A Corpus for Entity Profiling in Microblog Posts

Damiano Spina*, Edgar Meij†, Andrei Oghina†, Minh Thuong Bui†, Mathias Breuss† , Maarten de Rijke†

* UNED NLP & IR Group
Juan del Rosal, 16
28040 Madrid, Spain
damiano@lsi.uned.es

†ISLA, University of Amsterdam
Science Park 904
1098 XH, Amsterdam, The Netherlands
edgar.meij@uva.nl, {oghina,mbui,mbreuss}@science.uva.nl, derijke@uva.nl

Abstract

Microblogs have become an invaluable source of information for the purpose of online reputation management. Streams of microblogs are of great value because of their direct and real-time nature. An emerging problem is to identify not only microblog posts (such as tweets) that are relevant for a given entity, but also the specific aspects that people discuss. Determining such aspects can be non-trivial because of creative language usage, the highly contextualized and informal nature of microblog posts, and the limited length of this form of communication. In this paper we present two manually annotated corpora to evaluate the task of identifying aspects on Twitter, both of them based upon the WePS-3 ORM task dataset and made available online. The first is created using a pooling methodology, for which we have implemented various methods for automatically extracting aspects from tweets that are relevant for an entity. Human assessors have labeled each of the candidates as being relevant. The second corpus is more fine-grained and contains opinion targets. Here, annotators consider individual tweets related to an entity and manually identify whether the tweet is opinionated and, if so, which part of the tweet is subjective and what the target of the sentiment is, if any.

1. Introduction

Online Reputation Management (ORM) deals with monitoring and handling the public image of entities, such as people, products, organizations, or brands, on the Web. In the field of ORM, much of the effort is focused towards analyzing mentions on social web streams that are relevant to the entity of interest. An emerging problem in this area is to identify not only microblog posts (such as tweets) 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—the principal vectors that coagulate the public interest regarding the company. Aspects can cover a wide range of notions and they include, without being limited to, company products, services, key people, and events. They can change over time as public attention shifts from some aspects to others. For instance, when a company releases its quarterly earnings report, this can become, for a certain period of time, a topic of discussion and, hence, an aspect. Although aspects have been investigated in the context of, e.g., discussion fora (Thet et al., 2010), automatically determining aspects on streams of microblog posts is still an unsolved problem.

A well-known application in the context of ORM on social web streams is sentiment analysis (Jansen et al., 2009), with numerous online demos and tools. Since state-of-the-art methods for sentiment analysis still yield noisy results, it is common to measure aggregate sentiments, i.e., aggregating sentiment scores for a set of microblog posts. While measuring such "overall" sentiment has its merits, it also has obvious limitations. Especially in the context of entities such as large companies—which typically have many products or services to offer—a more fine-grained approach is needed.

Some current ORM tools such as UberVU[1] allow online reputation managers to monitor sentiment regarding a predefined set of keywords, such as product names (Amigó et al., 2010). However, the fluidity of microblogging streams renders this method too rigid, since aspects can have a dynamic nature, changing and emerging over time. Therefore, a better approach would be to extract the relevant, most discussed aspects of an entity in an automatic fashion.

To the best of our knowledge, there are no readily available datasets suitable to evaluate the task of identifying either aspects or opinion targets in the context of ORM on social web streams. In this paper we present two manually annotated corpora to fill this gap. Both of them are based upon the WePS-3 ORM task and will be made available online.[2]

The first dataset is created using a pooling methodology. Here, we have implemented various methods for automatically extracting aspects from tweets that are relevant for an entity. We subsequently generate a ranked list of aspects using each method, take the highest ranked aspects, and pool them. Then, human assessors consider each aspect and determine whether it is relevant in the context of the entity or not. The second dataset that we present is similar, but more fine-grained. Here, annotators consider individual tweets related to an entity and manually identify whether the tweet is opinionated and, if so, which part of the tweet is (i) sub-

[1] http://www.ubervu.com/walkthrough/

[2] http://nlp.uned.es/~damiano/datasets/entityProfiling_ORM_Twitter.html

jective and (ii) what the target of the sentiment is, if any.
In the next section, we briefly discuss related work and datasets. In Section 3. we describe the WePS-3 ORM task dataset, upon which our annotated corpora are based. In Sections 4. and 5. we introduce the corpus containing the entity aspects and the one containing the opinions, respectively. Section 6. briefly compares the two corpora, including an analysis of the overlap between them. We end with a concluding section.

