Papers on Social Representations
Volume 20, 38.1-38.7 (2012)
Peer Reviewed Online Journal
ISSN 1021-5573

[http://www.psych.lse.ac.uk/psr/]

Text Mining Methods: An answer to Chartier and Meunier

SAADI LAHLOU
London School of Economics

Almost 20 years ago, in a paper introducing the text mining (TM) technique to my fellow statisticians, I expressed the fear that: "it would be unfortunate that this technique, because it is apparently so easy to use, would be abused by incompetent analysts" (Lahlou, 1994, my translation). And therefore I urged expert statisticians to engage in this issue and circumscribe abuses.

Chartier and Meunier's paper in this issue is an echo to this ancient plea. Just as for any other statistical technique used in psychology, it appears that TM is sometimes reduced to using a software without clear understanding of the underpinnings. This is the usual destiny of statistical techniques in social sciences.

Fortunately, it seems that for TM, this stage is transitory. While in the first years of the technique, many publications using TM in Social Science and Humanities would have justified Chartier and Meunier's criticism that "the software is confused with the method", this is now more rarely the case. I think and hope that social psychology simply falls slightly behind in this evolution.

Therefore Chartier and Meunier's paper in this issue is useful in at least two respects. First (and although understated by the authors, this warning deserves our full attention) they make us aware that too many publications in our domain still tend to confuse the software with a method. This is especially the case for those publications using the Alceste software developed by Max Reinert (Reinert, 1983, 1987, 1990, 1993, 1998), because this software is so easy to use.

Second, they provide a good presentation of the underlying theoretical framework that enables us to apply TM in Social Representations (SR) studies (Moscovici, 1961, 2008). Their approach is focused on unsupervised classification techniques in TM (as opposed to supervised techniques of machine learning which rely on a pre-existing model); this is relevant since it is indeed so far the most used approach in SR research. Chartier and Meunier provide a mathematized presentation of how a text is transformed into numerical data with the Vector Space Model, therefore enabling it to be processed with mathematical techniques such as classification; and they connect the model to the Harrisian approach of constructing meaning from statistical distributions (Harris, 1952, 1991).

This detailed presentation is very useful for social psychological research: the absence of such formalisation in English has always been a major obstacle to publishing papers in English-speaking scientific journals. This was due to the fact that most reviewers cannot access the classic literature on these techniques, which is in French for historical reasons, all these techniques stemming from the mathematic school of Jean-Paul Benzecri (Benzecri, 1973, 1981). Now, at last, those who use the technique will have a reference in English to help enlightening those reviewers who are not yet familiar with this technique.

Chartier and Meunier provide a clear three-step model for the TM method itself (data collection, data modeling, data analysis). The last step is part of the first step of the SR analysis *per se*, as described by Abric (2003): (1) SR content and category identification, (2) SR structure identification, and (3) SR core identification. As they pertinently note, most TM on SR so far have only reached this first step -or at best the second one. In this respect, I must admit that my own attempt (Lahlou, 1993) to delimit the core of a SR through analysis of material coming only from first-order associations (*vs*. those coming from first- and second-order associations) is indeed limited and lacks experimental validation with other techniques. Let me take this opportunity to highlight that triangulation of methods is always recommended.

The authors conclude, and I concur, that these TM methods are still underused in SR research and have a bright future; I will come back to this point.

It is a useful paper, and I must refrain from making a very long response. The gist is: Yes, TM is of great value for SR research, because it is adapted to processing the large amounts of data that a truly social approach, with many sources/participants, requires. Yes, "the method must not be confused with the software that implements it", and researchers should be more aware of which operations they are actually performing in their "analysis".

The idea that I will try to elaborate on below is the following: Just as the software is not the method, the software outputs are not the analysis. Interpretation is an abductive process; it emerges through a series of trial and errors where the researcher tries to match the outputs of the analysis with a model. This is some kind of triangulation, which can be done by trying various parameters with the same technique or software, using several softwares, using different techniques (e.g., TM and another investigation technique like interviews), and more generally comparing several views of the empirical material. For this reason, basing a model on a single run of analysis with one software is not enough.

Multivariate analysis –and TM classification techniques described in Chartier and Meunier's paper belong to this family- should be used as an *exploratory* technique to construct a model. In the case of SR, we are interested in finding a psychological model, while our data are expressions in natural language of the object being represented. Therefore, these data reflect the psychological model, but also the way by which the data have been obtained and projected into language. TM provides findings regarding both, and these two layers have to be untangled by the analyst. For example, some aspects of the social representation will not be present in the data because they have not been well projected into the discourse, because they are difficult to express, are politically incorrect in the context of enunciation, or because the language has its own structure which also appears in the co-occurrences. Only the "external knowledge" of the analyst (outside of the data set itself) enables to interpret the data. This means that the knowledge that the analyst has of the language, of the topic, and of the software are limiting factors.

