A hybrid framework for scalable Opinion Mining in Social Media: detecting polarities and attitude targets

Carlos Rodríguez-Penagos
Barcelona Media Innovació
Av. Diagonal 177
Barcelona, Spain
carlos.rodriguez@barcelonamedia.org

Jens Grivolla
Barcelona Media Innovació
Av. Diagonal 177
Barcelona, Spain
jens.grivolla@barcelonamedia.org

Joan Codina Fibá
Barcelona Media Innovació
Av. Diagonal 177
Barcelona, Spain
joan.codina@barcelonamedia.org

Abstract

Text mining of massive Social Media postings presents interesting challenges for NLP applications due to sparse interpretation contexts, grammatical and orthographical variability as well as its very fragmentary nature. No single methodological approach can be expected to work across such diverse typologies as twitter micro-blogging, customer reviews, carefully edited blogs, etc. In this paper we present a modular and scalable framework to Social Media Opinion Mining that combines stochastic and symbolic techniques to structure a semantic space to exploit and interpret efficiently. We describe the use of this framework for the discovery and clustering of opinion targets and topics in user-generated comments for the Telecom and Automotive domains.

1 Introduction

Social Media (SM) postings constitute a messy and highly heterogeneous media that nonetheless represent a highly valuable source of information about the attitudes, interests and expectations of citizens and consumers everywhere. This fact has driven a trove of recent research and development efforts aimed at managing and interpreting such information for a wide spectrum of commercial applications, among them: reputation management, branding, marketing design, etc. A diverse array of techniques representing the state of the art run the gamut from knowledge-engineered rule-and lexicon-base approaches that (when carefully crafted) provide high precision in homogeneous contexts, to wide-coverage machine learning approaches that (when suitable development data is available) tackle noisy text with reasonable accuracies in some genres.

As SM channels are as different from each other as, say, spoken text from essay writing, we believe that no single technique, powerful as it may be, is capable of interpreting all domains, genres and channels in the vast universe of SM conversations. Faced with an industrial demand for simultaneous monitoring of heterogeneous opinion sources, our approach has evolved into combining diverse NLP technologies into a robust semantic analysis framework to create a high-granularity representation of user-generated commentaries amenable to machine interpretation.

Analysis of Telecom-related social postings has shown how a modular and scalable analysis framework can combine a veritable arsenal of NLP and data mining techniques into a hybrid application that adapts well to the unique challenges and demands of different Social Media genres.

Section 2 will present the UIMA-Solr framework and components used to process opinionated text, as well as discuss the representational choices made for analysis. Section 3 will frame our approach within the State-of-the-Art of Sentiment analysis and Opinion mining as we interpret it, while Sections 4 and 5 describe data and results of the application of our proposed approach in the context of opinion topic detection and clustering of SM postings in the Telecoms and Automobile domains respectively, and with different textual genres. Finally, Section 6 will focus on the conclusions and future work that presents to us at this point.

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 46–52,
Avignon, France, April 23 - 27 2012.

2 A modular toolset for SM processing

For semantic processing of our data we use a UIMA [1] (Ferrucci & Lally, 2004) architecture plus Solr-based clustering and indexing capabilities. Our choice of UIMA is guided in part by our wish to achieve good scalability and robustness, and that all components can be implemented modularly and in a distributed manner using UIMA-AS (Asynchronous Scale out). Also, UIMA's data representation as CAS objects allows preserving the documents integrity since annotations are added as standoff metadata, without modifying the original information.

