Monitoring science and technology in Italian news, 1993-2013

Andrea Lorenzet
Observa Science in Society
University of Padua
andrea.lorenzet@unipd.it

Federico Neresini
Observa Science in Society
University of Padova
federico.neresini@unipd.it

Abstract

In this paper we describe work of corpus construction and analysis for science and technology coverage for the period 1993-2013. Artificial week schedule has been applied to build a corpus of articles from two main Italian newspapers (La Repubblica and Il Corriere della Sera) and to analyze the presence of science and technology topics over the whole period; data collected has been stored in a database and a user interface/search engine built; the Observa Science in Society *Science in the Media Monitor* (SMM) machine system has been used to automatically detect presence of science and technology in the corpus, peaks of coverage and main associated issues, and LDA algorithm is used for automated topic detection. Results show that there were two periods in which science and technology has been peaking, and precisely 1997-2000 and 2008-2010 in Italian newspapers and that issues related to digital, renewables and stem cells tend to be more covered in the second peak of coverage; at the same time, 6 main topics regarding science and technology have been detected within the relevant articles of the whole period of coverage, and precisely economics, policy, education, health, environment, and web/ict.

In order to analyze the presence and the content of scientific news over a consistent period of time we used the methodology of artificial week taken from London School of Economics sample for the MACAS (MApping the Cultural Autority of Science)[1] project, covering even years in the period 1990-2013 and applied it to both even and odd years, in order to prepare a corpus of articles for this timespan, and at the same select relevant science and technology articles through the automatic selection on the basis of a weighted keywords classifier developed as part of the Science in the Media Monitor project at Observa Science in Society, Italy[2].

To do this, the websites of two main Italian newspapers have been scraped through dedicated software, in order to collect all the articles published in the selected dates. The system collected in total 167472 articles starting from year 1990, and a user interface has been created in order to store, analyze and make the material available. In this paper we do not consider the first three years of the sample (1991, 1992, and 1993, years for which data were very few and results were therefore not consistent with the following period), and from some local news and video pages present above all in the latter years of the sample.

At the end of this procedure, we ended up with a sample of a total of 141819 articles, of which 61669 (43.5%) were collected from the Corriere della Sera database, and 80150 (56.5%) from the La Repubblica database.

To this corpus we applied the SMM thesaurus for automatic detecting science and technology content in the news (Di Buccio et al. 2014, Giardullo and Lorenzet 2013, Lorenzet and Neresini 2012, Lorenzet and Neresini 2011), in order to automatically assess science and technology content of news articles.[3]

The system found a total of 8424 relevant articles with science a technology content, of which 44.6% were belonging to Corriere della Sera (n=3758), and 55.4% to La Repubblica (n=4666).

For this articles, the measure of salience, that is the percentage of science and technology content relevant articles out of the total has been calculated; this measure tells

[1] For more information about MACAS see http://www.macas-project.com
[2] For more information about the Science in the Media Monitor project at Observa Science in Society see http://www.observa.it/science-in-the-media-monitor/?lang=en and the link pointing to the demo preview.
[3] The thesaurus of keywords is at the moment under update phase, data here presented should therefore be considered subject to changes after the process of updating and publication of results.

us the degree to which over the considered period the media discuss S&T contents in quantitative terms; for the period 1990-2010 salience for the two newspapers Il Corriere della Sera and La Repubblica has been 6%, that is 6 articles out of 100 had consistent science and technology content.

Peaks detection

Chart 1 describes the curve of salience for each sampled year and for the two newspapers together, telling that over the considered period there were two sub-periods in which science and technology topics headed the top of the news.

