Journal of Internet Technology

Volume 9 (2008) No.4

網際網路技術學刊

Journal of Internet Technology

Special Issue on "TAAI 2008"

ISSN 1607-9264

Journal of Internet Technology

Electrical Engineering
National Dong Hwa University

Journal of Internet Technology

Volume 9 (2008) No.4

網際網路技術學刊

Journal of Internet Technology

Special Issue on "TAAI 2008"

ISSN 1607-9264

Journal of Internet Technology

Electrical Engineering
National Dong Hwa University

Journal of Internet Technology

Volume 9, No. 4 October 2008

Contents

Introduction to the Special Issue on Intelligent Agent and Knowledge Mining

The intersection between Computational Intelligence and Agent technology opens new significant scenarios in many application fields. In the formulation of Agent-based systems, the role of uncertainty is crucial for an efficient and coherent resolution of complex problems. In recent years there has been a growing awareness that Computational Intelligence handling of uncertainty in agents is equally important as other features of agent paradigm. In addition, the knowledge mining plays an important role in the intelligent agents. Knowledge mining is to extract desirable knowledge or interesting patterns from data with different formats for specific purposes. It has become a process of considerable interest in recent years, as the amounts of data in many databases have grown tremendously large. Many types of knowledge and technology have been proposed for knowledge mining.

This issue collects fully revised and extended versions of three contributions initially presented at the special session on (1) Intelligent Agent, held in Taiwan, on September 16-17, 2007, in the 12th Conference on Artificial Intelligence and Applications (TAAI 2007), (2) Ontology Application and Knowledge Management, held in Taiwan, on December 20-21, 2007, in 2007 National Computer Symposium, and (3) Intelligent Agent and Knowledge Mining, held in Singapore, on October 12-15, 2008, in the 2008 IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2008). It covers various facets of recent research at the intelligent agent, ontology, and knowledge mining.

This volume contains eleven papers that consider different aspects of perception based intelligent agent, ontology, and knowledge mining. The first three papers deal with ontology to apply to different research fields such as investment knowledge management, semantic web, and Capability Maturity Model Integration (CMMI). Applying Fuzzy Candlestick Pattern Ontology to Investment Knowledge Management by Chiung-Hon Leon Lee and Alan Liu proposes a fuzzy candlestick pattern based on ontology to assist in the candlestick pattern representation, storage, and reuse. Adaptable Distributed Ontology Alignment System by Chih-Hao Liu, Meng-Shium Tzou, Yong-Feng Lin, and Jason Jen-Yen Chen proposes an ontology alignment system with adjustable matching strategy and distributed processing to share the different ontologies across platforms and languages. The third paper A Novel Fuzzy CMMI Ontology and Its Application to Project Estimation by Mei-Hui Wang, Chang-Shing Lee, Zhi-Rong Yan, Hao-Han Chuang, Chi-Fang Lo, and Yi-Chen Lin proposes a novel fuzzy ontology for project planning and an ontology-based fuzzy agent, including a project planning ontology and a Takagi-Sugeno-Kang (TSK)-based project cost estimation, to estimate the total project cost.

The next five papers describe the application about the agent. In the fourth paper Single-Occupancy Simulator for Ambient Intelligent Environment by M. Javad Akhlaghinia, Ahmad Lotfi, Caroline Langensiepen, and Nasser Sherkat addresses the simulation of an occupant's behavior in a single-occupant ambient intelligent environment. The fifth paper Applying a Case-based Reasoning System Development Tool in the Design of BDI Agents by Ken Yen-Ru Cheng, Chiung-Hon Leon Lee, and Alan Liu, designs a Java Case-Based Reasoning Development Tools (JCBRDT) to make the system developers use CBR to export a Belief-Desire-Intention (BDI) agent, create a CBR system easily, and save the time on designing and maintaining the CBR system. The sixth paper Designing of an Autonomous Reinforcement Learning Agent for Dynamic Power Management in Embedded System by Cheng-Ting Liu and Roy Chaoming Hsu, describes a dynamic power management mechanism based on reinforcement learning agent to adaptively manage power consumption and service

achievability of an embedded system device. In Customized Advertising in E-Commerce Services Provision, Vincenzo Loia, Sabrina Senatore, Maria I. Sessa, and Mario Veniero introduce a web-centric system to provide a straightforward support to e-commerce market mediation through an agent-based architecture and fuzzy techniques. The paper Sentiment Classification for the Italian Language: a Case Study on Movie Reviews of Paolo Casoto, Antonina Dattolo, and Carlo Tasso evaluates the performance obtained by a set of high performance opinion polarity classifiers for the Italian language and a multi-agent is exploited to offer graph-centric views and navigation of the results.