Under the UIMA architecture, a hybrid NLP analysis framework is possible, combining powerful Machine Learning modules like Maximum Entropy (ME, OpenNLP) [2] or Conditional Random Fields (CRF, JulieLab), [3] with gazetteer and regular expression matchers and rule-based Noun Phrase chunkers. The basic linguistic processing has a sentence and token identifier, a POS tagger, a lemmatizer, a NP chunker and a dependency parser. In addition, we employ gazetteers to match products, companies, and other entities in text, as well as a hand-crafted lexicon of polar terms created from corpus exploration of Telecom domain text, as well as a regular expression module to detect emoticons when available. Also, two models for Named-Entity recognition were applied using CRF: one trained on conventional ENAMEX Named Entity Recognition and Classification entities, and another trained using data from customer reviews from various domains (Cars, Banking, and Mobile service providers), in order to detect opinion targets and cues. One of the objectives of this relatively straightforward processing (although by no means the only one), was to select candidates for classifiers that could identify both the specific subject of each opinion expressed in text, as well as capture a more general topic of the whole conversation (which conceivably could coincide or not with one of the specific opinion targets). Targets and topics are usually expressed as entity names, concepts or attributes, and thus can appear in language as noun, adjectival, adverbial or even verbal phrases. Opinion cues (or Q-elements) are words, emoticons and phrases that convey the actual attitude of the speaker towards the topics and targets, and a strength and polarity can be attributed to them, both *a priori* and in context.

Our modular processing approach allows customizing the annotation for each domain or genre, since, for example, regular expressions to detect emoticons will be useful for twitter micro-blogging, but less so for more conventional blogs where such sentiment-expression devices are less frequent; Also pre-compiled lists of known entities can provide good target precision while customised distributional models will help discover unlisted names and concepts in text.

The output of the semantic and syntactic processing pipeline is indexed using the Apache Solr framework,[4] which is based on the Lucene engine. This setup allows the implementation of clustering and classification algorithms, allowing us to obtain reliable statistical correlations between documents and entities.

We also developed or adapted a number of visualization components in order to present the data stored in Solr in an interactive page that is conducive to data exploration and discovery by the system's corporate users. At the same time, Carrot2 is connected to Solr and is used to test clustering conditions and algorithms, providing a nice visualization interface. Carrot2 is an open source search results clustering engine (Osiński & Weiss, 2005). It can automatically organize collections of documents into thematic categories.

3 Previous work

Two good overviews of general Opinion Mining and Sentiment Analysis challenges are Pang & Lee (2008) and, focused specifically on customer reviews, Bhuiyan, Xu & Josang (2009). Detecting the subject or targets of opinions is one of the main lines of work within Opinion Mining, and considerable effort has been put into it, since it has been shown to be a highly-domain specific task (consumer reviews will focus on specific products and features, tweets have hashtags to identify topics, blogs can talk almost about anything, etc.).

Outside of user-generated content, Coursey, Mihalcea, & Moen (2009) have suggested using indirect semantic resources, such as the Wikipedia, to identify document topics. For Opinion Mining genres, and extending on Hu & Liu (2004), Popescu & Etzioni (2005) use a combination of Pointwise Mutual Information,

[1] Unstructured Information Management Architecture
[2] http://maxent.sourceforge.net
[3] http://www.julielab.de
[4] http://lucene.apache.org/solr/

relaxation labeling and dependency analysis to extract possible targets and features in product reviews. Kim & Hovy (2006), for example, use thematic roles to establish a relation between candidate opinion holders and opinion topics, while exploiting clustering to improve coverage in their role-labeling. Recent approaches have included adaptation of NER techniques to noisy and irregular text, either by using learning algorithms or by doing text normalization (Locke & Martin, 2009; Ritter, Clark & Etzioni, 2011).

4 Exploring the semantic space of Telecom-related online postings

We collected close to 200,000 postings from various SM sources in a 4 month timeframe, including fairly carefully-written product-oriented forums, blogs, etc., as well as more casually-drafted Facebook and twitter micro-blogging, that discussed Spanish Telecom's services and products. Of these, we randomly sub-selected a representative 190-document sample that was manually marked-up (for a test involving machine learning of cue-polarity-target relationships) by two different human annotators with a 20-document overlap, using simplified annotation guidelines focused on opinion targets, topics, cues and polarities. An interesting observation about the interannotator agreement (but one we can't discuss in detail here) is that with regard to targets one of the human annotators tended more towards complete syntactic units (noun phrases), while the other chose more conceptual and semantic extensions as subjects for the opinions. The 20-document overlap was meant to help us evaluate this guideline development process, but the misalignment of guideline interpretation by the two human annotators made it very difficult to measure any kind of true interannotator agreement. Also, single annotation adjudication was made difficult due to the fact that both interpretations presented valid aspects, and we chose to use each set as an independent evaluation set to detect any unnoticed patterns that could emerge from using one of the other in our training and validation, but those results are inconclusive and merit further research. Since no adjudicator was incorporated in the process to resolve disagreements, the final annotated sets do not constitute a true Gold Standard, but each human-annotated set was used in turn as a benchmark against automatic annotators.

