Identifying Entity Aspects in Microblog Posts

Spina, D.; Meij, E.; de Rijke, M.; Oghina, A.; Bui, M.T.; Breuss, M.

DOI
10.1145/2348283.2348483

Publication date
2012

Document Version
Final published version

Published in
SIGIR'12: the proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval: August 12-16, 2012: Portland, Oregon, USA

Identifying Entity Aspects in Microblog Posts

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

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

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

Andrei Oghina
ISLA, University of Amsterdam
oghina@science.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., discussion fora [9], automatically determining aspects in streams of microblog posts remains an unsolved problem.

We study the following scenario. Given a stream of microblog posts related to a company [1, 7], we are interested in a ranked list of aspects that are being discussed with respect to the company. We formulate our scenario as an information retrieval (IR) task, where the goal is to provide a ranking of terms, extracted from tweets that are relevant to the company.[1] We compare different methods that address this task with three main goals: (i) to analyze how state-of-the art IR approaches perform, (ii) to see how methods tailored specifically to identifying opinion targets perform, and (iii) to create a publicly available, humanly annotated dataset to facilitate follow-up research [8].[2]

2. IDENTIFYING ENTITY ASPECTS

We evaluate four models for identifying aspects, given an entity and a stream of microblog posts related to that entity. All models work according to the same principle: comparing a pseudo-document D built from entity-specific tweets with a background corpus C. This comparison allows us to score a term t using a function $s(t, D, C)$.

We compare four methods for identifying entity aspects: TF.IDF, the log-likelihood ratio (LLR) [2], parsimonious language models (PLM) [3] and an opinion-oriented method (OO) [5] that extracts targets of opinions to generate a topic-specific sentiment lexicon; we use the targets selected during the second step of this method. Table 1 describes how the scoring function is computed by each method. As usual, $tf(t, D)$ denotes the term frequency of term t in pseudo-document D; $cf(t)$ denotes the term frequency in the collection C and $df(t)$ denotes the total number of pseudo-documents $D_i \in C$ in which the term t occurs at least once.

3. EXPERIMENTS

Determining aspects of an entity in streams of microblog posts involves two tasks. In the first task, tweets relevant to a given entity need to be identified; in the second, these tweets need to be analyzed in order to identify aspects. We focus on the second task and base our annotations on the data used for the WePS-3 ORM Task [1]. Here, the task that participating systems need to solve is to decide which tweets containing a company name are actually related to the company. In total, 99 companies are used for testing, with around 450 tweets (manually annotated for relevance) on average for each company. In our experiments we only consider the tweets that are related to a company, adding up a total of 94 companies and 17,775 tweets with an average of 177 tweets per company.

[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`

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

Table 1: Scoring functions for identifying entity aspects.

Method	Scoring function
TF.IDF	$s(t,D,C) = tf(t,D) \cdot \log \frac{N}{df(t)}$ N = number of pseudo-documents D_i in C $df(t) = \sum_i^N tf(t,D_i) > 0, D_i \in C$
LLR	$s(t,D,C) = 2 \cdot ((a \cdot \log(\frac{a}{E_1})) + (b \cdot \log(\frac{b}{E_2})))$ $E_1 = \frac{c\cdot(a+b)}{c+d}$ $E_2 = \frac{d\cdot(a+b)}{c+d}$ $a = tf(t,D)$ $b = cf(t)$ $c = \sum_i tf(t_i,D)$ $d = \sum_i cf(t_i)$
PLM	$s(t,D,C) = P(t\|D)$ (when model converges) E-step: $e_t = tf(t,D) \cdot \frac{\lambda\cdot P(t\|D)}{(1-\lambda)\cdot P(t\|C)+\lambda\cdot P(t\|D)}, \lambda = 0.1$ M-step: $P(t\|D) = \frac{e_t}{\sum_t e_t}$ initial $P(t\|D) = \frac{tf(t,D)}{\sum_i tf(t_i,D)}$ $P(t\|C) = \frac{cf(t)}{\sum_i cf(t_i)}$
OO	$s(t,D,C) = \chi^2(target(t,D), target(t,C))$ $\chi^2(o,e) = \frac{(o-e)^2}{e}$ $target(t,D)$ = freq. of potential target t in tweets D $target(t,C)$ = freq. of potential target t in background C

We lowercase, remove punctuation, and tokenize the tweets. We do not perform stopword removal or stemming, but only keep terms occuring 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., *haha*, *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).

Table 2: Aspect identification results. Best results in boldface; significant changes are w.r.t. the TF.IDF All words baseline.

	Method	MAP	P5	P10	MRR
All words	TF.IDF	**0.3953**	**0.6957**	**0.6426**	0.7908
	LLR	0.3879	**0.6957**	0.6309	**0.7979**
	PLM	0.3685▼	0.6723▽	0.6096▼	**0.7979**
	OO	0.1537▼	0.4596▼	0.2915▼	0.7021
Noun filter	TF.IDF	0.4015	**0.7213**	0.6436	**0.7979**
	LLR	0.4055	0.7128	0.6511	**0.7979**
	PLM	**0.4097**	0.7106	**0.6617**△	**0.7979**
	OO	0.1635▼	0.4809▼	0.3000▼	0.7021

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 nrs 612.-061.814, 612.061.815, 640.004.802, 380-70-011, 727.011.005, the Center for Creation, Content and Technology (CCCT), the Hyperlocal 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

[1] E. Amigó, J. Artiles, J. Gonzalo, D. Spina, B. Liu, and A. Corujo. WePS-3 evaluation campaign: Overview of the online reputation management task. In *CLEF '10*, 2010.

[2] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19, 1993.

[3] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *SIGIR '04*, 2004.

[4] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent twitter sentiment classification. In *ACL '11*, 2011.

[5] V. Jijkoun, M. de Rijke, and W. Weerkamp. Generating focused topic-specific sentiment lexicons. In *ACL '10*, 2010.

[6] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *ACL '03*, 2003.

[7] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *WSDM '12*, 2012.

[8] D. Spina, E. Meij, A. Oghina, M. T. Bui, M. Breuss, and M. de Rijke. A Corpus for Entity Profiling in Microblog Posts. In *LREC Workshop on Language Engineering for Online Reputation Management*, 2012.

[9] T. Thet, J. Na, and C. Khoo. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Inf. Sci.*, 36(6):823, 2010.

[10] E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005.