ISSN 2499-4553

# IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 1, Number 1
december 2015

## Emerging Topics at the First Italian Conference on Computational Linguistics

aA accademia university press

registrazione in corso presso il Tribunale di Trento

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)


direttore responsabile
Michele Arnese




ISSN 2499-4553
ISBN 978-88-99200-63-3

www.aAccademia.it/IJCoL_01

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it

# IJCoL

**Volume 1, Number 1**
**december 2015**

## Emerging Topics at the First Italian Conference on Computational Linguistics

*a cura di*
Roberto Basili, Alessandro Lenci,
Bernardo Magnini, Simonetta Montemagni

CONTENTS

# Geometric and Statistical Analysis of Emotions and Topics in Corpora

Francesco Tarasconi *
CELI S.R.L.

Vittorio Di Tomaso **
CELI S.R.L.

*NLP techniques can enrich unstructured textual data, detecting topics of interest and emotions. The task of understanding emotional similarities between different topics is crucial, for example, in analyzing the Social TV landscape. A measure of how much two audiences share the same feelings is required, but also a sound and compact representation of these similarities. After evaluating different multivariate approaches, we achieved these goals by applying Multiple Correspondence Analysis (MCA) techniques to our data. In this paper we provide background information and methodological reasons to our choice. MCA is especially suitable to analyze categorical data and detect the main contrasts among them: NLP-annotated data can be transformed and adapted to this framework. We briefly introduce the semantic annotation pipeline used in our study and provide examples of Social TV analysis, performed on Twitter data collected between October 2013 and February 2014. The benefits of examining emotions shared in social media using multivariate statistical techniques are highlighted: using additional dimensions, instead of "simple" polarity of documents, allows to detect more subtle differences in the reactions to certain shows.*

## 1. Introduction

Classification of documents based on *topics* of interest is a popular NLP research area (see, for example, Hamamoto et al. (2005)). Another important subject, especially in the context of Web 2.0 and social media, is the sentiment analysis, mainly meant to detect polarities of expressions and opinions (Liu 2012). Sentiment Analysis (SA) is both a topic in natural language processing which has been investigated for several years and a tool for social media monitoring which is used in business services. A recent survey that explores the latest trends is Cambria (2013). While the first attempts on English texts date back to the late 90s, SA on Italian texts is a more recent area of research (probably the first scientific publication is Dini and Mazzini (2012)). A sentiment analysis task which has seen less contributions, but of growing popularity, is the study of *emotions* (Wiebe et al. 2005), which requires introducing and analyzing multiple variables (appropriate "emotional dimensions") potentially correlated. This is especially important in the study of the so-called Social TV (Cosenza 2012): people can share their TV experience with other viewers on social media using smartphones and tablets. We define the empirical distribution of different emotions among viewers of a specific TV show as its *emotional profile*. Comparing at the same time the emotional profiles of several formats requires appropriate descriptive statistical techniques. During the research we conducted, we evaluated and selected geometrical methods that satisfy these requirements and provide an easy to understand and coherent representation of the results. The methods we used can be applied to any dataset of documents classified based on

* Via San Quintino 31 - 10121 Torino, Italia. E-mail: tarasconi@celi.it.
** Via San Quintino 31 - 10121 Torino, Italia. E-mail: ditomaso@celi.it.

topics and emotions; they also represent a potential tool for the quantitative analysis of any NLP annotated data.

We used the BlogMeter platform[1] to download and process textual contents from social networks (Bolioli et al. 2013). Topics correspond to TV programs discussed on Twitter. Nine emotions are detected: the basic six according to Ekman (1972) (*anger, disgust, fear, joy, sadness, surprise*), *love* (a primary one in Parrot's classification) and *like/dislike* expressions, quite common on Twitter.

Topics and emotions are detected using a rule-based system. In the case of TV episodes, the mention of a show or its characters in the context of a tweet is the most important factor in assigning it to a specific topic. To improve precision in identifying posts connected to the Social TV, the temporal range of analysis can be reduced to a set of windows centered around relevant episodes.

We examined the emotional landscape of the Italian Social TV during December 2013, treating each show as a different topic. The analysis evidenced a strong negative mood associated with politics and the programs that tackled this subject. We then focused on two popular formats: the music talent show X Factor and the competitive cooking show MasterChef. Each episode was considered as a different topic. Whereas the progression of the season through emotional phases (from selections to finals) was clearly visible in the case of X Factor, MasterChef was much more erratic and strongly influenced by scripted events taking place in each episode. By comparing directly X Factor and MasterChef in the same analysis, we concluded that the subject of the show strongly influences the reactions of its viewers, in a way that goes beyond the simple expression of positive/negative judgements. This supports the claim that the analysis of emotions can provide additional information and detect deeper differences than polarity in the study of social media.

The paper is organized as follows: section 2 describes the tools used for topic and emotion detection, section 3 introduces the mathematical model used to analyze NLP-annotated data, section 4 focuses on the choice of statistical methods adopted to represent and extract the most relevant structures in our datasets and section 5 presents the case studies.

This research was originally presented in reduced form at CLiC 2014, the First Italian Conference on Computational Linguistics.

