DEEPEST IN THE SEA: FALSE INFORMATION ABOUT CAM IN ITALIAN WEB CONVERSATIONS

Carlo Bianchini[1], Ivana Truccolo[2], Ettore Bidoli[2], Mauro Mazzocut[2]

[1] University of Pavia, Department of Musicology and Cultural Heritage, Cremona, Italy
[2] CRO Aviano National Cancer Institute, Scientific and Patients library, Aviano, Italy

Corresponding author: Carlo Bianchini, carlo.bianchini@unipv.it

Introduction

Internet is frequently used by patients to google and to exchange a relevant amount of medical information in oncology. A undefined part of the available online information is false, and a major function of the medical library is to inform patients about risks of misinformation in the internet. Patients often have low health literacy and related skills and they have a critical need to be helped by health librarians (1–3).
Libraries and scholars have developed tools to evaluate web sites, but there are not researches that deal with deep content analysis. In particular, there are many different meanings for the term "false": it can refer to the presence of wrong words or grammatical or syntactical errors, to the absence of evidences or support to the information or to the statement of scientifically unproven facts (4–6). When any kind of the previous falsehood is intentional, it is also fraudulent.

Semantic value of the term "false"

Linguistic approach	Methodologic approach	Scientific approach	Intentional approach
error	Unfounded	Untrue	Fraudulent

As intention can be established only after a comprehensive analysis of the website from the point of view of its layout, scope, aims and whole content, it will be treated in future studies.
So, to better understand the process of production and dissemination of false information on the web and to better help users in the identification of false information, this study aims to prove that a lot of different false information is present in CAM conversations, to qualify the kind of falsehood present in websites, to quantify falsehood within each website and to explore relations among kinds of falsehood and to the general content of the websites.

Methods

This project is one of the outcomes of the "How deep is the sea: web intelligence for patient education" presented at the EAHIL Conference 2014 in Rome (7). We analyzed a sample of 15 CAM websites taken from the previous study to verify if they contain examples of false information, which degree of falsehood (erroneousness, foundationless or scientific incorrectness) and in which proportion they are present within each conversation.
Many kinds of websites were included and categorized based on their approach: general websites, blogs, e-journals, forums etc.
For each topic, two or three websites were chosen, for a total amount of 15 websites. Among these, 8 are health related websites (3 CAM, 2 Health and fitness, and 3 Oncology), while 7 are not heath related (3 online newspaper, 2 pseudoscience and conspiracy, and 2 debunking. Lastly, an authoritative Italian webpage was added to be valued and used as benchmark (8). The complete list will be published as Annexure to the paper.

Document analysis

First of all, a macro analysis of each website was conducted, recording data such as author of the text (if any), publication date, last access date, URL address, etc., mainly to highlight eventual presence of evident false information with relation to its various falsehood.
After this broad exam, websites needed to be comprehensively scored; HONcode grid was manually applied to rank sites on the basis of their supposed reliability (9) and with respect to HONcode characteristics (authoritative, complementarity, privacy, attribution, justifiability, transparency, financial disclosure, and advertising policy).
Only 3 webites out of 15 (20%) presented more than half of the requested characteristics; further only two websites have the HONcode seal, but one of these didn't obtain a full score, because it didn't include the complementary disclaimer, the links to the sources, and the advertising policy required by HONcode.

Text analysis

To deepen analysis and to get more data to establish scientific quality of websites, each exposition and each conversation was subdivided in 1355 items. Items were defined as a sentence going from dot to dot, but phrases with more than two subordinates were separated in a new item, and items with more than one scientific content were separated to contain just one.
Each item was examined in relation to its possible falsehood. For this purpose, three classes of falsehood were used: 1) erroneous (e.g. lexical, grammatical or syntactical mistaken); 2) unfounded (not based on an evidence), 3) scientifically incorrectness.
To grant uniformity in the analysis, a flowchart was discussed and used to identify the class to which assign each item.
Analysis allowed to identify immediately erroneous, unfounded or not valuable items, while items with relevant scientific content needed to be evaluated by expert scholars. An error occurred in this passage and 24 item on 1355 (1,77%) were not sent to experts for the analysis.
A team of experts chosen for their with different disciplinary profiles with no potential conflicts of interest declared with the content of the items (pharmacist, psychologist, oncologist, nutritionist, biomedical researcher) was created; a layman was asked to evaluate items too.
For each item, results from the first analysis and from experts' answers were coded into one table and values were represented by a code, to be statistically treated.

Results

Item analysis allowed to determine a general overview on falsehood in the text of the analyzed websites. As linguistic correctness, items are distributed as follows: 89,8% correct, 6,1% lexically or grammatically incorrect, 4,1% syntactically incorrect (Figure 1).

