# A New Machine Learning Based Approach for Sentiment Classification of Italian documents

Paolo Casoto, Antonina Dattolo, Paolo Omero, Nirmala Pudota, Carlo Tasso
Department of Mathematics and Computer Science
University of Udine
Via delle Scienze, 206 - Loc. Rizzi
Udine, Italy
Email: $\{paolo.casoto, antonina.dattolo, paolo.omero, nirmala.pudota, carlo.tasso\}$@dimi.uniud.it

***Abstract*—Several sites allow users to publish personal reviews about products and services available on the market. In this paper, we consider the problem of applying classification techniques to identify, in terms of positive or negative degree, the overall opinion polarity expressed in these documents, written in natural language. In particular, we are interested in evaluating the performance obtained by applying machine learning techniques, based on n-gram selection and originally developed for English, to documents written in Italian.**

**In order to obtain results comparable to those presented in the literature for English, we use the same evaluation procedure applied in the majority of the works in this field. We have developed a specific framework for experimentation in Italian. The research is ongoing and we present some preliminary results, a comparison with results presented in literature and an overview of our future work.**

## I. Introduction

The information and knowledge sharing represents a meaningful goal for many current websites; in particular, some of these sites (e.g. the *Ciao.it*[1] repository) allow users to publish their own reviews about products and services available on the market. A feature common to each posted review is the presence of an *overall opinion polarity* (called OvOP) [1], [2], which describes the positive, neutral or negative opinion of the author with respect to judged items. OvOP is defined as the overall opinion polarity of a review because it does not depends on the opinions expressed by the single sentences of the review, which may also be contradictory, but is related to the opinion that arises from the document seen as a whole.

The amount of reviews available on the Web and the set of described items are enormously increasing every day. To handle such information in an efficient way a partially or full automated approach to OvOP analysis is required [1]–[3].

An automated approach may be very useful in many application's fields [4], like in advertising, political campaigns, financial markets and *business intelligence*. Analysis of the customers' opinion allows the system extracting information useful for strategic marketing and brand monitoring. An example of this application is represented by the system developed by Tong [5]; the use of keyword matching generates *sentiment timelines* able to track the opinions expressed by customers over time about a specific item.

In this paper we are interested in developing a corpus of reviews in Italian and using it to evaluate the performance obtained by some known methodologies introduced in the literature in the area of sentiment analysis. More specifically, we studied in which way such techniques, originally developed for English, may be used for Italian without loss in performance. In particular, we focused our attention on the effects generated by stemming on the performances of the OvOP analysis task. Our main goal is the development of a module of sentiment analysis that can be integrated in a wider and more general architecture for personalized information and knowledge extraction.

The next of this paper is so organized: Section II presents an overview of related works; section III proposes the development of an Italian corpus of reviews related to movies, while section IV describes our framework devoted to OvOP analysis; section V describes experimental results and the following section VI discusses and compares our results with those obtained by Pang et al. [1]. Conclusions and future works end the paper.

## II. Previous Work

Two works mainly inspire our research: the experience of Pang et al. [1] and the work of Salvetti et al. [2]; both works are aimed at evaluating the performance, in terms of precision and recall, obtained by the OvOP analysis task, applying machine learning techniques and feature selection to a corpus of movie reviews. In [1] authors studied differences in precision using three different machine-learning techniques: Bayesian Networks, Support Vector Machines and Maximum Entropy. They show that OvOP identification is a task harder than topic classification or generalization, both performed mainly by keyword identification, because sentiment tends to be expressed in more subtle ways.

The work [2], on the other hand, focuses attention on feature selection and feature generalization, integrating WordNet [6] as a repository of lexical relations.

Works like [7] and [8] study how domain dependency and time dependency are related to the performance achieved by the classifiers described in [1]; in particular, these two works aim at evaluating how the effects of limitations can be reduced introducing in document representation features

[1] www.ciao.it

loosely coupled with the domain of the documents on which OvOP is performed.

In [9] and in other works the OvOP analysis of a review is defined as the evaluation of the co-occurrence between a series of n-grams, extracted from the text using a set of patterns, and a seed set of well oriented terms; they used AltaVista and its indexed corpus to investigate the amount of such co-occurrences.