The last three papers deal with the genetic algorithm and knowledge mining. In A GA-based Document Clustering Method for Search Engines Chun-Wei Tsai, Ming-Chao Chiang, and Chu-Sing Yang present a multiple search genetic algorithm to cluster the web pages returned by a search engine and provide a taxonomy of those web pages to the user. In the tenth paper A Modified Three-Phased Object-Oriented Mining Approach for Association Rules by Tzung-Pei Hong, Jun-Song Dong, and Wen-Yang Lin proposes a modified mining algorithm to derive association rules from object-oriented data with more pruning effects. In the last paper The Step Similarity Comparisons on Method Patents, Cheng-Yen Chen and Von-Wun Soo establish the technologies of automatic similarity analysis and comparison of two method patents in order to reduce the cost and human efforts.

As guest editors of this special issue, we thank the authors for their contributions. We also would like to thank Miss Mei-Hui Wang and Mr. Wei-Chun Sun, members of the Ontology Application & Software Engineering (OASE) Lab at National University of Tainan, Taiwan, for their supports of this special issue. We are most grateful to the referees for spending their valuable time in reviewing the manuscripts and providing kind cooperation and help. Finally, we greatly appreciated Professor Han-Chieh Chao, the Executive Editor of JIT, and the JIT for providing us with the opportunity to edit and publish this special issue, as well as for their valuable instructions in the editorial process.

Chang-Shing Lee, Guest Editor
Department of Computer Science and Information Engineering
National University of Tainan
Tainan, Taiwan

Vincenzo Loia, Guest Editor
Department Mathematics & Computer Science
University of Salerno
Salerno, Italy

Tzung-Pei Hong, Guest Editor
Department of Electrical Engineering
National University of Kaohsiung
Kaohsiung, Taiwan

Sentiment Classification for the Italian Language: a Case Study on Movie Reviews

Paolo Casoto, Antonina Dattolo, Carlo Tasso
Department of Computer Science
University of Udine
Italy
{paolo.casoto, carlo.tasso}@dimi.uniud.it, antonina.dattolo@uniud.it

Abstract

We consider the problem of tracking the opinion polarity, in terms of positive or negative orientation, expressed in documents written in natural language and extracted from a heterogeneous set of Web sources. More specifically, we focus our attention on the movie reviews domain. We are interested in evaluating the performance obtained by a set of high performance opinion polarity classifiers for the Italian language. Classification of polarity expressed by the input documents is achieved by means of several sets of specialized autonomous or interacting agents, devoted, respectively, to document gathering, classification and visualization. In particular the results of opinion analysis are represented by means of a graphical interface, where a multi agent based implementation of zz-structures is exploited to offer graph-centric views and navigation of results. The specific experimental evaluation performed so far shows an accuracy level, which is higher than previous results reported in the literature.

Keywords: Opinion Analysis, Sentiment Analysis, Cognitive Agents, zz-Structures, Machine Learning.

1 Introduction

The *subjectivity* [1] of a text is defined as the set of elements describing the private state of the writer. Assumptions, beliefs, thoughts, experiences, opinions, and judgments expressed in texts are typical clues of subjectivity. *Sentiment* is defined as the subset of subjective clues that can be measured in terms of positive, neutral or negative orientation.

Many rating indicators [1, 2, 3, 4, 5] have been introduced in literature to measure the orientation of the sentiment of a text; in this work, we adopt the *Opinion Polarity* (OP) indicator, defined in [2] as *"the classification that the author of a review would assign to it, if requested, with values expressed as positive or negative"*. More specifically we are interested in evaluation of the *Overall Opinion Polarity* (OvOP), defined as the opinion that arises from the document seen as a whole; OvOP does not depends on the opinions expressed by the single sentences of the text, which may also be contradictory.

Opinion Polarity Analysis (OPA) is the process aimed at identifying the OP of an input text and classifying such text accordingly to the polarity it expresses.

The collaborative generation of contents, which is one of the cornerstone of the Web 2.0 philosophy, is increasing enormously the amount of available subjective information. More specifically many sites allow users to publish opinions and reviews related to products or services, such as, for example, cars, cell phones or movies.

In order to handle such information in an efficient way a partially or fully automated approach to OPA is needed. OPA may be proficiency seen as a specific instance of the more general and well-known textual classification problem; instead of looking at topicality of a text, defined in terms of expressed topics, documents are assigned to the same class if they are characterized by the same opinion polarity. Works like [3, 5] show how OPA is more difficult than topic classification, because, as Pang observed [3], *"sentiment may be expressed in more subtle ways"* than topicality, which is strongly related with the presence of few specific concepts or pattern of concepts.

Subjective information expressed by writers may be very useful in many application fields; the most important one being business intelligence. The analysis of customers' opinion helps at identifying information useful for strategic marketing and brand monitoring, advertising, political campaigns and financial markets.

Many others application fields are widely described by Turney [6].

This work describes a multi-agent system devoted to OvOP analysis (OvOPA) on movie reviews extracted and aggregated from a set of heterogeneous Web sources. Classification is achieved by training in a supervised way a group of domain dependent high-precision classifiers. Each classifier acts as an autonomous agent evaluating the OvOPA of the input documents with respect to its specific domain and classification features. In this way, documents are extracted from the Web sources, enriched with the knowledge inferred by agents devoted to OvOPA-based classification, and visualized in an innovative way, adopting an agent-based extension [7, 8] of the Nelson's zz-

structures [9]. OvOPA is performed on Italian language texts, implementing and tuning some of the representation models presented in the literature for English language [3, 10], improved by introducing feature selection criteria.