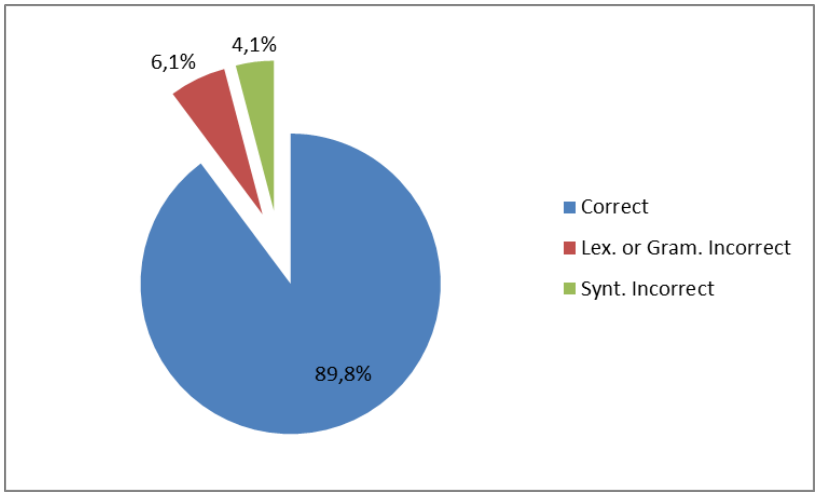

Figure 1 – Linguistic errors

The total amount of incorrectness is 10,2% and it could be relatively high. In fact, the larger part of mistakes are due to automatic and very poor translations from other languages (mainly from English to Italian).
A very interesting information comes from the analysis of attribution of items: apart from a relevant amount of not valuable items (35,65%), only 11,88% of items have references to bibliographic resources or valid links. More than the half of items with a semantic content doesn't show any references (Figure 2).

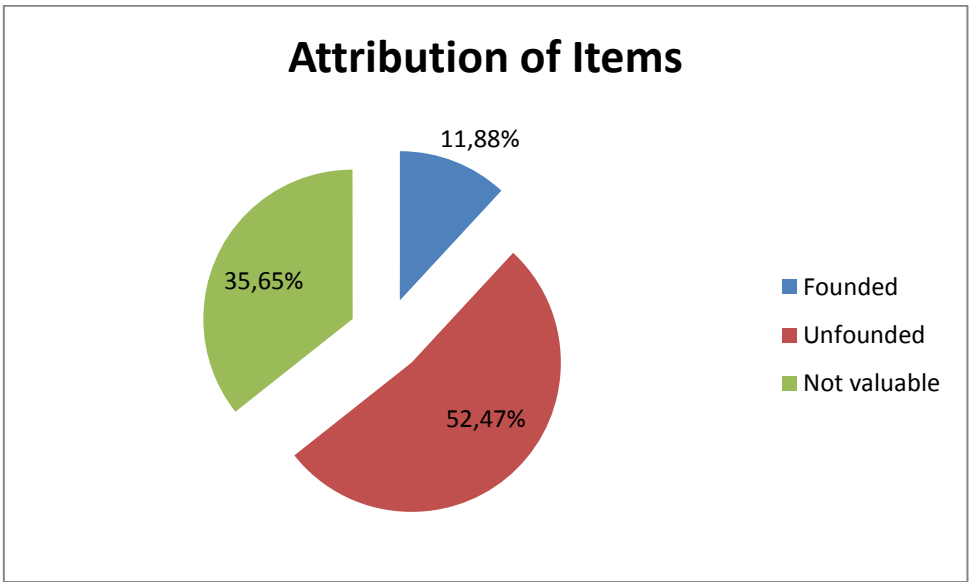

Figure 2 – Attribution of items

The synthesis of evaluation of items allowed to find out general and scientific reliability of websites information. The largest amount of items was not valuable (50,55%); 22,3% of the items were context (correct or incorrect), while scientifically evaluated items were distributed as follows: scientifically correct 7,97%; scientifically incorrect 5,39%; scientifically controversial 11,14%; and probably scientifically incorrect 2,66% (Figure 3).