Content elicitation was combined with activity and network mining for an enriched overview of the social conversation ecosystems, but the second aspect won't be discussed here for the sake of brevity. For the same reason, although other aspects of sentiment analysis were performed on this data (cue and polarity detection, for example), we will also restrict the scope of these discussions on the detection and clustering of specific targets and general topics of the opinions expressed in such SM channels. Obviously, a deeper and more textured view of opinionated text is needed to be of any real use, but the overall features, shortcomings and advantages of our chosen approach are adequately discussed even if we restrict this paper to these very specific tasks.

The first series of experiments about clustering using semantics explored the above-mentioned corpus of SM posting that discussed a Spanish Telecom, one of the aims being detecting and aggregating the topics and targets of online opinions. Different processing modules geared towards topic and target detection were compared against each human annotator's choices, but also against each other and to the combined output of each. The main modules involved were: (A) generic NERC, (B) a target and topic NERC model (StatTarg), (C) a Noun Phrase Chunker, and (D) a Gazetteer matcher (Taxonomy). Figures 1 through 4 show, respectively, recall (1) and precision (2) with regard to human annotated topics, and recall (3) and precision (4) with regard to human annotated targets.

The results presented here are the overall performance across genres and domains, since the 190 documents annotated covered the whole range from forums to tweets.

