10.12871/CLICIT2014167

An Italian Corpus for Aspect Based Sentiment Analysis of Movie Reviews

Antonio Sorgente
Institute of Cybernetics
National Research Council
Via Campi Flegrei 34,
Pozzuoli (Naples) Italy
a.sorgente@cib.na.cnr.it

Giuseppe Vettigli
Institute of Cybernetics
National Research Council
Via Campi Flegrei 34,
Pozzuoli (Naples) Italy

Francesco Mele
Institute of Cybernetics
National Research Council
Via Campi Flegrei 34,
Pozzuoli (Naples) Italy

Abstract

English. In this paper we will present an Italian corpus focused on the domain of movie reviews, developed in order to support our ongoing research for the development of new models about Sentiment Analysis and Aspect Identification in Italian language. The corpus that we will present contains a set of sentences manually annotated according to the various aspects of the movie that have been discussed in the sentence and the polarity expressed towards that particular aspect. In this paper we will present the annotation guidelines applied, some statistics about the corpus and the preliminary results about the identification of the aspects.

Italiano. *In questo lavoro presenteremo una nuova risorsa linguistica sviluppata per la creazione di nuovi modelli per la Sentiment Analysis Ascpect Based in lingua Italiana. Di seguito saranno introdotte le linee guida adottate per l'annotazione del corpus ed alcuni risultati preliminari riguardanti l'identificazione di aspetti.*

1 Introduction

Nowadays, on the Web there is a huge amount of unstructured information about public opinion and it continues growing up rapidly. Analysing the opinions expressed by the users is an important step to evaluate the quality of a product. In this scenario, the tools provided by Sentiment Analysis and Opinion Mining are crucial to process this information. In the particular case of movie reviews, we have that the number of reviews that a movie receives on-line grows quickly. Some popular movies can receive hundreds of reviews and, furthermore, many reviews are long and sometimes they contain only few sentences expressing the actual opinions. This makes hard for a potential viewer to read them and make an informed decision about whether to watch a movie or not. In the case that one only reads a few reviews, the choice may be biased. The large number of reviews also makes it hard for movie producers to keep track of viewer's opinions. The recent advances in Sentiment Analysis have shown that coarse overall sentiment scores fails to adequately represent the multiple potential aspects on which an entity can be evaluated (Socher et al., 2013). For example, if we consider the following review from Amazon.com about the movie *Inception*:

> *"By far one of the best movies I've ever seen. Visually stunning and mentally challenging. I would recommend this movie to people who are very deep and can stick with a movie to get the true meaning of the story."*

One can see that, even if the review is short, it not only expresses an overall opinion but also contains opinions about other two aspects of the movie: the photography and the story. So, in order to obtain a more detailed sentiment, an analysis that considers different aspects is required.

In this work, we present an Italian corpus focused on the domain of movie reviews developed in order to support our ongoing effort for the development of new models about Sentiment Analysis and Aspect Identification in Italian language.

The paper is structured as follows. In the Section 2 we present the motivations that led us to the creation of a new corpus and a short survey about related resources that already exist. Section 3 describes the guideline used to annotate the corpora, while Section 4 presents some statistical information about it. In section 5 we present some preliminary experiments about the identification of the as-

pects. Finally, in section 6 some conclusions will be offered.

2 Motivations

During the last years many studies focused on how to combine Sentiment Analysis and Aspect Identification.

The first attempt to combine Sentiment Analysis and Aspect Identification was made in (Hu and Liu, 2004), where a system to summarize the reviews was proposed. The system extracts terms that are related to various aspects of the textual comments and they tested their approach using 500 reviews about five types of electronics products (digital cameras, DVD players, mp3 players and mobile phones). The reviews were taken from Amazon.com and Cnet.com.

In (Ganu et al., 2009) a corpus of about 3400 sentences, gathered from a set of reviews about restaurants, have been annotated according to specific aspects of the restaurant domain with the related sentiment. The authors used the corpus to develop and test a regression-based model for the Sentiment Analysis. The same data and a set of about 1000 reviews on various topics collected from Amazon.com were used in (Brody and Elhadad, 2010). In this work the authors presented an unsupervised model able to extract the aspects and determine the related sentiments.

In the SemEval-2014 challenge, a task with the aim to identify the aspects of given target entities and the sentiment expressed towards each aspect has been proposed. The task is focused on two domain specific datasets of over 3000 sentences, the first one contains restaurant reviews extracted from the same data used in (Ganu et al., 2009) and the other one contains laptop reviews. Regarding the movie reviews, in (Thet et al., 2010) a method to determine the sentiment orientation and the strength of the reviewers towards various aspects of a movie was proposed. The method is based on a linguistic approach which uses the grammatical dependencies and a sentiment lexicon. The authors validated their method on a corpus of 34 reviews from which 1000 sentences were selected.