Many works [10]–[12] *generate* sets of words or synsets, in particular adjectives, with a prior known positive or negative polarity. During OvOP classification process, the generated lists are used as features in representation of documents.

Other works, like [13], are aimed at merging the previously introduced approaches; more specifically, large datasets of documents are used to adjust the polarity of a set of terms with prior known polarity depending on the context in which such terms are used. Attention is focused on syntactical structures, like negation or hypothetical speaking, which can modify the prior known polarity of the words.

Previously described works show how the OvOP identification (or extraction) task may be seen both as a classification and as an *information extraction* process.

The classification approach assigns to each input document a label describing its sentiment rating using, as training set, a set of manually labelled reviews.

On the other hand, the information extraction approach identifies the orientation of a given text using a set of extraction patterns and seed lists to highlight sentiment-bearing patterns.

OvOP analysis in Italian is still an almost unexplored research field; many unsolved problems arose during our research, mainly related to the lack of freely available tools for natural language processing of Italian documents; other open problems are generated by the specific characteristics of Italian (irregular forms, adjective declination, et. al.). There are not available sets of Italian words with previously estimated polarity, which can be used during feature selection task. A last family of problems we faced is related to the lack of quality, in terms of syntactical and grammatical soundness, of the input documents.

## III. Development of an Italian movie-reviews corpus

In order to compare our results to the results obtained in [1] and in [2], we focus our attention on the same domain of movie reviews, but, instead of working with English texts, we collected reviews written in Italian; these last are hand-labelled by authors or readers using an overall rating indicator in order to summarize the document's OvOP.

The source of our corpus is the *FilmUp*[2] website, which collects a wide set of visitors' opinions on more than 4500 movies.

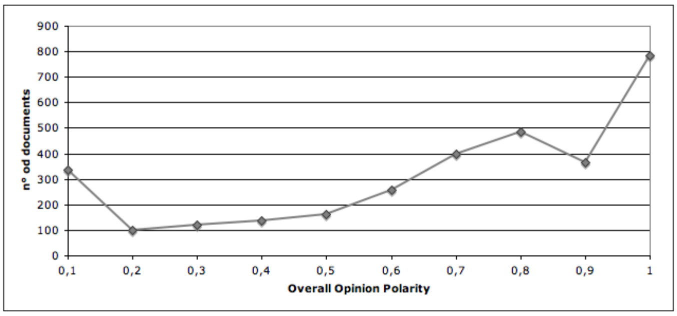

Fig. 1. OvOP distribution of reviews present in the corpus.

Using a crawler written in Java, we downloaded a subset of site's pages, referring to the most recent movies. In particular, by means of an ad hoc developed parser, we extracted from each review the following fields:

1) the title assigned by author to the review and used to give emphasis and to summarize its content;
2) the body of the review, which consists of a short natural language text;
3) the overall polarity rating indicator, expressed in terms of stars with a range between 0 and 10, where 0 stars represent a high negative polarity, while 10 stars represent an high positive polarity;
4) the review publishing date;
5) personal data about author, like name, age and city of residence.

Only first three fields are effectively used during the OvOP analysis process, while the last two ones have been stored for future use; in the next future, for user modeling and content personalization, we will identify *personalized sentiment timelines*, which can be seen in terms of a graph showing the attitude expressed by a user over time towards a specific item.

In this paper we consider only two classes of reviews: positive and negative: we do not consider the class of neutral reviews, removing from our evaluation all those reviews with a overall polarity rating indicator greater than 4 stars and lower or equal to 6 stars.

We collected more than 3000 reviews referring to 300 different movies or fictions. The distribution of polarity between reviews is not fair, as observed [1] for the English corpus. The positive reviews are 2038 (64.7% of the entire corpus), while the negative reviews are only 694 (22% of the collection). The distribution of the overall polarity indicator pre-assigned to reviews is reported in figure 1.

Polarity has been normalized between 0 and 1 to allow integration with different kind of ranking methods.

## IV. The OvOP Analysis Framework

We have implemented and extended the techniques described in [1] and [2] developing a framework, specialized for Italian, able to perform the tasks of OvOP analysis and classification on a given training set.