## 2. A social media monitoring platform

The processing tools which we will describe are implemented in a social media monitoring service called BlogMeter, operating since 2009. The monitoring process includes three main phases:

- Listening: thanks to purpose-developed data acquisition systems, the platform detects and collects from the web potentially interesting data;
- Understanding: a semantic engine is used to structure and classify the conversations in accordance to the defined drivers (topics and entities mentioned in the texts, but also emotions of interest);
- Analysis: through the analysis platform the user can navigate the conversations in a structured way, aggregate the drivers in one or more dashboards, discover unforeseen trends in the concept clouds and drill down the data to read the messages inside their original context.

1 www.blogmeter.it

It is of particular interest for our research the understanding phase, which includes automatic classification and sentiment analysis. It can be further divided into:

- creation of a domain-based taxonomy (i.e. an ontology of topics such as brands, products or people);
- identification and automatic classification of relevant documents (according to the taxonomy);
- polarity and emotion detection.

The monitored sources are typically user-generated media, such as blogs, forums, social networks, news groups, content sharing sites, sites of questions and answers (Q&A), reviews of products / services, which are active in many countries and in different languages. The overall number of sources is more than 500,000 blogs (of which approximately 70,000 active, with a post in the last three months) and 700 gathering places (forums, newsgroups, Q&A sites, content sharing platforms, social networks). This computation considers Facebook and Twitter as single sources, but in fact, they are the largest collectors of conversations.

### 2.1 Semantic annotation pipeline

Documents extracted from the web in the form of unstructured information are made available to the semantic annotation pipeline which analyzes and classifies them according to the domain-based taxonomies defined for the client. The annotation pipeline uses the UIMA framework (the Unstructured Information Management Architecture originally developed by IBM and now by the Apache Software Foundation [2]).
UIMA annotators enrich the documents in terms of linguistic information, recognition of entities and concepts, identification of relations between concepts, entities and attitudes expressed in the text (opinions, mood states and emotions). Some linguistic resources and annotators are common to different application domains, while others are domain dependent. We will not describe here the pipeline modules in details, and we will focus on the main linguistic resource used in the sentiment analysis module, i.e. a concept-level sentiment lexicon for Italian.
The sentiment lexicon is used by the semantic annotator, which recognizes opinions and expressions of mood and emotions and associates them with the opinion targets. This component operates both on the sentence level (in order to treat linguistic phenomena such as negation and quantification) and on the document leve (in order to identify relations between elements that are in different sentences).

### 2.2 A concept-level sentiment lexicon for Italian

In this section we describe the *sentiment lexicon* used by the semantic annotator, i.e. the repository containing terms, concepts and patterns used in the sentiment annotation. Researchers have been building sentiment lexica for many years, in particular for the English language, and a review on recent results can be found for example in Cambria et al. (2013). The sentiment lexicon for Italian contains about 10.000 entries (6.200 single words and 3.400 multi-word expressions). Each entry has information about sentiment, i.e. polarity, emotions, and domain application (therefore it is a *contextualized sentiment lexicon*). It has been created and updated during the past three years, performing social media monitoring and SA in different application domains.
An important resource used in the creation of the lexicon is the WordNet-Affect project (Strapparava and Valitutti 2004).

2 UIMA Specifications: http://uima.apache.org/uima-specification.html

One aspect worth mentioning is that the valence of many words can change in different contexts and domains. The word "accuratezza" ("accuracy"), for example, has a default positive valence, just as it is for "affare d'oro" ("to do a roaring trade"). On the contrary, "andare a casa" ("going home") has no polarity in a neutral context, as long as it is not used in an area such as sentiment on Sanremo Festival, where it means instead being eliminated from the singing competition. Similarly, "truccato" ("to have make up on" or "to be rigged"), would not have negative polarity if the domain was a fashion show. Instead, in the field of online games or betting, the perspective changes.

### 2.3 Emotions

The interest for emotion detection in social media monitoring grew in 2011 after the publication of a paper by Bollen et al. (2011), where the authors argued that the analysis of mood in Twitter posts could be used to predict stock market movements up to 6 days in advance. In particular, they identified "calmness" as the predictive mood dimension, within a set of 6 different mood dimensions (happiness, kindness, alertness, sureness, vitality and calmness). The definition of a set of basic (or primary) emotions is a debated topic, and the study and analysis of emotions and their expression in texts obviously has a long tradition in philosophy and psychology (see for example Galati (2002)). In NLP tasks, Ekman's six basic emotions (anger, disgust, fear, joy, sadness, surprise) have often been used (e.g. in Strapparava and Valitutti (2004)). The platform we employed in our research adopts Ekman's list of emotions and "love", which is a primary emotion in Parrot's classification. Considering expressions of "like" and "dislike" as "emotional" was necessary to cover a large amount of social media documents, which clearly express a feeling towards a subject being discussed, but not an emotion in the common sense.
A similar approach is described in Roberts et al. (2012).
An argument could be made against adding arbitrary variables to a pre-existing model of basic emotions. However, from the perspective of an exploratory analysis of an unknown dataset, these variables can better capture specific features in social network communication. The issue of adding potentially correlated or even redundant variables is tackled in the dimension reduction framework we will define and employ in the following sections.
The manual annotation of emotions in a reference Italian corpus would be a useful advance for testing the accuracy of the automatic system.