The paper is organized as follows: Section 2 provides a description of the previous works concerning OvOPA and agent-based conceptual representations; Section 3 introduces and exploits the proposed system; Sections 4 and 5 describe and discuss the evaluation task and the obtained. Finally Section 6 closes the paper with a brief overview of our ongoing and future research activities.

2 Related Works

Two works mainly inspire our research: the experience of Pang and Lee [3, 5] and the work of Salvetti and others [2].

Both works aim to evaluate 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. Pang and Lee study the differences in precision using three different machine-learning techniques: Bayesian Networks, Support Vector Machines and Maximum Entropy (detailed references to these three methodologies can be found in [11]). They show that OvOP identification is a task harder than topic classification or generalization, both performed mainly by keyword identification. Salvetti, on the other hand, focuses attention on feature selection and feature generalization, integrating WordNet [12] as a repository of lexical relations.

Turney and others [6] defined OvOP analysis of a review 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. In particular they used AltaVista and its indexed corpus to investigate the amount of such co-occurrences. Other works, like [1], aim to merge 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. In particular attention is focused on syntactical structures, like negation or hypothetical speaking, which can modify the prior known polarity of the words.

OvOP analysis in Italian language is still an unexplored research field; many unsolved problems arisen during our research, mainly related with the lack of freely available tools for natural language processing of Italian documents. The lack of linguistic tools specialized in languages different from English has been previously analyzed, in the case of Romanian language, in [13]. Open issues are also related with the specific characteristics of Italian language (irregular forms, adjective declination, et. al.). No sets of Italian words with previously estimated polarity, which can be used during feature selection task, are available.

3 System Architecture

Figure 1 shows the general architecture of our system, dedicated to the automatic tracking of movie reviews written in Italian language.

System functionalities are grouped into a pipeline of three different modules, each one constituted by a set of agents devoted to different activities: the Harvesting Module, the OvOPA Module and the Navigation Module.

3.1 The Harvesting Module

The *harvesting module* is responsible for monitoring and crawling a set of Web sources and extracting from them the newly published texts containing movie reviews, which will be used as input of the OvOPA activity. Potential sources include Web sites, forums, blogs, and newsgroups. The extraction of the reviews is achieved, in the first prototype version of the proposed system, by using a set of autonomous agents devoted to Web crawling, each one of them implementing a parser specialized in extracting data from the specific Web source. The agents, constituting this module, continuously monitor the Web looking for new contents; when a new content is available, the source is accessed and the extraction activity takes place. Agents involved in the harvesting activity use the *Review Repository* as a centralized storage for the extracted contents.

For each review we extract the following information:

1. the title, assigned by the author to the review to summarize its content and give it emphasis;
2. the body of the review, which consists of a short natural language text;
3. the overall polarity rating indicator, already present in the review, when available: some of the selected sources

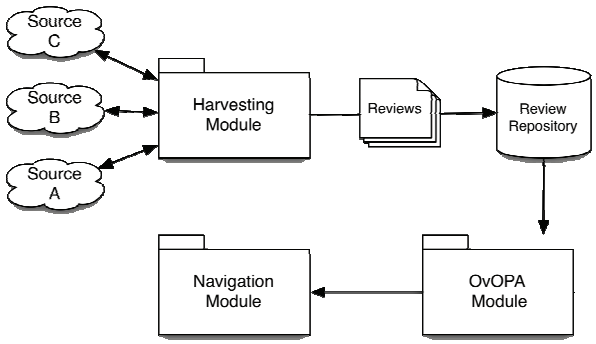

Figure 1 The Movie Reviews Classification Architecture

allow authors to summarize the polarity expressed in their reviews by means of different kinds of rating indicators, such as, for example, numeric values (from 0 to 5), marks (from A to F) or stars. In order to handle such heterogeneous indicators in a common way, all values have been normalized in a range between 0 (very negative opinion) and 1 (very positive opinion);

4. the publishing date of the review;
5. personal data about the author, like name, age and city of residence (if available).

Only the first three data (1. - 2. - 3.) are currently used during the OvOP analysis process, while the other two (4. - 5.) have been stored for future use.

In this work 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 0.4 and lower or equal to 0.6.

The main source of our corpus is the FilmUp (www.filmup.leonardo.it/opinioni) website, which collects a wide set of opinions, all in Italian language, about more than 4500 movies. Our choice is related with both the structure of the site, which allows us an easy extraction of reviews, and the presence, for each published review, of an overall polarity rating indicator. Data retrieved from this source can be proficiently used in training the OvOP classifiers.