2. Related Work

In other domains—such as product reviews or news—there exist various datasets to investigate aspects, typically in the form of opinion targets (Hu and Liu, 2004; Kim and Hovy, 2006; Wiebe et al., 2005). However, to the best of our knowledge, there are no manually annotated corpora to evaluate this task on microblog streams. Determining such aspects on streams of microblog posts can be non-trivial because of the creative language usage (including slang, emoticons, and acronyms), the highly contextualized and informal nature of microblog posts, and the limited length of this form of communication (Kaufmann and Kalita, 2010). This reduces the applicability of the techniques developed for other domains. Moreover, the amount of data produced on microblogging streams is substantially larger than that produced in customer reviews or news media, opening up opportunities for leveraging cross-post redundancy.
So far, most of the manually annotated corpora built upon Twitter are annotated at the level of individual tweets. For example, both the TwitterSentiment[3] and Sanders[4] corpora contain tweets labeled with subjectivity and polarity (i.e. positive, negative, and neutral).
In the TREC 2011 Microblog track,[5] the gold standard for the ad hoc real-time search task was built using a pooling methodology. The corpus used in this task was the Tweets2011 corpus[6]. Another recently released Twitter dataset contains semantic annotations, where each tweet is manually linked to a set of entities in the form of Wikipedia articles (Meij et al., 2012). Similarly, the WePS-3 ORM dataset links tweets to companies, as described in the next section.

3. WePS-3 ORM

Determining aspects of an entity in the context of streams of microblog posts such as tweets involves two tasks. In the first task, tweets relevant to a given entity need to be identified, while in the second these tweets need to be analyzed in order to identify aspects. In this paper we focus mainly on the second task and base our annotations on the data used for the WePS-3 ORM Task (Amigó et al., 2010). Here, the task that participating systems needed to solve was to filter tweets containing a given company name depending on whether the post is actually related to the company or not. This is challenging for ambiguous names, such as *Apple* or *Fox*. In total, 99 companies were used, with around 450 tweets on average for each, summing up to a total of 45,201 tweets. Mechanical Turk was used to perform the relevance assessments; each tweet is annotated as being either *related* or *unrelated* to a given company.
For the annotations presented in this work, only the tweets that are related to each company are considered. For our first dataset pertaining to the identification of aspects, a total of 94 companies have been considered. This adds up to 17,775 tweets in total, with an average of 177 tweets per company. From this set, all the related tweets for 59 companies have been annotated in a second round, where we identify opinion targets and subjective phrases. The latter corpus constitutes our second dataset and includes 9,396 tweets in total, i.e., an average of 159 tweets per company.

4. Annotating Aspects

Let us consider the following profiling scenario: given a stream of tweets that are related to a company, we are interested in a ranked list of aspects representing the hot topics that are being discussed with respect to the company. Examples of aspects include products, services, key people, events, or entities that are associated with the company in a certain time frame.
This scenario can be formulated as an information retrieval task, where the goal of a system implementing a solution to this task is to provide a ranking of terms, extracted from tweets that are relevant with respect to the company.[7] We have implemented various methods addressing this task. For each company, each method returns a ranked list of terms associated with each company. The underlying principle for all methods is a comparison of the contents of the relevant tweets—henceforward, the *foreground* corpus—with a common *background* corpus, e.g., the whole WePS-3 collection. Using this comparison we identify and score terms based on their relative occurrence. Our methods include TF.IDF (Salton and Buckley, 1988), the log-likelihood ratio (Dunning, 1993) and parsimonious language models (Hiemstra et al., 2004). Since aspects can be opinion targets, we also applied an opinion-oriented method (Jijkoun et al., 2010) that extracts potential targets of opinions to generate a topic-specific sentiment lexicon. We use the targets selected during the second step of this method.
This dataset is then created using a pooling methodology (Harman, 1995): the 10 highest ranking terms from each method are merged and randomized. Then, human assessors consider each term and determine whether it is relevant in the context of the company or not.

