Data Min Knowl Disc (2012) 25:603–634
DOI 10.1007/s10618-012-0259-9

Using EmotiBlog to annotate and analyse subjectivity in the new textual genres

Ester Boldrini · Alexandra Balahur · Patricio Martínez-Barco · Andrés Montoyo

Received: 19 January 2011 / Accepted: 13 February 2012 / Published online: 7 March 2012

Abstract Thanks to the increasing amount of subjective data on the Web 2.0, tools to manage and exploit such data become essential. Our research is focused on the creation of EmotiBlog, a fine-grained annotation scheme for labelling subjectivity in non-traditional textual genres. We also present the EmotiBlog corpus; a collection of blog posts composed by 270,000 tokens about 3 topics and in 3 languages: Spanish, English and Italian. Additionally, we carry out a series of experiments focused on checking the robustness of the model and its applicability to Natural Language Processing tasks with regards to the 3 languages. The experiments for the inter-annotator agreement, as well as for feature selection, provided satisfactory results, which have given an impetus to continue working with the model and extend the annotated corpus. In order to check its applicability, we tested different Machine Learning models created using the annotation in EmotiBlog on other corpora in order to see if the obtained annotation is domain and genre independent, obtaining positive results. Finally, we also applied EmotiBlog to Opinion Mining, proving that our resource allows an improvement the performance of systems built for this task.

Keywords Sentiment analysis · Annotation model · Feature selection · Opinion Mining · New textual genres

Responsible editor: Fei Wang, Hanghang Tong, Phillip Yu, Charu Aggarwal.

E. Boldrini (✉) · A. Balahur · P. Martínez-Barco · A. Montoyo
University of Alicante, GPLSI, Apartado de Correos 99,
03080 Alicante, Spain
e-mail: eboldrini@dlsi.ua.es

1 Introduction

In the last few years, alternative ways of communication have arisen—blogs, forums and reviews—that people from all over the world employ as source of information in addition to traditional textual genres such as newspaper articles. As a consequence, there has been an exponential growth in the amount of subjective information on the Web 2.0, which has led to the need for innovative Natural Language Processing (NLP) tools, resources and methods, able to properly analyse and manage such data. In this context, blogs (one of the new textual genres) have peculiar features, which distinguish them from other types of documents (Paquet 2003). They have personal editorship, a hyperlinked posting structure, frequent updates, free public access to the content via the Internet, and archived posts. Generally, one single author creates a blog. However, there are exceptions for collective blogs where contributors post and debate short essays and opinion pieces. The comments represent the users' point of view and their interpretations of an event/product, the content and context.

This phenomenon is also described by "The State of the Blogosphere 2009", a survey published by Technorati,[1] which proves the growing influence of the blogosphere on the methods via which information on a wide range of topics is disseminated through communities. Analyzing the statistics presented in this study, there is evidence that users are blogging with high frequency. Thus, this phenomenon makes the State of the Blogosphere robust across geographical regions, age and even gender. Blogs are becoming a point of reference for professionals who decide to employ this means of communication, contradicting the common belief about the predominance of an informal style (Balahur et al. 2009a).

The subjective content of the Web is increasing exponentially, reflecting people's opinion about a wide range of topics (Cui et al. 2006). Technorati's survey also demonstrates that self-expression and sharing expertise by means of opinions are the primary motivation for bloggers. They describe significant, positive impacts on their personal lives, as well as their business.

Bearing this in mind, we can deduce that today blogs represent an important source of real-time, unbiased information, which can be exploited to develop many practical applications for diverse users with different profiles and needs. Examples of such applications could be the business of brand image monitoring by means of which a company analyses the opinions of its clients and external people. Further applications could consist in the detection of opinions regarding the competitors or the evaluation of the clients' opinion on a product depending on their needs and experience. Another possible scenario could be the existing social network where blogs can be exploited for monitoring attitudes for behavioural or psychological purposes or studies. Finally, an innovative way of exploiting such data could be focused on the prediction of relevant social or economic tendencies by means of information monitoring. For example, the recent worldwide economic crisis could be monitored via opinion gathering, helping the people in charge to take the appropriate measures. These are just a few examples of the potential options for real applications. As we can see, the number of situations

[1] http://technorati.com/blogging/article/state-of-the-blogosphere-2009-introduction/.

in which NLP techniques for subjective information processing can be employed is very high, due to the wide range of topics bloggers discuss daily.

Parallel to the growth of this new textual genre, there has been an increasing interest from researchers to develop methods to extract data from the subjective information available from these new sources. The data published by bloggers represents an opportunity to carry out interdisciplinary studies, related to fields such as Sociology, Economics, Law, Computational Linguistics, etc. Part of this research is intended for the development of technologies to organise textual information, not just in terms of topical content, but also taking into account the emotions and opinions embedded, as well as the source of the discourse.

The NLP task in charge of dealing with the treatment of subjective data is called Sentiment Analysis (SA) or Opinion Mining (OM). It can be briefly defined as the task that automatically detects the opinion expressed from texts on an entity and classifies it depending on its polarity (positive, negative or neutral).

The main challenges of this field are diverse: mixture of text styles, a wide range of topics and sources, multiple languages, grammar and spelling mistakes, informal language, use of colloquialisms or slang, extensive amounts of data, continuous updating of information, etc.

In order to lessen and counteract the abovementioned challenges, raised by the large quantity of subjective data available and the lack of resources to exploit such data, the general purpose of the research we carried out was to:

- Design a fine-grained annotation schema able to capture the linguistic means of affective expression in non-traditional textual genres.
- Annotate a collection of blog posts using the resulting schema.
- Evaluate the robustness of the scheme creating Machine Learning models using the annotated elements. In this way, by means of widely employed and versatile techniques we evaluate the usefulness of the resource and its application to a concrete NLP task: Opinion Mining.

The abovementioned objectives were achieved as follows:

- We collected a multilingual corpus of blog posts and analysed the linguistic features of subjectivity in text. Subsequent to this analysis, we proposed the list of elements for our annotation model and labelled the corpus.
- Later we assessed the robustness of the model calculating the inter-annotator agreement.
- Starting from the results obtained in the previous point, we carried out feature selection experiments, to check the impact of the elements that compose the model.
- After having performed these evaluations, we tested if the elements annotated in our corpus (Boldrini et al. 2009a) were useful to build Machine Learning models that can spot subjectivity in other corpora (pertaining to other domains and genres). In order to carry out this assessment we employed the elements annotated with EmotiBlog to build different Machine Learning models which we then employed to detect subjectivity in the ISEAR (Scherer and Wallbott 1997), NTCIR 8 MOAT[2] and SEMEVAL 2007 Task 14 corpora (Strapparava and Milhacea 2007).

[2] http://research.nii.ac.jp/ntcir/ntcir-ws8/permission/ntcir8xinhua-nyt-moat.html.

- We then employed our resource to carry out Opinion Mining and Opinion Mining in real-time.

This paper is organized as follows. Section 2 presents the related work and the added value of our research. Section 3 is focused on the motivation of our research and Sect. 4 illustrates the EmotiBlog annotation scheme and corpus. In Sect. 5 we carry out the feature selection experiments and in Sect. 6 the Opinion Mining and polarity classification tests. Finally, Sect. 7 summarises our conclusions and future work.

2 Related work

In recent years, the study of affect-related phenomena as an NLP subtask has constantly grown in importance. There is a significant increase of interest in the automatic identification and extraction of opinions, emotions, sentiments and points of view in the new textual genres. Different researchers have addressed the needs and possible methodologies from the linguistic, theoretical and practical points of view.

Sentiment Analysis can be divided in two main research areas:

2.1 Creation of resources

The first is composed by corpora and annotation schemes. An example of these resources—one of the pioneer researches and the one by which we were mainly inspired—is the MPQA, an annotation model created by Wiebe et al. (2005). Its main goal is to represent internal mental and emotional states from material presented as facts. Therefore, they concentrate on the notion of ’private states’, a general term that includes opinions, beliefs, thoughts, feelings, emotions, etc. Hu and Liu (2004) present a fine-grained model for sentiment analysis, which they denote as “feature-based Opinion Mining”. The ISEAR corpus consists in a collection of phrases where people describe a situation when they felt a certain emotion. Another related research done by Wilson et al. (2004a; 2004b) presents a classification for measuring the strength of opinions and other types of subjectivity and also classifies the subjectivity of deeply nested clauses. Craggs and Wood (2005) propose an annotation scheme for emotion in dialogue using categorical labels. Somasundaran et al. (2007) designed an annotation scheme for manual labelling of opinion categories in meetings. Somasundaran et al. (2008) developed genre-specific lexicons using interesting function word combinations for detecting opinions. Mathieu (2005) built up a semantic lexicon in the field of feelings and emotions and a system to annotate emotions in texts. Other relevant resources in the field can be WordNetAffect (Strapparava and Valitutti 2004), an extension of WordNet Domains, including a subset of synsets suitable to represent affective concepts correlated with affective words. SentiWordNet (Esuli and Sebastiani 2006), is a lexical resource in which each WordNet synset is associated to three numerical scores—Obj(s), Pos(s) and Neg(s)—describing how objective, positive, and negative the terms contained in the synset are. Micro-WNOP (Cerini et al. 2007), is a corpus composed by 1105 WORDNET synsets or emotion triggers (Balahur and Montoyo 2008) which are words or concepts expressing an idea which, depending on the reader's interest, cultural, educational and social factors, leads to a possible emotional interpretation of the text content. They are lexicons, which contain single words, whose polarity and associated emotion are not necessarily the same in a larger context. The underlying difference between the abovementioned studies and our work resides in the fact that we annotate larger text spans to be able to consider the undeniable influence of context.