The framework has been developed in Java and is constituted by four main modules:

1) the *persistency module* is responsible for data storage on

[2] www.filmup.leonardo.it/opinioni

different available repositories, like relational databases, sets of inverted indexes and native XML databases;

2) the *language processing module* is an API able to perform several task related to natural language processing like multilingual stemming, n-gram and language models generation and Part-Of-Speech tagging (POS-Tagging). The API also defines the data structures used during OvOP analysis to represent the documents. To implement an efficient and high quality Italian POS-Tagging a properly trained version of TreeTagger [14] has been integrated into the API;
3) the *classification module* performs OvOP analysis. Based on the WEKA [15] library for machine learning, this module allows us to abstract from a specific implementation of a machine learning technique and focus attention mainly on the feature of selection task. The classification tools are used to generate WEKA classifiers that can be easily reused, whenever properly trained, in other application environments, independently from the OvOP analysis system;
4) the *evaluation module* generates pipelines constituted by resources implemented in the previously defined modules. The pipelines can be executed inside this module, which works as runtime environment, and their results are statistically analyzed and presented to users. Each evaluation experience is implemented with a specific pipeline.

Likewise in [1] and in [2], in this work we adopted Naive Bayes (NB) classifiers, a typical approach in text classification. Let C be a set of labelled classes

$$C = \{positive, negative\} \tag{1}$$

and $d$ the document we want to classify.

The NB classifier assigns the document $d$ to the class $c \in C$ that maxims the equation

$$P_{NB}(c|d) := \frac{P(c)\prod_{i=1}^{m} P(f_i|c)^{n_i(d)}}{P(d)}. \tag{2}$$

where P($f_i|c$) is the probability to find a given feature $f_i$ in the documents labelled with class $c$. The training set, represented by a set of labelled documents, is used to evaluate the conditional probabilities of each feature. To perform our work we have used the implementations of this technique included into the WEKA library [15].

## V. Experimental Results

Our experimental evaluation is based on a subset of our corpus containing the same number of positive and negative reviews: more in detail, we randomly chosen 500 reviews with polarity greater than 0.6 and 500 with polarity lower or equal to 0.4. Body and title of each reviews have been loaded from the database and merged together as a unique field; in this way, we uniformed our data with the English reviews proposed in [1] and [2]. This solution looks very simple and does not take care of the strength assigned by the author to the words that appear in the title of the documents.

The work described in [1] uses unstemmed words in the document representation; perhaps, considering the differences between Italian and English, with particular reference to adjectives, we stemmed all words using an Italian stemmer. No stopword list has been used.

Each review is tagged using the Part-of-Speech tagger integrated in our analysis system. We noticed that sometimes tagging of reviews was performed with low precision. This happens because part of the reviews contains lexical mistakes, lacks in punctuation, idiomatic expressions (wow, blah, ok) or emotive icons ( *;-)*, *:-(* ). We will consider this problem in the future works because such lexical exceptions are frequently used to express opinion and their analysis may lead to a potential improvement in OvOP analysis performance.

According to the procedure described in [1], we considered the set of 2122 unigrams that appears at least four times and the set of 2122 most frequent bi-grams.

Seven different sets of features have been analyzed; for each set we built and tested a specific classifier. Evaluation is achieved with a 3-cross folding methodology, implemented in the classification module.

Features used for representation are based on the following two concepts: *N-Gram Frequency* (NGF) and *N-Gram Presence* (NGP). The NGF of an n-gram $t$ in respect with a document $d$ is defined as the number of occurrences of $t$ in $d$; it is measured using integer values. The NGP of a given n-gram $t$ in respect with a document $d$ is defined as the presence of $t$ in $d$ and is measured with a boolean value. The choice of NGP is widely described in [1]; authors show how this feature leads to better performance with respect to NGF; this trend is confirmed also by our results, where precision is slightly boosted when we move from an NGF to an NGP representation.