### 2.4 Evaluation

The sentiment semantic annotator was partially evaluated on polarity classification of Twitter messages (with a focus on politics), which was conducted using the Evalita 2014 SENTIPOLC test set. As reported in Basile et al. (2014) it's a collection of 1,935 tweets derived from existing corpora: SENTI-TUT (Bosco et al. 2013) and TWITA (Basile and Nissim 2013).
We performed two runs of the analysis procedure: the first using only a generic lexicon, the second using a lexicon enriched specifically for the political domain. Both are pre-existing resources compared to the train and test set used for the SENTIPOLC task, which were not included in the creation of the lexicons.
Precision P, recall R and F-score were computed for the positive and negative predicted fields, separately for the different values that the field can assume (0 and 1). An average F-score for positive and negative polarities was then computed to calculate the final F-score F for the SENTIPOLC task. These metrics can be compared to the results achieved by the Evalita 2014 participants. Results for the CELI pipeline are given in Table 1. Our results are given for different lexicons used (generic/political).

**Table 1**
Precision, recall and F-score on the full test set, per class and combined

| | $\mathbf{CELI_{gen}}$ | $\mathbf{CELI_{pol}}$ |
|---|---|---|
| $\mathbf{prec_0^{pos}}$ | 0.7904 | 0.7944 |
| $\mathbf{rec_0^{pos}}$ | 0.8357 | 0.8533 |
| $\mathbf{F_0^{pos}}$ | 0.8124 | 0.8228 |
| $\mathbf{prec_1^{pos}}$ | 0.5419 | 0.5708 |
| $\mathbf{rec_1^{pos}}$ | 0.4674 | 0.4691 |
| $\mathbf{F_1^{pos}}$ | 0.5019 | 0.5150 |
| $\mathbf{F^{pos}}$ | 0.6572 | 0.6689 |
| $\mathbf{prec_0^{neg}}$ | 0.6664 | 0.6920 |
| $\mathbf{rec_0^{neg}}$ | 0.8643 | 0.8596 |
| $\mathbf{F_0^{neg}}$ | 0.7526 | 0.7667 |
| $\mathbf{prec_1^{neg}}$ | 0.7401 | 0.7565 |
| $\mathbf{rec_1^{neg}}$ | 0.4718 | 0.5328 |
| $\mathbf{F_1^{neg}}$ | 0.5762 | 0.6253 |
| $\mathbf{F^{neg}}$ | 0.6644 | 0.6960 |
| combined $\mathbf{F}$ | 0.6608 | 0.6824 |

## 3. Vector space model and dimension reduction

Let $\mathcal{D}$ be the initial data, a collection of $m_D$ documents. Let $\mathcal{T}$ be the set of $n_T$ distinct topics and $\mathcal{E}$ the set of $n_E$ distinct emotions that the documents have been annotated with. Let $n = n_T + n_E$. A document $d_i \in \mathcal{D}$ can be represented as a vector of 1s and 0s of length $n$, where entry $j$ indicates whether annotation $j$ is assigned to the document or not. The *document-annotation* matrix $\mathbf{D}$ is defined as the $m_D \times n$ matrix of 1s and 0s, where row $i$ corresponds to document vector $d_i$, $i = 1, \ldots, m_D$. For the rest of our analysis, we suppose all documents to be annotated with at least one topic and one emotion. $\mathbf{D}$ can be seen as a block matrix:

$$\mathbf{D}_{m_D \times n} = \left(\mathbf{T}_{m_D \times n_T}\ \mathbf{E}_{m_D \times n_E}\right),$$

where blocks $\mathbf{T}$ and $\mathbf{E}$ correspond to topic and emotion annotations.
The *topic-emotion* frequency matrix $\mathbf{T}_E$ is obtained by multiplication of $\mathbf{T}$ with $\mathbf{E}$:

$$\mathbf{T}_E = \mathbf{T}^T\mathbf{E},$$

thus $(\mathbf{T}_E)_{ij}$ is the number of co-occurrences of topic $i$ and emotion $j$ in the same document. In the Social TV context, rows of $\mathbf{T}_E$ represent emotional profiles of TV programs on Twitter. From documents we can obtain *emotional impressions* which are (*topic, emotion*) pairs.
Let us consider, for example, the following document (tweet):

> "*@michele_bravi sono star felice che tu abbia vinto xfactor :), cavolo telo meriti anche io ci vorrei andare ma ho paura :(* ",

which can be loosely translated as

> "*@michele_bravi I'm very happy that you won xfactor :), you really deserve it and I would like to participate too but I'm scared :(* ".

