

Validity and reliability of the automatic classification of texts according to the negative-positive criterion

Luca Giuliano ¹, Gevisa La Rocca ²

¹ Sapienza Università di Roma – Piazzale A. Moro – 00185 Roma – Italia

² Università degli Studi di Palermo – Viale delle Scienze – 90128 Palermo – Italia

Riassunto

Il criterio di classificazione positivo-negativo è uno dei più rilevanti per la qualificazione di un testo. I primi studi risalgono a Osgood et al. (1957), con riferimento alle dimensioni principali del significato identificate in valutazione, potenza e attività. Successivamente Boucher e Osgood formularono la nota Pollyanna hypothesis (1969). Più di recente Bolasco e della Ratta-Rinaldi (2004), riprendono il lavoro su questa ipotesi con lo scopo di costruire uno strumento di analisi che permetta di categorizzare un testo a partire dalla presenza degli elementi valutativi che il testo contiene. Nella nostra comunicazione, intendiamo riprendere questo tema sotto il profilo della validità e affidabilità del criterio positivo-negativo quando si faccia ricorso a tecniche di classificazione automatica dei testi. Sotto il profilo della affidabilità si tratta di controllare l'effettivo potere discriminante dell'indice di negatività per il quale, in un'analisi automatica, non si può tenere conto del contesto complessivo della misura. Queste considerazioni metodologiche sono applicate a un corpus in lingua inglese costituito da articoli sulla Cina di quattro quotidiani del Regno Unito: *The Times*, *The Guardian*, *The Independent* e *The Daily Telegraph* dal 2000 al 2008.

Abstract

The positive-negative classification criterion is one of the most important to classify a text. The first studies date back to Osgood, Suci and Tannenbaum (1957), with reference to the main dimensions of meaning identified in evaluation, potency and activity. Later on Boucher e Osgood formulated the well known *Pollyanna hypothesis* (1969). Recently Bolasco and della Ratta-Rinaldi (2004) have taken up their work on this hypothesis with the aim to build an analysis tool capable of categorizing a text moving from the presence of the evaluative elements it contains. In our paper, we wish to develop this topic from the point of view of the validity and reliability of the positive-negative criterion when we make use of techniques of automatic text classification. From the reliability point of view we have to check the method of selection of the negativity index, according to which in an automatic analysis it is not possible to ignore the overall context of measurement. These methodological considerations have been applied to a corpus in English of articles about China taken from four UK newspapers: *The Times*, *The Guardian*, *The Independent* and *The Daily Telegraph* from 2000 to 2008.

Keywords: text mining, categorisation, evaluation, positive and negative

1. Background: in search of the meaning of words

The positive-negative classification criterion is one of the most important for the classification of a text; as such, over the years, it has aroused the interest of many scholars.

The Measurement of Meaning is, in fact, a problem which has already been discussed and solved by Osgood et al. (1957) using the idea of semantic differential. This will then be reused also by Holsti in a work published in 1969. The ambit in which Osgood *et al.* in 1957 faced the issue

of the measurement of the meaning of “words” considering human behavior is that of Content Analysis. The creation of categories and the attribution of terms to these categories is the core of this technique. When the researcher creates categories he makes use of his/her knowledge of the text or of the topic the text is about to be able to generate them. In the attribution of text units inside a specific category, it is possible to use extensional lists of terms which the researcher decides must be inserted in the category created. By using a software the researcher proceeds first to decide which words are to be placed in a specific category and he/she creates this way a sort of category dictionary of reference; secondly, he/she inserts the lists which have been created in the computer and asks for the KWIC (*Key Words In Context*¹). If the computer can easily solve the problems inherent to the calculation of occurrences, it must be instead the researcher to decide whether a word should be inserted in one category or in another (La Rocca, 2007). The authors mentioned above have noticed that there is no generalizable system of concept attribution to the categories and there is, therefore, no semantic differential test, still, in fact, to be developed.