In order to perform an evaluation of the quality achieved by the proposed classifiers, we collected more than 3000 reviews referring to 300 different movies or fictions. The distribution of polarity between reviews is not fair, as observed in [3] for the English corpus. The distribution of the overall polarity indicator pre-assigned to reviews is reported in Figure 2.

The positive reviews are 2038 (64.7% of the entire corpus), while the negative reviews are only 694 (22% of the collection).

3.2 The OvOPA Module

The OvOPA module is the core of our system; it implements the language processing and classification features. As shown in Figure 3 three different components constitute the OvOPA module: the Pre-processor, the Trainer and the Evaluator sub-modules.

The *Pre-processor* sub-module is constituted by a set of configurable and autonomous agents, aimed at identifying linguistic and statistical features appearing in the set of input documents and at building a representation, suitable for automatic classification of expressed overall polarity, of the collection of reviews. A feature is defined as a property of a document, which can be useful in discriminating the right class of the document.

Letting $F = \{f_1,...,f_m\}$ a fixed set of features, each document D retrieved from the review repository is transformed into a vector $\vec{d}$, where each component d_i represents the weight of the feature f_i evaluated on D. Each component d_i may assume several different kinds of values, such as, for example, boolean or numeric values, with respect to the specific semantic assigned to f_i. Several different kinds of linguistic and statistical features have been proposed and evaluated in the literature for the OvOP analysis of texts written in English language. In order to train a set of high-performance OvOP classifier for the Italian language, we implemented and evaluated some of the features described in [2, 3]; in particular, we aimed at trying to over perform the results exploited for the English language introducing linguistic tasks like stopword removal and stemming, aimed at reducing the side-effects typical of the grammar of the Italian language like, for example, the declination of adjectives. The agents devoted to the pre-processing activity adopt and integrate several language-independent resources, such as the n-gram generator and the punctuation analyzer, and/or language- dependent resources, like the Part-of-Speech tagger, the stemmer, and the list of stopwords, which have been properly specialized for the Italian language.

The vector representations of the document D is generated by the pre-processor and may be used in two different ways: when D is added to the Review Repository, it may be send to the *Evaluator* sub-module to be classified or, if no classifier agent has been already trained, to the

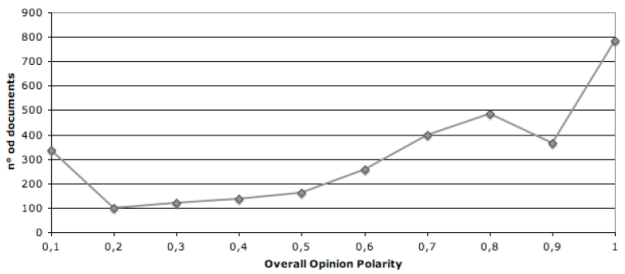

Figure 2 Distribution of Pre-assigned OvOP Values, Related to Automatically Extracted Reviews

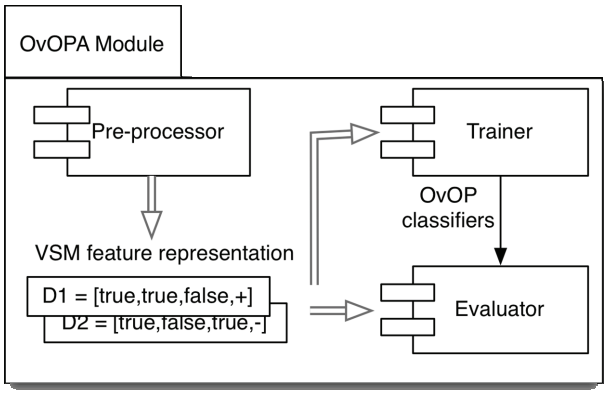

Figure 3 The OvOPA Module

Trainer sub-module. Only documents, characterized by a pre-assigned polarity, can be used as training set in the scenario described by the latter case.

The training task is performed by the *Trainer* sub-module; it uses a given training set of pre-judged documents, represented as a matrix M with dimension $n*m$, where n is the number of documents of the training set and m the number of features adopted in document representation, in order to train a set of OvOP classifiers. In this work only two algorithms for training of supervised classifiers have been exploited: Naïve Bayes (NB) and Support vector Machines (SVM), described in [11] Far from implementing such algorithms, we used the WEKA[1] library for machine learning. The *Trainer* sub-module may be seen as an agent generator, devoted to the creation of agents aimed at OvOP classification.

The output of the training activity is a set of OvOP classifiers; the trained classifiers are assigned to a set of autonomous agents, devoted to classify the incoming documents; these agents act into the third component of the OvOPA module: the *Evaluator* sub-module. According with the classification score estimated by each of trained OvOP agents, the *Evaluator* sub-module assigns newly added documents to one of two possible classes (positive or negative), describing the polarity expressed by the documents.

3.3 The Navigation Module

The navigation module supports users in two main functions:

1. dynamic and personalized access to movie's reviews;
2. creation of new interconnections among existing data and knowledge.