More specifically, we trained our classifiers evaluating, for each document of the training set, the seven following sets of features, previously introduced in [1]:

1) the NGF of the 2122 most frequent unigrams;
2) the NGP of the 2122 most frequent unigrams;
3) the NGP of the 2122 most frequent bi-grams;
4) the NGP of the 2122 most frequent bi-grams merged with the most frequent 2122 unigrams;
5) the NGP of the bi-grams composed by one of the most frequent 2122 unigrams and the POS tag associated with such unigram. This feature is strongly coupled with the execution of the POS tagging operation;
6) the NGP evaluated only on the unigrams tagged as adjectives by the POS tagger. This feature is based on the assumption, widely proven by many works [10], [11], that polarity is mainly carried by adjectives;
7) the NGP evaluated on the 1000 most frequent unigrams appearing in the training set. This feature aims to verify if a feature selection approach not based on syntactic roles may lead to an improvement in performance.

The performance of the built classifiers is measured in terms of precision, evaluated comparing label assigned by classifier during testing with the OvOP expressed by the author of the review. Table I shows the results obtained choosing each of the seven different features sets with and without stemming of the documents' terms. In order to understand how classification process is exploited, we propose two examples of reviews A e B, which are correctly classified (respectively as positive and negative) by our OvOP analysis framework. We consider the NGP-based classifier, which, as shown in Table I, achieves the best results in our experimental activity.

TABLE I
AVERAGE PRECISION OF THE BUILT CLASSIFIERS

| Feature | Precision (no stem.) | Precision (stem.) |
|---|---|---|
| Unigrams (NGF) | 75.5 | 76.4 |
| Unigrams (NGP) | 81 | 82.1 |
| Bigrams (NGP) | 68.6 | 70.1 |
| Unigram & Bigrams (NGP) | 79,2 | 79,9 |
| Unigram, POS (NGP) | 77.1 | —- |
| Adjectives | 72.7 | 76.1 |
| Top 1000 unigrams (NGP) | 80.3 | 81.6 |

Review A
*”Da Vedere. Ottimo film ed ottime interpretazioni, un film europeo di grandissimo valore . . . ”*[3].

Review B
*”Delusione! Non so cosa mi aspettassi da questo film, so per certo che è davvero orrendo! Mai visto niente di più scontato e noioso! Sinceramente sono senza parole . . . ”*[4].

Both reviews A and B are analyzed and converted into a vector-based representation, suitable for classification by means of Naïve Bayes classifiers. Such a representation is compared with the knowledge base acquired by the classifier during the training and, according to the classification effectiveness of each feature occurring into reviews, classified as positive or negative. More in detail, Table II lists occurrences of the 30 most significant stemmed unigrams used by the classifier. The significance of listed stems is estimated by means of *Information Gain* (IG) metric.

TABLE II
TOP 30 UNIGRAMS EXTRACTED FROM THE TRAINING SET WITH THE HIGHEST IG VALUE

| | | | | | |
|---|---|---|---|---|---|
| 1 | bellissim | 11 | bravissim | 21 | attor |
| 2 | brutt | 12 | pò | 22 | pir |
| 3 | bell | 13 | piac | 23 | grand |
| 4 | jack | 14 | noios | 24 | perfett |
| 5 | pessim | 15 | ridicol | 25 | sparrow |
| 6 | ottim | 16 | interpret | 26 | sonor |
| 7 | fantast | 17 | bast | 27 | orrend |
| 8 | evit | 18 | simpson | 28 | will |
| 9 | peggior | 19 | butt | 29 | schifezz |
| 10 | delusion | 20 | splendid | 30 | bel |

IG is defined as the number of bits of information obtained for category prediction by knowing the presence or absence of a feature in a document. Equation 3 reports the general formulation of IG with respect to the input unigram $t$:

$$IG(t) = -\sum_{i=1}^{m} Pr(c_i)logPr(c_i) + Pr(t)\sum_{i=1}^{m} Pr(c_i|t)logPr(c_i|t) + Pr(\not t)\sum_{i=1}^{m} Pr(c_i|\not t)logPr(c_i|\not t) \quad (3)$$

with $\{c_i\}_{i=1}^{m}$ set of available classes (positive and negative).

## VI. DISCUSSION

Our results can be compared with those described in [1], although evaluated on different corpora with specific dimensions and polarity distribution.

