ARTICLE IN PRESS

Information Processing and Management xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

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

On the impact of emotions on author profiling

Francisco Rangel [a,b,*], Paolo Rosso [a]

[a] NLE Lab, Universitat Politècnica de València, Camino de Vera, S/N, Valencia, Spain
[b] Autoritas Consulting, C/ Lorenzo Solano Tendero 7, Madrid, Spain

ARTICLE INFO

Article history:
Received 15 May 2014
Received in revised form 27 May 2015
Accepted 1 June 2015
Available online xxxx

Keywords:
Affective processing
Author profiling
Emotion-labelled graphs
EmoGraphs

ABSTRACT

In this paper, we investigate the impact of emotions on author profiling, concretely identifying age and gender. Firstly, we propose the EmoGraph method for modelling the way people use the language to express themselves on the basis of an emotion-labelled graph. We apply this representation model for identifying gender and age in the Spanish partition of the PAN-AP-13 corpus, obtaining comparable results to the best performing systems of the PAN Lab of CLEF.

1. Introduction

Seth Godin[1] says "Marketing is no longer about the stuff that you make, but about the stories you tell". In this sense, world is rapidly changing, social media are growing on daily basis and customers are becoming users looking for new experiences. Thus the need of automatically processing the affective content of social media is acquiring a growing importance in order to know what the users want and need.

The potentiality offered by social media from many perspectives such as marketing, security or health is undeniable. But the information users include about themselves, if they include it, may lack of credibility. Age, gender, affiliation, likes...many users simply make them up. Getting to know the demographic and psychosocial profile of users on the basis of their writing style is an opportunity for organizations and companies, and a challenge for natural language processing technologies, due to the fact that the unique certainty we can have is what we can obtain from what the users write and share in social media.

Studies like Koppel, Argamon, and Shimoni (2003) link the use of the language with demographics such as the gender of the author, but the vast majority of such investigations are limited to English. This investigation presents a method for automatically identifying emotions in social media, from a different perspective. Our hypothesis is that the way users express their emotions about topics depends on their age and gender. We aim at modelling the way they express them on the basis of a graph-based approach. The main motivation for using a graph-based approach is its capacity to analyse complex language structures. Pennebaker's (2011) investigations are our main inspiration, where the style of writing is associated with personal attributes, such as demographics in our case. He employed a set of psycho-linguistic features obtained from texts,

* Corresponding author at: NLE Lab, Universitat Politècnica de València, Camino de Vera, S/N, Valencia, Spain.
E-mail addresses: francisco.rangel@autoritas.es (F. Rangel), prosso@dsic.upv.es (P. Rosso).
URLs: http://www.kicorangel.com (F. Rangel), http://www.dsic.upv.es/~prosso (P. Rosso).

[1] http://heidicohen.com/seth-godin-7-truths-at-the-heart-of-marketing-how-to-use-them.

http://dx.doi.org/10.1016/j.ipm.2015.06.003
0306-4573/

such as parts-of-speech, sentiment words and so forth. Our aim is to go further analysing the writing style from the perspective people combine the different parts-of-speech in a text, the kind of verbs they employ, the topics they mention, the emotions and sentiments they express (and where they do in the text), etc. Based on the differences between genders and among ages pointed out by Pennebaker (e.g. men used more prepositions than women because they tend to describe more in depth their environment), our intuition is that men will use more prepositional syntagmas than women, writing about different topics and with different emotionality, and this will confer a special importance to the sequence of *preposition + determinant + noun + adjective*. In this vein, we build a graph with the different parts-of-speech of user's texts and enrich it with semantic information with the topics they speak about, the type of verbs they use and the emotions they express. We model the whole text as only one single graph, considering also punctuation signs in order to capture how a writer may start and end sentences (i.e., how she connects her concepts in sentences). Once the graph is built, we obtain several properties from the graph and used them as features for a machine learning approach. Although the focus of this work is on Spanish, the proposed methodology may be applied to other languages.

The rest of the paper is structured as follows. In Section 2 we describe the related work on affective processing, author profiling and graph theory applied to text processing. In Section 3 we present our proposal for modelling the writing style to automatically identify emotions, age and gender (in Appendix A we analyse the complexity of the feature extraction). In Section 4 the evaluation framework is presented. Results are presented in Section 5 and discussed in Section 6, where we aim at identifying specific differences among age groups and gender with respect to used words, emotions and topics. In Section 7 we draw some conclusions.

2. Related work

The main objective of this investigation is to show how the emotions expressed by users help us to profile them. We propose a novel approach based on graph theory for modelling the way people express their emotions. Therefore, the related work should be seen from three different perspectives: affective processing, author profiling and graph-based text processing.

2.1. Affective processing

Automatic processing of affectivity has been focused mainly on sentiment analysis. However, there are a series of studies oriented to classify documents in the corresponding emotional category, usually based on the six basic emotions of Ekman (anger, disgust, joy, surprise, sadness, fear) (Ekman, 1972). At SemEval 2007 a task on the identification of emotions in news headline was organized. In Chaumartin (2007) the authors used the Stanford syntactic parser for identifying what the main topic was about, estimating the polarity of each word with the help of Senti Wordnet Esuli and Sebastiani (2006) and Wordnet Affect Strapparava and Valitutti (2004). In Kozareva, Navarro, Vazquez, and Montoyo (2007) the authors utilised three search engines for searching all the words in the headline combined with each emotion, and then calculated the Pointwise Mutual Information according to the number of returned documents. In Katz, Singleton, and Wicentowski (2007) the authors used a supervised system based on unigrams and trained with another 1,000 news manually annotated by them. They used the Roget thesaurus to expand synonyms and extract the features. In Strapparava and Mihalcea (2008) the results of SemEval are compared with other proposals, for instance with an approach where Latent Semantic Analysis (LSA) is employed to calculate similarity between a text and each of the six basic emotions.

In Dhaliwal et al. (2007) the authors used the identification of imperative sentences, exclamation signs, the use of capital letters or the use of present and future, in order to identify polarity and emotional category. In a similar way, in Sugimoto and Yoneyama (2006) the authors used nouns, adjectives and verbs with the identification of keywords and types of sentences in Japanese in order to identify emotions. In García and Alías (2008) the authors used the ANEW affective dictionary. In Rangel (2013) a method based on the Spanish Emotion Lexicon (SEL) is described for identifying emotions in short stories in Spanish. SEL consists of 2,036 words associated with the measure of "Probability Factor of Affective use" (PFA) related to one of the six basic emotions of Ekman (1972): joy, disgust, anger, fear, sadness, surprise. It defines four possible degrees of relationship with each emotion (null, low, medium, high) and 19 annotators indicated these values for each word. The PFA was calculated as an average of the percentages assigned to each degree. Finally, in Mohammad and Yang (2011) emotions and gender are investigated in three kind of emails: love letters; hate emails; and suicide notes.

2.2. Author profiling

The study of how certain linguistic features vary according to the profile of their authors is a subject of interest for several different areas such as psychology, linguistics and, more recently, natural language processing. In Pennebaker, Mehl, and Niederhoffer (2003) the authors related language use with personality traits, studying how the variation of linguistic characteristics in a text can provide information regarding the gender and age of its author. In Argamon, Koppel, Fine, and Shimoni (2003) the authors analysed formal written texts extracted from the British National Corpus, combining function words with part-of-speech features for gender prediction.

Please cite this article in press as: Rangel, F., & Rosso, P. On the impact of emotions on author profiling. *Information Processing and Management* (2015), http://dx.doi.org/10.1016/j.ipm.2015.06.003