2.2 Opinion extraction and classification

Kim and Hovy (2004) built up a system that, given a topic, automatically finds people who hold opinions on that particular subject and the sentiment of each opinion expressed. This system contains a module for determining word sentiment and another for combining them within a sentence. Wilson et al. (2005) propose a new approach to phrase-level sentiment analysis that first determines whether an expression is neutral or polar and, following this distinction, later disambiguates the polarity of the subjective expressions. Using this approach, this system is able to automatically identify the contextual polarity for a large subset of subjective expressions. Wiebe and Mihalcea (2006) conclude that subjectivity is a property that can be associated to word senses and word sense disambiguation.

Different authors show that this initial discrimination is a crucial point for the sentiment task, as part of Opinion Information Retrieval (last three editions of the TREC Blog tracks[3] competitions, the TAC 2008 competition[4]), Information Extraction (Riloff and Wiebe 2003) and Question Answering (Stoyanov et al. 2005) systems. Once this discrimination is complete, or in the case of texts containing only or mostly subjective language (such as e-reviews), Opinion Mining becomes a polarity classification task.

Related work also includes:

- customer review classification at a document level, sentiment classification using unsupervised methods (Turney 2002),
- Machine Learning techniques including sentiment classification considering rating scales (Pang and Lee 2003),
- scoring of features (Dave et al. 2003),
- using PMI, syntactic relations and other attributes with Supporting Vector Machine (SVM) (Mullen and Collier 2004),
- supervised and unsupervised methods (Chaovalit and Zhou 2005) and semi-supervised learning (Goldberg and Zhu 2006).

Research in classification at a document level also included sentiment classification of reviews (Ng et al. 2006), sentiment classification on customer feedback data (Gamon et al. 2005), but also comparative experiments (Cui et al. 2006).

Other research has been conducted in analysing sentiment at a sentence level using bootstrapping techniques (Riloff and Wiebe 2003). This includes:

[3] http://trec.nist.gov/data/blog.html.

[4] http://www.nist.gov/tac/

- work focused on considering gradable adjectives (Hatzivassiloglou and Wiebe 2000),
- semi-supervised learning with the initial training some strong patterns and then applying NB or self-training (Wiebe and Riloff 2006)
- finding strength of opinions (Wilson et al. 2004a,b)
- sum-up orientations of opinion words in a sentence (or within some word window) (Kim and Hovy 2004; Wiebe and Wilson 2005),
- determining the semantic orientation of words and phrases (Turney and Littman 2003),
- identifying opinion holders (Stoyanov and Cardie 2006),
- comparative sentence and relation extraction and feature-based Opinion Mining and summing up (Turney 2002).

Finally, fine-grained, feature-based opinion summarisation is defined in Hu and Liu (2004) or Pang and Lee (2003). All these approaches concentrate on finding and classifying the polarity of opinion words, which are mostly adjectives, without taking into account modifiers or the context.

2.3 Our research

Our research represents a step forward to the abovementioned previous work. It is focused on the creation of a multilingual resource for the fine-grained detection of sentiment with a view towards reaching a contextual comprehension and a deep understanding of the linguistic elements that account for the emotion and opinion expressions in a text.

The formerly created resources are mainly focused on detecting some generic elements, such as polarity, level and source. In order to overcome this first limitation, EmotiBlog focuses on the annotation at different levels: document, sentence and element. It has been constructed for a more precise identification of the linguistic elements that give subjectivity to the text to be analysed. This multitude of levels also allows for different kinds of experiments, depending on the needs of the application that is being developed.

Given the fine granularity of EmotiBlog, our main objective is to confirm the hypothesis that a fine-grained model allows diverse kinds of text processing, from the most general to detailed and complex texts without compromising and performance results. Moreover, we carried out experiments to check if EmotiBlog can be employed with different topics and languages, thus representing a relevant added value for the automatic processing of written language and for concrete NLP tasks at a multilingual level and domains.

The EmotiBlog corpus constitutes, first of all, a contribution to the enhancement of the existing set of resources for NLP tasks in languages other than English. The corpus was also employed to improve the performance of Opinion Question Answering, Opinion Mining and Opinion Summarisation, presented in detail in (Balahur et al. 2010a,b,c, 2009a). As explained in these papers, traditional tools for the treatment of objective data have so far not allowed for high results, thus EmotiBlog provides a training collection for opinionated data useful for the improvement of such systems.

In summary, the proposed scheme provides a flexible framework for annotating subjective text. It improves the current labelling method by letting the annotator label various parts of text, whilst also providing detailed information advancing the annotation of emotions from word-level to multi-level apart from its deeper level of analysis.

3 Motivation

Given their worldwide availability, blogs have become a primary source of information (Balahur et al. 2009b) representing a point of reference to many users. For example, when making a decision about buying a product, people may search for information and opinions expressed on the Web on their subject of interest and base their final decision on the information found. At the same time, when using a product, people often write reviews about it, so that others can gain an idea of its performance and utility before purchasing. Therefore, it may be said that the growing volume of opinionated information available on the Web allows for better and more informed user decisions. However, the amount of information to be analysed requires the development of specialised NLP systems to automatically extract, classify and summarize the data available on different topics. Esuli and Sebastiani (2006) define Opinion Mining as a recent discipline at the crossroads of Information Retrieval and Computational Linguistics, which is concerned not only with the topic a document treats, but also with the opinion it expresses. The main challenge of processing opinionated text is that, unlike objective content, subjective information is generally difficult and complex to extract and classify employing static patterns and fixed rules. Different authors have addressed the problem of extracting and classifying opinion from different perspectives and at different levels, depending on a series of factors, which can be:

1. level of interest: (overall/specific).
2. querying formula ("Nokia E65"/"Why do people buy Nokia E65?").
3. type of text (review on forum/blog/dialogue/press article).
4. manner of expression of opinion—directly (using opinion statements, e.g. "I think this product is wonderful!"/"This is a bright initiative"), indirectly (using affect vocabulary, e.g. "I love the pictures this camera takes!"/"Personally, I am shocked one can propose such a law!") or implicitly (using adjectives and evaluative expressions, e.g. "It's light as a feather and fits right into my pocket!").

While determining the overall opinion on a movie is sufficient to decide to watch it or not, when buying a product, people are interested in the individual opinions on the different product features. When discussing a person, one can judge and give an opinion on the person's actions. Moreover, the approaches taken can vary depending on the manner in which a user asks for the data (general formula such as "opinions on X" or a specific question "Why do people like X?" and the text source that needs to be queried). Retrieving opinion information in newspaper articles or blogs posts is more complex; it involves the detection of multiple discussion topics and sources, the subjective phrases present and subsequently their classification according to polarity. Especially in the blog area, determining the points of view expressed in dialogues together with the mixture of quotes and pastes from newspapers on a topic can, additionally, involve determining the number of persons giving an opinion, and whether or not the opinion expressed is on the required topic or on a point previously made by another speaker. As a consequence, Opinion Mining requires the use of specialised data for system training and tuning to be gathered, annotated and tested within the different text spheres (and with these different elements). At the present moment, such data is almost inexistent. One reason is that the few currently existing corpora are annotated at a very coarse-grained level and the majority of them are in English. The resource we propose has been designed to overcome the abovementioned challenges. At the moment it may be employed for three languages (English, Spanish and Italian) and with multiple textual genres and domains. It may be considered a more lexico-semantic model in contrast to previous work, such as the MPQA, and it is finer-grained. In addition to the polarity, target and source of the discourse, EmotiBlog contemplates a wider series of linguistic elements (described in detail in Sect. 4). Moreover, the entire corpus is annotated because we discriminate between objective and subjective discourse, and later we label the linguistic elements that give the subjectivity to the discourse (further details are presented in the Sect. 4.1). Consequently, we can offer a more exhaustive analysis of subjectivity expressions that could be exploited to automatically extract the required features in plain texts and use them for real life applications. We will use our corpus to extract the subjective information we are looking for and the approach adopted is different to those based on rules. Due to the changing nature and complexity of the new textual genres style, rules will cover a minimum number of the linguistic phenomena, not assuring an exhaustive discourse analysis. Additionally, rules are strictly dependent to domain and language. As we can deduce, the approach we employed assures an in-depth analysis of the discourse, multilingualism and no domain restrictions.

4 EmotiBlog

This section describes how we created the EmotiBlog annotation scheme (Boldrini et al. 2009b, 2010c), but also the corpus collection and the annotation process we carried out.

4.1 EmotiBlog annotation scheme

The main objective of our annotation model is to capture the most relevant linguistic elements that constitute subjectivity expressions. "Subjectivity" is a general concept, which can be expressed by means of different linguistic strategies. It can be expressed in texts by means of expressions of feelings[5] (an idea or belief, especially a vague or irrational one), opinions (a thought or belief about something or someone), emotions (a strong feeling such as love or anger, or strong feelings in general) or points of view (a way of considering something)[6]. The following step will consist in checking if an extremely fine-grained annotation model could be more effective. As we can observe

[5] http://oxforddictionaries.com/.

[6] http://dictionary.cambridge.org/.

