ESWC'14 Challenge on Concept-Level Sentiment Analysis

Diego Reforgiato Recupero[1(✉)] and Erik Cambria[2]

[1] CNR, Via Gaifami 18, 95028 Catania, Italy
diego.reforgiato@istc.cnr.it
[2] School of Computer Engineering, Nanyang Technological University, 50 Nanyang Ave, Singapore, Singapore
cambria@ntu.edu.sg

Abstract. With the introduction of social networks, blogs, wikis, etc., the users' behavior and their interaction in the Web have changed. As a consequence, people express their opinions and sentiments in a totally different way with respect to the past. All this information hinders potential business opportunities, especially within the advertising world, and key stakeholders need to catch up with the latest technology if they want to be at the forefront in the market. In practical terms, the automatic analysis of online opinions involves a deep understanding of natural language text, and it has been proved that the use of semantics improves the accuracy of existing sentiment analysis systems based on classical machine learning or statistical approaches. To this end, the Concept Level Sentiment Analysis challenge aims to provide a push in this direction offering the researchers an event where they can learn new approaches for the employment of Semantic Web features within their systems of sentiment analysis bringing to better performance and higher accuracy. The challenge aims to go beyond a mere word-level analysis of text and provides novel methods to process opinion data from unstructured textual information to structured machine-processable data.

1 Introduction

As the Web rapidly evolves, people are becoming increasingly enthusiastic about interacting, sharing, and collaborating through social networks, online communities, blogs, wikis, and so forth. In recent years, this collective intelligence has spread to many different areas, with particular focus on fields related to everyday life such as commerce, tourism, education, and health, causing the size of the social web to expand exponentially.

The opportunity to capture the sentiment of the general public about social events, political movements, company strategies, marketing campaigns, and product preferences has raised growing interest both within the scientific community, leading to many exciting open challenges, as well as in the business world, due to the remarkable benefits of marketing prediction. However, the distillation of knowledge from such a large amount of unstructured information is so difficult

V. Presutti et al. (Eds.): SemWebEval 2014, CCIS 475, pp. 3–20, 2014.
DOI: 10.1007/978-3-319-12024-9_1

that hybridizing different methods from complementary disciplines facing similar challenges is a key activity.

Various Natural Language Processing (NLP) techniques have been applied to process texts to detect subjective statements and their sentiment. This task is known as sentiment analysis, and overlaps with opinion mining. Sentiment analysis over social media faces several challenges due to informal language, uncommon abbreviations, condensed text, ambiguity, illusive context, etc. Much work in recent years focused on investigating new methods for overcoming these problems to increase sentiment analysis accuracy over Twitter and the other social networks [5].

Mining opinions and sentiments from natural language involves a deep understanding of most of the explicit and implicit, regular and irregular, syntactical and semantic rules proper of a language. Existing approaches mainly rely on identifying parts of text in which opinions and sentiments are explicitly expressed such as polarity terms, affect words and their co-occurrence frequencies. However, opinions and sentiments are often conveyed implicitly through latent semantics, which make purely syntactical approaches ineffective [6].

To this end, concept-level sentiment analysis aims to go beyond a mere word-level analysis of text and provide novel approaches to opinion mining and sentiment analysis that allow a more efficient passage from (unstructured) textual information to (structured) machine-processable data, in potentially any domain. Indeed, semantics can play an important role in enhancing our ability to accurately monitor sentiment over social media with respect to specific concept and topics. For example, using semantics will enable us to extract and distinguish sentiment about, say Berlusconi, in politics, business, criminal investigations, soccer, or for different events that involve him. When moving from one context to another, or from one event to another, opinions can shift from positive to negative, or neutral.

Semantics can capture this evolution and differentiate its results accordingly, whereas most existing sentiment analysis systems provide an analysis that can be too coarse-grained, due to missed contextualization.

Concept-level sentiment analysis focuses on a semantic analysis of text through the use of web ontologies or semantic networks, which allow the aggregation of conceptual and affective information associated with natural language opinions. By relying on large semantic knowledge bases, concept-level sentiment analysis steps away from blind use of keywords and word co-occurrence count, but rather relies on the implicit features associated with natural language concepts [4].

It has been proved that the quality of sentiment analysis algorithms improves when considering semantic features [8,12,18]. The natural direction is therefore to provide existing sentiment analysis systems and algorithms based on machine learning techniques with semantic capabilities in order to increase their accuracy.

The Concept-level sentiment analysis challenge[1] has provided breeding ground for this process. In particular, the challenge has focused on the introduction, presentation, and discussion of novel approaches to concept-level sentiment analysis. Participants had to design a concept-level opinion-mining engine that exploited

[1] http://challenges.2014.eswc-conferences.org/index.php/SemSA

common-sense knowledge bases, e.g., SenticNet[2], and/or Linked Data and Semantic Web ontologies, e.g., DBPedia[3], to perform multi-domain sentiment analysis.