It is from here that what could be defined as “an analysis of word intention” started off; research through the years has shown us how difficult it can be to identify the univocal meaning of a word and its classification, whether it is a noun, a verb, an adjective or any other part of speech. One of the most used techniques is the one based on semantic scales; the logic behind this operation is the same of a level scale where on the two opposite poles you can find two adjectives. It is necessary to establish – by assigning a score – if the unit of analysis is closer to one pole or to the other. The three original dimensions identified by Osgood et al. and used also by Holsti (1969) are: *evaluation (positive versus negative effect)*, *potency (strength versus weakness)* and *activity (active versus passive)*.

Although the use of scales in the creation of categories can sometimes be complicated, it is important to remember that every scale is semantically anchored to the attribution of a common meaning of an opposite pole of adjectives. In an approach based on quantitative semantics to establish/attribute meaning to data is to develop a way for their measurement. Once the measurement scale has been developed, it is then necessary to prepare a technique to analyze the results obtained. Osgood et al. consider both factorial analyses and three-dimensional representations, such as Multi Dimensional Scaling (MDS).

Years go by and it is Osgood himself in 1969 together with Boucher to go back on the issue formulating the well known Pollyanna hypothesis, according to which there is a tendency in communication affirming that we use more frequently words with a positive valence than words with a negative one. This hypothesis, in turn, generates a numerous combination of further hypotheses. In their work the authors test three specific ambits, that is: 1) that even working with 13 different linguistic communities the positive linguistic elements (E+) will have higher occurrences than their opposites (with the symbol E- the authors indicate the negative linguistic elements); 2) that also in the case of prefixes it is more frequent to add negative ones to words with a positive valence and not vice-versa; 3) that in the development of language in children positive qualifying elements are more frequent and appear before the negative ones. The results reported in their studies confirm the initial hypotheses and therefore also the validity of the *Pollyanna hypothesis*.

¹ The KWIC lists show the context in which we can find a word and they perform mainly three functions: They indicate the variability and the coherence in the meaning or the use of the words; They offer structural information in order to determine if the meaning of some words is decided by their use within specific sentences; They provide information on the occurrence of words.

The topic of semantic classification of documents is returned actually thanks to the work on taxonomy language belonging at Appraisal framework (Martin and White, 2005), This approach organizes language around the significance of three categories: *attitude* (propositions that express feelings), *engagement* (the context in which the proposition is inserted), *graduation* (the scaling mechanism that regulates the intensity and quantification of attitudes). The *affect* is the primary *attitude* and develops along the positive-negative polarity with different intensities.

Recently, in the context of computational linguistics have received increasing attention the studies on “sentiment analysis” or “opinion mining”, also through the dissemination of digital text in web and consequently of importance that has taken the automatic classification of texts in terms of evaluative (Pang et al., 2002; Pang and Lee, 2008). In this area identified three main objectives: to assess the subjectivity / objectivity of a linguistic unit (word, phrase, sentence or document), specify the polarity of orientation and its intensity (Esuli and Sebastiani, 2005). Our work takes the approach of Bolasco and Ratta-Rinaldi (2004), which claim to be *Pollyanna hypothesis* in order to build an analysis tool that allows to categorize a text from the presence of evaluative elements that the text contains. Bolasco and della Ratta-Rinaldi have taken up their work on this hypothesis with the aim to build an analysis tool capable of categorizing a text moving from the presence of the evaluative elements it contains. To be able to compare the corpus of analysis with a list of positive and negative adjectives, Bolasco e della Ratta-Rinaldi used the tools produced within the *General Inquirer* (GI) system. They use one of the GI dictionaries and from this they extracted a list of adjectives capable of providing a first comparison tool. From this dictionary come from a total of 1.000 adjectives lemmatization subsequently translated and adapted in Italian by the authors. These 1.000 terms “exploded” become 6.500 graphic forms. For the evaluation of the text tone, the two authors, following Marchand’s lesson (1998), chose as category the adjectives. Marchand, in fact, stresses the importance of adjectives to evaluate the rating of a text. The positive-negative dictionary created this way by Bolasco and della Ratta-Rinaldi was then applied to different kinds of texts verifying the *Pollyanna hypothesis* and establishing a range per negative index ($\text{Occ. Neg/Pos} \times 100$) varying between 50% for big size texts and 40% for smaller size ones, such as, for example, the POLIF dictionary, which has been inserted as reference model in TalTac.