Table 1 List of EmotiBlog elements and attributes

Elements	Attributes
Objective speech	Confidence, comment, and source
Subjective speech	Confidence, comment, level, emotion, phenomenon, polarity, and source
Phenomenon	Confidence, comment, type (collocation, saying, slang, title, and rhetoric)
Adjectives	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, and source
Adverbs	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, and source
Prepositions	Confidence, comment, level, emotion, phenomenon, modifier/not, polarity, and source
Verbs	Confidence, comment, level, emotion, phenomenon, polarity, and source
Anaphora	Type, source and antecedent
Capital letter	Confidence, comment, level, emotion, phenomenon, polarity, and source
Punctuation	Confidence, comment, level, emotion, phenomenon, polarity, and source
Other languages	Confidence, comment, Latin or English, level, emotion, phenomenon, polarity, and source
Names	Confidence, comment, level, emotion, phenomenon, polarity, and source
Spelling mistakes	Confidence, comment, correction, level, emotion, phenomenon, modifier/not, polarity, and source
Emotions	Confidence, comment, emotion (accept, anger, anticipation, anxiety, appreciation, bad, bewilderment, comfort, compassion, confidence, consternation, correct, criticism, disappointment discomfort, disgust, despondency, depression, envy, enmity, excuse, force, fear, fright, good, grief...)

in Table 1, EmotiBlog allows different levels of annotation: at document, sentence and word level.

We were mainly inspired by the MPQA (Wiebe et al. 2005), which underlined the importance to the subjectivity expressed with its intensity, as well as the source and target of the discourse. Using this research as a basis, we conceived a finer-grained model for emotion detection for non-traditional genres. The elements proposed and their attributes are the result of a deep empirical analysis of the texts (blog posts) collected. We analysed the language and linguistic features bloggers employ to express subjectivity concluding that the MPQA elements are not sufficient to capture all the relevant ways of subjectivity expression in blogs. For this reason, the annotation model we designed is fine-grained and it includes elements that are representative of the linguistic phenomena we encountered in our empirical analysis (on the blog corpus we collected) encountered. The list of the EmotiBlog elements and attributes is presented in Table 1.

The elements, together with their corresponding attributes Table 1 presents are described below:

- Objective speech:
  - annotator's confidence (high, medium, low): which is the annotator's degree of certainly
  - comment (if necessary): simple text
  - source (writer): the source of the discourse, in general a name

```
<objective-speech-event source=“w” target=“ratification date of the
Kyoto Protocol”>The third Conference of Parties, more than 160
nations, adopted the Kyoto Protocol in 1997, in large part concerned
that the Intergovernmental Panel of Climate Change predicted that an
average global climate rise in temperature of up to 5.8°C (10.4°F) was
possible by 2100. </objective-speech-event>
```

In some cases, writers make use of rhetoric strategies to include objective information in support of a personal point of view. In order to be able to contemplate these cases we inserted the following elements in the model, initially put forward by Balahur and Montoyo (2008) that are presented below.

For subjective speech, the annotator has to specify the nature of the sentence to be labelled. In case the annotator is labelling a subjective sentence, the first step is to label the entire sentence underlining its nature, using the following tags:

- Phenomenon:
  - annotator’s confidence (high, medium, low): which the annotator’s degree of certain about the label that is assigning
  - comment (if necessary): simple text
  - type: the annotator has to choose among sentence or title

This element explains the nature of the sentence we are labelling. They can be collocation, saying, slang, title, and sentence. A saying is a well-known and often wise statement, which has a meaning different from the simple meanings of the words it contains[7]. A “collocation” is a word or phrase, which is frequently used with another word or phrase, in a way that sounds correct to native speakers, but might not be expected from the individual words’ meanings[8].

- Subjective speech:
  - annotator’s confidence (high, medium, low): which is the annotator’s degree of certain
  - comment (if necessary): simple text
  - polarity: positive or negative
  - level: the polarity degree (high, medium or low)
  - emotion: the subjectivity category (see Table 2)
  - phenomenon: sentence, title, saying, collocation, or slang
  - source: the source of the discourse, generally his name

```
<phenomenon polarity=“negative” degree=“high” category=“phrase”
emotion=“anger” source=“w” target=“opinion about Bush”>It amazes me
how much president Bush gets what he wants</phenomenon>
```

In the case of a subjective sentence, the annotator labels elements which give subjectivity to the discourse. EmotiBlog contemplates the following:

[7] Definition according to the Cambridge Advanced Learner’s Dictionary.

[8] Definition according to the Cambridge Advanced Learner’s Dictionary.

Table 2 Complete list of emotions used in EmotiBlog

Group	Emotions
Criticism	Sarcasm, irony, incorrect, criticism, objection, opposition, scepticism
Happiness	Joy, joke
Support	Accept, correct, good, hope, support, trust, rapture, respect, patience, appreciation, excuse
Importance	Important, interesting, will, justice, longing, anticipation, revenge
Gratitude	Thank
Guilt	Guilt, vexation
Fear	Fear, fright, troubledness, anxiety
Surprise	Surprise, bewilderment, disappointment, consternation
Anger	Rage, hatred, enmity, wrath, force, anger, revendication
Envy	Envy, rivalry, jealousy
Indifference	Unimportant, yield, sluggishness
Pity	Compassion, shame, grief
Pain	Sadness, lament, remorse, mourning, depression, despondency
Shyness	Timidity
Bad	Bad, malice, disgust, greed

- Adjective/Adverbs:
  - annotator’s confidence (high, medium, low): which is the annotator’s degree of certain
  - comment (if necessary): simple text
  - polarity: positive or negative
  - level: the polarity degree (high, medium, low)
  - emotion: the subjectivity category (see Table 2)
  - phenomenon: sentence, title, saying, collocation, slang or other language
  - source: the source of the discourse, generally his name
  - modifier (yes/no): if this element is a modifier of the following word

```
The Kyoto Protocol is a <adjective target=“Bush’s opinion about the
Kyoto Protocol” phenomenon=“collocation” source=“Bush” ismodifier=
“yes”>thinly-veiled </adjective> attack on industrialized nations
```

- Verbs/Noun: annotator’s confidence, comment, level, emotion, phenomenon, polarity, mode, and source.

```
A recollection of all the reasons why we <verb phenomenon=“idioms”
degree=“high” source=“w” polarity=“negative” tense=“Indicative”
target=“people’s opinion about Bush’s decision” emotion=“bad”>hate
</verb> Bush.
```

- Anaphora:
  - annotator’s confidence (high, medium, low): which is the annotator’s degree of certain
  - comment (if necessary): simple text
  - type: pronominal, adverbial, ellipsis

  - source: the source of the discourse, generally his name
  - antecedent: the antecedent of the anaphora, generally a name or pronoun.

This element underlines the coreference phenomena at a cross-post level. Usually, blog posts can be a multi-party conversation and thus this element can be useful to follow the discourse in case of multiple posts or when it is interrupted with other posts about a subtopic or related topic.

```
From the meeting there has been no improvement for solving
<anaphora source=“discurso de Fidel Castro” pronominal=“dd”>this
situation </anaphora> which now is serious
```

- Capital Letter:
  - annotator’s confidence (high, medium, low): which is the annotator’s degree of certain
  - comment (if necessary): simple text
  - polarity: positive or negative
  - level: the polarity degree (high, medium, low)
  - emotion: the subjectivity category (see Table 2)
  - phenomenon: sentence, title, saying, collocation, slang or other language
  - source: the source of the discourse, generally his name

Bloggers generally produce genuine and spontaneous language and it is frequent to find complete words that are meant as a sign of a special user attitude.

```
I <capital letter=“phrase” source=“w” target=”Bush government”
polarity=”negative” level=”high” emotion=”criticism”>I HATE
</capital letter>the choices of this government
```

- Punctuation:
  - annotator’s confidence (high, medium, low): which is the annotator’s degree of certain
  - comment (if necessary): simple text
  - polarity: positive or negative
  - level: the polarity degree (high, medium, low)
  - emotion: the subjectivity category (see Table 2)
  - phenomenon: sentence, title, saying, collocation, slang or other language
  - source: the source of the discourse, generally his name

This phenomenon is similar to the previous one. An exceptional use of punctuation could mean a special feeling on the part of the writer.

```
Our environment will not survive for our sons <punctuation=“phrase”
source=“w” target=“environment” polarity=“negative” level=“high”
emotion=“criticism”>!!!!!!</punctuation>
```

- Other languages:
  - annotator’s confidence (high, medium, low): which is the annotator’s degree of certainly
  - comment (if necessary): simple text
  - polarity: positive or negative
  - level: the polarity degree (high, medium, low)

  - emotion: the subjectivity category (see Table 2)
  - phenomenon: sentence, title, saying, collocation, slang or other language
  - source: the source of the discourse, generally his name
  - Latin/English: if the expression is in Latin or English

This situation is a <phenomenon="other language" **polarity="negative" degree="high" category="phrase" emotion="anger" source="w" target=" climate change** >déjà vù</other language>…

- Emotions:
  - annotator's confidence (high, medium low): which is the annotator's degree of certainly
  - comment (if necessary): simple text
  - List of emotions: accept, anger, anticipation, anxiety, etc. (See Table 2).

Regarding the list of emotions employed, we grouped all sentiments into subgroups to facilitate the evaluation process. Emotions of the same subgroup will have less impact of disagreement on the inter-annotator agreement. In order to make this subdivision proper and effective, we were inspired by Scherer (2005) who creates an alternative dimensional structure of the semantic space for emotions. The graph below represents the mapping of the terminology Russell (1983) uses for his claim of an emotion circumflex in two-dimensional valence by activity/arousal space (upper-case terms). As can be noted in Fig. 1, 4 axes divide the circle. Moreover, Scherer distinguishes between positive and negative and between active and passive sentiments. Furthermore, emotions are grouped between obstructive and conductive, and finally between high power control and low power control. Using this classification, we grouped sentiments into positive and negative, high/low power control, obstructive/conductive and active/passive. Further on, we distributed the sentiments within our list into the Scherer slots, creating other smaller categories included in the abovementioned general ones.

We started from this classification, and subsequently created smaller sub-categories. The complete list of emotions we used for the corpus annotation is shown in Table 2.