Submitted and accepted systems had a semantics flavor (e.g., by making use of Linked Data or known semantic networks within their core functionalities) and authors showed how the introduction of semantics could be used to obtain valuable information, functionality or performance. Some of the submitted systems were based on natural language processing methods and statistical approaches and the authors pointed out how the embedded semantics played a main role within the core approach (engines based merely on syntax/word-count have been excluded from the challenge).

Concept-level sentiment analysis research benefited also from the First Workshop on Semantic Sentiment Analysis[4], held at ESWC2014 concurrently with the challenge. The workshop focused on the introduction, presentation, and discussion of novel approaches to semantic sentiment analysis even if the approaches were still at early stage and no evaluation had been conducted. The audience of the workshop included researchers from academia and industry as well as professionals and industrial practitioners to discuss and exchange positions on new hybrid techniques, which use semantics for sentiment analysis.

Similar initiatives and papers related to the semantic sentiment analysis are listed and mentioned in the Sect. 2. Section 3 describes in detail the five tasks of the Concept-level sentiment analysis challenge that the challengers' systems had to face. Details on the creation of the annotated dataset where the challengers' systems have been tested is explained in Sect. 4. Section 5 includes details on the evaluation measures performed on each submitted system and each task. Section 6 presents the submitted systems whereas Sect. 7 shows the results of each of them for each addressed task. Section 8 ends the paper with comments and experiences gained from this challenge.

2 Related Work

The 2014 edition was the first ESWC to include a challenge call and session within its program, and the first time for an event on semantic sentiment analysis at ESWC.

The concept of challenges related to the Semantic Web domain is not new within the most prestigious international conferences.

For example, SemEval (Semantic Evaluation) is an ongoing series of evaluations workshops of computational semantic analysis systems which evolved from the Senseval word sense evaluation series. The goal is to evaluate semantic analysis systems in a wide range of domains and in a different set of tasks. The semantic sentiment analysis task was introduced in SemEval2007 and had a presence in 2010 and 2013 editions (the reader notices that between 2007 and 2013 there were only four SemEval events; it was the 2012 edition where the

[2] http://sentic.net/
[3] http://dbpedia.org
[4] http://ontologydesignpatterns.org/wiki/SemanticSentimentAnalysis2014

sentiment analysis task was missed). Reflecting the importance of this problem in social media, the current edition, SemEval2014[5], includes two different tasks for semantic sentiment analysis: (i) the aspect-based sentiment analysis and (ii) sentiment analysis on Twitter.

One more example is constituted by the International Semantic Web Conference, ISWC[6], that with a slightly broader coverage than ESWC, each year hosts a Semantic Web challenge whose central idea is to extend the current human-readable web by encoding some of the semantics of resources in a machine-processable form. Its target is quite general and the goals are:

- to show to the society what the Semantic Web can provide,
- to give researchers an opportunity to showcase their work and compare it to others,
- and to stimulate current research to a higher final goal by showing the state-of-art every year.

Semantic Web challenge at ISWC has not detailed tasks but only an Open Track and a Big Data track. As a consequence, the overall evaluation of the submitted systems is not based on precision/recall analysis or similar but a group of judges decide the finalists and the winners according to a set of requirements that the systems have to fulfill.

The 2013 edition of the ISWC challenge call included 17 systems to be evaluated[7]. One of them, *Sentilo: Semantic Web-based Sentiment Analysis*, represents the first semantic sentiment analysis system ever submitted for a Semantic Web challenge at ISWC. The challenger system was based on a Sentic Computing[8] method called Sentilo, [9], to detect holders and topic of opinion sentences. This method implements an approach based on the neo-Davidsonian assumption that events and situations are the primary entities for contextualizing opinions, which makes it able to distinguish holders, main topics, and sub-topics of an opinion. Besides, it uses a heuristic graph mining approach that relies on FRED [16], a machine reader for the Semantic Web that leverages NLP and Knowledge Representation (KR) components jointly with cognitively-inspired frames. Finally it developed a model for opinion sentences that was used for annotating their semantic representation. A more recent extension of this work is [17], where the authors have extended *OntoSentilo*, the ontology for opinion sentences, created a new lexical resource called *SentiloNet* enabling the evaluation of opinions expressed by means of events and situations, and introduced a novel scoring algorithm for opinion sentences which uses a combination of two lexical resources, SentiWordNet [1] and SenticNet [7], used among others as background knowledge for sentiment analysis.

