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

Multi-lingual opinion mining on YouTube

Aliaksei Severyn [a], Alessandro Moschitti [c,a,*], Olga Uryupina [a], Barbara Plank [b], Katja Filippova [d]

[a] DISI, University of Trento, Italy
[b] CST, University of Copenhagen, Denmark
[c] Qatar Computing Research Institute, Qatar
[d] Google Inc., Switzerland

ARTICLE INFO

Article history:
Received 15 May 2014
Received in revised form 4 February 2015
Accepted 8 March 2015
Available online xxxx

Keywords:
Natural Language Processing
Opinion mining
Social media

ABSTRACT

In order to successfully apply opinion mining (OM) to the large amounts of user-generated content produced every day, we need robust models that can handle the noisy input well yet can easily be adapted to a new domain or language. We here focus on opinion mining for YouTube by (i) modeling classifiers that predict the type of a comment and its polarity, while distinguishing whether the polarity is directed towards the product or video; (ii) proposing a robust shallow syntactic structure (STRUCT) that adapts well when tested across domains; and (iii) evaluating the effectiveness on the proposed structure on two languages, English and Italian. We rely on tree kernels to automatically extract and learn features with better generalization power than traditionally used bag-of-word models. Our extensive empirical evaluation shows that (i) STRUCT outperforms the bag-of-words model both within the same domain (up to 2.6% and 3% of absolute improvement for Italian and English, respectively); (ii) it is particularly useful when tested across domains (up to more than 4% absolute improvement for both languages), especially when little training data is available (up to 10% absolute improvement) and (iii) the proposed structure is also effective in a lower-resource language scenario, where only less accurate linguistic processing tools are available.

1. Introduction

The increasing prevalence of social media like Twitter, Facebook and YouTube, which enable millions of users to share information and opinions quickly, has urged the need for new tools that robustly and automatically process the sheer amounts of user-generated content produced every day. Of particular importance is the fact that such information units, e.g., *tweets* in the case of Twitter or *comments* in case of YouTube, often carry *opinions* (or *sentiment*), i.e., they express subjective opinions of a particular user. In particular, we estimated that roughly 60–80% of the YouTube comments do actually contain opinions. Therefore, social media provide a key source that raises the importance of automatic extraction of opinions affecting the reputation of a person, and organization, or a specific product.

In this study we focus on YouTube. It is a platform that hosts videos uploaded by users (companies, private persons, etc.). YouTube is a unique environment with many facets: it is multi-modal, multi-lingual, multi-domain and multi-cultural, since people from different regions of the world can upload videos including textual information about different topics, they can

[*] Corresponding author at: Qatar Computing Research Institute, Qatar.
E-mail addresses: severyn@disi.unitn.it (A. Severyn), amoschitti@qf.org.qa (A. Moschitti), uryupina@gmail.com (O. Uryupina), bplank@cst.dk (B. Plank), katjaf@google.com (K. Filippova).

http://dx.doi.org/10.1016/j.ipm.2015.03.002
0306-4573/© 2015 Elsevier Ltd. All rights reserved.

rate videos, and comments on videos in different languages. Therefore, work promoting the success of sentiment analysis systems in such an environment is of high interest for both the industry and the research community.

Most prior research on opinion mining has been carried out on well-edited texts (Pang & Lee, 2008), and there has been some recent effort for sentiment analysis on Twitter (Nakov et al., 2013). In contrast, YouTube comprises several new challenges, which need to be tackled: (i) words expressing polarity can refer to either the video content itself ("the girl is cute") or the advertised product ("I hate the G2"), or may even contain contrasting sentiment; (ii) many comments are unrelated and are spam ("go to http to win an iPad"); (iii) YouTube is a large resource covering a variety of domains, thus it is not clear how well a supervised system trained on, say, the *tablets* domain, fares on a different domain, e.g., videos about *automobiles*; and (iv) it covers a large variety of languages, both in terms of the video and the comments for a certain video, thus approaches that handle the multilinguality aspect are of particular interest.