4.2 Corpus collection

Following the MPQA (Wiebe et al. 2005) topics and structure, we collected texts about three topics of interest (two coinciding with the topics collected for the MPQA) in order to delimitate the collection process. The first topic is the Kyoto Protocol, the second is the election in Zimbabwe and thirdly texts about the last USA elections in 2008. We selected these subjects due to the fact that the MPQA corpus contained texts about these topics and thus the selection of the same issues allows a comparable study.

The blogs we chose have distinctive features, different from traditional texts. In fact, people writing a post may legitimately use informal language, colloquialism, emoticons, etc. to express their opinions. Furthermore, we can frequently find a mix of sources in the same post, with authors mentioning facts and giving their personal opinion about them. As we can deduce, the source and target detection represents one of the most complex tasks. For each of the abovementioned topics, we

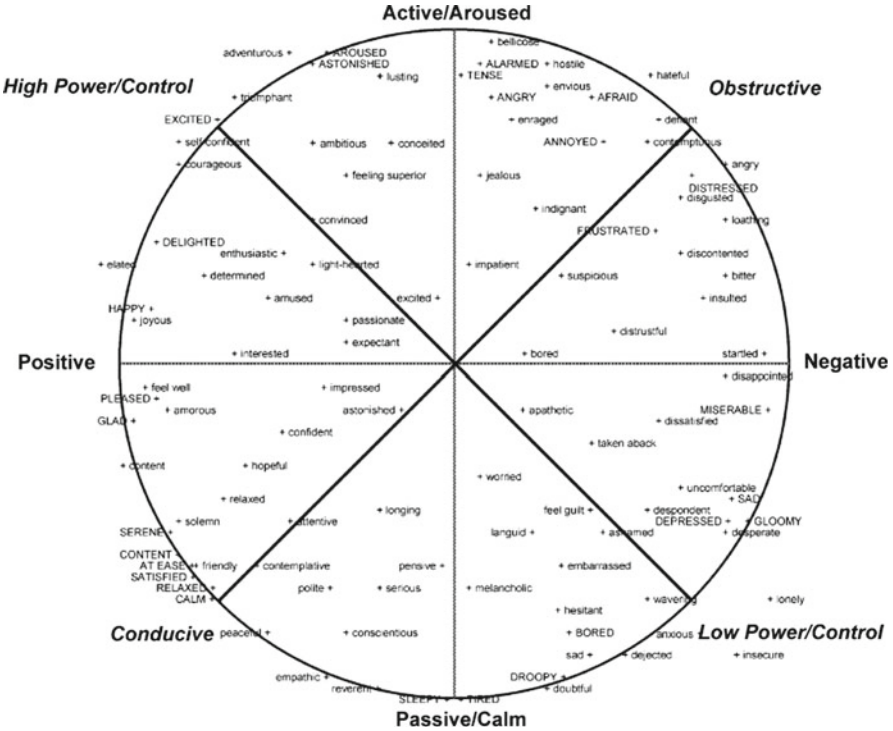

Fig. 1 Alternative dimensional structures of the semantic space for emotions. Image reproduced from: Scherer, K. R. (2005). What are emotions? And how can they be measured?". Social Science Information., 44(4), 693–727. Freely available online. Pag. 720

gathered 100 texts, summing up to a total of 30,000 words approximately for each language.

4.3 Inter-annotator agreement

We measured the inter-annotator agreement regarding the different elements and attributes of the annotation scheme. Two annotators (A and E) independently labelled (100 texts), a total of 30,000 words in Spanish about the ratification of the Kyoto Protocol.

Our objective was to verify that all annotators agree on which expressions should be marked. The evaluation process was extremely complex as when annotators identify the same expression in the text, they could differ in their marking of the expression boundary, as well as in other aspects, such as the emotion intensity or emotion itself. Apart from this, there was no guarantee that the annotators will identify the same set of expressions.

4.3.1 Measures used

The measure we wanted to employ for the inter-annotator agreement evaluation was the kappa value (when statistic classes are present), and the observed agreement (when non

Table 3 Inter-annotator agreement results for Spanish

Annotation	a	b	a||b	b||a	average
Noun	A	E	0.783	0.7532	0.765
Adjective	A	E	0.782	0.613	0.681
Verb	A	E	0.863	0.742	0.802
Adverb	A	E	0.831	0.764	0.794
Preposition	A	E	0.862	0.672	0.763
Punctuation	A	E	0.784	0.891	0.832
Capital letter	A	E	0.663	1	0.831
Spelling mist.	A	E	0.814	0.773	0.784
English	A	E	0.273	1	0.632
Latin	A	E	0.662	0.662	0.661
Phrase	A	E	0.524	0.662	0.592
Objective	A	E	0.762	0.734	0.745
Total average	0.740				

statistic classes are present). Kappa is computed according to Cohen method (Cohen 1960; Carletta 1996; Artstein and Poesio 2008):

$$Kappa = \frac{obscrved\ proportion\ of\ agreement - chance\ expected\ proportion\ of\ agreement}{1 - chance\ expected\ proportion\ of\ agreement}$$

Upon application, it was apparent that this measure was not the appropriate for our corpus, as it does not allow for measuring the inter-annotator agreement for each element and its corresponding attributes, given that the boundaries we considered in the annotation model were highly variable. Thus we decided to use the following measure:

$$agr(a||b) = \frac{|A\ matching\ B|}{|A|}$$

We evaluated the elements and attributes, which are part of the model, and they are listed below. The criteria used for matching is that the annotator's labelling should overlap and have the same annotated orientation, intensity and pertain to the same emotion category:

Discussion Looking at the results of Table 3, we can say that the elements with the best performance are *capital letter* and *punctuation* and the ones, which obtained the lowest average are *phrase* and *English*. With the remainder of the elements, we obtained results that are between 0.76 and 0.80. To gain a better understanding regarding the importance of each element of the model we will perform some feature selection experiments. We measured the relevance of each one of the annotated elements and checked if a fine-grained annotation model can be employed for supervised learning.

5 Feature selection experiments

We evaluated the consistency of our model and checked if the extended list of elements has a positive effect or not. To this end, we carried out a set of preliminary Machine Learning experiments in the Spanish, English and Italian part of our annotated corpus. It is worth mentioning that the Machine Learning techniques we employ are the most widely used. The algorithms we used are SVM and Multinomial Naïve Bayes (MNB). We decided to use them and not carrying out a deep and more detailed comparison of different Machine Learning approaches because in this way we can compare our results with similar research carried our in Sentiment Analysis (see Sect. 2). Moreover the techniques are robust, because their extended usage, and also their versatility. Using such approaches we believe our work will be more comparable. However in the future we will concentrate on carrying out a specific study focused on a deep Machine Learning experimentation and the comparison of the performance of different techniques. In this case, we aim to demonstrate that the validity of the annotation with EmotiBlog and the techniques we employ are the most adequate. We carried out a set of preliminary Machine Learning experiments with the Spanish part of our annotated corpus to evaluate the consistency of our model and the effect on the extended list of elements. We used each sentence as an individual instance (a sentence—The sentences have been extracted using the annotations of the corpus) in a classification task: its terms are the features (in our system each feature represents a word or a set of them, and can be used as they were found in the text or using their stem, depending on the experiment. They have been extracted splitting each sentence into individual words by means of a tokeniser) and its polarity is the category. The sentences and their polarity have been extracted from the corpus annotations, but the terms have been obtained splitting the sentences into words. The difference between each experiment consists on the set of terms used. Due to the size of the corpus we evaluated these experiments employing the well-known 10-fold cross-validation method. We employed the Weka[9] implementation of the widely used SVM algorithms (Gamon 2004; Vapnik 1995) and MNB (Lewis and Gale 1994; Sebastiani 2002) algorithms. We chose these algorithms because much of the research in SA demonstrated their effectiveness for this type of tasks: the first due to its robustness against noise and the second because of its simplicity and efficiency. We also used stemming techniques in some experiments (the Snowball[10] implementation for the Spanish language).

As a starting point, we extracted a bag of words from the corpus without taking into account the annotated elements. We will use this approach as baseline because it is the simplest one and it does not take into account fine-grained annotations. Note that Spanish is a language that uses accents, but users in informal registers, such as blogs often neglect to put them. For this reason we decided to eliminate them in all our experiments, as well as capital letters.

In some experiments we also eliminated stop words and negation words. In the first instance although they do not contain semantic information (Yang and Pedersen

[9] http://www.cs.waikato.ac.nz/ml/weka/.

[10] http://snowball.tartarus.org/.

Table 4 Results in terms of accuracy and F-measure for the MNB and SVM Machine Learning algorithms for each option in Spanish

Combinations	ML Features	MNB		SVM	
		Accuracy	F-measure	Accuracy	F-measure
BL	941	0.647	0.592	0.685	0.644
RS+RN	877	0.566	0.477	0.654	0.610
RS	878	0.532	0.420	0.625	0.572
SC	875	0.588	0.511	0.663	0.620
ST	819	0.672	0.625	0.714	0.683
RS+RN+ST	764	0.594	0.516	0.661	0.618
RS+ST	765	0.622	0.556	0.689	0.652
SC+ST	781	0.617	0.554	0.694	0.659

1997), was to check how their inclusion or exclusion could affect the results. In the second case, this will reverse the polarity of a sentence making it and important feature to take into account. Although the polarity changes depending on the position of these words inside the sentence, in these preliminary experiments the aim was to evaluate how much their presence could change the results.

Moreover, there are some groups of words that have a different meaning when grouped together, such as sayings and collocations. They are correspondingly annotated in the corpus. Thus, in some of our experiments we considered those groups as single features to check if these elements have an important impact in subjectivity detection. The abbreviations used in our experiments are the following: BL (baseline), RS (removing stopwords), RN (removing negation), SC (sayings and collocations as single features) and ST (stemming).