Knowledge inferred during the OvOPA activity is stored in zz-structures, a graph-centric model for organizing data and computing [7, 9]. A zz-structure can be represented as an edge-colored multi-graph, with the *restriction that every vertex*, called zz-cell, *has at most two incident edges of the same color*; each sub-graph, containing edge of a unique color, is called *dimension*. Given the previous restriction, the cells in a same dimension are linked into one or more linear and directed sequences, called *ranks*. In our model, a zz-cell is associated to each movie's review; colored edges represent semantic interconnections among reviews; examples of dimensions are the set of reviews related to a same movie, containing same keywords, expressing a positive (or negative) opinion.

Figure 4 shows a view related to the list of reviews found searching for "*Johnny Depp*"; each review is placed in a zz-cell.

[1] WEKA Web Site: www.cs.waikato.ac.nz/ml/weka

Figure 4 Set of Reviews Obtained by Search

Each cell is composed by the movie's title (e.g. Sweeney Todd, in first zz-cell), an emoticon for identifying the polarity of the review (positive or negative), a short reference to the review ("*Spettacolare è dire poco!…*"), the publishing date (e.g., 01-03-2008), the data source (e.g., FilmUp), the search tool and an advanced tool button (identified by the "+" symbol"). Title, emoticon, date, source and search tool are clickable and are associated to related dimensions; as highlighted (with red color associated to search tool icon) in Figure 4, current horizontal dimension contains the sequence of ten results obtained searching for "*Johnny Depp*".

If user clicks on title "*Pirati dei …*" present in third cell, dimension related to the same title reviews is visualized, as shown in Figure 5. The cell with red border represents the last user selection, while all reviews with same title are visualized in vertical dimension and are marked by red color associated to titles. In this way, user has access to a Cartesian view, on two semantic dimensions: *"Johnny Depp" and* title *"Pirati dei …"*. Next user clicks will propose new views on user chosen dimensions.

The advanced tool button enables users to add one or more reviews in a new dimension, tagging existing reviews.

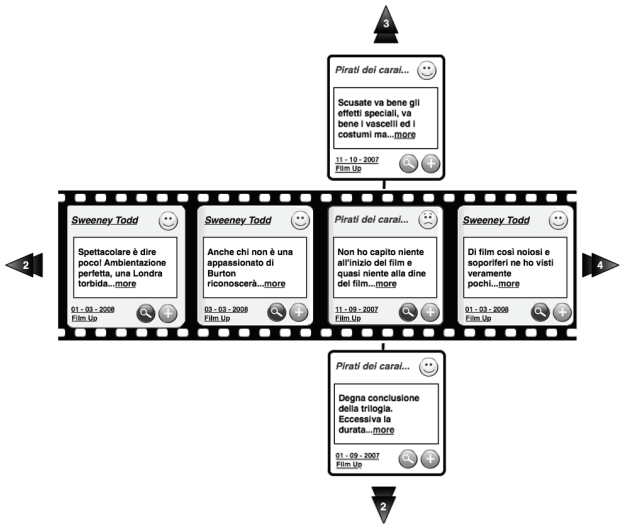

Figure 5 A View Related to Dimensions "Johnny Depp" and "Pirati dei Caraibi"

A new dimension will be composed by all reviews labeled with the same tag. In addition, it allows user to browse and visualize reviews selecting a dimension from the set of user created tags.

The *Navigation* module manages and visualizes the users' concept spaces. According to [15] a concept space is defined in terms of a multi-agent system composed of five types of agents: concept maps, dimensions, ranks, composite and atomic cells.

These five agent classes represent five abstraction levels of the concept space: more in details, concept maps know and directly manipulate dimensions and isolated cells; they include concepts and relationships between concepts, that are organized in dimensions. Dimensions, uniquely identified by their colors (or equivalently, tags), know and manipulate their connected components, i.e., their ranks. Ranks know and coordinate the cells and the

links that connect them; composite cells contain concept maps related to more specific topics, and finally atomic cells are primary entities and directly referenced documents.

In particular, in this work, the conceptual space represents the set of information extracted from the Web and properly enriched by means of an accurate OvOPA activity. The agents constituting the *Navigation* sub-module interact with each other and with the users, in order to perform the evolution of the set of visualized information. In order to proficiently implement the conceptual spaces described by the zz-structure model, agents are organized in the described hierarchy of levels, each one constituted by a set of agents specialized in a specific entity constituting the zz-structure.

4 Experimental Results

To perform our experimental evaluation, we extracted a subset of review repository by choosing randomly 500 reviews with polarity greater than 0.6 and 500 reviews with polarity lower or equal to 0.4, in order to obtain a training set with a balanced distribution of documents between classes.

For each selected review, the body and the title, when available, are loaded from the database and merged together as a unique field. 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. Stopword removal and stemming are applied to each of the selected reviews, in order to reduce the dimension of the vector representation and the sparseness of the matrix *M* associated to the training set.

In this work we focus our attention on the three following approaches to document representation:

1. *U3*: the set of unigrams occurring 3 or more time in the whole training set;
2. *UB3*: the set of unigrams and bigrams occurring 3 or more time in the whole training set;
3. *UBT3*: the set of unigrams, bigrams and trigrams occurring 3 or more time in the whole training set.