4.1. Annotations

The annotators were presented with an annotation interface, where they could select one of the companies from a list. Once a company is selected, the interface shows a randomized list of aspects. The interface also facilitated looking up

[3] http://twittersentiment.appspot.com
[4] http://www.sananalytics.com/lab/twitter-sentiment/
[5] http://sites.google.com/site/microblogtrack/2011-guidelines/
[6] http://trec.nist.gov/data/tweets/
[7] In our current setup, we only consider unigrams as aspects. When a unigram is an obvious constituent of a larger, relevant aspect, it is considered relevant.

a term; when clicked, the system would present all tweets that are relevant to the company and contain that particular term. The annotators could indicate one of the following labels for each aspect:

- **Relevant:** A relevant aspect can include, e.g., product names, key people, events, etc. Relevant aspects are in general nouns, but can also be verbs, and (rarely) adjectives. Relevant aspects can include terms from compound words, mentions or hashtags. Aspects should provide some insight into the hot topics discussed regarding a company, topics that would also differentiate it from other more general discussions, or its competitors.
- **Not relevant:** Common words and words not representing aspects or sub-topics are not relevant.
- **Competitor:** A term is (part of) a competitor name, including an opponent team name, a competing company or a product from a competing company.
- **Unknown:** If, even after inspecting the tweets were the term occurs, the judge still cannot use the other labels.

In this work we treat the label *Competitor* as being *Relevant*, although the data set contains this explicit label for possible follow-up work. Table 1 shows some examples of the aspects annotated in the corpus.

Entity	Aspects
A.C. Milan	milanello, ac, football, milan, galliani, berlusconi, brocchi, leonardo
Apple Inc.	ipad, iphone, prototype, apple, store, gizmodo, employee, gb
Sony	advertising, set, headphones, digital, pro, music, sony, xperia, dsc, x10, bravia, camera, vegas, battery, ericsson, playstation
Starbucks	coffee, latte, tea, frappuccino, starbucks, shift, pilot, barista, drink, mocha

Table 1: Examples of aspects annotated for some of the entities in the corpus.

4.2. Analysis

In order to determine the level of agreement between the three annotators J_i, we calculate *Cohen's kappa* and *Fleiss' kappa* (Landis and Koch, 1977) and compare the annotators both pairwise and overall. The results are given in table 2. All of the obtained kappa values are above 0.6, which indicates a substantial agreement.

Method	J_1-J_2	J_1-J_3	J_2-J_3	All
Cohen's κ	0.691	0.62	0.676	-
Fleiss' κ	0.69	0.62	0.676	0.662

Table 2: Inter-annotator agreement for the aspects dataset.

In the WePS-3 ORM dataset, the number of tweets relevant to each company is highly variable (Amigó et al., 2010). Thus, one could expect correlations between the ratio of relevant tweets and the ratio of relevant aspects annotated for each company.

Tweets	|C|	AvgTw	AvgTer	AvgRel	Rel%
0-10	19	4.05	12.47	2.79	22.36%
11-50	15	22.20	22.00	8.53	38.79%
51-150	12	97.67	26.75	13.58	50.78%
151-300	25	219.40	28.80	16.40	56.94%
301+	28	381.43	30.64	19.46	63.52%

Table 3: Distribution of relevant aspects, binned by the number of relevant tweets per company.

Table 3 shows the number of tweets, the number of extracted terms (*AvgTer*), and the number of identified relevant aspects (*AvgRel*) based on the annotations. For this, we consider all terms included in the pooling, and divide the entities in five groups, based on the number of tweets available for each company (0-10, 11-50, 51-150, 151-300, 301+). For each group C, we count how many companies are part of the group ($|C|$) and the average number of tweets for these entities (*AvgTw*). We also compute the percentage of the aspects that are relevant (*Rel%*).