Discussion Table 4 presents the results obtained in our experiments. They show that our starting point reaches an accuracy of 68.5% and an F-measure of 64.4% using SVM. With MNB the results are about 6% and 10% worse in terms of accuracy and F-measure respectively. The best results were obtained using stemming techniques achieving an accuracy of 71.4% and an F-measure of 68.3%. In fact the use of stemmer improves the results in every experiment we performed with this approach. As we can also see in Table 4 the use of a full list of stopwords negatively influences the results. This is probably due to the fact that some stopwords can be useful when detecting the polarity of a sentence. An additional aspect to underline is that negation is an important feature due to its inclusion raising the results between 4% and 7% with respect to the experiment without this feature.

Figure 2 represents the visualisation of the data obtained in the Table 4. As we explained in the model element description, sayings and collocations (SC) are annotated as a whole. Generally they are composed of different words, but their meaning is not given by the normal sum of the meanings of the words that compose them. They are considered unique expressions of subjectivity and we did not pre processed them. The consequence is that, since we are using a limited corpus in terms of size, the number of occurrences of exactly the same SC is quite low and for this reason

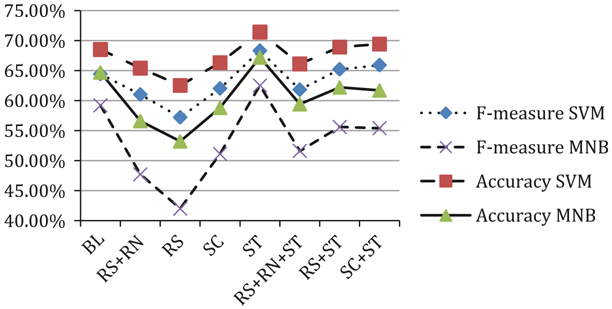

Fig. 2 Comparative results in terms of accuracy and F-measure for the MNB and SVM Machine Learning algorithms for each option in Spanish

hard to classify. We obtained a better BL because it has been calculated on the single words that have been annotated and previously pre processed (POS, lemmatiser, etc) in order to have a bigger data set. In the case of SC this was not possible since each one of them has been considered unique and not pre processed to increase the amount of the dataset. However in our future experiments we will also try to include a pre processing stage also for SC to be able to detect a higher number of occurrences and thus improve their automatic classification performance.

These initial results encourage us to continue improving the model. The first explanation for the performance obtained could be the size of the corpus. In fact, we worked with a small sample collection of texts, and as a consequence a significant number of examples of positive, negative and neutral sentences was extremely different.

5.1 Feature selection

In order to evaluate impact of the features described in the last section, we also conducted experiments using feature selection using dimensionality reduction, and more specifically, the Information Gain (IG) (Lewis and Gale 1994) algorithm. The goal of this dimensionality reduction is to obtain the best features for the polarity classification to improve and simplify the classification task. In these experiments we employ the global feature selection, which measures the relevance of each term of the corpus taking into account the three categories of polarity. The results are shown in Table 5 and are based on the baseline feature set.

Discussion Analysing the results in Table 5, we observed that the features selected by the algorithm are not the opinionated ones. This occurs because the corpus we are using is unbalanced (the categories do not have a similar number of sentences, especially for the positive polarity) and the non-neutral polarities (especially positive ones) have fewer instances, so the frequency of the positive terms is decreased noticeably. This makes the feature selection algorithm remove these kind of terms as a first step. This can be considered to be the reason why the application of feature selection does not improve our previous evaluation.

After having carried out the experiments, we also evaluated the impact of each element for the classification system, thus employing the annotation. In the Table 6

Table 5 Results of the global feature selection for Spanish

Percentage of selected features	ML features	MNB		SVM	
		Accuracy	F-measure	Accuracy	F-measure
80	752	0.487	0.322	0.496	0.341
85	799	0.485	0.320	0.525	0.399
90	846	0.488	0.325	0.539	0.429
95	893	0.510	0.371	0.598	0.528
99	931	0.564	0.470	0.650	0.602

Table 6 Impact of the EmotiBlog elements on the system for Spanish

Element	Effect%
Verb	2.998
Phrase	2.664
Adjective	2.244
Noun	1.756
Preposition	0.338
Pronoun	−0.323
Onomatopoeic	−0.784
Adverb	−0.914

we show the results of the measurement of the impact of the annotated elements using the EmotiBlog Model.

Discussion Table 6 shows the percentage of improvement for each element, thus the proportion of experiments that have been improved by including each feature (their impact). It has been calculated by taking the number of experiments improved by adding an element, and dividing this by the number of experiments that did not see an improvement. The majority of the elements have a beneficial effect on the system, but the pronoun, onomatopoeic and adverb have lower improvements. This measurement of the impact of each element is very useful because, depending on our needs, we may decide which features to include for the Machine Learning model training.

5.2 Evaluation after the reclassification

After having analysed the previous experiments, we labelled the English part of the blog corpus about the Kyoto Protocol and we refined the model according to our previous findings and subsequently we carried out the cross-fold evaluation for the categories subjective and objective speech, but also positive and negative polarity in case of subjective discourse. Table 7 shows the results obtained for English after the reclassification of our corpus in terms of precision, recall and F-measure, whilst Table 8 illustrates the impact of the EmotiBlog elements.

Discussion The results of the reclassification are more satisfactory using the refined version of EmotiBlog (if compared with Table 5), improved after having calculated

Table 7 Reclassification results for English

	Precision%	Recall%	F-measure%
Subj.	92.2	75.4	83
Obj.	75.6	72	73.8
Posit.	72.1	82	76.7
Neg.	92.4	92.4	92.4

Table 8 Effect of the EmotiBlog elements on the system for English

Element	Effect %
Phrase	2.951
Verb	0.560
Pronoun	0.337
Adjective	0.221
Noun	−0.177
Onomatopoeic	−0.278
Preposition	−0.283
Adverb	−0.525

the impact of the elements in Spanish. It represents the proportion of experiments that have been improved by including each feature. We can deduce that a deep analysis of the errors emerged from the evaluation process is fundamental to increase the performance of our tools.

Table 8 shows that, concerning English, the elements that have beneficial effect on the system are less than for Spanish language.

In order to show that EmotiBlog is indeed a useful resource for all the languages contemplated (English, Italian and Spanish), in the subsequent experiments we tested the ability to detect and classify opinion in blog posts in Italian. Coherent with the previous tests for English and Spanish, we used the annotation of the EmotiBlog posts in Italian. Subsequently, each of the annotated elements was extracted from the corpus. In the following experiments we tested, through a ten-fold cross-validation, if we are able to correctly classify the sentences in the blog posts in Italian. In order to achieve this, we computed the Lesk (Salton and Lesk 1971) similarity score between each of the sentences in the corpus and each of the annotated elements. The Lesk similarity of two concepts is defined as a function of the overlap between the corresponding definitions, as provided by a dictionary (Lesk 1986). Thus, each sentence was represented as a vector of features, each of which corresponding to the similarity scores to all annotated elements. Subsequently, we performed two types of classification. The first one was geared towards assessing the accuracy of distinction among subjective and objective sentences. The second classification was conducted among the 3 considered classes of sentiment polarity: positive, negative and neutral. The result of the cross-validation is presented in Table 9.

As it can be seen in Table 9, the results of the experiments with the part of the EmotiBlog corpus in Italian classification stay in line with the previous evaluations. Nevertheless, a slight drop in performance when classifying the blog sentences can

Table 9 Cross-validation of the sentence-level opinion classification for Italian

	Precision %	Recall %	F1 %
Subj.	73.1	56.3	63.6
Obj.	86.1	67.5	75.4
Posit.	71.2	73.2	72.2
Neg.	89.4	95.1	92.2

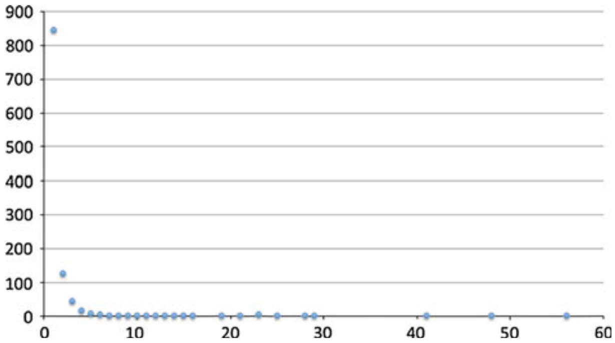

Fig. 3 Feature vector scarcity for Italian

be observed, both among subjective and objective, as well as among the three classes of sentiment polarity. The explanation for this phenomenon is the higher language variability in this subset of the corpus, which was observable from the feature vectors' scarcity, presented in Fig. 3.

6 Opinion Mining and polarity classification experiments

In order to evaluate the appropriateness of the EmotiBlog annotation scheme and to check if the fine-grained level it aims for a positive impact on the performance of the systems employing it for training we performed several experiments (Balahur et al. 2009c; Balahur and Montoyo 2009). Given that (a) EmotiBlog contains annotations for individual words, as well as for multi-word expressions and at a sentence level, and (b) they are labelled with polarity, but also emotion, our experiments show how the annotated elements can be used as training for Opinion Mining and Polarity Classification tasks, as well as for emotion detection. Moreover, taking into consideration the fact that EmotiBlog labels the intensity level of the annotated elements, we performed a brief experiment on determining the sentiment intensity, measured on a three-level scale: low, medium and high. In order to perform these evaluations, we chose three different corpora. The first one is a small collection of quotes (reported speech) from newspaper articles presented in (Balahur et al. 2010d), enriched with the manual fine-grained annotation of EmotiBlog[11]; the second is the collection of newspaper titles in the test set of the SemEval 2007 task number 14—Affective Text. Finally, the third one

[11] Freely available on request to the authors.

is a corpus of self-reported emotional response—ISEAR. The intensity classification task is evaluated only on the second corpus, given that it is the only one in which scores between −100 and 0 and 0 and 100, respectively, are given for the polarity of the titles.

