

Medical communication quality in the Italian pharmaceutical industry: measurement and analysis by ‘NOOS’

P. Mariani¹ & G. Ventre²

¹*Department of Statistics, Faculty of Statistics, Milano Bicocca University, Italy*

²*ASM, Milan, Italy*

Abstract

Methodology represents an original application of text mining aimed at utilising Broadbent’s model in the pharmaceutical ambit. This approach offers a quantitative measure of the message scatter. From communication from the emitter, the receiver absorbs partly or wholly the communication, but can adapt the contents by its experience or by “communication filters”. Measurement can be performed by asking the receiver to describe the communication contents by means of verbalisation of the communication itself. The use of multiple correspondence analysis and cluster analysis has permitted a quantification of the main dimension beneath the data and a division in patterns on the basis of different linguistic models to be obtained. A test has been performed on a sample of 51 articles extracted from “NOOS”, a magazine of psychiatric argument. Four articles were chosen like the model and compared with the rest of the sample, verifying textual correlation and related with the four articles model. In this way a simulation of the communication flow has been made from the Company (Medical Department) to the physicians, passing through different staff of the firm, particularly REP, and it has been possible to measure the quota of correct non-scattered communication between companies and physicians.

Keywords: text mining, communication scatter.

1 Introduction

The number of physicians enrolled in the Italian Physician Roll are 347.759 (Federazione Nazionale degli Ordini dei medici chirurghi e degli Odontoiatri,



2004), the number of health organisations like hospitals, public and private, in convention or not with SSN (Health National Service), is 1.410 (www.ministerosalute.it, November 2004); pharmacy both private and hospital are 16.808 (www.federfarma.it, November 2004). Pharmaceutical enterprises that in any different way contact doctors are, more or less, 262 with a distributor's network fast supplying the pharmacy for the 5.019 drugs on sale for an amount of 8.742 packages (Farmindustria [7–9]). Pharmaceutical enterprises have direct contact with the doctors by the REP (Representative, Scientific Informer about Drugs) by means of these representatives the companies communicate and inform about their products. The flow of communication starts from the “middle” of the firm and arrives outside across different steps. During the path towards the final receiver it is possible that a message is liable to variations and, often awkward transformations. In other words, it may happen that a part of the message is lost or misunderstood. This work proposes a technique to use the Broadbent's model [4] by the quote of scattered communication between companies and physicians

Medical-scientific information, in general terms advertising and promotion, has, as a goal, to improve brand and product notoriety, to improve the perception of product characteristics and, last but not least, to augment the prescriptive propensity.

There is a three steps process that represents the answer of the consumer, in this case the physician, to the promotional/informative action: a first cognitive step where the message is perceived, elaborated and, at least in part, memorised; a second step where a communication effect changes the perception related to a product or a brand and the third step where we can aspect a change of behaviour. In this article we discuss the aspects linked to the first step or rather the aspects concerning the measure of the efficacy of the communication on the cognitive answer.

Particular importance assumes the measure of the memory of the medical scientific information communication in the pharmaceutical market where the choice of means is limited and there are law enforcements about the ways and the contents of the communication.

2 The model

In literature there are cited several models for the measurement of the effects of advertising in the mass market, among the more utilised the models of Zielske, Morgenstern, Broadbent (Brasini et al. [3]). In general terms in these models the memory is a function of variables correlated with the number of exposure. In the Broadbent's model the memory is a function of a latent variable indicating in a time “t” the share of advertising memory like effects of past and present activity: this variable is called “adstock”.

At time t :

$$S_t = bAds_t \quad (1)$$

$$Ads_t = (rA_t + r^2 A_{t-1} + r^3 A_{t-2} + \dots + r^n A_{t-n-1}) \quad (2)$$

where:

r is the decay of information action

A represents the promotional pressure

Ads is the adstock variable

The proposed technique measures, by means of textual analysis, the memory of the communication and calculates an index that considers the exposure at the communication. This index creates the parameter expressing the decay or the memory in a model that relay the physician's behaviour with the exposure at the scientific-medical information.