In this paper we recall this topic from the point of view of the validity and reliability of the positive-negative criterion when one uses techniques of automatic text classification. Our attention focuses also on marker categories because they have a different weight when addressing the positive or negative tones of speech.

1.1. A preliminary tool: The Inquirer dictionary

In the development of an analysis aiming at evaluating the validity and the reliability of the automatic classification of texts according to the negative-positive criterion, a useful resource is, what has been developed over time by the *General Inquirer system*. In the *Inquirer dictionary* (Id) used in the analysis here carried out are present categories extracted from four different sources: the Harvard dictionary IV-4; the Laswell dictionary developed considering some categories of value; other categories of recent creation and then some marker categories. Moving from this dictionary our objective was to hone the positive-negative category applying it to a specific corpus of articles which appeared in: *The Guardian*, *The Independent*, *The Daily Telegraph*, *The Times* from 2003 to 2008 and with China as subject. This work of building up a *specific dictionary* starting from a *generic one* is what Stone et al. (1966) had already pointed out in *The General Inquirer: A Computer Approach to Content Analysis*. The positive-negative category present in the *Inquirer dictionary* descends from the work of Osgood et al. (1957)

on the three different semantic dimensions, then adopted by Holsti in the *Stanford Political Dictionary* of 1965. At present this dictionary contains 1.915 words considered positive and 2.291 words considered negative. In this list of nouns structured in categories it is possible to find nouns present in more than one category, and this because as Stone et al. already affirmed: «words, like other objects, can be classified in many different ways. If we have some lemons, some bananas, and some apples, we can classify the lemons and bananas together on the ground that they are yellow and the apples red; we can put the lemons and apple together on the ground that they are round and the bananas long; or we can put the apples and bananas together because they are sweet and the lemons sour. The number of possible classification of words in a language on a semantic basis is infinite» (1966, p.137).

To classify a semantic category automatically means to have at one's disposal a list of graphic forms capable of avoiding as much as possible the problem of semantic ambiguity. In this case the ambiguity could be caused by a few polysemic words and, in particular by bisemic words when referred to the negative-positive polarity: These are words which are positive or negative according to the context. The first thing to do is therefore to examine and reduce the list of the Id, to a list of words exclusively classifiable in the positive-negative categories.

The list of the negative-positive words, examined from this point of view, brings to the following observations (Tab. 1). In the list of the Id the polysemies identified (278 positive headwords and 286 negative ones) are not classifiable according to "evaluative" semantics, but according to other meanings which are independent from it. Only 16 headwords are bisemic from the evaluative point of view. Their presence is therefore of no influence in the overall attribution of tone to a corpus with a rather extended lexicon. The list reduced to the headwords with one only entry consists of 1.637 positive headwords and 2.005 negative ones.

<i>Headwords Id</i>	<i>NEGATIVE</i>	<i>POSITIVE</i>
Monosemic	1,989	1,621
Bisemic (Pos&Neg)	16	16
Polysemic	286	278
Total	2,291	1,915
Unique headwords	2,005	1,637

Table 1: Headwords of the Inquirer dictionary per semantic category

<i>Categories Id</i>	<i>NEGATIVE</i>	<i>% NEG</i>	<i>POSITIVE</i>	<i>% POS</i>
Nouns	901	44.94	716	43.74
Adjectives	643	32.07	587	35.86
Verbs	437	21.79	294	17.96
Adverbs	22	1.10	40	2.44
Personal pronouns	1	0.05	-	-
Prepositions	1	0.05	-	-
Total	2,005	100.00	1,637	100.00

Table 2: Marker classification of Positive-Negative headwords in TreeTagger ²

² The grammatical tagging on the terms of the IG was performed with TreeTagger, <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>.

If we examine the list considering the marker categories, it seems more informative (Tab. 2). The adjectives, which are certainly decisive in classifying as positive or negative the tone of a speech, are present in the list with respectively 35.86 and 32.07% of the headwords. The presence of verbs and nouns is too significant to be neglected, but we must stress the different relevance verbs have inside each of the two semantic categories: 17.96% among the positive headwords and 21.79% among the negative ones. The choice to include or not the verbs in the list is decisive, as we will see, when classifying the tone of speech.