This document can be annotated with {*topic = X Factor, emotion = fear, emotion = love*}. When represented as a vector, its non-zero entries correspond to *X Factor, fear, love* indices. It generates distinct emotional impressions (*X Factor, fear*) and (*X Factor, love*).
Let $\mathcal{J}$ be the set of all $m_J$ emotional impressions obtained from $\mathcal{D}$. Then we can define, in a manner similar to $\mathbf{D}$, the corresponding *impression-annotation* matrix $\mathbf{J}$, a $m_J \times n$ matrix of 0s and 1s. $\mathbf{J}$ can be seen as a block matrix as well:

$$\mathbf{J} = \left(\mathbf{T}_J \; \mathbf{E}_J\right),$$

where blocks $\mathbf{T}_J$ and $\mathbf{E}_J$ correspond to topics and emotions of the impressions.
In our previous example, the emotional impression (*X Factor, fear*) can be represented as a vector with only two non-zero entries: one corresponding to column *X Factor* in $\mathbf{T}_J$ and one to column *fear* in $\mathbf{E}_J$.
We can therefore represent documents or emotional impressions in a vector space of dimension $n$ and represent topics in a vector space of dimension $n_E$. Our first idea was to study topics in the space determined by emotional dimensions, thus obtaining emotional similarities from matrix representation $\mathbf{T}_E$. These similarities can be defined using a distance between topic vectors or, in a manner similar to information retrieval and Latent Semantic Indexing (LSI) (Manning et al. 2008), the corresponding cosine. Our first experiments highlighted the following requirements:

1. to reduce the importance of (potentially very different) topic absolute frequencies (e.g. using cosine between topic vectors);
2. to reduce the importance of emotion absolute frequencies, giving each variable the same weight;
3. to graphically represent, together with computing, emotional similarities, as already mentioned;
4. to highlight why two topics are similar, in other words which emotions are shared.

In multivariate statistics, the problem of graphically representing an *observation-variable* matrix can be solved through dimension reduction techniques, which identify convenient projections (2-3 dimensions) of the observations. Principal Component Analysis (PCA) is probably the most popular of these techniques. See Abdi and Williams (2010) for an introduction. It is possible to obtain from $\mathbf{T}_E$ a reduced representation of topics where the new dimensions better explain the original variance. PCA and its variants can thus define and visualize reasonable emotional distances between topics. After several experiments, we selected Multiple Correspondence Analysis (MCA) as our tool, a technique aimed at analyzing categorical and discrete data. It provides a framework where requirements 1-4 are fully met, as we will show in section 4. An explanation of the relation between MCA and PCA can be found, for example, in Gower (2006).

## 4. Multiple Correspondence Analysis

(Simple) Correspondence Analysis (CA) is a technique that can be used to analyze two categorical variables, usually described through their *contingency table* $\mathbf{C}$ (Greenacre 1983), a matrix that displays the frequency distribution of the variables.
CA is performed through a Singular Value Decomposition (SVD) (Meyer 2000) of the matrix of *standardized residuals* obtained from $\mathbf{C}$. Residuals represent the deviation from the expected distribution of the table in the case of independence between the two variables. SVD of a matrix finds its best low-dimensional approximation in quadratic distance. CA procedure yields new

axes for rows and columns of $\mathbf{C}$ (variable categories), and new coordinates, called *principal coordinates*. Categories can be represented in the same space in principal coordinates (symmetric map). The reduced representation (the one that considers the first $k$ principal coordinates) is the best $k$-dimensional approximation of row and column vectors in *chi-square* distance (Blasius and Greenacre 2006). Chi-square distance between column (or row) vectors is a Euclidean-type distance where each squared distance is divided by the corresponding row (or column) average value. Chi-square distance can be read as Euclidean distance in the symmetric map and allow us to account for different volumes (frequencies) of categories. It is therefore desirable in the current application, but it is defined only between row vectors and between column vectors.
CA measures the information contained in $\mathbf{C}$ through the *inertia* $I$, which corresponds to variance in the space defined by the chi-square distance, and aims to explain the largest part of $I$ using the first few new axes. Matrix $\mathbf{T}_E$ can be seen as a contingency table for emotional impressions, and a representation of topics and emotions in the same plane can be obtained by performing CA. Superimposing topics and emotions in the symmetric map apparently helps in its interpretation, but the topic-emotion distance doesn't have a meaning in the CA framework. We have therefore searched for a representation where analysis of topic-emotion distances was fully justified.
MCA extends CA to more than two categorical variables and it is originally meant to treat problems such as the analysis of surveys with an arbitrary number of closed questions (Blasius and Greenacre 2006). But MCA has also been applied with success to positive matrices (each entry greater or equal to zero) of different nature and has been recast (rigorously) as a geometric method (Le Roux and Rouanet 2004). MCA is performed as the CA of the *indicator matrix* of a group of respondents to a set of questions or as the CA of the corresponding *Burt matrix* (Greenacre 2006). The Burt matrix is the symmetric matrix of all two-way crosstabulations between the categorical variables. Matrix $\mathbf{J}$ can be seen as the indicator matrix for emotional impressions, where the questions are which topic and which emotion are contained in each impression. The corresponding Burt matrix $\mathbf{J}_B$ can be obtained by multiplication of $\mathbf{J}$ with itself:

$$\mathbf{J}_B = \mathbf{J}^T\mathbf{J} = \begin{pmatrix} \mathbf{T}_J^T\mathbf{T}_J & \mathbf{T}_J^T\mathbf{E}_J \\ \mathbf{E}_J^T\mathbf{T}_J & \mathbf{E}_J^T\mathbf{E}_J \end{pmatrix}.$$