The proposed model is:

$$S_t = b_d Ads_t \quad (3)$$

$$_d Ads_t = ({}_d r {}_d A_t + {}_d r^2 {}_d A_{t-1} + {}_d r^3 {}_d A_{t-2} + \dots + {}_d r^n {}_d A_{t-n-1}) \quad (4)$$

where:

${}_d r$ is the decay of information action declined by:

- ${}_d r_{pr}$ product scatter
- ${}_d r_{pa}$ pathology scatter
- ${}_d r_{to}$ general scatter

${}_d A$ represents the promotional pressure

${}_d Ads$ is the adstock variable corrected for the pharmaceutical market.

The general rule to define the decay is not necessary a Geometric distribution but it is linked to the different r defined by the model.

3 The technique

Methodology represents an original application of text mining. Connected with the possibility of considering every single word like a qualitative modality of a positional variable within a text, this approach offers a quantitative measure of the message scatter. There are two hierarchical levels between emitters and receivers. The first indicates the flow from the Company (Medical Department) to the physicians, passing through different staff of the firm particularly REP. The second level indicates the measure and moves re-ascending the previous flow and is defined from doctors to the enterprises. In front of a communication from the emitter, the receiver absorbs partly, or in the whole, the communication but can adapt the contents by his experience (Williams [13]) or by "communication filters" (Kanizsa et al. [10]). Measurement can be performed asking the receiver to describe the contents communication by means of a verbalisation of the communication itself. The use of multiple correspondence analysis and cluster analysis (Bouroche and Saporta [2]) offers a quantification of the main dimension beneath the data and a division in patterns on the basis of different linguistic models (Lebart and Salem [12]).

Statistical analysis concerning textual data divided the text in graphical construction. We can therefore define:



- OCCURRENCE: every word that appears in the text·
- WORD TOKEN: the number total of occurrences·
- WORD TYPE: the various number of words that compose the text·
- DICTIONARY: the group of the lemma of a language
- RANK: when the word type is ordered

The text can be used as in the example, or can be subordinate to two procedures: the first considers grouping the occurrences that correspond to the different declinations of the same lemma (lemmatisation); the second that distinguishes the various uses or means that can have identical occurrence (disambiguation). The disambiguation increases the content of the corpus while the lemmatisation, even though necessary, loses information like the subject, the gender, etc.

Confronting different corpus or parts of it's, the dictionary can be distinguished in:

- banal: that is not typical and it's present in both
- specific: that it introduces strong differences in the two corpus
- original: or highly specific exclusive of the corpus in examination.

Both the specific dictionary and the original dictionary constitute an object of analysis. The result of this approach to the text is directed to the textual analysis by means of the preparation of matrixes of data then elaborated with factorial analysis and cluster analysis.

The multiple correspondences analysis method (Carroll [5]) identifies the structure of inner dependency into a table of n rows and p columns. This technique is used for categorical variable and allows also the use of qualitative variable. The main advantages can be synthesized in two points:

- the technique does not impose a metric. The measure of the relations between the variable ones is the chi-squared that analyses the variable like categorical; in this way hypothesis on the distribution and on the variable scale are not requested;
- big data amount from n -dimensional spaces can be reduced to less dimensions with a relatively little loss of information.

In the correspondences analysis the descriptive statistical aspect is pre-eminent. The cluster analysis whose objective is to characterize groupings of objects that have two complementary characteristics: the maximum homogeneity in each cluster, the maximum difference between the clusters (Everitt [6]). The advantage to base the analysis of the cluster on the factorial analysis is to classify the elements starting from the position that they have in the factorial space characterized previously. In this way result of the relations evidenced from the main factors is kept meaningful and result of relations of smaller importance is not. Moreover we use the factorial scores to calculate the distances between the elements of the cluster. These elements, originally qualitative, are expressed in distances in quantitative data and used with an appropriated metric.

4 The simulation

For the simulation, a sample of 51 articles of psychiatric subject extracted from the review “NOOS - Aggiornamenti in Psichiatria” (NOOS – Update in Psychiatry) has been used. This is a quarterly review published by “Il Pensiero Scientifico Editore” in collaboration with a pharmaceutical Company. Every number is monographic. Noos started publishing in 1995 and is already on course. In this application articles from number 1 to number 18 of the review have been used. Four articles have been chosen like communication models representing, as emitter, the communication of the Medical Department. From these articles have been extracted the exclusive and the more significant words, in order to compare them with the result from the whole sample. The subject of the four model articles are: borderline disease, obsessing compulsive disease, schizophrenia, and phobia. In this way, comparing a single article with the whole sample, it is possible to verify the textual correspondences and link them with a specific issue/article which represent a specific communication source.