The raw results of TM usually fall into one of five categories: the trivial, the classic, the unexpected, the artefact, and the residue (Schonhardt-Bailey, Yager, & Lahlou, 2012). The *trivial* are those which are so obvious as to be uninteresting (although in the case of SR, still, they are usually worth to mention since SR are precisely common sense). The *classic* are the ones that are consistent with previous research. The *unexpected* are new findings that the analyst can back up as "solid" with some other source of explanation or data. The *artefact* is what is due to technical issues with the data processing, e.g., in the case of Alceste that some repeated chains or idiomatic expressions (like "God Bless America") may generate a cluster in the classification process because of the strong association between these words; therefore sentences containing these words may be aggregated together wrongly, "pulled" in the class by this strong association between the words in the repeated chain. Of course the analyst

should understand the way the software proceeds in order to spot and correct such artefacts. Finally the *residue* is what the analyst is unable to interpret.

Therefore a good analysis requires understanding the underlying theoretical framework of the technique and the software, and awareness of the abductive processes at work in interpretation. This is needed in order to determine what is indeed interpretable, to construct a model, and to enrich the classic knowledge with explanation for the unexpected. Because this process is abductive, it happens through loops of successive data processing producing a software output (which we abusively tend to call "analyses"), where, by varying the parameters in the software, the analyst gradually understands what are the artefacts, as the analyst tries to match the results in the ouput with her own understanding of the results. For example, the artefacts coming from fixed expressions like "God Bless America" as mentioned earlier can be suppressed by transforming such fixed expression into one single word unit: God_bless_America, hereby suppressing the excessive association between "God" and "America" in the corpus. This is why it is so dangerous to operate only a single run with the default settings of the software.

We have described elsewhere in detail (Beaudouin & Lahlou, 1993; Lahlou, 1995a, 1996, 2003) how the analyst intervenes in the analytic process, at all the three phases described by Chartier and Meunier. Farr's paper (1984) reprinted in this issue with Jovchelovitch's comment insightfully points at how the social representations of the scientists themselves frame the situation they observe. More caution is thus needed in distinguishing the SR as we, scientists, construct it as a model, and the SR as it is *per se* in the wild, from the participants' perspectives. The former will always be a biased simplification of the latter.

TM are definitely a progress on simple quantitative variables, because they cover a larger array. But TM techniques are still limited by the very nature of the verbal material they use, and SR are multimodal. SR study should rely on a larger "praxeo-discursive" corpus, including practice and discourse (Flament, 1989, 1994). Motor aspects, emotional connotations, and other embodied aspects are poorly projected in verbal material (Lahlou, 1995a: 241, 283, 302) and I would strongly recommend that TM be used only as part of a triangulation approach combining several methods and tapping into activity-in-context and not only discourse

I share the authors' frustration that TM is still applied so scarcely in SR research. This may be partly due to the cost of the software (see below). I am surprised and disappointed that the technique which I set up for SR, using electronic dictionaries as a source, which is so easy

and handy, is not yet systematically applied as a first screening technique to explore SR in every research.

I have no space to expand here on technical considerations; I have attempted to describe elsewhere in detail the nature of the interpretation process, and how one should proceed in my view (Beaudouin & Lahlou, 1993; Lahlou, 1995a, 2003); the reader will also find in the appendix of a paper cited above (Schonhardt-Bailey et al., 2012) a description in English of how Alceste actually processes the data which may be helpful when one wants to publish in English journals.

To conclude, I also concur with Chartier and Meunier on the general lack of creativity in using TM techniques: too many papers are simply using Alceste as a default solution. I must admit some share of responsibility in this state of affairs, because I initially advocated for Alceste and trained my colleagues to use it; but I could have never imagined that it would lead to such a limitation. Even if Alceste is a great software, thanks to the genius of Max Reinert, there is a wealth of statistical software available to match specific needs: e.g., see (Brugidou et al., 2000; Quatrain et al., 2003). And counting! There are now many web platforms to orient the users, and lists (e.g., at http://www.kdnuggets.com/software/text.html). The Text Mining research seminar set up at LSE by Martin Bauer and Aude Biquelet is an effort in the same direction.

Regarding cost issues, a new software using the same algorithms as Alceste, in an open-source and free version, has now been programmed by Pierre Ratinaud using the R statistical programming language. This program, IRAMUTEQ, is available for download at: http://www.iramuteq.org/. All one needs to do is to install the R software (http://www.r-project.org/) and then IRAMUTEQ, which incorporates some new interesting features that were not included in Alceste. Ratinaud has an in-depth understanding of these techniques and his work is brilliant. No doubt this should foster the development of TM in the SR research field which Chartier and Meunier call for.