Diagonal blocks $\mathbf{T}_J^T\mathbf{T}_J$ e $\mathbf{E}_J^T\mathbf{E}_J$ are diagonal matrices and all the information about correspondences between variables is contained in the off-diagonal blocks. From the CA of the indicator matrix we can obtain new coordinates in the same space both for respondents (impressions) and for variables (topics, emotions). From the CA of the Burt matrix it is only possible to obtain principal coordinates for the variables. MCAs performed on $\mathbf{J}$ and $\mathbf{J}_B$ yield similar principal coordinates, but with different scales (different singular values). Furthermore, chi-square distances between the columns/rows of matrix $\mathbf{J}_B$ include the contributions of diagonal blocks. For the same reason, the inertia of $\mathbf{J}_B$ can be extremely inflated.
Greenacre (2006) solves these problems by proposing an adjustment of inertia that accounts for the structure of diagonal blocks. Inertia explained in the first few principal coordinates is thus estimated more reasonably. MCA of the Burt matrix with adjustment of inertia also yields the same principal coordinates as the MCA of the indicator matrix. Finally, in the case of two variables, CA of the contingency table and MCA yield the same results. Thus the three approaches (CA, MCA in its two variants) are unified.
When analyzing *topic* and *emotion* variables in this framework, we are ignoring co-occurrences of multiple topics or multiple emotions in the same documents. Discounting interactions between topics is desiderable, as our aim in this analysis is to focus on emotional similarities between subjects of online conversation. Discounting interactions between emotions can potentially discard useful information, because emotions that often co-occur in the same span of text might

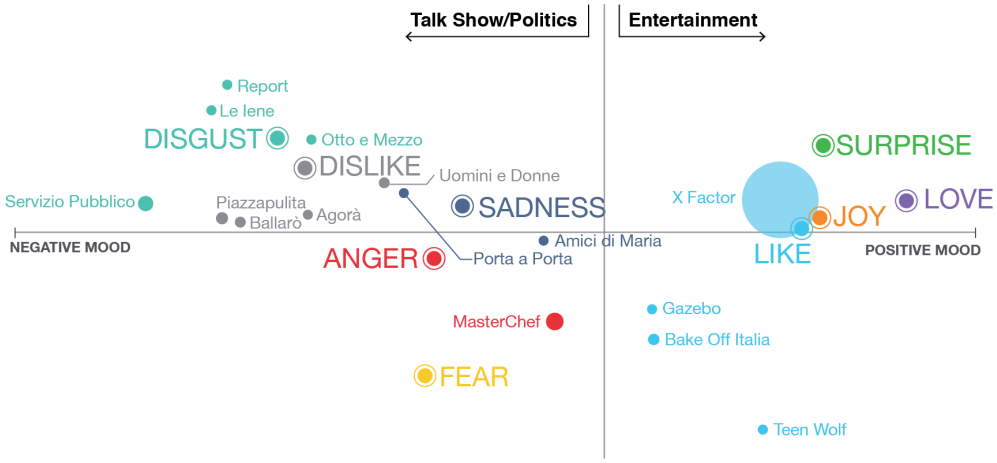


**Figure 1**
MCA of most emotional Italian TV programs discussed on Twitter during December 2013.

be considered closer in an ideal emotional space (for example *love* and *joy*). However, the amount of tweets that contain more than one annotation of type *emotion* is very small (less than 1% in the considered datasets). Moving to the analysis of emotional impressions allows us to adopt the MCA framework and, in particular, to better estimate the explained inertia of our dataset: considering interactions between *emotion* variables would instead change the structure of one diagonal block in the Burt matrix and the adjustment proposed by Greenacre could not be applied. MCA offers possibilities common to other multivariate techniques. In particular, a measure on how well single topics and emotions are represented in the retained axes is provided (*quality* of representation).
Symmetric treatment of topics and emotions facilitates the interpretation of axes. Distances between emotions and topics can now be interpreted and, thanks to them, it is possible to establish why two topics are close in the reduced representation. An additional (and interesting) interpretation of distances between categories in terms of *sub-clouds of individuals* (impressions) is provided by Le Roux and Rouanet (2004).

## 5. Case studies

### 5.1 One month of Twitter TV

Data were collected during December 2013 (1,2 million tweets). Tweets were aggregated to generate monthly TV show profiles. We selected the 15 "most emotional" shows to analyze. MCA was performed using programs and emotions as variables in a vector space model as described in sections 3 and 4. Results are shown in Figure 1. Size of programs' points is proportional to the number of distinct emotional impressions for that category. As explained in section 4, distances between emotions and programs have a mathematical interpretation and can serve as a measure of correlation. Thanks to this fact we were able to perform a straightforward classification of TV shows, based on the closest emotion in the MCA subspace. This classification is represented by programs' colors in Figure 1. We can see, for example, that Italian talk shows about politics (second quadrant) are similar and share the most negative emotions. Instead, entertainment shows are characterized by better mood overall, although they do not share the full emotional spectrum. For example, MasterChef's public is dominated with *anger*. *Fear*, despite not being dominant, is