Besides SentiWordNet and SenticNet, current approaches for concept-level sentiment analysis use other affective knowledge bases such as ANEW [3], WordNet-Affect [19], and ISEAR [22]. In [20], a two step method integrates iterative regression

[5] http://alt.qcri.org/semeval2014/
[6] Check http://iswc2014.semanticweb.org/ for the current edition
[7] http://challenge.semanticweb.org/2013/submissions/
[8] http://sentic.net/sentics/

and random walk with in-link normalization to build a concept-level sentiment dictionary. The approach, based on the assumption that semantically related concepts share a common sentiment, uses ConceptNet [13] for the propagation of sentiment values.

A similar approach is adopted in [14], which presents a methodology to create a resource resulting from automatically merging SenticNet and WordNet-Affect. Authors trained a classifier on the subset of SenticNet concepts present in WordNet-Affect and used several concept similarity measures as well as various psychological features available in ISEAR.

One more recent work that exploits an existing affective knowledge base is [11], which extracts from SentiWordNet the objective words and assess the sentimental relevance of such words and their associated sentiment sentences. A support vector machines classifier is adopted for the classification of sentiment data. The resulting method outperforms the traditional sentiment mining approaches where the objectivity of opinion words in SentiWordNet is not taken into account.

In [2] the authors survey existing works related to the development of an opinion mining corpus. Moreover the authors present Senti-TUT, an ongoing Italian project where a corpus for the investigation of irony within the political and social media domain is developed.

Other existing works exploit the combined advantages of knowledge bases and statistical methods. For example, in [21], the authors introduced a hybrid approach that combines the throughput of lexical analysis with the flexibility of machine learning to cope with ambiguity and integrate the context of sentiment words. Ambiguous terms that vary in polarity are identified by the context-aware method and are stored in contextualized sentiment lexicons. These lexicons and semantic knowledge bases map ambiguous sentiment terms to concepts that correspond to their polarity.

Further works based on machine-learning include [10], which develops a new approach for extracting product features and opinions from a collection of free-text customer reviews about a product or service. The approach exploits a language-modeling framework that, using a seed set of opinion words, can be applied to reviews in any domain and language. The approach combines both a statistical mapping between words and a kernel-based model of opinion words learned from the seed set to approximate a model of product features from which the retrieval is performed.

3 Proposed Tasks of the Challenge

The Concept-Level Sentiment Analysis challenge was defined in terms of five different tasks (Elementary Task 0 Polarity Detection, Advanced Task 1 Aspect-Based Sentiment Analysis, Advanced Task 2 Semantic Parsing, Advanced Task 3 Topic Spotting, The Most Innovative Approach Task). Participants had to submit a description of their system indicating which tasks their system was going to target. One of the five tasks, the **most innovative approach task**,

took into account all the submitted systems and gave a deep analysis on each of them. Within this task, a mixture of innovation and the employment of semantics were taken into account for the evaluation.

The first task was elementary whereas the second, third and fourth were more advanced. The input units of these four tasks were sentences. Sentences were assumed to be in grammatically correct American English and had to be processed according to the input format specified at http://sentic.net/challenge/sentence.

Following we will describe in detail each task.

3.1 Elementary Task 0: Polarity Detection

The main goal of task 0 was the classical polarity detection. The proposed systems were assessed according to precision, recall and F-measure of detected binary polarity values (1 = positive; 0 = negative) for each input sentence of the evaluation dataset, following the same format as in http://sentic.net/challenge/task0. As an example, considering the sentence of the above URL, *Today I went to the mall and bought some desserts and a lot of very nice Christmas gifts*, the correct polarity that a system should identify is positive (related to the Christmas gifts) and therefore it should write 1 in the polarity tag of the output. The problem of subjectivity detection was not addressed within this challenge, hence participants could assume that there were no neutral sentences. Participants were encouraged to use the Sentic API or further develop and apply sentic computing tools.

3.2 Advanced Task 1: Aspect-Based Sentiment Analysis

The output of this task was a set of aspects of the reviewed product and a binary polarity value associated to each of such aspects, in the format specified at http://sentic.net/challenge/task1. So, for example, while for the elementary task an overall polarity (positive or negative) was expected for a review about a mobile phone, this task required a set of aspects (such as speaker, touchscreen, camera, etc.) and a polarity value (positive or negative) associated with each of such aspects. Systems were assessed according to both aspect extraction and aspect polarity detection. As an example, the sentence *The touchscreen is awesome but the battery is too short* contains two aspects, *touchscreen* and *battery*, and a sentiment for each of them, positive for the former and negative for the latter.

