Information Processing and Management 53 (2017) 595–607

Contents lists available at ScienceDirect

Information Processing and Management

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

Supervised sentiment analysis in multilingual environments

David Vilares*, Miguel A. Alonso, Carlos Gómez-Rodríguez

Grupo LyS, Departamento de Computación, Universidade da Coruña, Campus de A Coruña s/n, 15071, A Coruña, Spain

ARTICLE INFO

Article history:
Received 13 July 2016
Revised 26 October 2016
Accepted 9 January 2017

Keywords:
Sentiment analysis
Multilingual
Code-Switching

ABSTRACT

This article tackles the problem of performing multilingual polarity classification on Twitter, comparing three techniques: (1) a multilingual model trained on a multilingual dataset, obtained by fusing existing monolingual resources, that does not need any language recognition step, (2) a dual monolingual model with perfect language detection on monolingual texts and (3) a monolingual model that acts based on the decision provided by a language identification tool. The techniques were evaluated on monolingual, synthetic multilingual and code-switching corpora of English and Spanish tweets. In the latter case we introduce the first code-switching Twitter corpus with sentiment labels. The samples are labelled according to two well-known criteria used for this purpose: the *SentiStrength scale* and a *trinary scale* (*positive, neutral* and *negative* categories). The experimental results show the robustness of the multilingual approach (1) and also that it outperforms the monolingual models on some monolingual datasets.

1. Introduction

Automatically understanding all the information shared on the Web and transforming it into knowledge is one of the main challenges in the age of Big Data. In terms of natural language processing (NLP), this usually involves comprehending different human languages such as English, Spanish or Arabic, which are implicitly related with relevant human aspects such as cultures, countries or even religions. A very simple example of these real differences can be illustrated by the concept *dragon*, which has a positive perception in Chinese, but not necessarily in other languages such as English or Spanish.

In this context, Twitter has become one of the most useful social networks for trending analysis, given the amount of data and its popularity in different countries (Cambria, Rajagopal, Olsher, & Das, 2013a; Cambria, Schuller, Liu, Wang, & Havasi, 2013b). Some of these trends are global (e.g. the Oscars, Superbowl, Rihanna or the recent Volkswagen scandal) and so their trending topics are also global (e.g. '#oscars2016', '#superbowl2016', ...). However, the public perception of these trends often changes from one culture to another and the task becomes even harder when tweets are written in different languages. This is a challenge for global companies and organizations that need to make specific business and marketing decisions depending on their target population. However, if their monitoring processes are focused on a single language (usually English) the knowledge that they acquire might be incomplete, or even worse, inaccurate. There are even more difficult and unexplored multilingual variants, such as code-switching texts (i.e. texts that contain terms in two or more different languages). Colloquial creole languages such as *Spanglish* (a mix of Spanish and American English) or *Singlish* (English-based creole from Singapore) or even official languages such as the *Haitian creole* (which merges Portuguese, Spanish, Taíno, and West African languages), are some of the best-known situations.

* Corresponding author.
E-mail addresses: david.vilares@udc.es (D. Vilares), miguel.alonso@udc.es (M.A. Alonso), carlos.gomez@udc.es (C. Gómez-Rodríguez).

http://dx.doi.org/10.1016/j.ipm.2017.01.004
0306-4573/

As a result, there is a need to provide effective support for analyzing user-generated content that lacks structure and is created in different languages (Dang et al., 2014). In this context, *sentiment analysis* (SA) techniques have been successfully applied to this social network in order to monitor a wide variety of issues ranging from the perception of the public with respect to popular events (Thelwall, Buckley, & Paltoglou, 2011) to political analysis, determining the political opinion of users (Cotelo, Cruz, Enríquez, & Troyano, 2016) or showing whether the sentiment expressed in messages is positive, negative or neutral (Vilares, Thelwall, & Alonso, 2015d). However, most of the existing research on sentiment analysis is either monolingual or cross-lingual: models intended for purely multilingual or code-switching messages are scarce. This article fills this gap, describing a novel method for multilingual polarity classification that relies on fusing existing monolingual corpora, instead of applying MT techniques or language-specific pipelines.

