Information Fusion 31 (2016) 54–64

Contents lists available at ScienceDirect

Information Fusion

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

Tweet categorization by combining content and structural knowledge

J.M. Cotelo*, F.L. Cruz, F. Enríquez, J.A. Troyano

Department of Languages and Computer Systems, University of Seville, Avda. Reina Mercedes s/n, Seville 41012, Spain

ARTICLE INFO

Article history:
Received 18 June 2015
Revised 10 November 2015
Accepted 6 January 2016
Available online 13 January 2016

Keywords:
Twitter
Tweet categorization
Ensemble learning
Knowledge combination

ABSTRACT

Twitter is a worldwide social media platform where millions of people frequently express ideas and opinions about any topic. This widespread success makes the analysis of tweets an interesting and possibly lucrative task, being those tweets rarely objective and becoming the targeting for large-scale analysis. In this paper, we explore the idea of integrating two fundamental aspects of a tweet, the proper textual content and its underlying structural information, when addressing the tweet categorization task. Thus, not only we analyze textual content of tweets but also analyze the structural information provided by the relationship between tweets and users, and we propose different methods for effectively combining both kinds of feature models extracted from the different knowledge sources. In order to test our approach, we address the specific task of determining the political opinion of Twitter users within their political context, observing that our most refined knowledge integration approach performs remarkably better (about 5 points above) than the textual-based classic model.

1. Introduction

Twitter is a successful worldwide social media platform where millions of people frequently express ideas and opinions about a myriad of topics. Texts written in Twitter, called tweets, are characterized by having a very short length (140 characters) and often written using devices like smartphones with almost no revision before sending them, trading redaction quality and/or correctness for speed. Aside from the textual content, tweets (and Twitter itself) offer many other data and information that may serve as knowledge sources for solving different tasks of interest. In this work, we explore how the integration of these heterogeneous knowledge sources improves the overall performance when it is used for addressing the automatic tweet categorization task.

The widespread success of Twitter makes the analysis of the tweets a very interesting (and possibly lucrative) task. The amount of information in these texts is huge and tweets are rarely objective, becoming the target of large-scale analysis that could be really useful for marketing campaigns, public opinion determination or even inferring how a population responds for specific events. Branding [1,2], political analysis [3–5] or user profiling for market analysis [6] are examples of actual applications for the analysis of Twitter and other social media texts. Protection and detection against malware [7] is also an application of interest, as Twitter trending topics are also vulnerable to scamming, phishing or spamming.

Categorizing twitter messages is an interesting and valuable task. In this paper, we address the categorization of tweets, focusing on determining the political opinion of Twitter users within their political context. A collection of tweets not only provides textual information, but also provides structural information due to the relationship between users and messages, forming an underlying network. We discuss the analysis of both types of content, applying different approaches and yielding different feature models for each of them, and we propose several methods of combining these feature models (both structural an textual ones) in the classification process.

Although the approach presented in this article was originally designed for the tweet categorization task in mind, it can be applied to other social networks (e.g. Facebook, Google+, …), as long as those social networks on which this approach is applied, may exhibit relevant structural features in its messages.

Fig. 1 shows a diagram of the overall process of our proposal, which can be described as follows. It starts with the retrieval phase, in which we collect tweets using an automatic topic-related retrieval method that ensures that collected tweets are politically relevant. Before continuing with the experimental process, those tweets are manually annotated by two independent annotators. The retrieval and annotation process, along with the reference to retrieval method, are described in Section 3. These politically relevant tweets are fed to two distinct pipelines in order to generate different feature models, being each pipeline focused on analyzing

* Corresponding author. Tel.: +34 676582325.
E-mail addresses: jcotelo@us.es (J.M. Cotelo), fcruz@us.es (F.L. Cruz), fenros@us.es (F. Enríquez), troyano@us.es (J.A. Troyano).