The evaluation of the proposed sets of features, achieved with 3-cross folding evaluation process, shows several results similar to those obtained for English. In particular unigram presence (NGP), accordingly to evaluation provided by [1], is the feature that leads to the best performances, when applied to both stemmed and not-stemmed documents. The results show how information provided by unigram presence overperforms clearly information provided by unigram frequency (NGF). The differences in precision between the two features, both applied to unigrams, has been estimated between 5 and 6 percent.

Bigrams selection performs as the worst feature in our evaluation process, in contrast with the results obtained in [1] with the English corpus; the loss in precision moving from English to Italian has been estimated in more then 10 percent. This may not be explained looking at the number of bigrams extracted from our test corpus: 2122 against the 16000 identified in [1] for the English corpus. Increasing the number of considered bigrams respectively to 7500 and 15000 does not lead to significant improvements in precision.

Features based on Part-of-Speech tagging, as the (POS-Token, Unigram) coupling and the adjectives selection, show a loss in performance too with respect to the precision measured on the English corpus. This is mainly caused by two factors: the accuracy of POS tagging process and the quality, defined in terms of lexical and syntactical soundness, of the input documents. Adjective selection shows an improvement in precision when applied to stemmed documents; such improvement is related to the declination of adjectives, which is one of the most significant differences between Italian and English.

The results obtained by adjective selection confirm the assumption, widely proven in literature and discussed previously, that adjectives are an important clue in determining the polarity of a given text but not enough to achieve alone the best performances in classification.

The classifier trained on the 1000 most occurring unigrams shows an average precision comparable with the one obtained

[3] ”It shall be seen. Great movie and great acting, a high value European movie. . . ”

[4] ”Delusion! I do not know what I was looking for from this movie, but I know it is really horrid. I have never seen something so boring and predictable. Sincerely I have no words to describe it . . . ”

using all the 2122 unigrams occurring more than 4 times in the training set. The increase in dimension of document representation, defined in terms of unigrams considered as features, does not lead to significant improvement, when applied to both stemmed and not-stemmed documents set.

Table III shows results achieved in [1] on the English movie reviews corpus with Bayesian networks as classification algorithm.

The average performance estimated for our seven classifiers is lower than the one calculated in [1], although our unigram-based classifier achieves the best performance of both evaluations, even if the representation is limited to the 1000 most occurring unigrams. Perhaps it is necessary to consider, when discussing about performance, the difference in dimensions of the training set too: our training set is composed by 333 documents (a 3-cross folding evaluation on a corpus of 1000 reviews), while the training set used in [1] is composed of 466 documents (a 3-cross folding evaluation on a corpus of 1400 reviews). Accordingly to the results described in [2], an increase in the dimension of the training set leads to significant increase in the performance achieved by the trained classifiers in OvOP analysis.

In the future we expect to improve the dimension of our collection in order to study how precision of the trained classifiers may change accordingly with the dimension of the training set.

TABLE III
CLASSIFICATION ACCURACY ON THE ENGLISH MOVIE REVIEW DOMAIN

| Feature | Precision ( [1]) |
|---|---|
| Unigrams (NGF) | 78,7 |
| Unigrams (NGP) | 81 |
| Bigrams (NGP) | 77,3 |
| Unigram & Bigrams (NGP) | 80,6 |
| Unigram, POS (NGP) | 81,5 |
| Adjectives | 77 |
| Top 2633 unigrams (NGP) | 80.3 |

## VII. CONCLUSIONS AND FUTURE WORK

Our work shows how OvOP classification can be achieved proficiently in Italian using machine learning techniques, originally developed for English, applied to a domain dependent corpus. In particular we proved experimentally that stemming may increase the accuracy of the trained OvOP classifiers, allowing them to achieve the same precision's rate of the English ones. The research is an ongoing work.

Our future activities will use the developed annotated corpus as a gold-standard in evaluating the accuracy of the trained classifiers.

In the future we will focus our attention on three different aspects of OvOP analysis:

1) the development and evaluation of new representation features;
2) the formalization of the classification process based on the representation of the OvOP of the single sentences of the reviews corpus;
3) the evaluation of the improvements to OvOP analysis offered by user modeling and sentiment timelines.