The weight assigned to a document d_i with respect to a selected n-gram n_j is defined as

$$\begin{cases} 1 & \text{if } \mathrm{occ}(n_j, d_i) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\mathrm{occ}(n_j, d_i)$ is the number of occurrences of the n-gram n_j in document d_i. The proposed weighting metric is simpler than other weighting metrics, such as the TF-IDF, usually adopted in text classification; it has been proven [3], that occurrence based representation is more effective for OvOPA than term frequency. Preliminary results, obtained comparing the results produced by classifiers based on TF with classifiers based on n-gram occurrence, show how the second weighting metric out-perform the first one when applied to all U3, UB3 and UBT3. In this work we add an extended set of features to the ones included in U3, UB3 and UBT3; additional features are used to analyze and measure some punctuation and the statistical properties of each input review. More specifically, for each document, we evaluate:

1. the number of question and exclamation marks;
2. the number of sentences;
3. the number of long words (7 characters or more);
4. the average length of a sentence;
5. the average length of a word.

Some works [1, 14] show that these fields are good clues in identifying the right polarity of an input document, independently from the specific domain of the training set.

Six different classifiers have been trained, applying U3, UB3 and UBT3 to both the NB and the SVM algorithm. Evaluation is achieved with a 3-cross folding methodology, implemented in the evaluator component.

The performance of the built classifiers is measured in terms of *accuracy*, defined as the percentage of correct classification. In particular, we are interested in evaluating the $accuracy_+$ and the $accuracy\_$, measured with respect to the subset of positive and negative reviews respectively.

Table 1 shows the global *accuracy* of each of the trained classifiers, grouped with respect to the machine-learning algorithm used in training.

Table 1 Average Accuracy of the Six Classifiers

	U3	UB3	UBT3
Naïve Bayes	82.2	82.4	82.5
SVM	**84.9**	**84.4**	**84.4**

Table 2 Average and of the Six Classifiers

	U3	
	accuracy$_+$	*accuracy*$_-$
Naïve Bayes	85.4	79.0
SVM	85.2	84.6
	UB3	
	accuracy$_+$	*accuracy*$_-$
Naïve Bayes	86.2	78.6
SVM	83.8	85.0
	UBT3	
	accuracy$_+$	*accuracy*$_-$
Naïve Bayes	86.2	78.8
SVM	83.6	85.2

Table 2 reports the same data presented in Table 1 expressed in terms of measured *accuracy*$_+$ and *accuracy*$_-$; these two indexes allow us to verify if one of the two classification tasks, related to positive or negative reviews, is harder than the other.

The second part of our experimental evaluation has been aimed at computing the performance achieved by a new set of classifiers, trained after the introduction, into the training process, of a feature selection task. Feature selection is the task that aims at identifying the subset of features, which are more useful in assigning a set of documents to a group of classes. Reducing the noise introduced by sparsity of data, feature selection allows the trained classifiers to achieve better performance and to reduce the computational cost needed for their evaluation. In this work we adopt the *Information Gain* (IG) feature selection metric. 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. Following equation reports the general formulation of IG with respect to the input feature *t*:

$$IG(t) = -\sum_{i=1}^{m} \Pr(c_i)\log\Pr(c_i) +$$
$$+\Pr(t)\sum_{i=1}^{m} \Pr(c_i|t)\log\Pr(c_i|t) +$$
$$+\Pr(\bar{t})\sum_{i=1}^{m} \Pr(c_i|\bar{t})\log\Pr(c_i|\bar{t})$$

where $\{c_i\}_{i=1}^{m}$ represents the set of available classes.

Each of the identified features *t* can be ranked, accordingly with the respective, *IG(t)* value. Only the *n* best features are used for representation of the training set, where *n* is a parameter representing the number of features included into the document representation. Table 3 displays the list of the 50 features with the highest IG value when feature ranking is applied to the UBT3 representation model.

Table 3 Top 50 Features Extracted from the Training Set with the Highest IG Value

1	BELLISSIM	26	SPARROW
2	BRUTT	27	SONOR
3	BELL	28	ORREND
4	JACK	29	WILL
5	PESSIM	30	SCHIFEZZ
6	OTTIM	31	BEL
7	FANTAST	32	FAVOL
8	EVIT	33	HARRY
9	PEGGIOR	34	JOHNNY
10	DELUSION	35	COLONN
11	BRAVISSIM	36	INUTIL
12	PO'	37	STRAORDINAR
13	PIAC	38	"COLONN SONOR"
14	NOIOS	39	"FILM BELLISSIM"
15	RIDICOL	40	BRAV
16	INTERPRET	41	ECCEZIONAL
17	BAST	42	DEPP
18	SIMPSON	43	GRINDHOUS
19	BUTT	44	NOI
20	"JACK SPARROW"	45	STUP
21	SPLENDID	46	INSULS
22	ATTOR	47	"OTTIM FILM"
23	PIR	48	STUPID
24	GRAND	49	JONES
25	PERFETT	50	MOLT