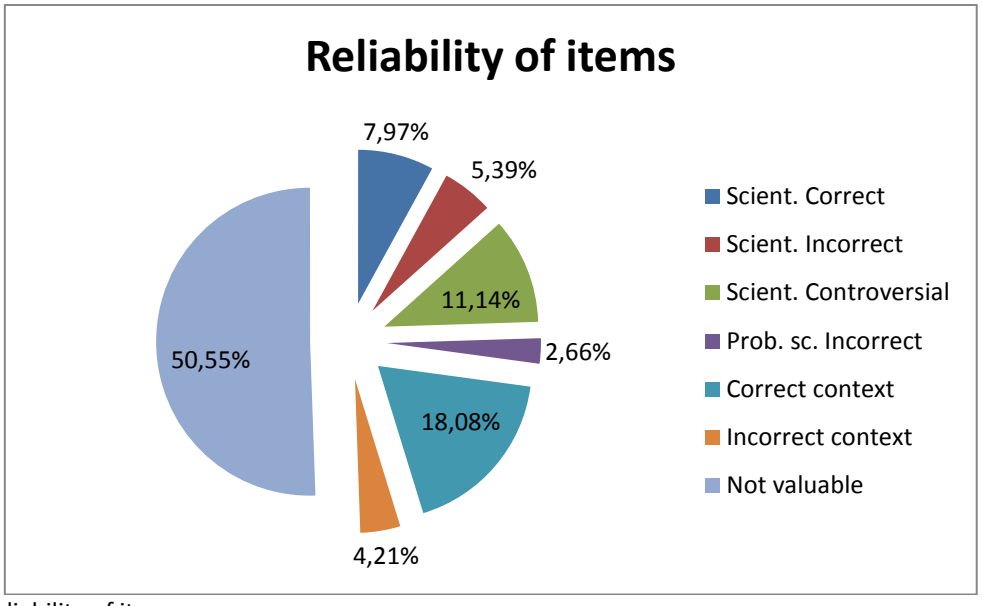

Figure 3 – Reliability of items

Discussion

In a second step, with the help of gathered data, we tried to define some indicators to express the value of quality of information contained in the websites. This value could be used to compare analysis results with other evaluation systems, such as HONcode, applied to the same websites, and with expected range of quality resulting from our macro website analysis.
On the basis of these eight indicators, medians were calculated. Websites values were confronted with the medians to rank their informational quality. 1 point was assigned to each website for each value higher than the median of a positive indicator (or 1 point for each value lower that the median of a negative indicator); otherwise 0 point.
Ranking of websites on the basis of the score per document obtained summing median values of the indicator is not satisfactory, as a document with 100% of true content but without scientific (correct or incorrect) content is ranked 6th, before than a document with a 56% of scientific correct content. Further, also a few websites with a relatively high scientifically correct percent are ranked lower (benchmark included).
So, a different ranking indicator had to be individuated.
Ordering websites by the different indicators shows that some indicators are more relevant and more effective than others to rank websites based on their information quality.
The most important indicator to establish information quality proved to be the percentage of scientifically correct items (ScC) on the total of scientifically examined items.
Websites sorted by scientifically correct indicator are ranked in a way that reflects expectations of information quality, but not completely. In fact, this ranking assigns a good position also to documents that have a relevant percentage of scientifically correct content and a relevant percentage of scientifically incorrect content. This is an issue, because often readers are deceived quite by mixing true and false sentences.
Documents with only scientifically correct content are better than documents with mixed scientifically correct and incorrect contents. So, the ideal ranking indicator should include values coming from both scientifically correct (ScC) and scientifically incorrect (ScI) contents, where the former is as higher as possible and the second as lower as possible; i.e., documents should be ranked by indicator

$$SQ = ScC/100*(1-ScI/100).$$

Websites ordered by SQ (scientific quality) from higher to lower values, rank as follows (Table 1):

Id. Doc	Title	Informative density (%)	True content (%)	Formal errors (No.)	Unfounded (%)	Scientifically Controversial (%)	Probably Scientifically Incorrect (%)	Scientifically Correct (%)	Scientifically Incorrect (%)	Doc Score	SQ
20111226	Tumore e rimedi natural	56,1	100	0	81,8	0	0	100	0	8	100,00%
20140220	Piante e cancro: come n	64,2	73,5	2	89,1	21,1	0	73,7	0	7	73,70%
20140221	Spiritualità nella cura de	73,5	86	6	55,3	27,8	0	72,2	0	7	72,20%
201312031	10 miti persistenti sul ca	66,7	81	7	48,1	37	0	55,6	3,7	6	53,54%
20140308	Dieta vegana, dieta alca	61	63,9	0	76,9	39,3	3,6	53,6	3,6	6	51,67%
20131231	Il cancro si cura con l'est	46,4	53,8	0	31,6	44,4	0	33,3	11,1	4	29,60%
20130129	Il cancro si cura con il bi	42,1	31,3	1	96	62,5	0	31,3	6,3	3	29,33%
201312032	Bufala! La guanabana cu	83,6	66,7	1	21,3	70,8	0	29,2	0	7	29,20%
20130211	Cancro al seno rimedi na	35,7	15	37	100	68,4	10,5	15,8	5,3	0	14,96%
20140515	I veri metodi di cura con	60,4	51,2	8	99,1	7,4	11,1	22,2	55,6	2	9,86%
20140320	L'inganno melatonina	66,7	21,9	2	89,7	52,4	4,8	9,5	14,3	1	8,14%
20140520	Il Ganoderma Lucidum r	51,5	16,3	40	98,3	46,3	25,9	7,4	14,8	0	6,30%
20131002	Ecco i prodotti, chi mi ai	35	11,3	25	97,7	10,3	2,6	2,6	82,1	1	0,47%
201303052	Funghi: alleati naturali r	33,3	100	0	55,6	0	0	0	0	6	0,00%
201303051	Cure alternative: "Il can	3,6	75	1	100	0	0	0	0	5	0,00%
20140126	Scappa di casa per evita	22,2	50	8	91,2	0	0	0	100	2	0,00%

