource for sentiment analysis. Common opinion mining tasks that can be applied to Twitter data are sentiment classification and opinion identification. Twitter messages are at most,

[1] http://www.twitter.com.

[2] This article extends a previous workshop paper [4] and provides a more thorough and detailed report.

Q1 Please cite this article in press as: F. Bravo-Marquez et al., Meta-level sentiment models for big social data analysis, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.05.016

UNCORRECTED PROOF

Download English Version:

https://daneshyari.com/en/article/6862533

Download Persian Version:

https://daneshyari.com/article/6862533

Daneshyari.com