We observe that the percentage of relevant aspects across increases with the amount of data available. For companies that have no more than 10 tweets each, only 22.36% of extracted aspects are annotated as being relevant. On the other hand, for entities with more than 300 tweets, 63.52% of all extracted aspects were annotated as being relevant. This suggests that the amount of data available plays an important role in the performance of the methods used for the pooling.

5. Annotating opinion targets

The second dataset we present consists of the tweets of 59 entities from the WePS-3 dataset, manually annotated at the phrase-level. Here, we aim to identify opinion targets in tweets, related to an aspect of a company. We define an opinion target as a phrase p that satisfies the following properties: (i) p is an aspect of the entity, (ii) p is included in a sentence that contains a direct subjective phrase (i.e. an expression that explicitly manifests subjectivity or an opinion) and (iii) p is the target of the expressed opinion.

5.1. Annotations guidelines

The annotators were asked to indicate the following.

- **Subjectivity:** Tweet-level annotation that indicates whether the tweet contains an explicit opinionated expression.
- **Subjective phrase:** If the tweet is opinionated, identify the phrase that express subjectivity. In our annotation schema, we only considered direct private states (Wiebe et al., 2005).
- **Opinion target:** If the tweet contains opinionated phrases, identify the target of the opinion expressed in that phrases.

Table 4 show some examples of opinionated tweets.

Phrase-level annotation require much more effort than tweet-level annotations or aspect assessments. In order to maximize the number of annotated entities, 59 entities were randomly distributed over seven different annotators, making a disjoint assignment of annotators to data.

Entity	Tweet
Linux	Lxer: A Slimline Debian Install: Its Easier Than You Might Think: There are some *superb* **desktop Linux distributions**... http://bit.ly/8ZSaF
MTV	@MTV has the *best* **shows** ever. i watch it all day every day (:
Oracle	IMHO, the *best part of* Oracle now **owning Java** is that whenever **Java** is *criticized* for something, Oracles name is attached.
Sony	@user Welll Im not getting one then. **Sony** is *expensive*
Starbucks	**The Dark Cherry Mocha** from @Starbucks is just *the best Mocha ever!*

Table 4: Examples of phrase-level annotated tweets, having subjective phrases (italic) and opinion targets (boldface).

5.2. Analysis

In total, 9,396 tweets were annotated. Only 1,427 (15.16%) tweets contain subjective phrases and 1,308 (13.82%) contain opinion targets. There are 119 tweets where the annotators identified subjective phrases but not opinion targets. Most of them are tweets containing either emoticons or phrases expressing subjectivity at tweet-level (e.g. `LOL`, `Yay!`, `#fail`).

Analogous to the first dataset, we divided the annotated entities in groups based on the number of annotated tweets and computed the average of tweets with subjective phrases (*AvgSubj*) and opinion targets (*AvgOT*). Table 5 reports these averages as well as the averaged percentage of subjective tweets (*Subj%*).

Tweets	|C|	AvgTw	AvgSubj	AvgOT	Subj%
0-10	7	3.57	0.85	0.85	35.11%
11-50	11	23.36	3.64	3.09	14.24%
51-150	9	96.22	11.77	10.33	11.88%
151-300	19	218.68	25.21	23.10	14.22%
301+	13	392.54	61.23	56.61	15.8%

Table 5: Distribution of subjective phrases and opinion targets, binned by the number of relevant tweets per company.

6. Aspects vs. Opinion targets

In this section we analyze the vocabulary overlap between the terms identified in the two corpora presented in this paper, i.e., between aspects and opinion target terms.

For the first dataset we consider a majority vote, labeling terms as relevant when they are annotated as such by two or more judges. We further restrict ourselves to the same 59 entities annotated with opinion targets in the second dataset. We tokenize the phrases identified as opinion targets, keeping the constituent terms that occur in them after removing stopwords and symbols. As an example, Table 6 shows opinionated aspects for some of the entities in the datasets.