**Table 2**
X Factor and MasterChef datasets: emotional impressions about the shows found on Twitter.

| X Factor 7 | |
|---|---|
| Date | Emotional impressions |
| 26/09/13 | 23,712 |
| 03/10/13 | 15,364 |
| 10/10/13 | 11,932 |
| 17/10/13 | 24,116 |
| 24/10/13 | 57,413 |
| 31/10/13 | 26,301 |
| 07/11/13 | 37,441 |
| 14/11/13 | 36,363 |
| 21/11/13 | 29,405 |
| 28/11/13 | 34,097 |
| 05/12/13 | 35,438 |
| 12/12/13 | 121,106 |
| TOT. | 452,688 |

| MasterChef Italy | |
|---|---|
| Date | Emotional impressions |
| 19/12/13 | 5,926 |
| 26/12/13 | 4,495 |
| 02/01/14 | 6,796 |
| 09/01/14 | 7,087 |
| 16/01/14 | 9,721 |
| 23/01/14 | 8,227 |
| 30/01/14 | 8,964 |
| 06/02/14 | 9,427 |
| TOT. | 60,643 |

an important component of dark comedy Teen Wolf's emotional profile. As many multivariate techniques, MCA also provides a measure of the quality of our representation (Blasius and Greenacre 2006). In this case 94% of statistical information (or *inertia*) was retained, so this can be considered an excellent approximation of the initial dataset.

### 5.2 Analyzing whole TV seasons

It is of interest not only to analyze the aggregated profile of a TV show, encompassing several weeks or months, but also to compare individual profiles of each episode. For example, the 7$^{th}$ edition of popular Italian music talent show X Factor consists of 12 episodes, including the auditions. We want to represent these 12 episodes and their emotional similarities with the highest precision in two dimensions. Another program we examined in detail is the competitive cooking show MasterChef Italy (3$^{rd}$ edition). See Table 2 for details on our datasets. Data were collected on a weekly basis, between 24 October and 12 December 2013 for X Factor, between 19 December 2013 and 6 February 2014 for MasterChef. X Factor obtained on average 47k emotional impressions for each episode; MasterChef an average of 8k impressions/episode.
Within the MCA framework, each episode can be considered as a separate category for the program variable we introduced in section 4. A representation similar to the one we obtained in section 5.1 can therefore be obtained for each show. See Figure 2 and 3 for results.
Emotional changes in the audience are reflected in the episodes' positions, numbered progressively.
As we briefly mentioned in section 4, MCA does not discount the weight of individual profiles, which in our case is the sheer number of emotional impressions for each episode. The origin of axes in an MCA map is also the weighted mean point of active variables' points (as shown in figure) and the mean point of emotional impressions' points (not represented). The origin (or barycenter) can then be taken as the average profile (an overall "summary") for the TV show in exam: a fact that we chose to highlight in our representation. Episodes are numbered progressively in each plot. As previously seen, the first axis expresses the contrast between