2. The corpus analyzed

The experimental control of the validity and reliability of the list of adjectives of the GI for the construction of a negativity indicator has been carried out on a corpus of 4.216 articles of four newspapers published in the United Kingdom from 2003 to 2008 (Tab. 3). The corpus has been normalized in TalTac 2.5 and entered as a headword for the English language with TreeTagger. On the corpus already lemmatization were made all the analysis of semantic tagging using the list of positive and negative terms extracted from the GI. The indices of negativity were calculated for each grammatical category, the uotput of TreeTagger has been imported into TalTac as external resource.

Token words (N)	2,558,606
Type words (V)	69,981
type/token ratio = (V/N)*100	2.73
% hapax	43.97
Mean = N/V	36.56

Table 3: Lexico-statistic characteristics of the corpus* (* Software used: TalTac 2.5 ³)

3. Towards the construction of a negativity indicator

The list of positive-negative headwords applied to the semantic tagging of the corpus China-UK allows us to observe how the negativity index (Occ. Neg/Pos*100) is very sensitive to the headwords used from the grammatical point of view (Tab. 4). Using the complete vocabulary we obtain a negativity index of 66.36 whereas with the list of adjectives only the index is less negative (59.75). The overall score of the negativity index is decisively influenced by the presence of the verbs which, by themselves, have a negativity index of 83.19.

<i>Grammatical Category</i>	<i>NEGATIVE</i>	<i>POSITIVE</i>	<i>N/P*100</i>
Adjectives	20,319	34,085	59.61
Verbs	27,623	24,484	112.82
Nouns	63,362	67,055	94.49
Others	4,675	7,717	60.58
All categories	115,979	133,341	86.98

Table 4: Negativity index of the corpus China UK

From Tab. 5 we can observe how the evaluation of the newspapers brings to different results according to the composition of the indicator. The ADJ indicator reveals that The Guardian

³ S. Bolasco (2000); Bolasco et al. (2004).

is the newspaper with the most negative tone towards China, while the The Independent appears to be the least negative one; The VERBS indicator indicates The Daily Telegraph as the most negative and The Times as the least negative; The NOUNS indicator reveals that The Independent is the most negative and that The Times is the least negative; Finally, the ALL indicator presents The Guardian as the most negative and The Times as the least negative.

<i>Newspapers</i>	<i>Adjectives</i> <i>N/P*100</i>	<i>Verbs</i> <i>N/P*100</i>	<i>Nouns</i> <i>N/P*100</i>	<i>Others</i> <i>N/P*100</i>	<i>All cat.</i> <i>N/P*100</i>
The Guardian	62.36	117.18	96.62	62.40	89.79
The Independent	55.20	112.82	96.86	55.30	85.79
The Daily Telegraph	59.17	118.32	96.25	61.50	88.71
The Times	60.47	106.42	90.40	62.36	84.43

Table 5: The negativity index of the corpus China UK by considering the newspapers

The evaluation based on the years is stabler. In all the three main marker categories the year with the highest negativity index is 2003. A greater variability can be found instead in the lowest index: 2006 for the ADJ indicator; 2004 for the VERBS and NOUNS indicators; 2008 for the ALL indicator.

<i>Years</i>	<i>Adjectives</i> <i>N/P*100</i>	<i>Verbs</i> <i>N/P*100</i>	<i>Nouns</i> <i>N/P*100</i>	<i>Others</i> <i>N/P*100</i>	<i>All cat.</i> <i>N/P*100</i>
2003	65.63	121.80	100.15	59.12	92.99
2004	58.61	108.00	90.51	57.97	83.63
2005	60.40	109.82	95.44	65.76	87.60
2006	57.25	109.84	91.56	61.56	84.18
2007	58.83	109.35	95.11	60.17	86.35
2008	59.50	121.59	95.64	57.84	89.06

Table 6: Negativity index of the corpus China/UK by considering the years

The first result obtained considering each newspaper and the years encourages us to ask some questions.