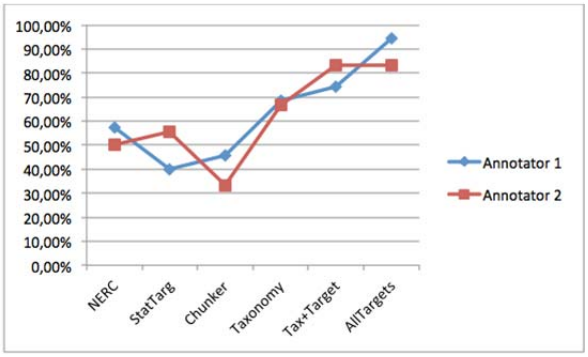

Figure 1. Topic recall

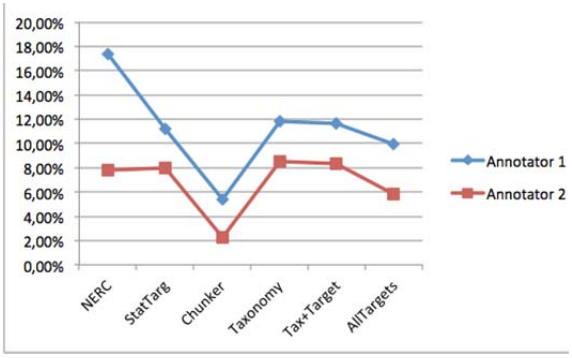

Figure 2. Topic precision

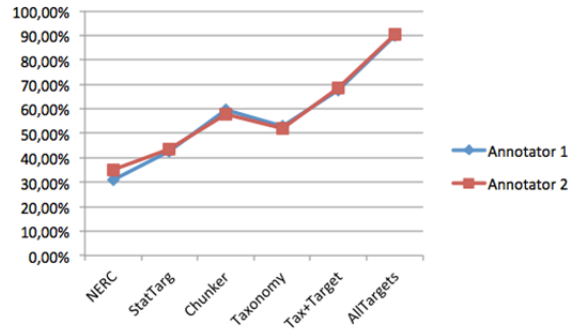

Figure 3. Target recall

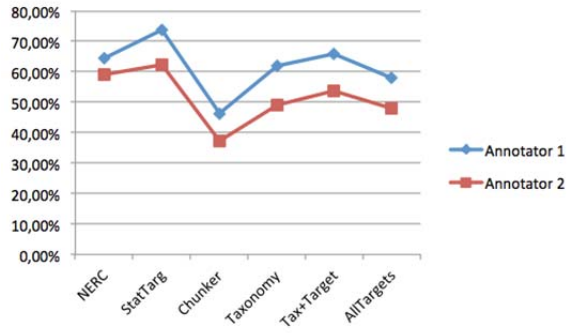

Figure 4. Target precision

For this experiment, and as a guideline for the human annotators, targets were roughly defined as occurrences in the text of objects of opinion, whereas topics where to represent the main focus of the document or message. The annotators usually marked one topic per document, which was almost always also one of the targets.

The customized taxonomy has a good precision with regard to target and topic identification, while the NERC and NP Chunk approaches improve the recall but suffer a bit on precision. Generic NER models have a moderately high precision (63%) with regard to manually annotated targets but rather low recall (specially in genres where capitalization is irregular which hinders NER detection), while NP Chunks present the opposite case: moderately (56%) high recall with low precision. This can be explained in part by the "greediness" of each methodology, with the chunker annotating extensively while the NERC model being much more selective. Another noteworthy result is the strong domain bias of target annotators trained on a Ciao customer reviews for Banking, Automotive and Mobile Service markets. The models implemented through training from multi-domain review sites were found to have medium precision, but very low recall.

The combination of all modules (*AllTargets*, a combination of NERC, Chunker, Taxonomy and StatTarget) had a very high recall of around 90%. With regard to topic detection, the combination of all modules had a recall of 94% and 83%, depending on which gold standard it is compared to (the one created by one expert human annotator or the other), which is an excellent recall level. The precision obtained on topic detection is very low. This, however, is expected as the evaluation is done using all candidates given by the different annotation layers, with no selection process. Since most of the topics are already identified as targets, the key issue here is to identify which of the comment targets is the main topic.

It is important to note that merging the *Chunker* output with that of the rest of the modules improves the recall of the system but the precision becomes low. The main reason is that most targets and topics are noun phrases, but not all noun phrases are targets or topics.

It is important to note that combining the output of different annotation layers (except for the NP chunker) does not reduce overall precision, while greatly increasing recall.

For the clustering experiments, we chose Carrot2's Lingo, a clustering algorithm based on Singular Value Decomposition. We envisioned the content-based clustering as an interactive exploratory tool, rather that providing a single "correct" and definitive set of groupings. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery that involves trial and failure. It will often be necessary to modify the preprocessing and adjust parameters until the result achieves the desired properties.

The query "*problem*", for example, sent to some of the telecom forums in May produced groupings suggestive of complaints relating to rates, internet access, SIM chips, SMS, as well as with regard to specific terminal models and companies. Even this limited capability can be helpful for some of our user's market analysis purposes.

The visualization of query-based clustering with detection of target, cues and topics, and the possibility of tracking trends over time, provided a very powerful overview of how consumer attitudes, expectations and complaints about products and services are reflected in dynamic automakers. The most relevant nouns, adjectives, bigrams and named entities from a given query, are projected into a polarity versus time dynamic map. The clustering was performed by the combined use of vector space reduction techniques and the K-means classification