6.1 Creation of training models

For the Opinion Mining and polarity classification task, we first extracted the Named Entities (NE) contained in the annotations using Lingpipe[12] and united through a "_" all the tokens pertaining to the NE. All the annotations of punctuation signs that had a specific meaning together were also united under a single punctuation sign. Subsequently, we processed the annotated data, using Minipar. We compute, for each word in a sentence, a series of features (some of these features are used in (Choi et al. 2005):

- the part of speech (POS)
- capitalization (if all letters are in capitals, if only the first letter is in capitals, and if it is a NE or not)
- opinionatedness/intensity/emotion—if the word is annotated as opinion word, its polarity, i.e. 1 and −1 if the word is positive or negative, respectively and 0 if it is not an opinion word, its intensity (1.2 or 3) and 0 if it is not a subjective word, its emotion (if it has, none otherwise)
- syntactic relatedness with another opinion word—if it is directly dependent upon and opinion word or modifier (0 or 1), plus the polarity/intensity and emotion of this word (0 for all the components otherwise)
- role in 2-word, 3-word and 4-word annotations: opinionatedness, intensity and emotion of the other words contained in the annotation, direct dependency relations with them if they exist and 0 otherwise.

We computed the length of the longest sentence in EmotiBlog. The feature vector for each sentence contains the feature vectors of each of its component words, and 0s was assigned to the corresponding feature vectors of the missing words (which the current sentence has less than the longest annotated sentence). Finally, to each sentence as a feature we added binary features for subjectivity and polarity, the value corresponding to the intensity of opinion and the general emotion. These feature vectors are fed into the Weka[13] SVM SMO Machine Learning algorithm and a model is created (Emoti Blog I). A second model (EmotiBlog II) is created by adding to the collection of single opinion and emotion words annotated in EmotiBlog, the Opinion Finder lexicon[14] and the opinion words found in MicroWordNet[15], the General Inquirer[16] resource and WordNet Affect (Strapparava and Valitutti 2004).

[12] http://alias-i.com/lingpipe/.

[13] www.cs.waikato.ac.nz/ml/weka/.

[14] www.cs.pitt.edu/mpqa/opinionfinderrelease/.

[15] www-3.unipv.it/wnop/.

[16] www.wjh.harvard.edu/~inquirer/.

Table 10 Results for polarity and intensity classification using the models built from the EmotiBlog annotations

Test corpus	Eval. type	Precision %	Recall %	F1 %
JRC quotes I	Polarity	32.2	54	40.3
	Intensity	36	53.2	42.9
JRC quotes II	Polarity	36.4	51	42.9
	Intensity	38.7	57.8	46.4
SemEval I	Polarity	38.6	51.3	44
	Intensity	37.4	50.9	43.1
SemEval II	Polarity	35.8	58.7	44.5
	Intensity	32.3	50.4	39.4

6.2 Evaluation of models on test sets

In order to evaluate the performance of the models extracted from the features of the annotations in EmotiBlog, we performed different tests (Balahur and Montoyo 2010). The first regarded the evaluation of the polarity and intensity classification tasks using EmotiBlog I and II constructed models on two test sets—the JRC quotes collection and the SemEval 2007 Task Number 14 test set. Since the quotes often contain more than one sentence, we considered the polarity and intensity of the entire quote as the most frequent result in each class, corresponding to its constituent sentences. Also, given the fact that the SemEval Affective Text headlines were given intensity values between -100 and 100, we mapped the values contained in the Gold Standard of the task into three categories: $[-100, -67]$ is high (value 3 in intensity) and negative (value -1 in polarity), $[-66, 34]$ medium negative and $[-33 - 1]$ is low negative. The values between [1 and 100] are mapped in the same manner to the positive category. 0 was considered objective (i.e. it does not contain any sentiment) and was its intensity will be 0. The results are presented in Table 10 (the values I and II correspond to the models EmotiBlog I and EmotiBlog II):

Discussion The results presented in Table 10 show a significantly high improvement over the results obtained in the SemEval task in 2007, in which the SWAT system obtained: Prec. 45.71, Rec. 3.42, F1 6.36, while the UPAR7 system obtained: Prec. 57.54, Rec. 8.78, F1 15.24.

In this competition, fine-grained evaluations were conducted using the Pearson measure of correlation between the system scores and the gold standard scores, averaged over all the headlines in the data set. In the coarse-grained evaluation each emotion was mapped to a 0/1 classification ($0 = [0, 50)$, $1 = [50, 100]$), and each valence was mapped to a $-1/0/1$ classification ($-1 = [-100, -50], 0 = (-50, 50), 1 = [50, 100]$). For the coarse-grained evaluations, we calculated accuracy, precision, and recall. On the one hand the accuracy is calculated taking into account all the possible classes, and thus it can be artificially high in case of working with unbalanced datasets (as some of the emotions are, due to the high number of neutral headlines). On the other hand, the precision and recall figures exclude the neutral annotations. The table

Table 11 Results for emotion classification using the models built from the EmotiBlog annotations

Test corpus	Eval. type	Precision %	Recall %	F1 %
JRC quotes EmotiBlog Model I	Emotions	24.7	15	18.7
JRC quotes EmotiBlog Model II	Emotions	33.6	19	24.2
SemEval EmotiBlog Model I	Emotions	29	18.9	22.9
SemEval EmotiBlog Model II	Emotions	33	18.4	22.6
ISEAR EmotiBlog Model I	Emotions	22.3	15	17.9
ISEAR EmotiBlog Model II	Emotions	25.6	17.8	21

show both the fine-grained Pearson correlation measure and the coarse-grained accuracy, precision and recall figures. Moreover, R represents the margin within which the result is correct.

Our improvements on the Semeval competition is explainable, on the one hand, by the fact that systems performing the opinion task did not have at their disposal the lexical resources for opinion employed in the EmotiBlog II model, but also because they did not use apply supervised learning on a corpus comparable to EmotiBlog (as seen from the results obtained when using solely the EmotiBlog I corpus). Compared to the NTCIR 8 Multilingual Analysis Task 2010, we obtained significant improvements in precision, with a recall that is comparable to most of the participating systems. In the second experiment, we tested the performance of emotion classification using the two models built using EmotiBlog on the three corpora—JRC quotes, SemEval 2007 Task No.14 test set and the ISEAR corpus.

The "JRC quotes" have also been labelled using EmotiBlog. However, the remaining two are labelled with Eckman's set of emotions—6 in the case of the SemEval data (joy, surprise, anger, fear, sadness, disgust) and 7 in ISEAR (joy, sadness, anger, fear, guilt, shame, disgust). Moreover, the SemEval data contains more than one emotion per title in the Gold Standard, therefore we consider as correct any of the classifications containing one of them. R is the margin within which the result is correct. This margin has been proposed in the framework of the SEMEVAL competition, thus we decided to maintain it because we based our experiments in already established ways of evaluating the performance of our system.

In order to unify the results and obtain comparable evaluations, we assessed the performance of the system using the alternative dimensional structures defined in Fig. 1. Those not overlapping with the category of any of the 8 different emotions in SemEval and ISEAR are considered as "Other" and are not included either in the training, nor test set. The results of our evaluation are presented in Table 11. Again, the values I and II correspond to the models EmotiBlog I and II. The "Emotions" category contains the following emotions: joy, sadness, anger, fear, guilt, shame, disgust, and surprise.

In our case, the best results for emotion detection were obtained for the "anger" category, where the precision was around 35 percent, for a recall of 19 percent, improving the results obtained by the Semeval Participating systems presented in Table 12. The lowest results obtained were for the ISEAR category of "shame", where precision was around 12 percent, with a recall of 15 percent. We believe this is due to

Table 12 SEMEVAL System results for emotion annotations

Emotions and systems	Acc.	Prec.	Rec.	F1
Anger				
Swat	92.10	12.00	5.00	7.06
Ua	86.40	12.74	21.6	16.03
Upar7	93.60	16.67	1.66	3.02
Disgust				
Swat	97.20	0.00	0.00	–
Ua	97.30	0.00	0.00	–
Upar7	95.30	0.00	0.00	–
Fear				
Swat	84.80	25.00	14.40	18.27
Ua	75.30	16.23	26.27	20.06
Upar7	87.90	33.33	2.54	4.72
Joy				
Swat	80.60	35.41	9.44	14.91
Ua	81.80	40.00	2.22	4.21
Upar7	82.20	54.54	6.66	11.87
Sadness				
Swat	87.70	32.50	11.92	17.44
Ua	88.90	25.00	0.91	1.76
Upar7	89.00	48.97	22.02	30.38
Surprise				
Swat	89.10	11.86	10.93	11.78
Ua	84.60	13.70	16.56	15.00
Upar7	88.60	12.12	1.25	2.27

the fact that the latter emotion is a combination of more complex affective states and it can be easily misclassified to other categories of emotion. Moreover, from the analysis performed on the errors, we realized that many of the affective phenomena presented were more explicit in the case of texts expressing strong emotions such as “joy” and “anger”, and were mostly related to common-sense interpretation of the facts presented in the weaker ones. As it can be seen in Table 3, results for the texts pertaining to the news category obtain better results, most of all news headlines. This is due to the fact that such texts, although they contain few words, have a more direct and stronger emotional charge than direct speech (which may be biased by the need to be diplomatic, find the most suitable words etc.). Finally, the error analysis showed that emotion that is directly reported by the persons experiencing it is more “hidden”, in the use of words carrying special meaning or related to general human experience. This fact makes emotion detection in such texts a more complex task. Nevertheless, the results in all corpora are comparable, showing that the approach is robust enough to handle different text types. All in all, the results obtained using the fine and coarse-grained annotations in EmotiBlog increased the performance of emotion detection as compared to the systems in the SemEval competition.