By means of multiple correspondence analysis the main dimension beneath the data are defined. The first dimension ANXIETY (22,9 %) is qualified for the opposition between the negative side characterised by phobia and the positive side characterised by different pathology like obsessing compulsive disease, schizophrenia and others. This factor divides the anxiety dimension and other pathology and therapy. The second dimension MOOD (18,7 %) has on the opposite side different aspects of mood: on the first side the article referring to the pathology maniacal; on the other side pathology and therapy related to the psychotic mood are present. In the third dimension THERAPY (17,7 %) the opposition is between pharmaceutical therapy and psychotherapeutic issues.

The second step was the application of a hierarchical classification using the Ward algorithm. Clusters generated in this way are compared with the four articles chosen like communication models.

PHARMACEUTICAL APPROACH (59.8 % of cases): cluster groups lemma related to the pharmaceutical approach, a general approach in the medical culture not related with a specific communication model. **PHOBIA AND ANXIETY** (6.0 % of cases): specifically linked to the lemma of the model article about phobia is possible to consider this cluster related with its model.

BORDERLINE AND PERSONALITY DISEASE (26.1 % of cases): it is possible to link the cluster with the correspondent model article by means of lemma having high frequencies both in the cluster and in the article. Particularly significant the correspondence of some lemma, like “borderline”, relevant in the cluster and exclusives in the model article. **MANIACAL MOOD** (8.1 % of cases): this cluster has poor affinity with the model articles too and, as in the first cluster, the language is general of medical culture.

The third step consists in determining the values of r and $\mathbf{A}$ of the Broadbent's model.

$\mathbf{A}$ represents the promotional pressure measured as ratio between the calls (the frequencies of the visit) for the product of the company and the total of the calls for the product of the same ATC (Anatomical Therapeutical Classification)

done by the concurrent company on the same target. The value of r (the decay of the memory) may be defined in three different ways in relation to:

- the product ($_{DPr}r$): the ratio between key words referred to the products and found within the responder's clusters and the whole of the products key words;
- the pathology ($_{DPa}r$): the ratio between key words referred to the pathology and found within the responder's clusters and the whole of the pathology key words;
- the total of the communication ($_{Dtot}r$): the ratio between the sum of key words the products and referred to the pathology and found within the responder's clusters and the whole of the key words.

The cross between the cluster and the communication produces the r value reported in table 1.

Table 1: Decay of information.

	PHARMACEUTICAL APPROACH	PHOBIA AND ANXIETY	BORDERLINE AND PERSONALITY DISEASE	MANIACAL MOOD	TOTAL
% of cases	59,8	6,0	26,1	8,1	100,0
EXCLUSIVE	$_{DPr}r$				
REP 1	0,0%	0,0%	30,0%	0,0%	30,0%
REP 2	6,4%	2,6%	0,0%	0,0%	9,0%
REP 3	4,2%	0,0%	0,0%	4,2%	8,3%
REP 4	2,0%	6,0%	6,0%	0,0%	14,0%
TOTAL	12,6%	8,6%	36,0%	4,2%	61,3%
CHARACTERISTIC	$_{Dpa}r$				
REP 1	1,8%	0,0%	12,3%	1,8%	15,8%
REP 2	8,2%	0,0%	8,2%	0,0%	16,3%
REP 3	9,0%	0,0%	4,5%	0,0%	13,4%
REP 4	6,5%	4,3%	2,9%	0,0%	13,8%
TOTAL	25,4%	4,3%	27,8%	1,8%	59,3%
TOTAL	$_{Dtot}r$				
REP 1	1,5%	0,0%	14,9%	1,5%	17,9%
REP 2	7,1%	1,6%	3,1%	0,0%	11,8%
REP 3	7,7%	0,0%	3,3%	1,1%	12,1%
REP 4	5,3%	4,8%	3,7%	0,0%	13,8%
TOTAL	21,6%	6,4%	25,1%	2,6%	55,6%

For instance the REP 1 had a major effect of his communication on the cluster "BORDERLINE AND PERSONALITY DISEASE" representing 26,1% of the cases and obtain:

- $_{DPr}r$ = cluster's products key words (exclusive) / total products key words
 $\% = 30\%$,

- $D_{Pa}r = \text{cluster's pathology key words (characteristics)} / \text{total pathology key words} \% = 12,3\%$,
- $D_{tot}r = (\text{cluster's products key words} + \text{cluster's pathology key words}) / \text{total key words} \% = 14,9\%$.