3.3 Advanced Task 2: Semantic Parsing

As suggested by the title, the challenge focused on sentiment analysis at concept-level. This means that the proposed systems were not supposed to work at word/syntax level but rather work with concepts/semantics. Hence, this task evaluated the capability of the proposed systems to deconstruct natural language

text into concepts, following the same format as in http://sentic.net/challenge/task2. SenticNet could be taken as a reference to test the efficiency of the extracted concepts of the proposed systems, but they did not necessary have to match SenticNet concepts. The proposed systems, for example, were supposed to be able to extract a multi-word expression like *buy christmas present* or *go mall* or *buy desserts* from sentences such as *Today I bought a lot of very nice Christmas presents*. The number of extracted concepts per sentence were assessed through precision, recall and F-measure against the evaluation dataset.

3.4 Advanced Task 3: Topic Spotting

Input sentences were about four different domains, namely: books, DVDs, electronics, and housewares. This task focused on the automatic classification of sentences into one of such domains, in the format specified at http://sentic.net/challenge/task3. All sentences were assumed to belong to only one of the above-mentioned domains. The proposed systems were supposed to exploit the extracted concepts to infer which domain each sentence belonged to. Classification accuracy was evaluated in terms of precision, recall and F-measure against the evaluation dataset. As an example, the sentence *The touchscreen is awesome but the battery is too short* should be classified in the domain of electronics.

3.5 The Most Innovative Approach Task

This task looked for the most innovative system, how the semantics was employed and the overall innovation brought by the adopted method.

4 Dataset Generation

4.1 Data Collection

We arbitrarily chose 50 *electronics*, *book*, *housewares* and *dvd* reviews from the Blitzer dataset[9]. Reviews were then split into sentences and each of these was labeled by a pool of four annotators (two native English speakers, 1 Chinese and 1 Indian). The dataset can be freely downloaded[10]; the compressed file contains the annotated dataset for each of the four tasks.

4.2 Task 0: Polarity Detection

Annotators were asked to label sentences according to their polarity, i.e., positive or negative (neutral sentences were removed). This yielded 2,322 sentences bearing either positive or negative sentiment. Specifically, annotators were asked to empathize with the speaker. So, in a sense, the polarity associated with each sentence does not reflect the conveyed emotions but rather is an inference about

[9] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/
[10] http://sentic.net/eswc14.zip

the speaker's sentiments. This is key to disambiguate sentences that refer to more than one actor, e.g., "I love the movie that you hate". For each sentence, the polarity with the highest inter-annotator agreement was selected. We obtained 1,420 negative sentences and 902 positive (Table 1).

Table 1. Example sentences with polarity scores

Sentence	Polarity
The cheapest option I found at the time but an excellent pen drive	positive
What a useless thing	negative
They are very sharp and of high quality	positive
I've used this kettle for more than 1 year and it's still working perfectly	positive
The book is disproportionally focused on single and multilayer feedforward networks	negative
Its a shame to be forced to give this novel a one star rating	negative
Great product, I use it every day	positive

4.3 Task 1: Aspect Extraction

For the aspect extraction task, annotators were asked to infer aspects and label the sentiment associated with each of them. For this task, we liaised on majority voting for the selection of extracted aspects and their sentiment labels. It was notable that for most sentences the inter annotator agreement was greater than 2, i.e., most of the times, at least 3 annotators extracted same aspects and labeled them with the same sentiment. Sentences that did not have any aspect were removed from the final corpus. Table 2 shows the top 15 aspects extracted according to their occurrence in the corpus. 1725 sentences have been generated for such a task. The statistics on number of sentences having n number of aspects are shown in Table 3. Finally, Table 4 shows example sentences with aspects.

4.4 Task 2: Semantic Parsing

For semantic parsing task, we manually selected 2,398 sentences and asked annotators to extract the most useful concepts from them. Majority voting technique was applied on the extracted concepts to come up with a final list of concept for each sentence. The guideline was to choose multiword expressions richer in semantics so that in a sentence like "I went to the mall to buy food" the parsed concepts would be `go_mall` and `buy_food` rather than simply `go`, `mall`, `buy`, and `food`. Table 5 shows some statistics about the semantic parsing dataset.

