10.12871/CLICIT2014170

Geometric and statistical analysis of emotions and topics in corpora

Francesco Tarasconi
CELI S.R.L. / Turin, Italy
tarasconi@celi.it

Vittorio Di Tomaso
CELI S.R.L. / Turin, Italy
ditomaso@celi.it

Abstract

English. 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 adapting Multiple Correspondence Analysis (MCA) techniques to our data. In this paper we provide background information and methodological reasons to our choice. We also provide an example of Social TV analysis, performed on Twitter data collected between October 2013 and February 2014.

Italiano. *Tecniche di NLP possono arricchire dati testuali non strutturati, individuando topic di interesse ed emozioni. Comprendere le somiglianze emotive fra diversi topic è un'attività cruciale, per esempio, nell'analisi della Social TV. E' richiesta una misura di quanto due tipi di pubblico condividano le stesse sensazioni, ma anche una rappresentazione compatta e coerente di queste somiglianze. Dopo aver valutato diversi approcci multivariati, abbiamo raggiunto questi obiettivi adattando tecniche di Multiple Correspondence Analysis (MCA) ai nostri dati. In questo articolo presentiamo background e ragioni metodologiche dietro tale scelta. Forniamo un esempio di analisi di Social TV, effettuata su dati Twitter raccolti fra ottobre 2013 e febbraio 2014.*

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). 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 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 (Ekman, 1972) (*anger, disgust, fear, joy, sadness, surprise*), *love* (a primary one in Parrot's classification) and *like/dislike* expressions, quite common on Twitter.

[1] www.blogmeter.it

2 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:

$$\mathbf{D}_{m_D \times n} = \begin{pmatrix} \mathbf{T}_{m_D \times n_T} & \mathbf{E}_{m_D \times n_E} \end{pmatrix},$$

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. For example, a document annotated with $\{$*topic = X Factor, emotion = fear, emotion = love*$\}$ 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:

$$\mathbf{J} = \begin{pmatrix} \mathbf{T}_J & \mathbf{E}_J \end{pmatrix},$$

where blocks $\mathbf{T}_J$ and $\mathbf{E}_J$ correspond to topics and emotions of the impressions.

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 to obtain 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 3. An explanation of the relation between MCA and PCA can be found, for example, in Gower (2006).

3 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 $\mathbf{C}$. 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 an 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.

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).

4 Comparison between MasterChef and X Factor

Among the studies we conducted, we present a comparison between two popular Italian formats: X Factor (music talent show, seventh edition) and MasterChef (competitive cooking show, third edition). Each episode is considered as a different topic. Results are shown in figure 1. 82% of total inertia (after adjustment) is preserved in two dimensions, making the representation accurate.

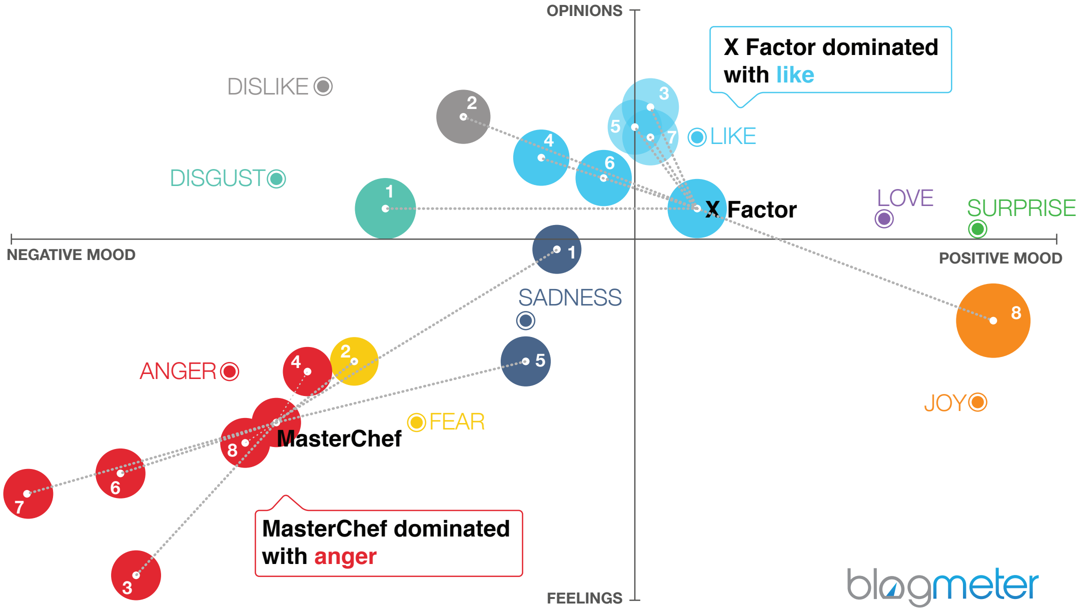

Figure 1: Comparison via MCA between X Factor and MasterChef formats, 2013-2014 editions.

Origin of axes in a MCA map acts as the barycenter, or weighted mean based on the number of emotional impressions, of all topics. In a similar way, we can consider the barycenters of X Factor/MasterChef episodes, highlighted in figure, as representants of the whole shows. Episodes are numbered progressively within each show: 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. This difference in volume is reflected in the distances from the origin, which can be considered as the average profile, and therefore closer to X Factor.

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.

5 Conclusions and further researches

By applying carefully chosen multivariate statistical techniques, we have shown how to represent and highlight important emotional relations between topics. 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 (**J** and **D** matrices); authors could also be considered as mean points of well-defined subclouds.

Acknowledgements

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.

References

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

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

Andrea Bolioli, 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, pp. 156-163. CEUR-WS.org.

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

Paul Ekman, Wallace V. Friesen and Phoebe Ellsworth. 1972. *Emotion in the Human Face*. Pergamon Press.

Dario Galati. 2002. *Prospettive sulle emozioni e teorie del soggetto*. Bollati Boringhieri.

John C. Gower. 2006. *Divided by a Common Language: Analyzing and Visualizing Two-Way Arrays*, Multiple Correspondence Analysis and Related Methods, Chapter 3. CRC Press.

Michael Greenacre. 1983. *Theory and Applications of Correspondence Analysis*. Academic Press.

Michael Greenacre. 2006. *From Simple to Multiple Correspondence Analysis*, Multiple Correspondence Analysis and Related Methods, Chapter 2. CRC Press.

Masafumi Hamamoto, 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.

Ian T. Jolliffe. 2002. *Principal Component Analysis*. Springer.

Brigitte Le Roux and Henry Rouanet. 2004. *Geometric Data Analysis: From Correspondence Analysis to Structured Data*. Kluwer.

Bing Liu. 2012. *Sentiment Analysis e Opinion Mining*. Morgan & Claypool Publishers.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Carl D. Meyer. 2000. *Matrix Analysis and Applied Linear Algebra*. Siam.

Janyce Wiebe, Theresa Wilson and Claire Cardie. 2005. *Annotating expressions of opinions and emotions in language*, Language Resources and Evaluation, Volume 39, Issue 2-3, pp. 165-210.