From the works mentioned, it follows that the approaches based on sentence-level analysis are predominant for the detection of the aspects in the reviews.

There are few studies that use Italian language because Italian lacks resources for sentiment analysis of natural language, although, some interesting resources have been produced using Twitter as data source. For example, in (Basile and Nissim, 2013) a dataset that contains 100 million tweets was proposed. This dataset contains 2000 tweets annotated according to the sentiment they express, 1000 of them regard general topics, while the remaining 1000 regard politics. In the Senti-TUT project (Bosco et al., 2013), an Italian corpus which consists of a collection of texts taken from Twitter and annotated with respect to irony was created. EVALITA (Evaluation of NLP and Speech Tools for Italian) has provided many interesting activities and resources, until now it has not hosted any activity or task about Sentiment Analysis aspect based. However, to the best of our knowledge, there is only one Italian corpus, which has been used in (Croce et al., 2013), where both the aspects and their polarity are taken in account. In particular, the corpus is focused on review of Italian wine products and it has been used to build a model able to classify opinions about wines according to the aspect of the analyzed product, such the flavor or taste of a wine, and the polarity.

3 Annotation Guidelines

To build our corpus, the annotators were instructed through a manual with the annotation guidelines. The guidelines were designed to be as specific as possible about the use of the aspects and the sentiment labels to be assigned.

3.1 Aspects

The aspects of the movies that we have considered for the annotation were suggested by some movie experts. For each of them we have provided a short guideline that helps the annotator with the identification of the aspect in the sentences:

- **Screenplay**: the sentence describes one or more scenes of a movie, their temporal distribution during the story, the quality and, the type or the complexity of the plot of the story. For example: *"Invictus è certo un film edificante di buone volontà, ma anche un bel film di solida struttura narrativa"* (*"Invictus is certainly an enlightened film of goodwill, but also a good movie with a solid narrative structure."*).

- **Cast**: the sentence expresses the importance

of the actors and their popularity. For example: *"Sono andato a vedere il film perchè c'era il mitico Anthony Hopkins"* (*"I went to see the movie because there was the legendary Anthony Hopkins"*).

- **Acting**: the sentence expresses an opinion on the actors' performances. For example: *"Grande come sempre la Bullock!"* (*"Bullock is always great!"*).
- **Story**: the sentence references to the storyline of the film. For example: *"Il finale di questo film è particolarmente amaro."* (*"The ending of this movie is particularly bittersweet."*).
- **Photography**: the sentence is about the colors, close-ups, countershot and shot used in the movie. For example: *"La fotografia è magistrale."* (*"The photography is great."*).
- **Soundtrack**: the sentence refers to the soundtrack and music used in the movie. For example: *"Fantastiche le basi musicali usate durante il film e l'indimenticabile sigla di chusura"* (*"The backing tracks used during the film and the unforgettable theme song in the ending are fantastic"*).
- **Direction**: the sentence is on the work of the director. For example: *" Tim Burton è pazzo ma anche un genio!"* (*"Tim Barton is crazy, but he is also a genius!"*).
- **Overall**: this aspect is used when the sentence doesn't report the description of any particular aspect of the movie but a general opinion or description. For example: *"Un film veramente bello, da vedere!!!"* (*"A really nice movie, to watch!!!"*).

3.2 Polarity Labels

We used 5 sentiment labels to represent the polarity: *strongly negative*, *negative*, *neutral*, *positive* and *strongly positive*. The description of each label follows:

- **Strongly Negative**: there is an extremely negative opinion.
- **Negative**: there is a negative opinion.
- **Neutral**: there is no opinion or if it expresses an opinion boundary between positive and negative.
- **Positive**: there is a positive opinion.
- **Strongly Positive**: there is an extremely positive opinion.

Aspect	Count
Overall	1370
Screenplay	226
Cast	165
Acting	338
Story	647
Photography	55
Soundtrack	40
Direction	235

Table 1: Distribution of the aspects in the Corpus.

4 Corpus description

The corpus contains 2648 sentences. Each sentence has been manually annotated according to the various aspects of the movie that have been discussed in the sentence; then, each aspect found has been annotated with the polarity expressed towards that particular aspect. So, for each sentence annotated we have a set of aspect-polarity pairs. Also, the sentences that were extracted from the same review are linked by an index in order to enable the study of the contex.