1. How valid is the dictionary taken from the *General Inquirer system* here used?
2. How reliable is the negativity indicator here employed?
3. Moving from this first result is it possible to identify the articles with a more negative tone among those inserted in each newspaper?

For what concerns the first two questions we must take into account that the list of the positive-negative words of the GI dates back to the studies on semantic differential, in particular to the three dimensions identified by Osgood et al. and then used also by Holsti (1969). In the original classification given by the author, we have: favorable, neutral, unfavorable (ivi: 107). This is a semantic scale. We must remember to this regard that “the semantic differential is proposed as an instrument for measuring meaning; ideally, therefore, we should correlate semantic differential scores with some independent criterion of meaning – but there is no commonly accepted quantitative criterion of meaning” (Osgood et al., 1957: 140). This affirmation opens the section on *The measurement of meaning* dedicated to the validity of the instrument. We must remember that the validity of a measuring instrument refers to the conformity of the instrument to the properties of the object it intends to measure. In this case we are talking

about a problem of “content validity”: Does what is observed/classified/measured really correspond to the positive and negative tone of the articles which make up the corpus? The content validity is mainly theoretical. The list of words coming from the GI is the result of a selection carried out beforehand on the basis of semantic criteria, which, in the operative phase of the classification, loses any sensitivity to the context. The adjective “cold” (inserted in the list of the negative words) has thirteen meanings referring mostly to temperature and color and these do not necessarily have negative connotations; Only the context makes it possible for psychological and emotional connotations to emerge. In an automatic analysis the context can derive only from a complex of classifications, from a “semantic field” which is stabler when more extended are its characteristics. The validity of the evaluative classification, in the list of the GI is however sufficiently guaranteed by limited presence of polysemies (Tab. 1; 12.5% in the list of the negative words and 14.5% in the list of the positive ones).

Different is the case of the reliability which concerns the precision, the stability and the reproducibility of the measurement. Let’s put aside the precision (which is always approximative in a complex ambit as is that of the meaning of words) and the reproducibility (not applicable in an instrument of automatic classification in which the measurement is exclusively procedural). Our attention focuses on the stability or reliability of the data collector, the one which can be attributed to the constitutive modalities of the measuring instrument. From this point of view the elements which make up the indicator (the list of words) present some problems with reference to the marker categories and to the levels of frequency.

All the measurements indicated in the tables 5 and 6 converge in assigning to the sections of the corpus a negative tone (a negativity index over 40-50%). However, the negativity index attributed taking into consideration only the verbs is considerably higher while the index based on the adjectives is lower.

The negativity index behaves differently also if we consider the levels of frequency (Tab. 7) and in particular the levels of frequency based on the marker categories. In the medium-high frequency level the negativity index of the adjectives reaches its lowest value, while the highest value is reached among the verbs belonging to the low frequency level. If, when calculating the negativity index, we consider only the adjectives, the importance of those with a positive valence in the medium-high frequency becomes decisive for the result.

<i>Grammatical Category</i>	<i>Alta Media f. N/P*100</i>	<i>Bassa f. N/P*100</i>
Adjectives	48.19	78.00
Verbs	96.83	158.11
Nouns	82.12	128.74
Others	64.21	67.04
All categories	76.48	114.59

Table 7: Negativity index of the corpus China UK based on levels of frequency

4. Identification and application of the evaluation techniques

The semantic tagging carried out on the corpus China-UK was necessary to calculate a negativity index per each printed article and to order them from the most to the least negative on the basis of their general tone. In this classification list we calculated the distribution in quartiles and we