6.3 Opinion Mining in real time

Apart from evaluating the model for the OM task, we also evaluated OM in real time. This takes into account the inferred information on the polarity of the sentiment, offering the corresponding feedback (Balahur et al. 2009c). We used as an extrinsic source a set of sentences on "recycling" (a subject directly related to environmental issues and which shares large set of vocabulary) and our annotated corpus of blog posts about the Kyoto Protocol in Spanish. We used the labelled elements to train our OM system and after this process, we classified new sentences regarding "recycling". Furthermore, we manually created a collection of 150 sentences on recycling, 50 for each of the positive, negative and neutral categories. The purpose of the first experiment is to demonstrate that the corpus is a robust resource, and that our annotation could be useful for the training of our OM system.

6.3.1 Cross-fold evaluation of the annotation

As we presented in Sect. 4, the inter-annotator agreement obtained is 0.59 and 0.745 respectively for subjective and objective discourse. However, in order to carry out this evaluation, we considered only the sentences upon which we agreed on the phrases whose annotation length was not more than four tokens of the type noun, verb, adverb or adjective. Each sentence is POS-Tagged and lemmatized using FreeLing[17]. Further on, we represented each sentence as a feature vector, composed by unigram features containing the positive and negative categories of nouns, verbs, adverbs, adjectives, prepositions and punctuation signs (having 1 in the corresponding position of the feature vector for the words contained and 0 otherwise), the number of bigrams/trigrams and 4 g overlapping with each of the phrases we annotated as positive and negative or objective, respectively, and finally the overall similarity given by the number of overlapping words with each of the positive, negative or objective phrases from the corpus, normalized by the length of the given phrase. Thus, each of the sentences to be classified was compared to each of the annotations in the EmotiBlog corpus, and the similarity score (computed with the Lesk distance) was used as a feature in the SVM classifier. Subsequently, we tested our method with the sentence classification among subjective and objective, for which the vectors contain "subjective" or "objective" as final values. Next we performed a classification of subjective sentences into positive and negative, for which case the classification vectors contain the values "positive" and "negative". Finally, we perform a ten-fold cross validation of the corpus for each of the two steps. The results, taking into account precision and recall are presented in Table 13. As a baseline we have a 33% of possibility of good classification. In fact we use 3 classes (positive, negative, objective) with the same number of examples, thus we have a one in three chance of classifying in the most adequate way.

[17] http://www.lsi.upc.edu/~nlp/freeling/.

Table 13 Classification using ten fold cross validation

	Precision	Recall	F1
Subj.	0.988	0.632	0.771
Obj.	0.682	0.892	0.773
Posit.	0.799	0.511	0.623
Neg.	0.892	0.969	0.929

Table 14 Classification results using all n g

	Precision	Recall	F1
Subj.	0.977	0.619	0.758
Obj.	0.442	0.954	0.604
Posit.	0.881	0.769	0.821
Negat.	0.923	0.962	0.942

In order to carry out the classification using all n-gram features of our test data, we ran FreeLing[18] on the set of positive, negative and neutral sentences on recycling to lemmatize and tag each word on part of speech. Subsequently, we represented sentences as feature vectors in the same manner as in the first conducted experiment, and we carried out two experiments on this data. The first one aims at training a SVM classifier on the corpus phrases pertaining to the "subjective" versus "objective" categories, and to test it on answers for the recycling topic pertaining to the positive or negative versus neutral categories. The second consists in classifying the instances according to three polarity classes: positive, negative and neutral. The results of the two experiments are presented in Table 14 (Subj/Obj represents the classification of phrases among subjective and objective and the Pos/Neg the classification of phrases according to their being positive or negative. The classification was performed on the 150 sentences on recycling new examples.

In these experiments, we test the importance of the annotating affect in texts at the token level. Thanks to our corpus, we have a large number of nouns, verbs, adverbs and adjectives annotated as positive or negative and also at the emotion level. We used these words when classifying examples using n g, with n ranging from 1 to 4. In order to test their importance, we removed the vector components accounting for their presence in the feature vectors and re-classified, both at the level of objective versus subjective, as well as at the positive, negative, neutral level. In Table 14, we can see the results obtained.

As we can see, removing single words with their associated polarities from the training data resulted in lower scores. Therefore, fine-grained annotation helps at the time of training the opinion mining system and is well worth the effort (Table 15).

[18] http://nlp.lsi.upc.edu/freeling/.

Table 15 Classification results using n g, $n > 2$

	Precision	Recall	F1
Subj.	0.933	0.601	0.731
Obj.	0.432	0.743	0.546
Posit.	0.834	0.642	0.726
Negat.	0.902	0.910	0.906
Neu.	0.901	0.963	0.931

7 Conclusions and future work

In this paper we presented how we built up a fine-grained annotation scheme for detecting emotion in non-traditional textual genres and how we collected a corpus of blog posts about three topics in three languages (Spanish, English and Italian). In order to demonstrate that the model is not ambiguous and it does not present significant difficulties for annotators we calculated the inter-annotator agreement. After that, with the objective of testing the impact of the EmotiBlog elements we carried out a series of Machine Learning experiments based on feature selection techniques and concluded that the totality of elements have a beneficial impact on the system.

After having performed these evaluations, we tested the EmotiBlog validity for its application to standard corpora and if it is domain-independent. We tested our model, with the ISEAR (Scherer and Wallbott 1997), NTCIR 8 MOAT[19] and SEMEVAL 2007 Task 14 corpora (Strapparava and Milhacea 2007) obtaining promising results. Last but not least, we applied the annotation and the consequential Machine Learning techniques to the OM task due to its high level of applicability and usefulness in the present Information Society.

As seen in the previous sections and in the description of the experiments carried out we can deduce that, due to its special structure, EmotiBlog can be employed either for basic tasks of sentiment polarity classification, but also for emotion detection, either in very fine-grained categories or as psychology-based emotion classes. We also proved that given its structure, EmotiBlog could further be employed for the treatment of the subjective data in different languages and domains.

We also addressed the presence of "copy and paste" items from news articles or other blogs, as well as the frequent presence of quotes. To solve this possible ambiguity, we included the annotation of both the directly indicated source, as well as the anaphoric references at cross-document level. We performed several experiments on three different corpora, aimed at detecting and classifying both the opinion as well as the expressions of emotion contained within; we showed that the fine and coarse-grained levels of annotation that EmotiBlog contains offer relevant information on the structure of affective texts, leading to an improvement of the performance of systems that are based on models trained on it.

Moreover, we carried out supervised learning experiments and a feature selection process with the objective of checking if the elements of the model have a beneficial

[19] http://research.nii.ac.jp/ntcir/ntcir-ws8/permission/ntcir8xinhua-nyt-moat.html.

impact on systems. Despite of the small size of the corpus and to high granularity of the model the results obtained were promising.

From the OM experiments we performed, we can also deduce that combining the features extracted from the EmotiBlog fine and coarse-grained annotation help to balance the results obtained for precision and recall. The use of additional resources that contain opinion words has let to an increase in the recall of the system, at the cost of a slight drop in precision, which proves that the approach is robust enough so that additional knowledge sources can be added. Although the corpus is small, the results obtained show that the phenomena it captures are relevant for the OM task, not only for the blogosphere, but also for other types of text (newspaper articles, self-reported affect).

Finally, we evaluated the corpus through a ten-fold cross validation and further on we described a method to mine opinion from user input in Spanish, using n-gram and phrase level similarity with the annotated elements, obtaining high precision results. We proved that the use of the fine-grained annotations leads to the improvement of the results in comparison to the use of only coarse-grained annotations.

Further work is foreseen for the annotation of the remaining corpus, of similar corpora for other languages (due to the high flexibility of the employed model, this is easily achievable) and of text belonging to different genres. On the application side, it is well known that OM is an extremely challenging task and a relatively young discipline, thus there is room for improvement above all to solve linguistic phenomena such as the co-reference resolution at a cross document level, temporal expression recognition, etc.

Regarding the feature selection experiments, we will divide the problem and create three different binary classifiers, one for each polarity, instead of a multi-class one in order to obtain a different list of selected features for each polarity and solve the problems encountered related to the corpus balancing. We also contemplate the possibility of using a local feature selection instead of the global one, to extract the best features for each polarity and other well-known feature selection. Our results show that using only a stemmed bag of words with a SVM classifier will produce the best performance. Thus we will explore alternative Machine Learning algorithms with different features. Furthermore, in order to improve the OM task, we will study the effect of including linguistic tools such as WSD systems or lemmatisers, together with a study focused on finding what the most relevant stopwords are to be included in our experiments, and also to contemplate more complex negation and comparative structures. It is worth mentioning that the binary classification will also be effective in determining in which kind of polarity the negation is determinant. We will carry out the integration of Semantic Roles in order to detect the source of the discourse, and also plan to perform a more detailed NE recognition in order to understand which is the most relevant topic of our documents.