Table 4 Average Accuracy of the U3 Based Classifiers after Feature Selection

	U3		
	50 feat.	**100 feat.**	**250 feat.**
Naïve Bayes	81.0	83.8	83.8
SVM	83.2	86.2	85.5
	500 feat.	**1K feat.**	**2K feat.**
Naïve Bayes	85.6	86.7	85.4
SVM	86.8	**87.5**	85.7

Table 5 Average Accuracy of the U3 Based Classifiers after Feature Selection

	UBT3		
	50 feat.	**100 feat.**	**250 feat.**
Naïve Bayes	80.1	84.4	84.5
SVM	82.4	85.5	85.9
	500 feat.	**1K feat.**	**2K feat.**
Naïve Bayes	85.4	86.6	86.6
SVM	87.2	87.9	**89.0**

Tables 4 and 5 show the accuracy of the trained classifiers, based respectively on the U3 and UBT3 representation, at varying of dimension of the relevant features' set.

The number of unigrams appearing at least 3 times into the training set is lower than 3000, so we decided, for such family of classifiers, to limit the evaluation of the feature selection improvements to a set of 2000 potential features.

Figure 6 displays the accuracy curves obtained from data in Table 4 and 5.

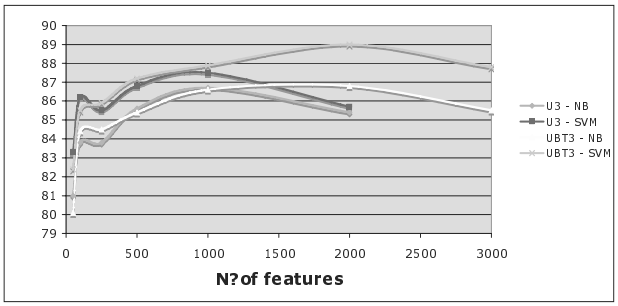

Figure 6 Feature Selection and Accuracy for Both NB and SVM Classifiers

5 Discussion

The obtained results confirm that the introduced assumptions [3], related with the ability of unigrams to express the OP of a given text, can be also applied to the Italian language. More specifically representation models based entirely or partially on the occurrence of unigrams lead to the best performances, with both the adopted training approaches NB and SVM. Although evaluated on different corpora, constituted by documents written in English language and with specific dimensions and polarity distribution, the classifiers trained by [2, 3] shows performances comparable to those presented in this work.

Our work shows that SVM classifiers clearly over perform NB ones in OvOP classification task, as previously stated by [3, 5]. In particular, analyzing values reported in Table 2, it is clear that NB classifiers tend to classify positive reviews better than negative ones, while SVM classifiers tend to be more fair, showing similar accuracy values for instances of both positive and negative classes.

Feature selection may improve the accuracy achieved by the trained classifiers; in particular the IG selection metric shows an average improvement in accuracy between 2 and 4.6 percent. The highest improvement in accuracy is achieved when IG is applied to the UBT3 representation model including the SVM strategy. Feature selection may also be useful to distinguish between features that are effective in OvOP classification and features that are not so effective and that can introduce noise in the document representation. In particular, the results reported in Table 4 and 5, obtained by varying the number of features used to build the vector representation of the documents constituting the training set, can be used to identify the set of features that achieve the best performance. Looking manually at the set of the top features, included in Table 3 and ranked with respect to the IG metric, it is possible to identify the stems of strongly polarized adjectives like, for example *bellissim*=wonderful, *brutt*=ugly, *pessim*=worst, the stems of adverbs used as adjective amplifier, such as *po'*=quite, and the stems related with domain specialized terms or multi-terms, like, for example *"colonn sonor"*=soundtrack or *interpret*=actor.

Accuracy achieved by the classifiers trained on the selected set of features is better than all results previously described in [2, 3].

In the future we expect to increase the size of our collection by monitoring and crawling the set of selected Web sources; a larger collection of labeled documents will allow us to study how accuracy of the trained classifiers may change accordingly with the size of the training set.

6 Conclusions and Future Works

Our work shows how OvOP classification can be achieved effectively in the Italian language using machine-learning techniques, originally developed for English, applied to the movie domain. In particular we proved experimentally that stemming, stopword removal and feature selection may increase the accuracy of the trained OvOP classifiers, allowing them to achieve and out perform the accuracy of some of the systems described in the literature for the English language. The research is ongoing and in the next months we will focus our attention on following aspects of OPA:

1. the formalization of new features useful if document representation and their evaluation;
2. the evaluation of a series of classifiers based on sentence level OP instead of OvOP.

The first goal will be achieved by introducing new features able to identify OvOP; more specifically, we are interested in studying patterns suitable for extracting emotive icons and idiomatic expressions. The development of a parser able to identify and spread the negation between terms of a sentence may also be useful, as proven in [3], to increase the precision of the classification process.