4.5 Task 3: Topic Spotting

The topic spotting dataset was also built at sentence level. For each sentence, annotators labeled the topic and a majority voting technique determined the

Table 2. Top 15 aspects

Aspect	Frequency	Aspect	Frequency	Aspect	Frequency
player	188	camera	99	software	90
size	61	phone	54	picture	47
price	42	sound	41	battery	37
battery life	35	feature	34	use	31
weight	31	dvd	29	sound quality	29

Table 3. Number of sentence having n number of aspects

No. of aspects = 1	No. of aspects = 2	No. of aspects = 3	No. of aspects $\geq$ 4
1453	203	52	17

Table 4. Example sentences with aspects

Sentence	Aspects
but , if you 're looking for my opinion of the apex dvd player, i love it!	dvd player
for the price it is a well spent investment!	price
customer service and technical support are overloaded and nonresponsive - tells you about the quality of their products and their willingness to stand behind them.	customer service, technical service

Table 5. Number of sentence having n number of concepts

No. of concepts ≤ 5	No. of concepts > 5	No. of concepts ≤ 10	No. of concepts > 10
1037	1361	1845	553

final topic label for that sentence. It is notable that for almost every sentence annotator agreement was 4 (but this is mainly due to the fact that topics were predefined). The final dataset contains 1,122 sentences about *electronics*, 442 sentences about *books*, 1104 sentences about *dvds* and 1088 sentences about *housewares*. Table 6 shows example sentences and their topic.

5 Evaluation Measures

To evaluate the accuracy of the challenge tasks we analyzed each task and came up with a measure scheme for each of them. We wrote a Python script which automatically read the output of each system for each task and computed the accuracy according the scheme we adopted. In general, we followed

Table 6. Example sentences and their topic

Sentence	Topic
I love these speakers and the price was great	electronics
This dvd system is sweet and the sound system is off the hook its worth your Dollar	dvd
Nicely printed and bound - If you like James Allen you'll like this book	books
Though I have not tried the juicer yet, but i could not pass off the price	housewares

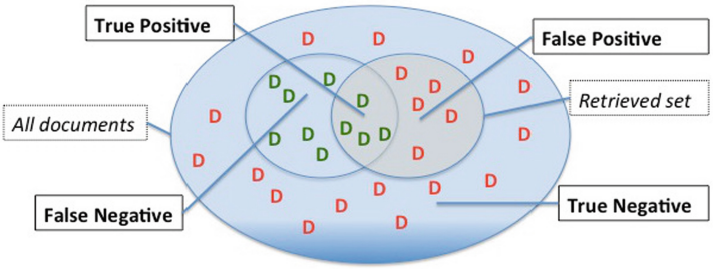


Fig. 1. Precision/Recall reference image.

the precision/recall study[11] with the observations and analysis defined in [15]. Figure 1 shows a general view of the precision/recall analysis where retrieved documents (true positive and false positive) are a subset of all the documents containing false negative and true negative. In general and where otherwise mentioned, the winner of a task was the resulting system with the highest F1 measure.

5.1 Evaluating Task 0

This task was pretty straightforward to evaluate. A precision/recall analysis was implemented to compute the accuracy of the output for this task. A true positive (tp) was defined when a sentence was correctly classified as positive. On the other hand, a false positive (fp) is a positive sentence which was classified as negative. Then, a true negative (tn) is detected when a negative sentence was correctly identified as such. Finally, a false negative (fn) happens when a negative sentence was erroneously classified as positive. With the above definitions, we defined the precision as

$$precision = \frac{tp}{tp + fp}$$

[11] http://en.wikipedia.org/wiki/Precision_and_recall

the recall as

$$recall = \frac{tp}{tp + fn}$$

and the F1 measure as

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

5.2 Evaluating Task 1

Task 1 was a bit more tricky than the previous one as it hindered two different subtasks, the extraction of the aspects/features and the polarity of each of them. A precision/recall analysis, similar to the one adopted for Task 0, has first been applied to the extraction subtask. Therefore, when a system detected a correct feature we marked that as true positive (*tp*); if the detected feature was not into the annotation dataset, then that was classified as false negative (*fn*). All the features present into the annotation dataset but not retrieved by the system constituted the false positive (*fp*) set. The precision, recall and F1 measure were then straightforward computed by using the formulas above.

As we have not taken the polarity information into account yet, we had to perform one more step and we decided to implement another precision/recall analysis as follows. If the extracted feature was correct and its associated polarity was also rightly spotted then we counted it as a true positive (*tp*), otherwise we counted it as a false negative (*fn*). The false positive (*fp*) set remained unchanged as in the previous subtask. At the end, for Task 1, we had two different F1 measures for each system. We simply took the average of those in order to establish the winners.

5.3 Evaluating Task 2

For Task 2, the annotated dataset we built, provided a set of concepts for each sentence. A concept might be written in several ways, using prepositions, articles and so on. That is why, when we built the annotated dataset for Task 2, we tried to generate as many different grammatical forms of a concept as possible. When performing the precision/recall analysis for Task 2 we classified as true positive (*tp*) a given concept of a certain system that was also included into the annotation dataset. The false negative (*fn*) set was constituted by all the defined concepts that were not present into the annotation dataset; finally, the concepts present into the annotations but not included into the system output were classified as false positive (*fp*). The precision, recall and F1 measure were then computed with the formulas above. The reader notices that the recall for this task was much lower than the other tasks because the presence of a large amount of concepts we wrote in different forms in our annotated dataset that increased the size of the false negative set.