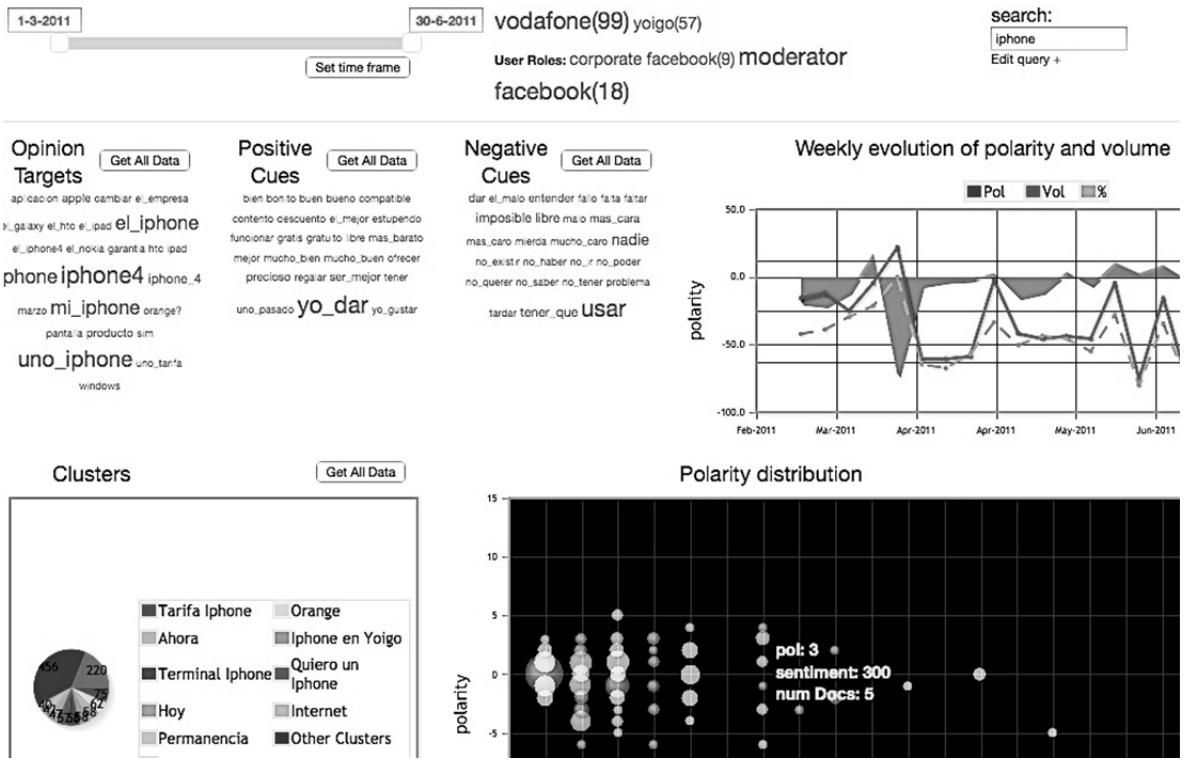

Figure 5. Facebook's "Iphone" semantic exploration (screenshot)

interchanges in various SM channels. These results are available through an online demo[6] (Figure 5, shown for Facebook postings).

5 Visualizing the evolution of customer opinion

In addition to exploring SM data for the Telecom domain, we performed some experiments using clustering without directly using annotated semantics, but instead using the semantics only for data interpretation. We crawled more than 10,000 customer reviews in the automotive domain in Spanish, along with some metadata that included the numerical ratings added by the reviewers themselves. Using our modular pipeline, we did shallow document clustering followed by linguistic processing that included lemmatization, POS tagging and Named Entity Recognition, in order to allow for analytical exploitation of the community-driven discussion on automobiles, product features and paradigm in a completely unsupervised manner. Clusters thus obtained were represented by sets of words that best described them to obtain a view of the emerging terms, trends and features contained in the opinions, with the aim of providing a representation of their collective content. Since evaluating clustering techniques *per se* was not the objective of these experiments, and since a gold standard was not available, the purpose of the system was (A) to validate the coherence of the groupings according to the review's content, and (B) assess if those clusters also aggregate as well along declared global polarity. Although inconclusive from a quantitative point of view, those experiments show the feasibility of leveraging existing Social Media resources in order to develop applications that can visualize and explore the semantic ecosystem of consumer opinions and attitudes, in a cost-effective and efficient manner. A demo of the functionalities of the system described here is also publicly

[6] http://webmining.barcelonamedia.org/Orange/

available.[7]. One cluster, a very positive one (based on the average user rating), is represented by the terms *land-terreno-todoterreno-rover-campo-4x4* (*off-road, field, ground, land, Rover*), while another one, *aceite-garantía-servicio-problemas-años* (*oil-warranty-service-problems-years*), in the lower right side might indicate unhappy reviewers.