The 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 training set. High performance classifiers, based on domain dependent and independent resources, are described in [10].

The proposed multi-agent system provides the flexibility we require to train and evaluate a set of agents devoted to OvOP classification. A new set of features can be added to the system by providing a new agent to the

pre-processor module able to extract the required features from the set of extracted documents. In the same way, a new source can be easily added to the set of monitored Web sources by defining a new specialized agent devoted to data extraction and adding it to the harvesting module.

References

[1] J. Wiebe, T. Wilson, R. F. Bruce, M. Bell and M. Martin, *Learning subjective language*, *Computational Linguistics*, Vol. 30, No. 3, 2004, pp. 277-308.

[2] F. Salvetti, S. Lewis and C. Reichenbach, *Impact of lexical filtering on overall opinion polarity identification*, *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Stanford, US, 2004.

[3] B. Pang, L. Lee and S. Vaithyanathan, *Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-02)*, Philadelphia, Pennsylvania, 2002, pp. 79-86.

[4] P. D. Turney, *Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews*, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002),* Philadelphia, July, 2002, pp. 417-424.

[5] B. Pang and L. Lee, *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts*, *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, Spain, July 21 - 26, 2004, pp. 271-278.

[6] P. D. Turney and M. L. Littman, *Measuring praise and criticism: Inference of semantic orientation from association, ACM Transactions on Information Systems*, Vol. 21, No. 4, 2003, pp. 315-346, .

[7] A. Dattolo and F. Luccio, *A new actor-based structure for distributed systems, Proceedings of the IEEE MIPRO International conference on Hypermedia and Grid Systems (HGS07),* Opatija, Croatia, May 21-25, 2007, pp. 195-201.

[8] A. Dattolo and F. Luccio, *Formalizing a model to represent and visualize concept spaces in e-learning environments, Proceedings of the 4th International conference on Web Information Systems and Technologies*, Volume 1, Funchal, Madeira, Portugal, May 4-7, 2008, pp. 339-346.

[9] T. Nelson, *A cosmology for a different computer universe: data model mechanism, virtual machine and visualization infrastructure*, *Journal of Digital Information: Special Issue on Future Visions of Common-Use Hypertext*, Vol. 5, No. 1, 2004.

[10] A. Aue and M. Gamon, *Customizing sentiment classifiers to new domains: a case study*, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-05)*, Borovets, Bulgaria, September 21-23, 2005.

[11] T. Hastie, R. Tibshirani and J. H. Friedman, *The Elements of Statistical Learning*, Springer, August 2001.

[12] C. Fellbaum, *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, MIT Press, 1998.

[13] R. Mihalcea, C. Banea and J. Wiebe, *Learning Multilingual Subjective Language via Cross-Lingual Projections*, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, Prague, Czech Republic, June 23 – 30, 2007, pp. 976-983.

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

[15] A. Dattolo and F. Luccio. Visualizing Personalized Views in Virtual Museum Tours. *International Conference in Human Computer Interaction (HSI),* Krakow, Poland, May 25-27, 2008, pp. 109-114.

Biographies

Paolo Casoto is a PhD student in the Department of Mathematics and Computer Science at the University of Udine, Italy. He has the MSc in Computer Science from the University of Udine (2006).

His research interests include artificial intelligence, machine learning, natural language processing and software testing.

Antonina Dattolo is assistant professor in Computer Science at the Department of Mathematics and Computer Science of the University of Udine. She received the MSc in Computer Science with full marks from the University of Salerno in 1990 and the Ph.D in Applied Mathematics and Computer Science from University of Naples Federico II in 1997.

She is author of more than 60 original research papers in international journals, book chapters, and in international conference proceedings. Her current research interests in-

clude concurrent architectures for distributed hypermedia models, adaptive hypermedia, new generation Web, multi-agent systems, conceptual maps and authoring tools in Web 2.0.
Dr. Dattolo serves as a reviewer for International Journals, is member of Technical Committees of International Conferences, and has coordinated some European research projects related to e-learning, computer-science and cultural heritage fields.

Carlo Tasso is Full professor in Computer Science at the Department of Mathe- matics and Computer Science of the University of Udine, where he is also the Dean of the Faculty of Sciences. He has been one of the founders of the Italian Association for Artificial Intelligence and, in 2001, founder of the infoFACTORY Group, the first ICT spin-off company of the University of Udine. He is author of more than 120 scientific publications. His current research interests include personalized information filtering and knowledge management within Web 2.0.
Prof. Tasso is member of AAAI and ACL; program chairman and general Chair of several scientific conferences, such as User Modeling (UM) conferences, Adaptive Hypermedia (AH) conferences, ACM SIGIR conferences. Member of the Editorial Board of the User Modeling and User Adapted Interaction Journal (UMUAI), Scientific Editor of the series CISM courses and Lectures published by Springer Verlag Wien - New York.

Journal of Internet Technology

Volume 9, No. 4 October 2008

Contents

ISSN 1607-9264

9 771607 926000