5.4 Evaluating Task 3

Task 3 was the easiest to measure. As each sentence of the output consisted of just one of the four possible domains (books, DVDs, electronics, and housewares), we simply counted the sentences with the correct detected domain and used this number as the final measure to identify the winners for this task. The system with the highest number of sentences whose domain was correctly identified was the winner.

5.5 Evaluating the Most Innovative Approach Task

A board of three judges, chosen among the challenge program committee, evaluated each system in more detail and gave their assessment on the employment of the semantics and the use of concept-level mechanisms of each system. In particular, an important aspect was related to the interaction between semantics and sentics and how the polarity was handled within the context. Minor points that were taken into account were the computational time and the easiness of utilization.

6 Submitted Systems

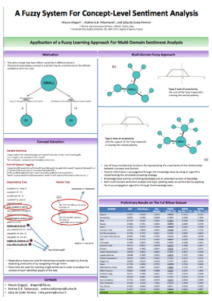
There were around 15 different intentional submissions to the Concept-Level Sentiment Analysis challenge. The challenge chairs had several discussions with many of the authors before the submission deadline about the requirements that the authors' systems had to satisfy. As each system had to have a semantic flavor using Linked Data, semantic resources, and so on, systems missing of semantics features were discouraged from the submission. Besides, the call for this challenge was launched at the end of December 2013 and the first deadline was for mid March 2014. Therefore time was not of help to authors with existing sentiment analysis systems for improving their systems with semantic resources and being able to satisfy the requirements of the challenge for the submission. However, six of them were able to ultimate their semantic sentiment analysis systems and those were submitted and accepted for the challenge. Participants were from very different countries: Italy, France, Israel, USA, Singapore, Mexico, UK, Taiwan. Only one system targeted and competed for all the tasks whereas the others participated for two, three or four tasks. Table 7 shows the title of the submitted systems, their authors and indicates the tasks that each of them targeted.

During the ESWC conference a poster and demo session was allocated for challengers to show their system by using either a poster or a demo (or both) to the public and explain the semantics their systems were based on. Table 8 shows a screenshot of the presented posters of four out of six systems participating to the Concept-Level Sentiment Analysis challenge whereas Table 9 shows a screenshot of five of them.

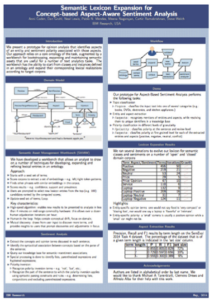
Table 7. The competing systems at the Concept-Level Sentiment Analysis challenge and the tasks they target.

System	Task 0	Task 1	Task 2	Task 3	Most Innovative
Mauro Dragoni, Andrea Tettamanzi and Celia Da Costa Pereira **A Fuzzy System For Concept-Level Sentiment Analysis**	X	X	X	X	X
Nir Ofek and Lior Rokach **Lechuzo: Weakly-Supervised System for Fine-Grained Sentiment Analysis**		X			X
Pablo Mendes, Anni Coden, Daniel Gruhl et al. **Semantic Lexicon Expansion for Concept-based Aspect-aware Sentiment Analysis**	X	X		X	X
Soujanya Poria, Nir Ofek **Sentic Demo: A Hybrid Concept-Level Aspect-Based Sentiment Analysis Toolkit**	X		X		X
Shafqat Mumtaz Virk, Yann-Huei Lee and Lun-Wei Ku **Sinica Semantic Parser for ESWC'14 Concept-Level Semantic Analysis Challenge**			X		X
Jay Kuan-Chieh Chung, Chi-En Wu and Richard Tzong-Han Tsai **Improve Polarity Detection of Online Reviews with Bag-of-Sentimental-Concepts**	X				X

Table 8. Four poster screenshots of the participants' systems.



Dragoni et al.



Mendes et al.

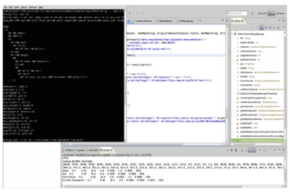
Virk et al.

Chung et al.

7 Results

During the challenge days, the evaluation dataset was revealed to the participants and the output of their systems was sent to the challenge chairs according to the same RDF format mentioned for each task description. In two cases, many of the sentences present within the output provided by the participants contained format errors and therefore they were excluded from that specific task. Following, the winners of each task and the evaluation measures results will be shown.

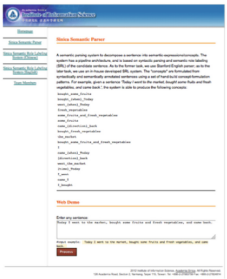
Table 9. Five screenshots of the running systems.