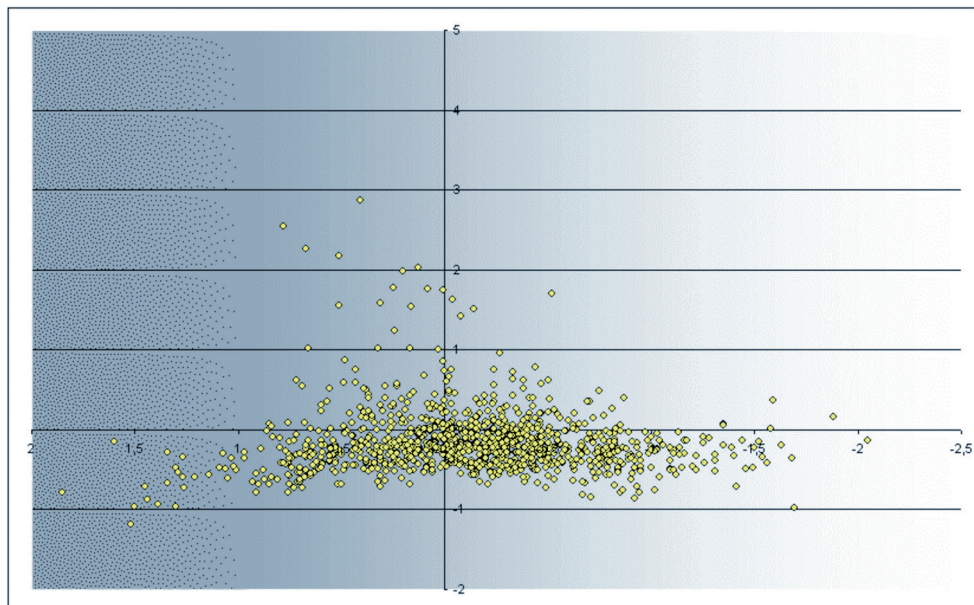
built a new evaluation variable with four modalities which were added to the previous variables of the corpus (Newspaper and Year):

- POS1= positive (first quartile of the articles with a negativity index from 0 to 40%)
- NEG2 = quite negative (second quartile with an index between 41% and 65%)
- NEG3 = negative (third quartile with an index between 66% and 100%)
- NEG4 = very negative (fourth quartile with an index over 100%).

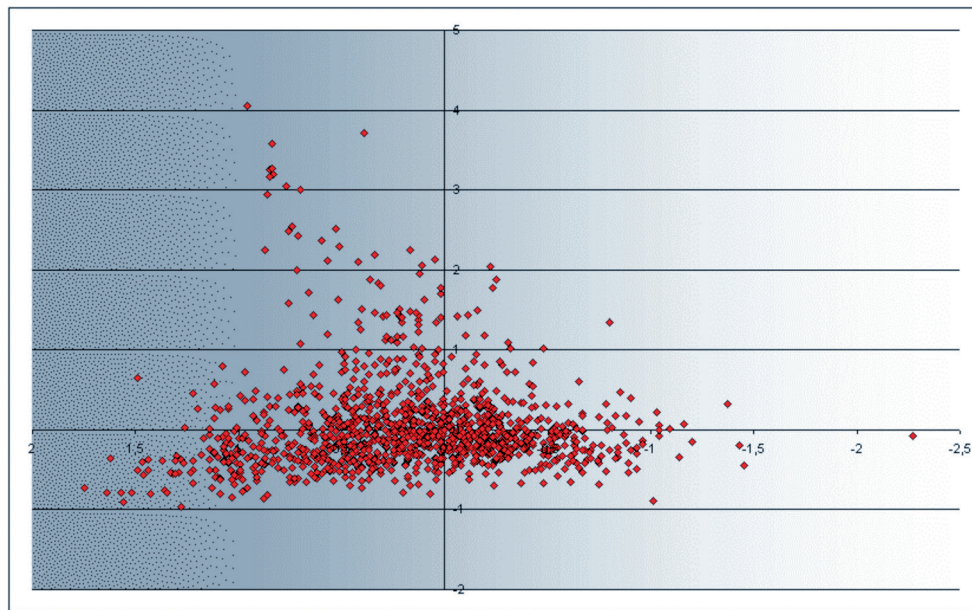
To the original corpus China-UK (the entering of words as headwords has not yet taken place) we applied the ASPAR (Correspondence analysis of the table Word x Responses) procedure where the responses represented by the newspaper articles and the variables of the categories (Year, Newspaper and Evaluation) were used by explanatory variables. The hypothesis at the basis of this analysis was that among the first factors extracted it was necessary to identify a dimension of negativity which made it possible to position in a linear way the negative and the positive words of the GI, around this were gathered other words not included in the list and which were positioned along the same dimension.