This article has the following research objectives:

1. To build the first code-switching corpus from Twitter for sentiment analysis. Each tweet collected in such a corpus will contain words written in at least two different languages.
2. To design a multilingual sentiment analysis system able to determine the sentiment present in texts written in different languages. To do this, we apply soft-data fusion (Khaleghi, Khamis, Karray, & Razavi, 2013) at the core level of the information fusion process applied to SA (Level 2 - Situation Refinement), as illustrated by Balazs and Velásquez (2016). In particular, existing monolingual corpora are fused to create such multilingual system.
3. To evaluate the performance of the multilingual sentiment analysis system on standard corpora and on the novel code-switching corpus, comparing its performance with respect to the combination of a language detection system and monolingual sentiment analysis systems.

For these purposes, we will consider English (*en*) and Spanish (*es*) as working languages throughout this article. Thus, the aim of the article is to show how current supervised approaches can address situations where monolingual, multilingual and code-switching texts appear.

The remainder of the paper is organised as follows: Section 2 discusses the state of the art regarding opinion mining on texts in diverse languages, including monolingual, cross-lingual and multilingual approaches. Section 3 describes the process and result of building the code-switching corpus. Section 4 introduces the main ideas and features of the proposed models. Section 5 defines the experimental framework and outlines the corpora used for evaluation, including both standard collections and the novel code-switching corpus. Section 6 presents the results obtained by the models on these corpora, which are discussed in Section 7. Finally, Section 8 draws our conclusions and outlines for future research.

2. Related work

We start by considering the issues we must face when mining opinions from non-English texts. We then focus on work applying a given opinion mining technique to corpora in different languages. Next, we review work on cross-language opinion mining and finally we consider work on multilingual subjectivity detection and polarity classification.

2.1. Mining opinions from non-English texts

There is recent work on the definition of language-specific methods for opinion mining in a wide variety of languages, including, among others, Arabic (Aldayel & Azmi, 2015), Chinese (Vinodhini & Chandrasekaran, 2012; Zhang, Zeng, Li, Wang, & Zuo, 2009), Czech (Habernal, Ptáček, & Steinberg, 2014), French (Ghorbel & Jacot, 2011), German (Scholz & Conrad, 2013), Hindi (Medagoda, Shanmuganathan, & Whalley, 2013), Italian (Neri, Aliprandi, Capeci, Cuadros, & By, 2012), Japanese (Arakawa, Kameda, Aizawa, & Suzuki, 2014), Russian (Medagoda et al., 2013), Spanish (Vilares, Alonso, & Gómez-Rodríguez, 2015c) and Thai (Inrak & Sinthupinyo, 2010). One of the problems we face when dealing with languages other than English is that many English language sentiment dictionaries are freely available, but such vocabulary lists are scarce for other languages. A current line of work is the automatic or semi-automatic generation of large non-English sentiment vocabularies (Steinberger, 2012). In this line, Kim, Jung, Nam, Lee, and Lee (2009) propose to create a sentiment lexicon for Korean using two sentiment lexicons for English, a bilingual dictionary and a link analysis algorithm. Hogenboom, Heerschop, Frasincar, Kaymak, and de Jong (2014) propose to project sentiment scores from the English SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010) to Dutch. In the same line, Cruz, Troyano, Pontes, and Ortega (2014) use MCR (Gonzalez-Agirre, Laparra, & Rigau, 2012) and EuroWordNet (Vossen, 1998) to transfer sentiment from the English SentiWordNet to the Spanish, Catalan, Galician and Basque WordNets.

Ghorbel and Jacot (2011) translate English SentiWordNet entries into French, finding that even if the translation is correct, in some cases two parallel words do not always share the same semantic orientation across both languages due to a difference in common usage. To deal with this issue, Volkova, Wilson, and Yarowsky (2013) propose to use crowdsourcing and bootstrapping for learning sentiment lexicons for English, Spanish and Russian from Twitter streams. Gao, Wei, Li, Liu, and Zhou (2013) found that the use of synonyms and word definitions does not improve the performance of their cotraining approach to learn a Chinese sentiment lexicon from existing sentiment lexicons for English and a corpus of parallel English-Chinese sentences. Chen and Skiena (2014) propose a method for building sentiment lexicons for 136 languages by integrating a variety of linguistic resources to produce a knowledge graph.