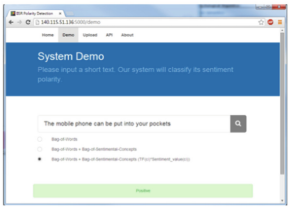
Dragoni et al.

Ofek et al.

Poria et al.



Virk et al.



Chung et al.

7.1 Task 0

Table 10 shows the precision-recall analysis for the output of the systems competing for Task 0 and the related winners. The system of *Chung et al.* had the best performing approach for this task and it was the winner of 100 euros award and a Springer voucher of the value of 150 euros.

Table 10. Precision-recall analysis and winners for Task 0.

System	Prec	Rec	F1	Pos
Chung et al.	0.78	0.57	0.66	1
Mendes et al.	0.66	0.59	0.62	2
Dragoni et al.	0.42	0.47	0.44	3
Poria et al.	Excluded for formatting errors			

7.2 Task 1

Table 11 shows the precision-recall analysis for the output of the systems competing for Task 1 and the related winners. The system of *Dragoni et al.* had the highest precision recall analysis and got an award of 100 euros and a Springer voucher of the value of 150 euros.

Table 11. Precision-recall analysis and winners for Task 1.

System	$Prec_1$	Rec_1	$F1_1$	$Prec_2$	Rec_2	$F1_2$	$F1_{avg}$	Pos
Dragoni et al.	0.25	0.26	0.26	0.12	0.11	0.11	0.19	1
Mendes et al.	0.24	0.15	0.18	0.12	0.06	0.09	0.14	2
Ofek et al.	0.12	0.06	0.08	0.09	0.04	0.06	0.07	3

7.3 Task 2

Table 12 shows the precision-recall analysis for the output of the systems competing for Task 2 and the related winners. The system of *Poria et al.* was the winner of an award of 100 euros and a Springer voucher of the value of 150 euros.

Table 12. Precision-recall analysis and winners for Task 2.

System	Prec	Rec	F1	Pos
Poria et al.	0.87	0.0.37	0.0.52	1
Virk et al.	0.05	0.003	0.005	2
Dragoni et al.	Excluded for formatting errors			

7.4 Task 3

Finally, Table 13 shows the results for the output of the systems competing for Task 3 and the related winners. The reader notices that some sentences have been taken out of the count when formatting errors were present. In the system of *Mendes* 3501 sentences were correctly evaluated whereas in the system of *Dragoni* 879 sentences have been taken out for problems with RDF specifications. Therefore, the system of *Mendes et al.* was the winner and got an award of 100 euros and a Springer voucher of the value of 150 euros.

Table 13. Results and winners for Task 3.

System	Number of sentences with correctly classified domain	Pos
Mendes et al.	1179 out of 3501	1
Dragoni et al.	458 out of 2622	2

7.5 The Most Innovative Approach Task

The Innovation Prize went to *Dragoni et al.* (a) for introducing the concept of fuzzy membership of multi-word expressions for dynamically detecting the polarity of natural language concepts according to different domains and contexts and

(b) for proposing the use of a two-level framework that nicely models the interaction between semantics and sentics for aspect-based sentiment analysis. These are two key elements for the advancement of sentiment analysis research because (a) polarity is not a static thing but rather a dynamic context-dependent measure and (b) semantic and affective relatedness are two different coefficients that need to be kept separate while used concomitantly. The most common mistakes in current sentiment analysis research, in fact, are (a) the a-priori definition of polarity, e.g., in the case of the "small" adjective which is neither positive nor negative but rather acquires a polarity according to the context, and (b) the (con)fusion of semantic and affective level, e.g., in the case of concepts like "joy" and "anger" which are highly semantically related (as they are both emotions) but have opposite affective relatedness.

8 Conclusions

The Concept-Level Sentiment Analysis challenge attracted several researchers mainly from two different domains: (i) those of the sentiment analysis area who have been pushed to explore the strengths and opportunities of the Semantic Web and tried to exploit it within their existing sentiment analysis systems which were based on traditional artificial intelligence, machine learning or natural language processing approaches. (ii) Those involved within the Semantic Web area, showing them the domain of the sentiment analysis and attracted them to develop their own systems with a strong base of Semantic Web features to solve some of the tasks of the challenge mentioned above. Besides, the concurrent execution of the First Workshop on Semantic Sentiment Analysis at ESWC on similar topics brought a process of cross-pollination of ideas among the attendees: researchers, editors of prestigious international journals and magazines, people from industry and key stakeholders in general. It is to highlight the number of attendees of the workshop which was around 30 including several participants of the challenge which had been asked to held a small session within the workshop briefly showing their system and giving tips on their learned experience about the technical development. During the challenge, all the participants were really active and we did not experience problems during the normal conduction of the challenge and its evaluation. Among the learned lessons we had, one is particularly important and to be shared as it is related to several other challenge even in different domains. We have noticed that it would have been much better to provide the participants not only an evaluation dataset where they have tested their systems but also the very same script we used for the precision/recall analysis. This could have given the participants further tips on the reasons related to the performance of their systems (e.g. the wrong format of the output of a few systems could have been spotted and fixed earlier). Overall, the Concept-Level Sentiment Analysis was successful and we aimed at reconsidering it again at the next edition of the ESWC.