The results of the ASPAR procedure, carried out with the software DTM of L. Lebart, can be seen in the graph 1 (positive words) and 2 (negative words) which represents the factorial plane made up of the axes 3 and 4. Along the third axis are positioned most of the words of the GI list. Its interpretation in terms of negativity-positivity is confirmed by the projection of the modalities of the variable Evaluation (Graph. 3). The fourth axis represents the temporal dimension.

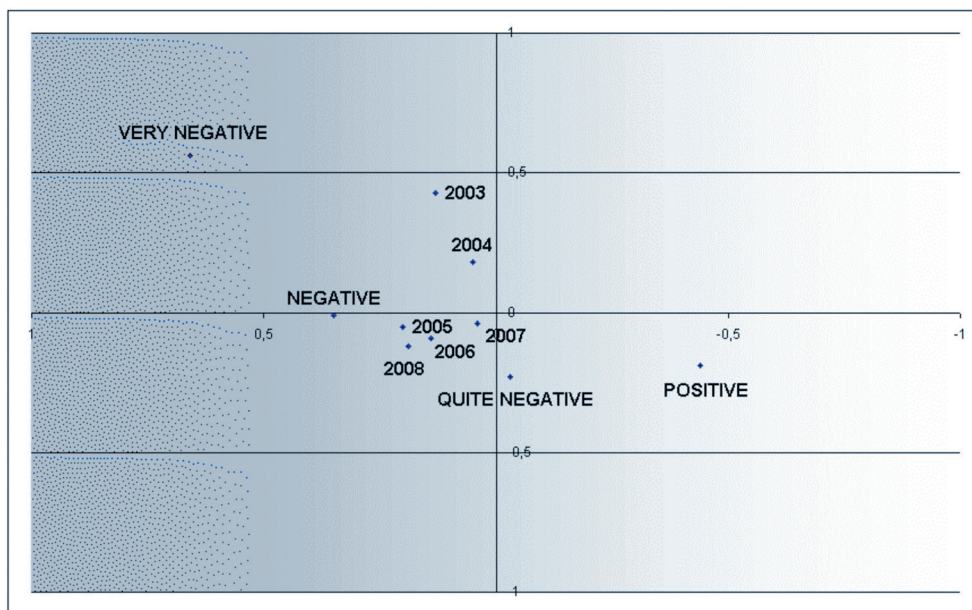
In the semantic field strongly characterized by a negative tone, as results from the quadrants III and IV, we find words such as: accused, aggression, arrested, atrocities, boycotted, crimes. In the quadrants I and II are positioned words less characterized by negativity or strongly positive such as: accommodation, admired, aesthetic, art, beautiful.



Graph 1: Correspondence Analysis of Corpus UK: factorial plane of positive words (Axis 3) and tempore dimension (Axis 4)



Graph 2: Correspondence Analysis of Corpus UK: factorial plane of negative words (Axis 3) and temporal dimension (Axis 4)



Graph 3: Correspondence Analysis of Corpus UK: factorial plane of negative-positive (Axis 3) and temporal dimension (Axis 4)

China and Beijing are positioned in the negative semantic field, confirming the validity of the negativity indicator, especially in its overall composition which considers terms selected in all the marker categories. In the negative semantic field we also find words such as bird flu, communist party, death penalty, Sars and virus which were not inserted in the list of the GI. The same can be said for Confucianism, evolution, glamor present in the positive semantic field. We must stress the fact that some words which in other contexts would be considered positive, in this corpus are instead considered negative, such as Dalai Lama, democratic and human rights remembered and mentioned to underline difficulties, conflicts and revendications.

5. Conclusion

«A content analysis dictionary is a collection of content analysis categories. It is not, however, simply a collection of any content analysis categories» (Stone et al., 1966: 139). It is obvious that on the possibility of a general application of the *Inquirer dictionary* already the first scholars who studied this system had expressed their doubts, due to the awareness that both dictionaries were born with specific purposes which brought them to include some categories and to exclude others. At the same time within the same category and therefore within the words selected and evaluated, there can be some ambiguous cases in which a word can be classified in more than one category. The analysis of the negativity index carried out in the previous paragraph shows the importance of the inclusion or exclusion of some words according to the grammatical categories differently represented among the levels of frequency.