The sentences of the corpus have been extracted from over 700 user reviews in the website FilmUp.it. The user reviews of this website have been also studied in (Casoto et al., 2008), where the authors focused on the classification of the reviews according to the sentiment without considering the specific aspects referred in the reviews and without providing a sentence level study.

The distribution of the aspects is reported in Table 1, while the distribution of the sentiment labels is reported in Table 2. We can see that 23% of the labels are Negative or Strongly Negative, the 21% are Neutral and that the 53% are Positive or Strongly Positive.

It is important to notice that, the size of the corpus is comparable to the size of the English corpora with the same purpose.

In order to evaluate the accuracy of the annotation, 800 sentences have been annotated by two different annotators and the agreement among the annotators has been evaluated using the Cohen's Kappa (κ) measure (Carletta, 1996). This metric

Polarity	Count
S. Negative	250
Negative	526
Neutral	706
Positive	1238
S. Positive	491

Table 2: Distribution of the polarity labels in the Corpus.

measures the agreement between two annotators taking into account the possibility of chance agreement. It is computed as

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)},$$

where $P(a)$ is the relative observed agreement between two annotators and $P(e)$ is the expected agreement between two annotators, using the observed data to calculate the probabilities of each observer randomly saying each category. If the annotators are in complete agreement then $\kappa = 1$. If there is no agreement among the annotators other than what would be expected by chance, $\kappa = 0$.

The inter-annotator agreement computed on our data is substantial for the aspect categories (0.7) and very good for the sentiment categories (above 0.8).

In case of disagreement the final annotation has been selected by a third annotator.

5 Preliminary experiments

In this section we report the results of a preliminary experiment on Aspect Identification. To do this, we have used Linear Discriminant Analysis (LDA) (Hastie et al., 2001) in order to build a set of classifiers, one for each aspect, able to recognize if a sentence is about or not a given aspect. The features used to train the classifiers were computed using the *tf-idf* (term frequency–inverse document frequency) model (Baeza-Yates and Ribeiro-Neto, 1999). In this model, the features extracted by a sentence are given by a set of terms, for each term t the value of the feature is computed as

$$tf(t, s) \times idf(t)$$

where $tf(t, s)$ is the count of t in the sentence s and $idf(t)$ is defined as

$$idf(t) = \log \frac{|S|}{1 + |\{s : t \in s\}|}$$

where S is the collection of sentences. Each classifier was trained on a different set of features (before the features extraction, stop-words were removed), and the terms for the features extraction were selected according to the χ^2 measure respect to the given aspect. This statistic measures the dependence between the features and a target variable and it is often used to discover which features are more relevant for statistical classification (Yang and Pedersen, 1997). Then, we have performed 5-fold cross validation and used accuracy, precision and recall to evaluate the quality of the classification. The Table 3 shows the error estimated using cross validation. With this basic model, we had a high accuracy (89% on average) and a good precision (70% on average). The recall was moderate (50% on average). In that Table, the evaluations with respect to the aspects *Photography* and *Soundtrack* are not reported because the samples for these categories are not enough to train and test a classification model.

Aspect	Accuracy	Precision	Recall
Overall	72%	70%	93%
Screenplay	92%	73%	42%
Cast	94%	71%	28%
Acting	90%	78%	52%
Story	82%	81%	49%
Direction	93%	78%	50%

Table 3: Aspect identification results.

6 Conclusion

We introduced an Italian corpus of sentences extracted by movie reviews. The corpus has been specifically designed to support the development of new tools for the Sentiment Analysis in Italian. We believe that corpus can be used to train and test new models for sentence-level sentiment classification and aspect-level opinion extractions.

In the paper, various aspects of the corpus we created have been described. Also, the results of some preliminary experiments about the automatic identification of the aspects have been showed.

7 Availability and license

The proposed Corpus is made available under a Creative Commons License (CC BY 3.0) and can be requested contacting one of the authors of this paper.

References

Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *roceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 804–812, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, 22(2):249–254, June.

Paolo Casoto, Antonina Dattolo, Paolo Omero, Nirmala Pudota, and Carlo Tasso. 2008. A new machine learning based approach for sentiment classification of italian documents. In Maristella Agosti, Floriana Esposito, and Costantino Thanos, editors, *IRCDL*, pages 77–82. DELOS: an Association for Digital Libraries.

Danilo Croce, Francesco Garzoli, Marco Montesi, Diego De Cao, and Roberto Basili. 2013. Enabling advanced business intelligence in divino. In *DART@AI*IA*, pages 61–72.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *WebDB*.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA, October. Association for Computational Linguistics.

Tun Thura Thet, Jin-Cheon Na, and Christopher S. G. Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Information Science*, 36(6):823–848.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.