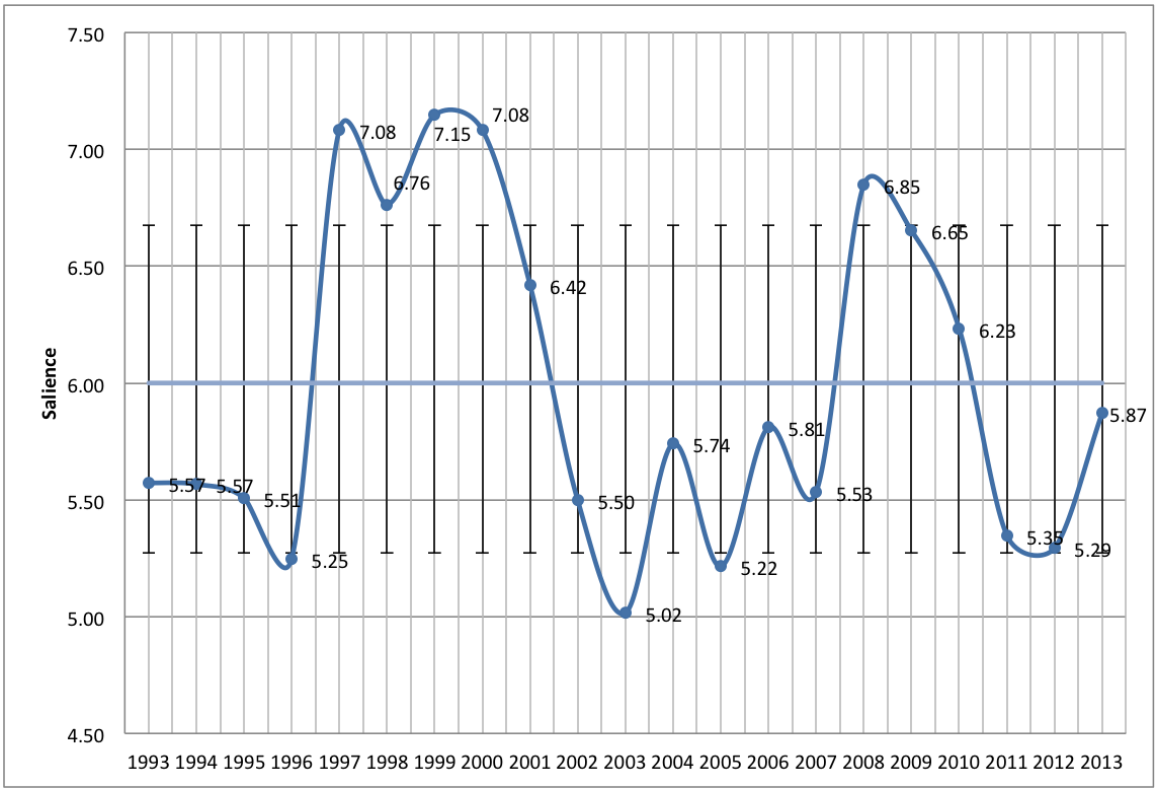

Figure 1 - S&T salience in Italian press, 1993-2013 (% of articles with science and technology content on total articles collected

The first peak period corresponds to years 1997-2000, and the second period corresponds to 2008-2009. While the first period can be considered a peak of *very high S&T coverage* in the Italian news with four years in a row in which salience has values that are higher than the average plus one standard deviation for the sampled considered years (indicated in the chart with the error bars), the second period is a shorter in time and less intense, and can be thus defined as a period of *high S&T coverage*. It is also worth noting that as many other phenomena, S&T coverage in the Italian news demonstrate to have an oscillatory trend/behavior, with periods of high or very high coverage that alternate to periods of *average S&T coverage* (within error bars span), and periods of *low S&T coverage* occurring in single years (in our sample in 1996, 2003, and 2005, when values are below error bars).

In order to have some information about the topics covered within the peaks periods we can use, among the others, indications offered by the most frequent keywords among those used in the classifier to automatically detect the articles; results of this analysis are shown in table 1.

As the data show, if we compare presence in the corpus of the relevant articles published in the two periods, words regarding the environment and digital innovation, together with the "stem cell" issue and the topic of researchers' funding within university became more relevant in time, while other topics such as biotechnology or physics got less attention in the coverage peaks.

	1997-2000 (n=1317)		2008-2009 (n=1141)		
	n	%	n	%	increase/ decrease (%)
renewables	15	1.1	54	4.7	*+3.6*
researcher	168	12.8	185	16.2	*+3.5*
stem cells	4	0.3	29	2.5	*+2.2*
digital	103	7.8	112	9.8	*+2.0*
eolic	2	0.2	20	1.8	*+1.6*
robot	31	2.4	45	3.9	*+1.6*
university	312	23.7	288	25.2	*+1.6*
biomass	3	0.2	20	1.8	*+1.5*
biology	120	9.1	111	9.7	*+0.6*
pollution	74	5.6	71	6.2	*+0.6*
kyoto	12	0.9	16	1.4	*+0.5*
genom*	30	2.3	15	1.3	*-1.0*
cancer	91	6.9	60	5.3	*-1.7*
software	106	8.0	70	6.1	*-1.9*
genetics	74	5.6	32	2.8	*-2.8*
biotech*	53	4.0	12	1.1	*-3.0*
drug	200	15.2	139	12.2	*-3.0*
engineering	90	6.8	42	3.7	*-3.2*
physics	166	12.6	91	8.0	*-4.6*
elettron	229	17.4	134	11.7	*-5.6*
computer	232	17.6	128	11.2	*-6.4*
medic*	401	30.4	274	24.0	*-6.4*

Figure 2 - Most frequent keywords in the relevant science and technology articles published during the peaks of S&T coverage in the Italian media.