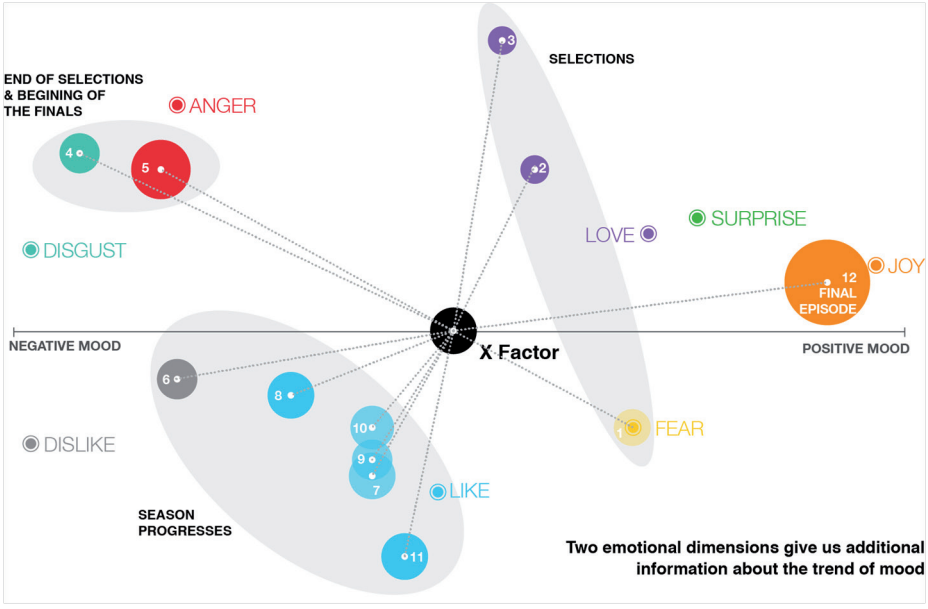


**Figure 2**
MCA of X Factor 7.

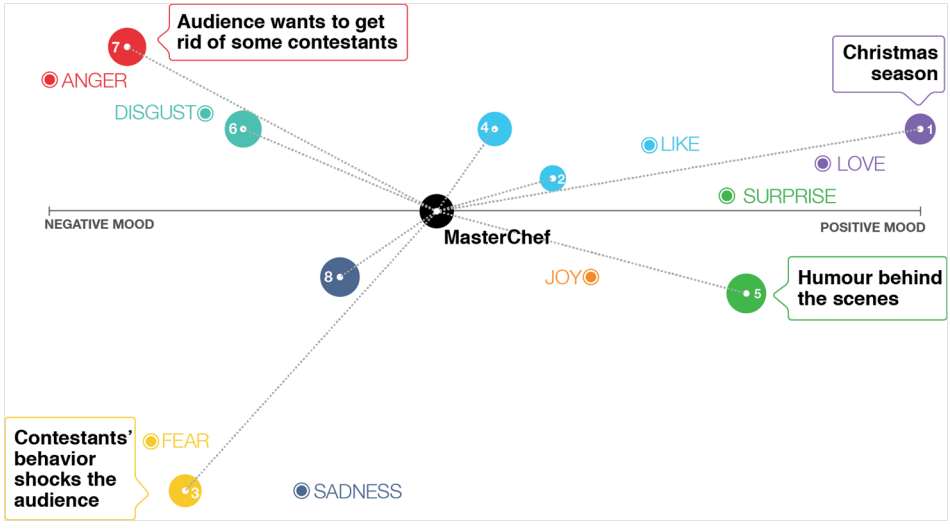


**Figure 3**
MCA of MasterChef Italy, first 8 episodes of 3rd season.

positive and negative mood.
Evolution phases are clearly visible in the X Factor plot (Figure 2). The selection process of the first three episodes is dominated by *love* and *fear* for the contestants. The beginning of the finals is marked by a strong and visceral disagreement about how the selections ended. Judgments dominates most of the season, as the audience is able to directly evaluate the contestants. The final episode is the most positive and emotional of the whole season. 73% of total inertia was

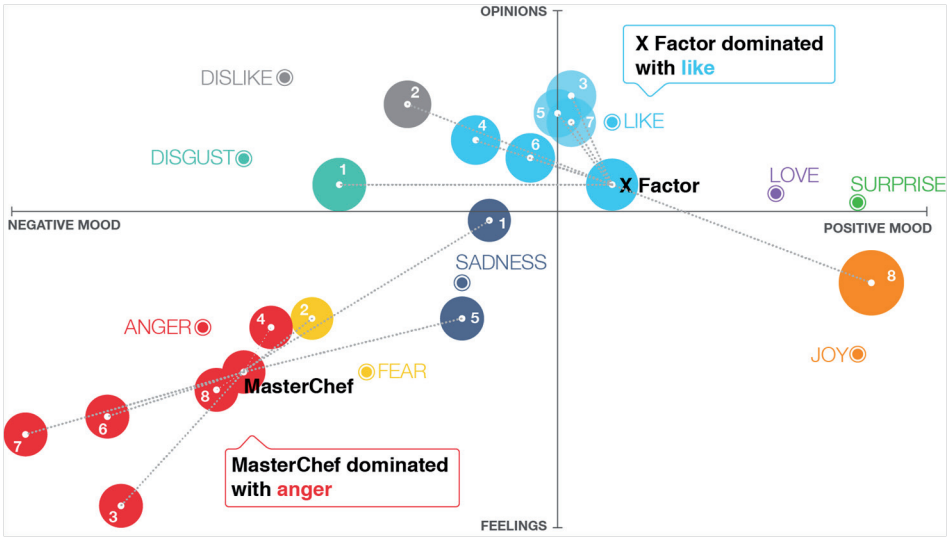


**Figure 4**
Comparison via MCA between X Factor and MasterChef formats, 2013-2014 editions.

retained in this map.
The MCA plot of the 3rd edition of MasterChef Italy (Figure 3) tells a different story (64% retained inertia). No trend emerges so there is a much greater dependence on single episodes, as described in the plot.

### 5.3 Comparison between MasterChef and X Factor

If we represent MasterChef and X Factor in the same space, individual episodes can still be used as categories for emotional impressions (Figure 4). In order to highlight differences between the two formats, we have plotted weighted mean points, obtained separately for each one of them. For example, the X Factor point corresponds to the (scaled) barycenter of the cloud of emotional impressions related to this talent show. Distances from the X Factor and MasterChef points have the same geometric and statistical interpretations as the distances between active variables' points. This type of analysis is strictly related to *structured data analysis*, where the dataset comes with a natural partition or structuring factor: in our case single episodes (original variables) are naturally grouped into their respective seasons. For more information on structured data analysis, see for example Rouanet (2006). Note that we are comparing X Factor's live show (last 8 episodes) with the first 8 episodes of MasterChef. In fact, at the moment our analysis was performed, MasterChef still had to reach its conclusion.
When MasterChef and X Factor are represented in the same MCA plot, we can clearly see how different these two shows are (82% retained inertia).
By looking at the position of emotions, the first axis can be interpreted as the contrast between *moods* (positive and negative) of the public, and this is therefore highlighted as the most important structure in our dataset. X Factor was generally perceived in a more positive way than MasterChef. The advantage of incorporating emotions in our sentiment analysis is more manifest when we look at the second retained axis. We can say the audience of X Factor lives in a world of opinion dominated by *like/dislike* expressions, while the public of MasterChef is characterized by true and active feelings concerning the show and its protagonists. This is coherent with the

fact that viewers of X Factor could directly evaluate the performances of contestants. This was not possible for the viewers of MasterChef, who focused instead on the most outstanding and emotional moments of the show. Reaching these conclusions would not have been possible by looking at simple polarity of impressions.
This difference in volume between the two shows is reflected in the distances from the origin, which can be considered as the average profile, and therefore closer to X Factor.
Other detailed examples on structuring an MCA analysis can be found in Rouanet (2006).