Still, the majority of systems for sentiment analysis rely on the simple bag-of-words (BOW) representation. That is, the input text is split into *n*-grams of words (or characters). These are used in machine learning algorithms, e.g., Support Vector Machines (SVM) or logistic regression, to induce a model that can classify new instances. In fact, the winning system (Mohammad, S, Kiritchenko, & Zhu, 2013) of the SemEval 2013 shared task (Nakov et al., 2013) used a BOW representation together with a sentiment lexicon in SVMs. However, opinions usually involve complex interactions between lexical items (e.g., variations in sentiment scope and target, modality and negations, etc.). The standard BOW representation cannot take those into consideration, since by definition it abstracts away from many important clues. For example, consider a comment from a YouTube video, where a person reviews a specific product, namely the *Motorola xoom* tablet:

> *this guy really puts a negative spin on this, and I'm not sure why, this seems crazy fast, and I'm not entirely sure why his pinch to zoom his laggy all the other* **xoom** *reviews*

The comment contains the product name (***xoom***) and a list of negative expressions, thus, a bag-of-words model would derive a negative polarity for this product. In contrast, the opinion towards the product is neutral as the negative sentiment is expressed towards the video. Similarly, the following Italian comment on an **iPad** video expresses positive sentiment about another product (*galaxy note*), but is neutral with respect to the topical product.

> *Questo video fa un presentazione interessante dell' iPad, ma ho preso il galaxy note:) ha uno schermo fantastico. veramente bello e fluido. te lo consiglio.* (This video gives a nice introduction of iPad but I took the galaxy note:) it has a fantastic screen. very nice and fluent. I recommend it).

The following short English comment illustrates the main problem even better:

> *iPad 2 is better. the superior apps just destroy the* ***xoom***.

It contains two positive and one negative word, yet the sentiment towards the product is negative (the negative token *destroy* refers to *Xoom*). Thus, it is important to distinguish if the sentiment on YouTube is directed either towards the source video itself, or the product described in that video or another product. This cannot be captured by a bag-of-words model, which lacks the needed structural information for linking the sentiment with the target product.

In this paper, we present the results of the first research effort on the systematic analysis of opinion mining (OM) for YouTube comments capitalizing on our previous work Uryupina, Plank, Severyn, Rotondi, and Moschitti (2014) and Severyn, Moschitti, Uryupina, Plank, and Filippova (2014). The contributions of our research are:

1. *User comment type and polarity classification*: to solve the issues outlined above, we devise a classification scheme that separates unrelated and spam comments from informative ones, which are, in turn, further categorized into product- or video-related (type classification). Moreover, we learn classifiers to assign polarity (positive, negative, neutral) to each type of informative comment. This allows us to filter out irrelevant comments, providing accurate OM distinguishing comments about the video and the target product.
2. *A novel structural representation*, based on shallow syntactic trees enriched with conceptual information, i.e., tags generalizing the specific topic of the video, e.g., *Fiat Panda*, *xoom*, *Toyota Aygo*. In particular, we define an efficient tree kernel derived from the Partial Tree Kernel (Moschitti, 2006), suitable for encoding structural representation of noisy user-generated comments into Support Vector Machines (SVMs).
3. *Creation and annotation of a corpus of YouTube comments*: it contains 50k manually labeled (by an expert coder) comments for two product domains: *tablets* and *automobiles*.[1] It is the first manually annotated corpus that enables researchers to use supervised methods on YouTube for comment classification and opinion analysis. The comments from different product domains exhibit different properties (cf. Section 5.2), which give the possibility to study the domain adaptability of the supervised models by training on one category and testing on the other (and vice versa).
4. *Multi-lingual experiments*: in contrast to our and other prior work focused exclusively on one language (mainly, English), we show that our structural representation also works well for a less-resourced language, namely, Italian. This is of particular interest since it tests the proposed representation under limiting conditions: the performance of linguistic

[1] The corpus and the annotation guidelines are publicly available at: http://projects.disi.unitn.it/iKernels/projects/sentube/.

Please cite this article in press as: Severyn, A., et al. Multi-lingual opinion mining on YouTube. *Information Processing and Management* (2015), http://dx.doi.org/10.1016/j.ipm.2015.03.002