But remember: as Chartier and Meunier say, the software is not the method; and as I tried to highlight above, the software outputs are not the analysis. The software is only an instrument for exploration; *interpretation* is performed by the analyst using her knowledge external to the text.

REFERENCES

Abric, J.-C. (2003). Abric, J.C. (2003). L'analyse structurale des représentations sociales. In. In S. Moscovici & F. Buschini (Eds.), *Les méthodes des sciences humaines* (pp. 375-392). Paris: P.U.F.

Beaudouin, V., & Lahlou, S. (1993). L'analyse lexicale : outil d'exploration des représentations. *Cahiers de Recherche CREDOC*.

Benzecri, J.-P. (1973). *L'analyse des données*. Paris: Dunod.

Benzecri, J.-P. (1981). Analyse statistique des données linguistiques. LA n° 1 (ANA. LING.). In J.-P. Benzecri (Ed.), *Linguistique et lexicologie, Vol. 3* (pp. 3-45). Paris: CNRS, Bordas.

Brugidou, M., Escoffier, C., Folch, H., Lahlou, S., Le Roux, D., Morin-Andreani, P., & Piat, G. (2000). Les facteurs de choix et d'utilisation de logiciels d'Analyse de Données Textuelles. *JADT 2000*.

Farr, R. M. (1984). Social Representations: their role in the design and execution of laboratory experiments. In R. M. Farr & S. Moscovici (Eds.), *Social Representations* (pp. 125-147). Cambridge: Cambridge University Press.

Flament, C. (1989). Structure et dynamique des représentations sociales. In D. Jodelet (Ed.), *Les représentations sociales* (pp. 204-219). Paris: P.U.F.

Flament, C. (1994). Structure, dynamique et transformation des représentations sociales. In J.-C. Abric (Ed.), *Pratiques sociales et représentations* (pp. 37-57). Paris: PUF.

Harris, Z. S. (1952). Discourse analysis. *Language*, 28, 1-30.

Harris, Z. S. (1991). A theory of language and information: A mathematical approach. Oxford: Clarendon Press.

Lahlou, S. (1993). *Penser manger : les représentations sociales de l'alimentation*. PhD thesis. Paris: Ecole des Hautes Etudes en Sciences Sociales,. http://tel.archives-ouvertes.fr/tel-00167257_v1/

Lahlou, S. (1994). L'analyse lexicale. *Variances*, 3, 13-24. http://www.ensae.org/gene/main.php?base=38&base2=2&detail_article=125

Lahlou, S. (1995a). Vers une théorie de l'interprétation en analyse des données textuelles. JADT 1995. 3rd International Conference on Statistical Analysis of Textual Data. S. Bolasco, L. Lebart, A. Salem (Eds). CISU, Roma, 1995, Vol I (pp. 221-228.).

Lahlou, S. (1995b). *Penser Manger. Alimentation et Représentations Sociales.* Paris: P.U.F

Lahlou, S. (1996). La modélisation de représentations sociales à partir de l'analyse d'un corpus de définitions. In E. Martin (Ed.), *Informatique textuelle*. INaLF, coll. "Etudes de sémantique lexicale" (pp. 55-98). Paris: Didier Erudition.

Lahlou, S. (2003). L'exploration des représentations sociales à partir des dictionnaires. In J.-C. Abric (Ed.), *Méthodes d'étude des représentations sociales* (pp. 37-58). ERES. Moscovici, S. (2008). *Psychoanalysis: Its image and its public* (p. xxviii, 384 p.). Cambridge: Polity.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Les Cahiers de l'Analyse des Données, Vol. VIII*, 187-198.

Reinert, M. (1987). Classification descendante hiérarchique et analyse lexicale par contexte: application au corpus des poésies d'Arthur Rimbaud. *Bulletin de Méthodologie Sociologique*, 13.

Reinert, M. (1990). ALCESTE: Une méthodologie d'analyse des données textuelles et une application: Aurélia de Gérard de Nerval. *Bulletin de Méthodologie Sociologique*, 26, 24-54.

Reinert, M. (1993). Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et Société*, 66, 5-39.

Reinert, M. (1998). Quel objet pour une analyse statistique du discours? Quelques réflexions à propos de la réponse Alceste. In S. Mellet (Ed.), *JADT* (pp. 557-569). Nice: Université de Nice.

Schonhardt-Bailey, C., Yager, E., & Lahlou, S. (2012). Yes, Ronald Reagan's rhetoric was unique — But statistically, how unique? *Presidential Studies Quarterly*, 42, 482-513.