The indication deriving from these considerations is to develop a dictionary for every context of application, therefore, for every specific research project. Considering these premises, but having as objective the evaluation of the reliability and validity of the positive-negative criterion in the automatic analysis of the texts, we proceeded considering our context, that is the articles of the four newspapers which talked about China from 2003 to 2008. From the point of view of our research, with the creation of a specific dictionary and the evaluation of a context equally specific, it has become necessary to identify a technique capable of singling out the words with a positive or a negative valence within each article.

Wanting to summarize the phases which brought us from the classification of some texts using the positive-negative criterion to the evaluation of the reliability of the technique chosen, we can describe this process identifying three main working stages: *ex ante*, *in itinere* ed *ex post*.

- **Ex ante:** we analyzed the *Inquirer dictionary* distinguishing and organizing the various categories it contains.
- **In itinere:** we carried out an experimental control on the validity and the reliability of the list of adjectives taken from the *Id* for the creation of a negativity index; this control was performed on the corpus of articles of the four newspapers that talked about China from 2003 to 2008. It is in this delicate moment that the negativity indicator is applied to our corpus and it is possible to see its strengths and weaknesses (See. §2).
- **Ex post:** Concluded the phase of construction and application of the negativity index, a further reflection became necessary for what concerns its reliability. This was tested by using the ASPAR procedure on the articles of the newspapers and on the variables of the categories. This application allowed us to identify, among the first factors extracted, a negativity dimension which made it possible to position the negative and the positive words of the *Id* in a linear way. Along this dimension we then noticed that other words not in the list, but belonging to the same dimension, were positioned (See. §3).

Following the path described above we were able – moving from a general list with its own semantic categories – to classify in an automatic way the corpus according to the criterion positive-negative and we created a specific sub-dictionary of the research subject.

References

- Bolasco S. (2000). TALTAC: un environnement pour l'exploitation de ressources statistiques et linguistiques dans l'analyse textuelle. Un exemple d'application au discours politique. *JADT 2000*, EPFL, Lausanne 9-11 marzo, tome 2, pp. 342-353.
- Bolasco S., Bisceglia B. and Baiocchi F. (2004). Estrazione di informazione dai testi. In *Mondo Digitale*, III, 1, 2004, pp. 27-43.
- Bolasco S. and della Ratta-Rinaldi F. (2004). Experiments on semantic categorisation of texts: analysis of positive and negative dimension. In *JADT 2004*, pp. 202-210.

- Boucher J. and Osgood C.E. (1969). The Pollyanna hypothesis. *Journal of Verbal Learning & Verbal Behavior*, vol. 8 (1): 1-8.
- Esuli A., Sebastiani F. (2005). Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. Bremen, Germany.
- Giordano R. and Voghera M. (2002). Verb system and verb usage in spoken and written Italian. In *Jadt 2002*, pp. 289-299.
- Holsti O.R. (1969). *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison-Wesley.
- Kelley H.M (2009). Naming on the bright side of life. University of Pennsylvania, <http://www.sas.upenn.edu/> (download 1/09/2009).
- La Rocca G. (2007). *Tecniche di analisi qualitativa online. Software, media, testi*. Roma: Edizioni Kappa.
- Marchand P. (1998). *L'Analyse du Discours Assistée par Ordinateur. Concepts, Méthods, Outils*. Paris: Colin.
- Martin J.R. and White P. (2005). *The Language of Evaluation*. New York: Palgrave.
- Osgood C.E., Suci G.J. and Tannenbaum H.P. (1957). *The Measurement of Meaning*. Urbana: University of Illinois Press.
- Pang B. and Lee L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2: 1-135.
- Pang B., Lee L. and Vaithyanathan S. (2002). Thumbs up? Sentiment classification using Machine Learning techniques. In *Proceedings of Conference on Empirical Methods in NLP*, pp. 79-86.
- Stone P.J., Dunphy D.C., Smith M.S. and Ogilvie D.M. (1966). *The General Inquirer. A Computer Approach to Content Analysis*. Cambridge (MA): The MIT Press.