Table 1 – Ranking of websites on the basis of SQ (Scientific Quality)

Values of SQ allows to divide websites in two sets: if SQ > 1, websites have a scientific relevance; if SQ < 1, websites have no scientific relevance. It must be noted that in the lowest positions, two websites 201303052 and 201303051 have a high percentage of TC (True content) and a good ranking as to the whole set of indicators (respectively 6 and 5), seemingly being valuable websites. Applying SQ instead, they correctly rank in the last positions.

Conclusion

SQ is a valid indicator to evaluate scientific quality of web information. In fact, sometimes formal characteristics are insufficient to assess the low quality of websites information. For instance, the absence of references does not mean necessarily a low quality content, and formal absence of mistaken doesn't necessarily relate to scientific quality.
Further, Informative DEensity (IDE), True Content (TC), and Unfounded items (UNF) cannot be used, singularly or together, as indicators to foresee quality of websites content.
This work should be continued along different search lines; the first one should be the analysis of a larger number of samples; a second one should be to investigate possible relationship of SQ and other online information evaluation methods; a third one should analyze all the items of the selected websites in order to identify possible logical fallacies and to explore relationships among fallacies and other measures (for example, IDE or TC) from a logical, not linguistic, point of view (10).

References

1. Eysenbach G, Powell J, Kuss O, Sa E-R. Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web. JAMA [Internet]. American Medical Association; 2002 May 22 [cited 2016 May 2];287(20):2691. Available from: http://jama.jamanetwork.com/article.aspx?articleid=194953
2. Powell JA, Lowe P, Griffiths FE, Thorogood M. A critical analysis of the literature on the Internet and consumer health information. J Telemed Telecare [Internet]. SAGE Publications; 2005 Jan 1 [cited 2016 May 2];11 Suppl 1(suppl 1):41–3. Available from: http://jtt.sagepub.com/content/11/suppl_1/41.abstract
3. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res [Internet]. Journal of Medical Internet Research; 2015 Jan 7 [cited 2016 Mar 1];17(5):e112. Available from: http://www.jmir.org/2015/5/e112/
4. Bianchini C. Il falso in Internet: autorevolezza del Web, information literacy e futuro della professione (Prima parte). AIB Stud [Internet]. 2014 Apr 2 [cited 2014 Jun 30];54(1):61–74. Available from: http://aibstudi.aib.it/article/view/9957
5. Bianchini C. Il falso in Internet: autorevolezza del Web, information literacy e futuro della professione (Seconda parte). AIB Stud [Internet]. 2014 Jun 16 [cited 2014 Nov 23];54(2/3):227–40. Available from: http://aibstudi.aib.it/article/view/10130
6. Bianchini C. Il falso in rete: il bibliotecario come antidoto. Convegno 'Bibliotecari al tempo di Google Profili, competenze, formazione' Relazioni. Milano: Editrice Bibliografica; 2016. p. 146–61.
7. Mazzocut M, Antonini M, Truccolo I, Omero P, Ferrarin E, Gandelli R, et al. How deep is the sea: web intelligence for patient education. 14th EAHIL 2014 Conference, Rome 11-13 June 2014. Rome: EAHIL; 2014. p. 51.
8. AIRC. AIRC. Associazione Italiana per la Ricerca sul Cancro [Internet]. 2016 [cited 2016 May 31]. Available from: http://www.airc.it/
9. Health On the Net Foundation. HONcode Site Evaluation Form [Internet]. 2014. Available from: https://www.hon.ch/cgi-bin/HONcode/Inscription/site_evaluation.pl?language=en&userCategory=providers
10. Hauch V, Blandón-Gitlin I, Masip J, Sporer SL. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. Pers Soc Psychol Rev [Internet]. 2015 Nov 1 [cited 2016 Apr 2];19(4):307–42. Available from: http://psr.sagepub.com/content/19/4/307