Moreover, it is worth of attention the fact that among the keywords that get more attention by the media in both periods we find “medic”, “university”, “researcher”, “drug”, “electron” and “computer”.

LDA – Automatic topic detection

On the overall corpus, a procedure of automated topic detection called LDA (Latent Dirichelet Allocation) has been applied (Blei 2012, Blei and Lafferty 2006).

Topic modeling algorithms are statistical methods allowing to automatically detect the most significant topics within a given set of documents. During the last decade several topic modeling algorithms have been proposed, differing mainly on their assumptions (for example on the basis of the relationships among the topics to be extracted). The method we used for this analysis is based on the algorithm *Latent Dirichelet Allocation* (LDA), on the basis of which we find the assumption that documents are characterized by a given set of topics, where a topic is defined as a distribution on a fixed set of words: for example within the topic "biomedical research and stem cells", the words regarding biomedical research and stem cells will be present with a high probability. Topics manifest within documents in different proportions: to do the analysis here described, we used the open source software called "Mallet", allowing to apply LDA to a set of documents, specifying the number of topics to be extracted and the number of keywords to be visualized (such words can be interpreted as the one that best describe each topic). To do the analysis, we extracted from the Science in the Media Monitor database 10 topics and set word visualization at 20 items for the whole sample of very relevant articles (those articles that obtained a score in the S&T classifier of the SMM machine higher than the arbitrary value of 50) that we collected in the period 1992-2010 (n=3293).

Results of our analysis are shown in table 1, where we see the presence of some of the topics highlighted by the procedure, and precisely of 6 topics that clearly indicate reference to specific domains of science and technology related issues. We decided to leave out from results 4 topics; besides the 6 topics displayed, the procedure actually found out a general frame for S&T articles, referring to the idea of future (weight 0.90), and other three topics for which words did not indicate direct reference to a coherent semantic domain. On the basis of the list of the 20 most relevant keywords within each topic, labels have been assigned to the topics in order to clearly identify them.

Topic Label	Weight	Keywords
ECONOMICS	0.23	million italy market billion year companies countries development firm sector world corporation europe research euro economics president industry innovation
POLICY	0.22	president law ministry yesterday government politics council jurisprudence italy commission safety fact national citizens union director healthcare rights case
EDUCATION	0.22	research university students year degree curricula sciences institute engineering school courier education study faculty young people courses deadline headquarters graduate
HEALTH	0.20	health cells disease diseases patients drugs risk cases doctor research doctors virus cancer kids hospital blood therapy women
ENVIRONMENT	0.18	water energy environment earth pollution gas nuclear environmental emission city km waste meters air millions planet year around page
WEB/ICT	0.14	internet computer network software technology web microsoft system market digital new google car services world site tv information online

Figure 3 - 6 science and technology topics in Italian newspapers 1992-2010 detected through the LDA procedure

In the table, weights indicate the degree to which each topic is present in the corpus of articles; higher the score, higher the probability to find that topic in an article. From the table, we see that economics, policy, and education are the main media issues regarding science and technology in the Italian press, and also that other major issues regard health, environment, and web/ict.

References

Blei David M. (2012), "Probabilistic topic models." Communications of the ACM 55.4, pp. 77-84.

Blei, David M., and John D. Lafferty (2006), "Dynamic topic models." Proceedings of the 23rd international conference on Machine learning, ACM.

Di Buccio Emanuele, Federico Neresini and Andrea Lorenzet (2014), Scienza e tecnologia nei media italiani: tendenze generali e dieci temi ricorrenti, in Bucchi M. e Saracino B., Annuario Scienza e Società 2014, Bologna, Il Mulino.

Giardullo P. e Lorenzet A. (2013), La ricerca emergente nei media: nanotecnologie, neuroscienze, biologia sintetica e proteomica, in Neresini F. e Lorenzet A., a cura di, Annuario Scienza e Società 2013, Bologna, Il Mulino.

Lorenzet A. e Neresini F. (2012), 'La scienza nei media italiani: tendenze e temi emergenti'. In: Bucchi M. e Pellegrini G. (a cura di.), Annuario Scienza e Società 2011, p. 39-53, Bologna, Il Mulino.

Lorenzet A. e Neresini F. (2011), 'Il dibattito sull'energia nei media italiani', in Neresini F. e Pellegrini G. (a cura di.), Annuario Scienza e Società 2012, Bologna, Il Mulino.