From a total of 783 aspects, 209 (26.69%) occur in opinion target phrases. Vice versa, the total number of terms extracted from the opinion target phrases is 1650; only 12.66% of those are also identified as relevant aspects. The overlap between aspects and opinion targets is lower than expected. The low overlap is probably given by the different methodologies used to annotate aspects and opinion targets. While aspects were annotated using a pooling methodology that considers the 10 highest ranking terms retrieved from each method, opinion targets were manually annotated inspecting the tweets related to each company.

We observe that, instead of an aspect, the actual name of the entity has a tendency to occur as a target. However, the remaining aspects occur only a few times, suggesting a power-law distribution. In fact, terms in opinion targets are very sparse. The average occurrence of a term in an opinion target equals 1.78 and more than 75% of all terms occur only once. This suggests that the WePS-based sample of around 150 tweets per entity might not be enough for opinion-based entity profiling. We leave verifying this hypothesis (and possibly creating a larger dataset) for future work.

Entity	Aspects in opinion targets
Jaguar Cars Ltd.	jaguar (0.26), xj (0.06), cars (0.02), rover (0.01), car (0.01), auto (0.01), xf (0.01)
Linux	linux (0.12), multitouch (0.02)
Sony	sony (0.05), music (0.04), vegas (0.03), headphones (0.02), battery (0.02), xperia (0.01), pro (0.01), ericsson (0.01), x10 (0.01), playstation (0.01), bravia (0.01), camera (0.01)
Starbucks	starbucks (0.33), coffee (0.11), tea (0.06), frappuccino (0.03), drink (0.03), latte (0.02)

Table 6: Examples of aspects that are included in opinion target phrases, with the frequency in opinion targets in parentheses.

7. Conclusions

An emerging problem in the field of online reputation management consists of identifying the key aspects of an entity commented in microblog posts. Streams of microblogs are of great value because of their direct and real-time nature and synthesizing them in form of entity profiles facilitates reputation managers to keep a track of the public image of the entity.

In this paper we have presented two manually annotated corpora to evaluate the task of identifying aspects on Twitter, both of them based upon the WePS-3 ORM task dataset and made available online. The first dataset we release contains aspects that are strongly related to a given company in a stream of tweets, while the second contains phrases in tweets that represent the targets and opinions expressed towards entities in those tweets. The low overlap between relevant aspects and terms occurring in opinion target phrases shows the different nature of the two corpora built. We believe that these resources will allow to evaluate different entity profiling systems in microblog posts and to make progress in the use of human language technologies for online reputation management.

8. Acknowledgements

This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the European Community's Seventh Framework Programme (FP7/ 2007-2013) under grant agreements nr 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe project), 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 Education Council of the Regional Government of Madrid, MA2VICMR (S-2009/TIC-1542), the Innovation project Holopedia (TIN2010-21128-C02-01) and by the ESF Research Network Program ELIAS.

We would like to thank Dr. Irina Chugur and Dr. Wouter Weerkamp for helping us with the definition and annotation of the opinion target corpus.

9. References

E. Amigó, J. Artiles, J. Gonzalo, D. Spina, B. Liu, and A. Corujo. 2010. WePS-3 evaluation campaign: Overview of the online reputation management task. In *CLEF 2010 Labs and Workshops Notebook Papers*.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19.

D. Harman. 1995. Overview of the Fourth Text REtrieval Conference (TREC-4). In *TREC-4*.

D. Hiemstra, S. Robertson, and H. Zaragoza. 2004. Parsimonious language models for information retrieval. In *Proceedings of the 27th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04)*.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (SIGKDD '04)*.

B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188.

V. Jijkoun, M. de Rijke, and W. Weerkamp. 2010. Generating focused topic-specific sentiment lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*.

M. Kaufmann and J. Kalita. 2010. Syntactic normalization of Twitter messages. In *International Conference on Natural Language Processing (ICNLP '10)*.

S.M. Kim and E. Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *ACL Workshop on Sentiment and Subjectivity in Text*.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33.

E. Meij, W. Weerkamp, and M. de Rijke. 2012. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining (WSDM '12)*.

G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513 – 523.

T.T. Thet, J.C. Na, and C.S.G. Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6):823.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.