http://dx.doi.org/10.1016/j.inffus.2016.01.002

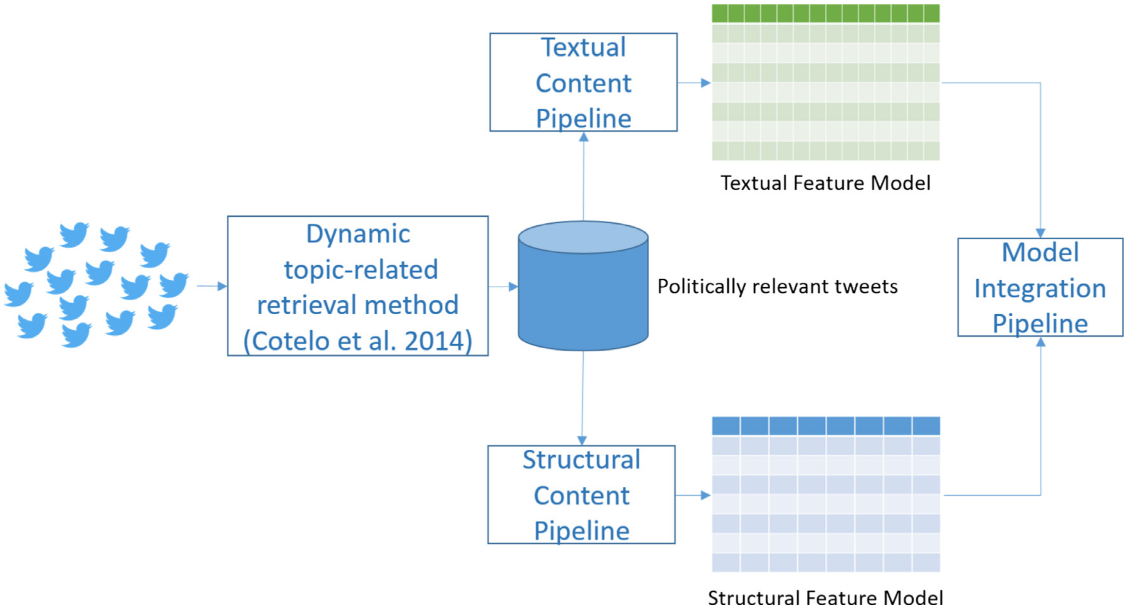
Textual Content Pipeline
Textual Feature Model
Dynamic topic-related retrieval method (Cotelo et al. 2014)
Politically relevant tweets
Model Integration Pipeline
Structural Content Pipeline
Structural Feature Model

Fig. 1. Overall process of our proposal.

textual or structural content. After each structural and textual feature models are generated, both are fed to the model integration pipeline, which combines both models and improves the performance results.

This article is structured as follows: In Section 2, we discuss several other works that are related with the one presented in this paper. In Section 3, we define the specific task addressed in this paper and we characterize the dataset, describing how it is generated and its particularities. In Section 4 we described the whole experimental setup developed for evaluating our proposal. In Section 5, we address the extraction of knowledge from the textual content of tweets, discussing the adaptation of models such as the *Bag-of-Words* and its caveats, proposing an automatic feature selection process for improving the performance of that model. In Section 6 we discuss how to extract information from the structural content of tweets, exploiting the topological features of the underlying network formed by the users and messages. We propose a topological approach, consisting in generating a bipartite friendship graph based on the identification of two major kinds of users (*content creators* and *content consumers*). From this graph, we propose two different community feature models based on the *Louvain method* and *Spectral Biclustering* technique respectively. In Section 7, we combine both structural and textual feature models in three different ways: directly combining both feature models, using *Stacking Generalization* and a variation of our own that we called *Multiple Pipeline Stacked Generalization*. Finally, in Section 8, we summarize our efforts, review the main points of our work and discuss the importance of combining structural and textual content for tweet.

2. Related works

There are many works dealing with the classification of tweets, especially in the area of sentiment analysis (considering the subjective content of many of the tweets). Most of these works face polarity classification, i.e., deciding whether a tweet expresses a positive, negative or neutral opinion toward a given topic. For example, in [8] the polarity classification is performed using polarity lexicons (resources consisting of lists of negative and positive words), and later tweaking this polarities from the semantic context in which the words appear. In [9] the lack of Arabic polarity lexicons is supplied by using emoticons appearing in the tweets to build a polarity classifier on Twitter; the idea is taken from [10], where the same technique was applied to a corpus of tweets in English. In [11], automatic summaries of the opinions of a Twitter user are generated by integrating the various negative and positive opinions expressed by users on various topics. In all these studies the classification of a tweet is performed based solely on their textual content: in no case structural information inherent to Twitter is used, as might be other user's tweets or other tweets using the same hashtags, for example. In [12], the authors transform the textual content to a graph representation where nodes are tweets, users, hashtags and words. Nodes are labeled with polarities, taken from lexicons and from a supervised classifier previously trained. After that, they apply a label propagation algorithm described in [13]. Their proposal is evaluated using three datasets (one of them is from the political domain). Although the graph representation of textual content proves to be effective and it yields better results than other approaches that directly handle the textual content, the proposal only takes into account knowledge from textual content and do not extract any information from the structural knowledge that the underlying network offers.

Many authors have been interested in studying the behavior of users on Twitter in relation to politics. In [14] an analysis of the hashtags used in politics in Canada was performed to distinguish the different objectives for which they are used in the political context. In [15] a study of the different types of Twitter users was conducted to characterize the so-called opinion leaders; they tend to seek information, mobilize, and express opinions publicly, and have a great influence on the political tendency of their followers (we will use this idea in our work, see Section 6). In [4] the LIWC2007 resource (Linguistic Inquiry and Word Count; Pennebaker [16]) was used to determine emotional and cognitive characteristics of tweets related to federal election of the national parliament in Germany 2009. They conclude that there exist high correlations between the above results and dataset statistical metrics such as concentration and share, and even the mere number of tweets mentioning each candidate is a good estimator of the election results. In [17] they also try to measure whether there is a correlation between activity in social networks (Facebook and Twitter) and the results of the US presidential election of 2012. Based on the tweets mentioning Obama and Romney, Facebook official pages