6 Conclusion and future work

The results obtained on the Telecom corpus with different automatic annotation layers suggest that a possible improvement in the system could come from researching which combinations of automatic annotators can enhance overall performance, as one module's strength might complement another weaknesses and *vice versa*, so that what one is missing another one can catch. An additional option to increase overall recall is to implement a weighted voting scheme among the modules, allowing calculation of probabilities from the combinations of various annotations that overlap a textual segment.

The fact that combination of annotation layers through simple merging of all annotations has such a great impact on recall while not reducing precision suggests that the different methods are very complementary. We expect to be able to trade off some of the gained recall for much improved precision by applying more sophisticated merging methods.

Another possibility to be explored is using top level dependencies (such as SUBJECT, SENTENCE, etc.) to rank and select the main topic and target candidates using sentence structure configuration. This approach would also ensure that once a polarity-laden cue is identified, the corresponding target could be uniquely identified. This linguistics-heavy approach is feasible only in texts whose characteristics more closely resemble the data used to train the parser.

Our work has helped us focus more clearly many of the challenges faced by any NLP system when used in a new user-generated content: scarce development data, novel pattern and form adaptability, tool robustness, and scalability to massive and noisy text.

One of the lessons learned during these experiences is that keeping a modular hybrid analysis framework can improve matching by either customizing the pipeline to each genre and task requirements, or by combining the results of different approaches to benefit from each one's strengths while minimizing each one's weaknesses. Extracting opinion centered information from highly heterogeneous text and from multitudes of authors will never be as straightforward as, say, doing IE on newswire or financial news, but it should be feasible and useful by using the right toolset. We are in the process of using crowdsourcing to fully annotate vast Spanish and English corpora of opinionated text, which will allow us to perform a better and more fine-grained quantitative analysis of our framework in the near future.

Another lesson learned is that even if high-precision opinion classification is not available (because not enough development data is available, or data is noisy, or for whatever other reason) doing even superficial semantic annotation of the text and unsupervised clustering can help industrial consumer of these technologies understand better what is being said in the Social Media ecosystems. Valuable objectives for a useful opinion mining system do not need to include all possible analyses or state-of-the-art performance.

Going forward, computational exploitation of Social Media and of community-based, data-driven discussions on diverse topics and products is definitely an important facet of future market and business intelligence competencies, since more and more of our activities as citizens, friends and consumers take place in an online environment, where everything seems possible but where also everything we do leaves a trace and has a meaning. Extracting the semantics of collective action enables us to access that meaning.

References

Ritter A, Clark S, Mausam, and Etzioni O (2011). Named Entity Recognition in Tweets: An Experimental Study. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)

Bhuiyan, T., Xu, Y., & Josang, A. (2009). State-of-the-Art Review on Opinion Mining from Online Customers' Feedback. Proceedings of the 9th Asia-Pacific Complex Systems Conference (pp. 385–390).

Coursey, K., Mihalcea, R., & Moen, W. (2009). Using encyclopedic knowledge for automatic topic identification. Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL '09 (pp. 210–218). Stroudsburg,

[7] http://webmining.barcelonamedia.org/cometa/index_dates

PA, USA: Association for Computational Linguistics.

Ferrucci, D., & Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering, 10(3-4), 327–348.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). Seattle, WA, USA: ACM. doi:10.1145/1014052.1014073

Kim, S. M., & Hovy, E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. Proceedings of the Workshop on Sentiment and Subjectivity in Text (pp. 1–8).

Locke, B., & Martin, J. (2009). Named entity recognition: Adapting to microblogging. University of Colorado.

Osiński and D. Weiss (2005), "Carrot 2: Design of a flexible and efficient web information retrieval framework," Advances in Web Intelligence, pp. 439–444, 2005.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1–135.

Popescu, A. M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. Proceedings of HLT/EMNLP (Vol. 5, pp. 339–346).