First goal will be pursued introducing new features able to identify OvOP; more specifically we are interested in studying pattern suitable for extraction of emotive icons and idiomatic expression. Some works, like [16], [17], presents such patterns applied to the specific application area of blogs. The development of a parser able to identify and spread the negation between terms of a sentence may also be useful, as proven in [1], to increase the precision of the classification process.

Second goal is more difficult to achieve, because it requires a huge amount of manual tagging of sentences and words that constitute the review corpus. In order to make tagging activity easier, we developed an intuitive and simple graphical user interface, which is actually used by human annotators. Our goal is the development of a corpus of evaluated sentences, useful in training of fine-grained classifier. These classifiers may lead to a better performance when applied to documents containing contradictory sentences, like some of the reviews included in our corpus. High performance classifiers, based on domain dependent and independent resources, are described in [13].

Our last goal is more related to the introduction of aggregated indicators useful in OvOP analysis; in particular we are interested in studying how sentiment analysis may improve user modeling and adaptive navigation, focusing mainly our attention on the knowledge representation structures introduced in [18].

## REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," *CoRR*, vol. cs.CL/0205070, 2002.

[2] F. Salvetti, S. Lewis, and C. Reichenbach, "Impact of lexical filtering on overall opinion polarity identification," in *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, US, 2004. [Online]. Available: http://www-plan.cs.colorado.edu/reichenb/papers/AAAI-SSS04.pdf

[3] M. Hearst, "Direction-based text interpretation as an information access refinement," 1992. [Online]. Available: citeseer.ist.psu.edu/270790.html

[4] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *CoRR*, vol. cs.CL/0309034, 2003.

[5] R. Tong, "An operational system for detecting and tracking opinions in on-line discussions," Workshop on Operational Text Classification Systems, 2001.

[6] G. Miller, "Wordnet: An on-line lexical database," *International Journal of Lexicography*, vol. 3(4), pp. 235–244, 1990.

[7] A. Aue and M. Gamon, "Customizing sentiment classifiers to new domains: a case study," in *RANLP-05, the International Conference on Recent Advances in Natural Language Processing*, Borovets, BG, 2005.

[8] C. Engstrom, "Topic dependence in sentiment classification," Master's thesis, University of Cambridge, 2004.

[9] P. D. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," in *ACL*, 2002, pp. 417–424.

[10] V. Hatzivassiloglou and K. McKeown, "Predicting the semantic orientation of adjectives," in *ACL*, 1997, pp. 174–181.

[11] J. Wiebe, "Learning subjective adjectives from corpora," in *AAAI/IAAI*. AAAI Press / The MIT Press, 2000, pp. 735–740.

[12] J. Kamps, M. Marx, R. o. Mokken, and M. de Rijke, "Using wordnet to measure semantic orientation of adjectives," in *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*, vol. IV, Lisbon, PT, 2004, pp. 1115–1118. [Online]. Available: http://turing.science.uva.nl/ kamps/publications/2004/kamp:usin04.pdf

[13] T. Wilson, J. Wiebe, and R. Hwa, “Just how mad are you? finding strong and weak opinion clauses,” in *AAAI*, D. L. McGuinness and G. Ferguson, Eds. AAAI Press / The MIT Press, 2004, pp. 761–769.

[14] “Treetagger project website: http://www.ims.uni-stuttgart.de/projekte/corplex/treetagger/.”

[15] “Weka project website: http://www.cs.waikato.ac.nz/ml/weka.”

[16] G. Mishne and M. de Rijke, “Capturing global mood levels using blog posts,” in *Proceedings ofAAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, Stanford, US, 2006. [Online]. Available: http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-028.pdf

[17] ——, “Moodviews: Tools for blog mood analysis,” in *Proceedings ofAAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, Stanford, US, 2006. [Online]. Available: http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-029.pdf

[18] A. Dattolo and F. Luccio, “Formalizing a model to represent and visualize concept spaces in e-learning environments,” in *Proceedings of the 4th Webist International conference*, Funchal, Madeira, Portugal, 2008, pp. 339–346.