Identifying Entity Aspects in Microblog Posts

Damiano Spina
UNED NLP & IR Group
damiano@lsi.uned.es

Andrei Oghina
ISLA, University of Amsterdam
oghina@science.uva.nl

Edgar Meij
ISLA, University of Amsterdam
edgar.meij@uva.nl

Maarten de Rijke
ISLA, University of Amsterdam
derijke@uva.nl

Minh Thuong Bui
ISLA, University of Amsterdam
mbui@science.uva.nl

Mathias Breuss
ISLA, University of Amsterdam
mbreuss@science.uva.nl

ABSTRACT

Online reputation management is about monitoring and handling the public image of entities (such as companies) on the Web. An important task in this area is identifying aspects of the entity of interest (such as products, services, competitors, key people, etc.) given a stream of microblog posts referring to the entity. In this paper we compare different IR techniques and opinion target identification methods for automatically identifying aspects and find that (i) simple statistical methods such as TF.IDF are a strong baseline for the task, significantly outperforming opinion-oriented methods, and (ii) only considering terms tagged as nouns improves the results for all the methods analyzed.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

Keywords

Microblog posts, entity profiling, aspects

1. INTRODUCTION

Online reputation management (ORM) deals with monitoring and handling the public image of entities such as people, products, organizations, companies, or brands on the web. In the field of ORM, much of the effort is focused on analyzing mentions on social web streams (such as tweets) that are relevant to the entity of interest. An important task in this area is to identify not only posts that are relevant for a given entity, but also the specific aspects that people discuss.

Aspects refer to “hot” topics that people talk about in the context of an entity and are of particular interest for companies. Aspects can cover a wide range of issues and include (but are not limited to) company products, key people, other entities, services, and events. They are typically nouns, but can also be verbs, and (rarely) adjectives. They can change over time as public attention shifts from some aspects to others. For instance, for a company releasing its quarterly earnings report, its earnings can become a topic of discussion for a certain period of time and, hence, an aspect. Identifying aspects not only helps reputation analysts in determining what people say about an entity of interest, but it also facilitates a more fine-grained sentiment analysis than is typically possible, since opinions pertaining to aspects rather than to the entity can be identified [4]. Although aspects have been investigated in the context of, e.g., dis-

Copyright is held by the author/owner.
SIGIR’12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

1 We only consider unigrams. When a unigram is an obvious constituent of a larger, relevant aspect it is considered relevant.
2 Available at http://bit.ly/profilingTwitter
Table 1: Scoring functions for identifying entity aspects.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scoring function</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>$s(t, D, C) = t f(t, D) \cdot \log \frac{N}{df(t)}$</td>
</tr>
<tr>
<td>LLR</td>
<td>$s(t, D, C) = 2 \cdot \frac{(a \cdot \log \left( \frac{a}{E_1} \right) + (b \cdot \log \left( \frac{b}{E_2} \right))}{a+b}$</td>
</tr>
<tr>
<td>PLM</td>
<td>$s(t, D, C) = P(t</td>
</tr>
<tr>
<td>OO</td>
<td>$s(t, D, C) = \chi^2(\text{target}(t, D), \text{target}(t, C))$ $\chi^2(o, e) = (o-e)^2$ $\chi^2(\text{target}(t, D), \text{target}(t, C)) = \text{freq. of potential target } t \text{ \ in \ tweets } D$ $\text{target}(t, C) = \text{freq. of potential target } t \text{ \ in \ background } C$</td>
</tr>
</tbody>
</table>

We lowercase, remove punctuation, and tokenize the tweets. We do not perform stopword removal or stemming, but only keep terms occurring at least 5 times in the corpus to remove noisy terms.

We evaluate the methods for ranking aspects using a pooling methodology [10]; the 10 highest ranked terms from each method are merged and randomized. Then, three human assessors consider each term and determine relevance in the context of the company: relevant aspects can include terms from compound words, mentions, or hashtags and should provide insight into the hot topics discussed regarding a company. We compute the inter-annotator agreement using both Cohen’s and Fleiss’ kappa and compare the annotators’ pairwise and overall. All obtained kappa values are above 0.6, indicating a substantial agreement.

Table 2 (upper part) shows the results of all methods for identifying aspects. Since TF-IDF is the simplest approach, it is considered as the baseline. We use Student’s t-test to test for statistical significance and indicate a significant difference with $\alpha = 0.01$ using \* (or \*) and \* (or \*) for $\alpha = 0.05$.

First, we observe that TF-IDF is a strong baseline. In terms of precision, it significantly outperforms PLM and OO, while differences between TF-IDF and LLR are not significant. The results for OO are much lower than for the other methods. Since terms that are (part of) the name of the entity were also annotated as aspects, and these terms are very frequent in the tweets related to the entity, they are often in the top of the ranking returned by the methods. This explains the high MRR values in the results.

When manually inspecting the results, we observe that the results for the frequency-based methods (TF-IDF, LLR and PLM) are very similar, while OO tends to return more subjective terms as aspects (e.g., hahaha, pls, xd, safety, win), probably because of errors in the syntactic parsing of tweets. Moreover, this approach has more difficulty to filter out generic terms (e.g., new, use, today, come).

Most of the true aspects are nouns (89.72\%); Hence, in addition to the preprocessing steps detailed above, we experiment with applying a part-of-speech filter and only consider terms tagged as nouns (Penn Treebank’s N\* tags) [6]. Table 2 (lower part) shows the results when non-noun terms have been filtered out from the vocabulary. For all methods, MAP and precision values are slightly higher than in the all words condition: considering only nouns helps to identify aspects. Interestingly, the relative order of the approaches (as determined by the scores they achieve) changes with respect to the upper part. PLM now outperforms TF-IDF for two of the four metrics (significantly so for P10).

4. CONCLUSION

We addressed the task of identifying aspects that people discuss in a stream of microblog posts related to an entity, a task at the heart of online reputation management. We modeled this task as a ranking problem and compared IR techniques and opinion target identification methods for automatically identifying aspects. We used a pooling methodology to evaluate the methods. Simple statistical methods such as TF-IDF are a strong baseline for the task. Moreover, it is difficult to identify aspects by extracting opinion targets mainly because the language used in tweets is often non-standard, hampering the performance of such techniques. Future work includes considering n-grams as aspects and applying topic modeling techniques.

5. ACKNOWLEDGMENTS

This research was supported by the European Union’s CIP ICT- PSP under grant agreement nr 250430, the European Community’s FP7 Programme under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSiNE), the Netherlands Organisation for Scientific Research (NWO) under project nr 612.-061.814, 612.061.815, 640.004.802, 380-70-011, 727.011.005, the Center for Creation, Content and Technology (CCCT), the Hyper-local Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, under COMMIT project Infiniti, the Spanish Ministry of Education (FPU grant nr AP2009-0507), the Spanish Ministry of Science and Innovation (Holopedia Project, TIN2010-21128-C02), the Regional Government of Madrid and the ESF under MA2VICMR (S2009/TIC-1542) and the ESF Research Network Program ELIAS.

References