Last but not least, our idea is to include the existing tools for a more effective semi-supervised annotation. After the training of the Machine Learning system we automatically obtain some markables which have to be validated or not by the annotator. The ideal option would be to connect these terms the system detects automatically with tools, such as zapping with an opinion lexicon based on WordNet (SentiWordNet, WordNet Affect, MicroWord-Net). This will be contemplated in order to automatically annotate all the synonyms and antonyms with the same or the opposite polarity respectively, and assign them some other elements contemplated in the EmotiBlog annotation scheme. This would result in time saving during the annotation process, while ensuring that the high quality of the annotation is maintained by human supervision.

Acknowledgements This work has been partially funded by the Spanish and Valencian Government through the projects "TEXTMESS 2.0: Las tecnologías del Lenguaje Humano ante los nuevos retos de la comunicación digital" (TIN2009-13391-C04-04), "Desarrollo de técnicas inteligentes e interactivas de minería de textos" (PROMETEO/2009/199) and the Generalitat Valenciana (ACOMP/2011/001) respectively.

References

Artstein R, Poesio M (2008) Inter-coder agreement for computational linguistics (survey article). Comput Linguist 34(4):555–596

Balahur A, Montoyo A (2008) Applying a culture dependent emotion triggers database for text valence and emotion classification. Procesamiento del Lenguaje Natural 40(40)

Balahur A, Montoyo A (2009) Semantic approaches to fine and coarse-grained feature-based opinion mining. In: Proceedings of the international conference on application of natural language to information systems, NLDB

Balahur A, Montoyo A (2010) OpAL: applying opinion mining techniques for the disambiguation of sentiment ambiguous adjectives in SemEval-2 evaluation exercises on semantic evaluation SemEval-2 task 18. Cophenagen, Sweden

Balahur A, Boldrini E, Montoyo A, Martínez-Barco P (2009a) Opinion and generic question answering systems: a performance analysis. In: Proceedings of ACL, 2009, Singapore

Balahur A, Boldrini E, Montoyo A, Martínez-Barco P (2009b) A Comparative study of open domain and opinion question answering systems for factual and opinionated queries. In: Proceedings of RANLP 2009

Balahur A, Boldrini E, Montoyo A, Martínez- Barco P (2009c) Cross-topic opinion mining for real-time human-computer interaction. In: Proceedings of the workshop on natural language and cognitive science, NLPCS, 2009

Balahur A, Boldrini E, Montoyo A, Martínez-Barco P (2010a) Opinion question answering: towards a unified approach. In: Proceedings of the ECAI conference

Balahur A, Boldrini E, Montoyo A, Martínez-Barco P (2010b) A unified proposal for factoid and opinionated question answering. In: Proceedings of the COLING conference

Balahur A, Boldrini E, Montoyo A, Martínez-Barco P (2010c) The OpAL system at NTCIR 8 MOAT. In: Proceedings of the NTCIR 8 MOAT conference, Tokyo, Japan

Balahur A, Steinberger R, Kabadjov M, Zavarella V, Van der Goot E, Halkia M, Pouliquen B, Belyaeva J (2010d) Sentiment analysis in the news. In: Proceedings of the 7th international conference on language resources and evaluation (LREC'2010), Valletta, Malta, 19–21 May 2010, pp 2216–2220

Boldrini E, Balahur A, Martínez-Barco P, Montoyo A (2009a) EmotiBlog: a fine-grained model for emotion detection in non-traditional textual. In: Proceedings of WOMSA 2009, Seville, Spain

Boldrini E, Balahur A, Martínez-Barco P, Montoyo A (2009b) EmotiBlog: an annotation scheme for emotion detection and analysis in non-traditional textual genres. In: Proceedings of the 5th international conference on data mining. Las Vegas, Nevada, USA

Boldrini E, Balahur A, Martínez-Barco P, Montoyo A (2010c) EmotiBlog: a finer-grained and more precise learning of subjectivity expression models. In: Proceedings of the fourth linguistic annotation workshop, association of computational linguistics, Copenhagen, Sweden

Carletta J (1996) Assessing agreement on classification tasks: the kappa statistic CoRR cmp lg/9602004

Cerini S, Compagnoni V, Demontis A, Formentelli M, Gandini G (2007) Micro-wnop: a gold standard for the evaluation of auto-matically compiled lexical resources for opinion mining. Franco Angeli Editore, Milano

Chaovalit P, Zhou L (2005) Movie review mining: a comparison between supervised and unsupervised classification approaches. In: Proceedings of HICSS-05, the 38th Hawaii international conference on system sciences

Choi Y, Cardie C, Riloff E, Patwardhan S (2005) Identifying sources of opinions with conditional random fields and extraction patterns. In: Proceedings of HLT/EMNL

Cohen J (1960) A coefficient of agreement for nominal scales. Edu Psychol Meas 20(1):37–46

Craggs R, Wood MM (2005) Evaluating discourse and dialogue coding schemes. Comput Linguist 31(3):289–296

Cui H, Mittal V, Datar M (2006) Comparative experiments on sentiment classification for online product reviews. In: Proceedings of the 21st national conference on artificial intelligence, AAAI

Dave K, Lawrence S, Pennock D (2003) Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of WWW-03

Esuli A, Sebastiani F (2006) SentiWordnet: a publicly available resource for opinión mining. In: Proceedings of the 6th international conference on language resources and evaluation

Gamon M (2004) Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In: Proceedings of COLING-04, the 20th international conference on computational linguistics, Geneva, CH, pp 841–847

Gamon M, Aue S, Corston-Oliver S, Ringger E (2005) Mining customer opinions from free text. Lecture notes in computer science

Goldberg AB, Zhu J (2006) Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In: HLT-NAACL 2006 workshop on textgraphs: graph-based algorithms for natural language processing

Hatzivassiloglou V, Wiebe J (2000) Effects of adjective orientation and gradability on sentence subjectivity. In: Proceedings of COLING 2000

Hu M, Liu B (2004) Mining opinion features in customer reviews. In: Proceedings of nineteenth national conference on artificial intellgience AAAI-2004

Kim SM, Hovy E (2004) Determining the sentiment of opinions. In: Proceedings of COLING 2004

Lesk M (1986) Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: proceedings of the SIGDOC conference 1986

Lewis D, Gale W (1994) A sequential algorithm for training text classifiers. In: Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval. Springer, New York

Liu B (2007) Web data mining. Exploring hyperlinks, contents and usage data, 1st edn. Springer, New York

Mathieu J (2005) Annotation of emotions and feeling in texts. Affectve computing and intelligent interaction. Bejing, China

Mullen T, Collier M (2004) Sentiment analysis using support vector machines with diverse information sources. In: Proceedings of EMNLP

Ng V, Dasgupta S, Arifin SM (2006) Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In: Proceedings 40th annual meeting of the association for computational linguistics

Pang B, Lee L (2003) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd annual meeting of the ACL, pp 115–124

Paquet S (2003) Personal Knowledge publishing and its uses in research. Knowledge Board, 10 January

Riloff E, Wiebe J (2003) Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 conference on empirical methods in natural language processing

Russell JA (1983) Pancultural aspects of the human conceptual organization of emotions. J Person Soc Psychol 45:1281–1288

Salton G, Lesk ME (1971) Computer evaluation of indexing and text processing. Prentice Hall, Englewood Cliffs, pp 143–180

Scherer K (2005) What are emotions? and how can they be measured? Soc Sci Inf 3(44)

Scherer K, Wallbott HG (1997) The ISEAR questionnaire and codebook

Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv (CSUR) 34: 1–47

Somasundaran S, Wilson T, Wiebe J, Stoyanov V (2007) Qa with attitude: exploiting opinion type analysis for improving question answering in on-line discussions and the news. In: Proceedings of the international conference on weblogs and social media, ICWSM

Somasundaran S, Wiebe J, Ruppenhofer J (2008) Discourse level opinion interpretation. In: The 22nd international conference on computational linguistics (COLING)

Stoyanov V, Cardie C (2006) Toward opinión summarization: linking the sources. In: Proceedingns of the COLINGACL 2006 workshop on sentiment and subjectivity in text

Stoyanov V, Cardie C, Wiebe J (2005) Multiperspective question answering using the opqa corpus. In: Proceedings of the human language technology conference and the conference on empirical methods in natural language processing (HLT/EMNLP)

Strapparava C, Milhacea R (2007) SemEval-2007 task 14: affective text

Strapparava C, Valitutti A (2004) WordNet-Affect: an affective extension of WordNet. In: Proceedings of the 4th international conference on language resources and evaluation, LREC, Lisbon, May 2004, pp 1083–1086

Turney P (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings 40th annual meeting of the association for computational linguistics

Turney P, Littman M (2003) Measuring praise and criticism: inference of semantic orientation from association. ACM Trans Inf Syst 21(4):315–346

Vapnik V (1995) The nature of statistical learning theory. Springer, New York

Wiebe J, Mihalcea R (2006) Word sense and subjectivity. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics. Sydney, Australia, July, pp 1065–1072

Wiebe J, Riloff E (2006) Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In: Proceedings of the 6th international conference on computational linguistics and intelligent text processing (CICLing-05)

Wiebe J, Wilson T (2005) Annotating attribution and private states. In: Proceedings of the ACL Workshop on frontiers in corpus annotation II: Pie in the Sky

Wiebe J, Wilson T, Cardie C (2005) Annotating expressions of opinions and emotions in language. Lang Resour Eval 1:2165–210

Wilson T, Wiebe J, Hwa R (2004a) Just how mad are you? Finding strong and weak opinion clauses. In: Proceedings of AAAI

Wilson T, Wiebe J, Hwa R (2004b) Just how mad are you? Finding strong and weak opinion clauses. In: Proceedings of AAAI 2004

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT-EMNLP 2005

Yang Y, Pedersen J (1997) A comparative study on feature selection in text categorization. In: Proceedings of ICML-97, 14th international conference on machine learning