References

1. Baccianella, A., Esuli, S., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta (2010)
2. Bosco, C., Patti, V., Bolioli, A.: Developing corpora for sentiment analysis: the case of irony and Senti-TUT. IEEE Intell. Syst. **28**(2), 55–63 (2013)
3. Bradley, M., Lang, P.: Affective norms for English words (ANEW): stimuli, instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida (1999)
4. Cambria, E., Hussain, A.: Sentic Computing: Techniques, Tools, and Applications, vol. 2. Springer, Heidelberg (2012)
5. Cambria, E., Schuller, B., Xia, Y., Havasi, C.: New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. **28**(2), 15–21 (2013)
6. Cambria, E., White, B.: Jumping NLP curves: a review of natural language processing research. IEEE Comput. Intell. Mag. **9**(2), 48–57 (2014)
7. Cambria, E., Olsher, D., Rajagopal, D.: Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In: Brodley, C.E., Stone, P. (eds.) Twenty-Eight AAAI Conference on Artificial Intelligence, pp. 1515–1521. AAAI Press, Palo Alto, July 2014
8. Chen, H., Wuand, Z., Cudré-Mauroux, P.: Semantic Web meets computational intelligence: state of the art and perspectives. IEEE Comput. Intell. Mag. **7**(2), 67–74 (2012)
9. Gangemi, A., Presutti, V., Reforgiato Recupero, D.: Frame-based detection of opinion holders and topics: a model and a tool. IEEE Comput. Intell. Mag. **9**(1), 20–30 (2014)
10. Garcia-Moya, L., Anaya-Sanchez, H., Berlanga-Llavori, R.: Retrieving product features and opinions from customer reviews. IEEE Intell. Syst. **28**(3), 19–27 (2013)
11. Hung, C., Lin, H.-K.: Using objective words in sentiwordnet to improve word-of-mouth sentiment classification. IEEE Intell. Syst. **28**(2), 47–54 (2013)
12. Johansson, R., Moschitti, A.: Relational features in fine-grained opinion analysis. Comput. Ling. **39**(3), 473–509 (2013)
13. Liu, H., Singh, P.: Conceptnet: a practical commonsense reasoning toolkit. BT Technol. J. **22**, 211–226 (2004)
14. Poria, S., Gelbukh, A.F., Hussain, A., Howard, N., Das, D., Bandyopadhyay, S.: Enhanced senticnet with affective labels for concept-based opinion mining. IEEE Intell. Syst. **28**(2), 31–38 (2013)
15. Powers, D.M.W.: Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation. Technical report SIE-07-001, School of Informatics and Engineering, Flinders University, Adelaide, Australia (2007)
16. Presutti, V., Draicchio, F., Gangemi, A.: Knowledge extraction based on discourse representation theory and linguistic frames. In: ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Acquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (eds.) EKAW 2012. LNCS, vol. 7603, pp. 114–129. Springer, Heidelberg (2012)
17. Reforgiato Recupero, D., Presutti, V., Consoli, S., Gangemi, A., Nuzzolese, A.: Sentilo: frame-based sentiment analysis. Cogn. Comput. (2014)

18. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of Twitter. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012, Part I. LNCS, vol. 7649, pp. 508–524. Springer, Heidelberg (2012)
19. Strapparava, C., Valitutti, A.: WordNet-affect: an affective extension of WordNet. In: LREC, Lisbon, pp. 1083–1086 (2004)
20. Tsai, A.C.-R., Wu, C.-E., Tsai, R.T.-H., Hsu, J.Y.-J.: Building a concept-level sentiment dictionary based on commonsense knowledge. IEEE Intell. Syst. **28**(2), 22–30 (2013)
21. Weichselbraun, A., Gindl, S., Scharl, A.: Extracting and grounding context-aware sentiment lexicons. IEEE Intell. Syst. **28**(2), 39–46 (2013)
22. Weigand, E. (ed.).: Emotion in Dialogic Interaction. Current Issues in Linguistic Theory, vol. 248. John Benjamins, Philadelphia (2004)