## 6. Conclusions and further researches

By applying carefully chosen multivariate statistical techniques, we have shown how to represent and highlight important emotional relations between topics. We presented some case studies, describing in detail the analyses of some live TV shows as they were discussed on Twitter. Further results in the MCA field can be experimented on datasets similar to the ones we used. For example, additional information about opinion polarity and document authors (such as Twitter users) could be incorporated in the analysis. The geometric approach to MCA (Le Roux and Rouanet 2004) could be interesting to study in greater detail the *clouds* of impressions and documents ($\mathbf{J}$ and $\mathbf{D}$ matrices); authors could also be considered as mean points of well-defined sub-clouds.

## Ancknowledgements

We would like to thank: V. Cosenza and S. Monotti Graziadei for stimulating these researches; the ISI-CRT foundation and CELI S.R.L. for the support provided through the Lagrange Project; A. Bolioli for the supervision and the essential help in the preparation of this paper. Last but not least, all colleagues for always giving their daily contributions.

## References

Abdi, Hervé and Lynne J. Williams. 2010. *Principal Component Analysis*, Wiley Interdisciplinary Reviews: Computational Statistics, Volume 2, Issue 4, pages 433-459.

Basile, Valerio and Malvina Nissim. 2013 *Sentiment analysis on Italian tweets*, Proceedings of WASSA 2013, pages 100-107.

Basile, Valerio, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014 *Overview of the Evalita 2014 SENTIment POLarity Classification Task*, Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014, pages 50-57.

Blasius, Jörg and Michael Greenacre. 2006. *Correspondence Analysis and Related Methods in Practice*, Multiple Correspondence Analysis and Related Methods, Chapter 1, pages 3-40. CRC Press.

Bolioli, Andrea, Federica Salamino, and Veronica Porzionato. 2013. *Social Media Monitoring in Real Life with Blogmeter Platform*, ESSEM@AI*IA 2013, Volume 1096 of CEUR Workshop Proceedings, pages 156-163. CEUR-WS.org.

Bollen, Johan, Huina Mao, and Xiao-Jun Zeng. 2011. *Twitter mood predicts the stock market*, Journal of Computational Science, 2(1):1-8.

Bosco, Cristina, Viviana Patti, and Andrea Bolioli. 2013. *Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT*, IEEE Intelligent Systems, Special Issue on Knowledge-based Approaches to Content-level Sentiment Analysis, 28(2):55-63.

Cambria, Erik, Björn Schuller, Yunqing Xia, and Catherine Havasi. 2013. *New Avenues in Opinion Mining and Sentiment Analysis*, IEEE Intelligent Systems, 28(2):15-21.

Cambria, Erik, Björn Schuller, Bing Liu, Haixun Wang, and Catherine Havasi. 2013 *Knowledge-Based Approaches to Concept-Level Sentiment Analysis*, IEEE Intelligent Systems, 28(2):12-14.

Cosenza, Vincenzo. 2012. *Social Media ROI*. Apogeo.

Dini, Luca and Mazzini Giampaolo. 2002 *Opinion classification Through information extraction*, Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields, pages 299-310

Ekman, Paul, Wallace V. Friesen, and Phoebe Ellsworth. 1972. *Emotion in the Human Face*. Pergamon Press.
Galati, Dario. 2002. *Prospettive sulle emozioni e teorie del soggetto*. Bollati Boringhieri.
Gower, John C. 2006. *Divided by a Common Language: Analyzing and Visualizing Two-Way Arrays*, Multiple Correspondence Analysis and Related Methods, Chapter 3. pages 77-105. CRC Press.
Greenacre, Michael. 1983. *Theory and Applications of Correspondence Analysis*. Academic Press.
Greenacre, Michael. 2006. *From Simple to Multiple Correspondence Analysis*, Multiple Correspondence Analysis and Related Methods, Chapter 2, pages 41-76. CRC Press.
Hamamoto, Masafumi, Hiroyuki Kitagawa, Jia-Yu Pan, and Christos Faloutsos. 2005. *A Comparative Study of Feature Vector-Based Topic Detection Schemes for Text Streams*, Proceedings of the 2005 International Workshop on Challenges in Web Information Retrieval and Integration, pages 122-127.
Jolliffe, Ian T. 2002. *Principal Component Analysis*. Springer.
Le Roux, Brigitte and Henry Rouanet. 2004. *Geometric Data Analysis: From Correspondence Analysis to Structured Data*. Kluwer.
Liu, Bing. 2012. *Sentiment Analysis e Opinion Mining*. Morgan & Claypool Publishers.
Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
Meyer, Carl D. 2000. *Matrix Analysis and Applied Linear Algebra*. Siam.
Roberts, Kirk, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. *EmpaTweet: Annotating and Detecting Emotions on Twitter*, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 3806-3813. European Language Resources Association (ELRA).
Rouanet, Henry. 2006. *The Geometric Analysis of Structured Individuals x Variables Tables*, Multiple Correspondence Analysis and Related Methods, CRC Press.
Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. *Annotating expressions of opinions and emotions in language*, Language Resources and Evaluation, Volume 39, Issue 2-3, pages 165-210.
Strapparava, Carlo and Valitutti, Alessandro. 2004 *"WordNet-Affect: an Affective Extension of WordNet"*, Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), pages 1083-1086, Lisbon.