5 Conclusions

This contribution represents an original application of text mining. In front of a communication from the emitter, the receiver absorbs partly or completely the communication but can adapt the contents by his experience or by “communication filters”. Measure can be performed asking the receiver to describe the contents communication by means of verbalisation of the communication itself. The use of multiple correspondence analysis and cluster analysis permits one to obtain a quantification of the main dimension beneath the data and a division in patterns on the basis of different linguistic models. A test has been performed on a sample article extracted from “NOOS”, four articles were chosen as a model and compared with the rest of sample, verifying textual correlation and the relation with the four articles model. In this way a simulation of the communication flow from Company (Medical Department) to the physicians has been made passing through different staff of the firm, particularly REP, and it has been possible to measure the quota of correct communication between companies and physicians. The high classification capability of the four articles represents the comprehension rate of the message. It offers the opportunity to try to define a value for the decay, for instance linking the key words of emitter with the key words used in his description by the receiver. The application of the scatter analysis model returns three different indicators: $D_{Pr}r$,

$D_{Pa}r$, $D_{tot}r$

$D_{Pr}r$: product scatter, the ratio between key words referred to the products and found within the responder's clusters and the whole of the products key words

$D_{Pa}r$: pathology scatter, the ratio between key words referred to the pathology and found within the responder's clusters and the whole of the pathology key words

$D_{tot}r$: general scatter, the ratio between the sum of key words referred to the products and referred to the pathology and found within the responder's clusters and the whole of the key words.

This index is inserted in a Broadbent's model, in order to estimate the “ r ” coefficient and the value of the adstock variable through the scatter of the communication and the decay of the memory.

The next step may be a pilot survey based on real cases.

Acknowledgements

The articles were supplied by Dr. Ettore Marre who deserves our thanks. This note though, is the result of a close collaboration and was specifically elaborated as follows: paragraphs 1,2,3 e5 by P. Mariani and 4 by G. Ventre.



References

- [1] AA. VV., *NOOS – Aggiornamenti in psichiatria*, n. 1-18, Il Pensiero Scientifico Editore, Roma, 1999-2000.
- [2] Bourroche J.M., Saporta G., *l'analyse des données*, Presse Universitaires de France, Paris, 1980.
- [3] Brasini S., Tassinari F. & Tassinari G., *Marketing e pubblicità. Metodi di analisi statistica*, Il Mulino, Bologna, pp.384-390, 1999.
- [4] Broadbent S., *Accountable advertising - A handbook for managers and analysts*, Admap Publication, Henley –on –Thames, pp. 153-183, 1997.
- [5] Carroll J.D. *Generalisation of canonical correlation analysis to three or more sets of variables*, Proceedings of the 76th Annual Convention of the American Psychological Association. 227-228, 1968.
- [6] Everitt B. S. *Cluster Analysis*. III edition, Arnold, London, 1993.
- [7] Farindustria, *Fatti e cifre 2004*, Centro Studi Farindustria, Roma, 2004.
- [8] Farindustria, *Indicatori farmaceutici*, Centro Studi Farindustria, Roma, 2004.
- [9] Farindustria, *Scenari per il settore farmaceutico*, Ufficio Economico Sanitario Farindustria, Roma, 2004.
- [10] Kanizsa G., Legrenzi & P. Sonino M., *Percezione, linguaggio, pensiero*, il Mulino, Bologna , 1983.
- [11] Lancia F, *Strumenti per l'analisi dei testi. Introduzione all'uso di T-LAB*, Franco Angeli, Milano, 2004.
- [12] Lebart L., Salem A., *Statistique textuelle*, Dunod, Paris, 1994.
- [13] Williams K. C., *Psicologia per il marketing* , il Mulino, Bologna, 1981.