Feasibility of automated detection of HONcode conformity for health-related websites

Célia Boyer, Ljiljana Dolamic
Health On the Net Foundation
Geneva, Switzerland

***Abstract*—In this paper, authors evaluate machine learning algorithms to detect the trustworthiness of a website according to HONcode criteria of conduct (detailed in paper). To derive a baseline, we evaluated a Naive Bayes algorithm, using single words as features. We compared the baseline algorithm's performance to that of the same algorithm employing different feature types, and to the SVM algorithm. The results demonstrate that the most basic configuration (Naive Bayes, single word) could produce a 0.94 precision for "easy" HON criteria such as "Date". Conversely, for more difficult HON criteria "Justifiability", we obtained precision of 0.68 by adjusting the system parameters such as algorithm (SVM) and feature types (W2).**

Keywords—internet content quality; health; machine learning

I. INTRODUCTION

Should I trust the medical information on an internet-based web page? Has it been written by a medical professional? Is the information up to date? Web users should ask these questions when accessing online health information. The Health on the Net Foundation (HON) addresses these issues by way of its certification program, the HON Code of Conduct (HONcode). The HONcode, launched in 1995 (HONcode,[9]) is a set of ethical principles defined by the HON Foundation with a consensus of health information editors in order to assess the quality of on-line health information. Currently, it is the oldest, most renowned and the most utilized quality code for online health information, with more than 8'300 certified health websites worldwide. Since 1996 HON has been working with Internet health editors as well as patients to guide web users to trustworthy health information.

The HONcode certification involves manual reviews that examine the processes used in creating and maintaining health-related websites. HON does not validate a site's content, per se.

The HONcode criteria used for certification are (see details at http://www.healthonnet.org/HONcode/Conduct.html):

- Authoritativness
- Complementarity
- Privacy protection
- Attribution
- Justifiability
- Transparency
- Financial disclosure
- Advertising policy

Given the overwhelming quantity of medical information currently available, internet users, such as patients, have difficulty finding information they judge to be trustworthy [1], [2]. The goal of HONcode certification is to enable patients to more easily distinguish websites that adhere to quality standards just by looking for the HONcode quality seal. Websites displaying this seal conform to the quality guidelines of the HONcode. Due to the ever-increasing number of health websites, reviews for HONcode certification now only occur when a site's webmaster requests such a review. This means that absence of the HONcode seal does not absolutely preclude a site from meeting HONcode criteria.

About 80% of searches carried out for health information start from a general search engine [3] that admixes quality health information with manipulated, biased or misleading health content in its displayed results [4]. Up to 35% of US adults used the Internet and did not visit a clinician to get a professional opinion [3] and only 41% of online diagnoses say a medical professional confirmed their online diagnosis. The ability to find credible information has a direct impact on the health of the public ([5]; [6]). Trust and quality are concepts that are difficult to define; [7] gives a rather exhaustive outline of what is trust on the Internet and within the context of the Semantic Web. Over 19 different factors have been listed as components of trust for websites [8].

Our goal in this study was to evaluate the feasibility of automatic detection of the eight HONcode quality principles. This is the goal of HON's research conducted within the European project KHRESMOI (2010-2014, project No. 257528). KHRESMOI helps to guide the general public to reach health websites which are HON certified or are otherwise selected based upon an automated system identifying the principles disclosed in a site.

The rest of this document is organised as follows; next section gives the insight to related work, the methods used in the experiments are described in the section III. The results are given in the section IV. Finally, Section V brings the conclusion and future work guidelines are given in the Section VI.

II. RELATED WORK

Most previous studies performed related to trust on the internet have focused on the e-commence domain for which certain criteria have been determined [16]. Despite the volume of research available in this domain, there remains a lack of basic consensus about the meaning of trust. The same problem exists in the domain of online health sites. To our knowledge, most studies of health-related trust have only tested individual specific points defined by a project [17][18]. The lack of a de facto comprehensive standard for trust of online health information precludes comparison among different studies.

The HONcode standard of conduct, establishes a process-based standard for trust of health information. This manual process requires large amounts of time and human resources. For this reason, automation of this process has become an important issue. In this effort a study has been conducted at HON in this purpose [10]. In this study, the extracts were separated into sentences which were then used as documents. As a result, the incorrect class attribution has occurred during the collection creation, since not each sentence conforms to a criterion if the document as whole does.

III. Methods

A. Collection acquisition

While HONcode certification process is performed for websites written in multiple languages. However the studies performed here were limited to those written in the English language only, since the data for this language is considered to be the most complete one. Table I gives the number of extracts per criteria available for this language

Unlike the previous studies conducted by HON in the domain of automatic detection of the HONcode principles [10], in this research we used the whole document as the classification unit. Indeed the statement about a certain criterion is spread within the whole document, and not concentrated into a single sentence, making the document a more suitable classification unit than sentences.

TABLE I. Number of Extracts Per Honcode Criteria

Criteria	Number of extracts
Authority	2812
Complementarity	2835
Privacy	2683
Reference	2349
Justifiability	872
Transparency	2861
Financial disclosure	2700
Advertising policy	1412
Date	2794

The authors created a database containing positive and negative examples of compliance with the eight HONcode principles. Human experts were asked to extract the text demonstrating whether a given website conformed to each HONcode criterion. Establishing negative examples was not possible, per se. What would comprise a negative example of privacy policy other than its mere absence? Therefore, the documents supporting other HON criteria were used as negatives for the criteria other than their targets. We divided the principle Attribution into two criteria: Reference and Date due to different requirements for these two elements. Each extract obtained in this manner represents one document within the training/test collection.

Documents could conform to support more than one criterion, so we classified the text into categories that were not mutually exclusive (any-of *classification).* If a document conformed to one criterion, it did not imply that it conformed (or did not conform) to any other. We took an approach described in [11]:

- Build a classifier for each class, where the training set consists of the set of documents in the class (positive labels) and its complement (negative labels).
- Given the test document, apply each classifier separately. The decision of one classifier has no influence on the decisions of the other classifiers.

We ran all the experiments using 10-fold cross-validation.

B. Machine learning algorithms

We used the machine learning algorithms described in the learning framework [12]: Naive Bayes (NB) and C_SVC Support Vector Machines with radial kernel (named in this document SVM).

We used the C_SVC SVM as it is the less time consuming compared to other SVMs. We applied the NB and the SVM using various features types, using different feature reductions levels.

C. Features

We pre-treated the documents linguistically, removing stop words and applying Porter stemming. The learning unit was then set to one of the following, to best identify which features suited the principle being automatically extracted:

- single word (W1) - “privacy information” → “privat”, “inform”
- two conjunct words (W2) - “privacy information” → “privat_inform”
- word co-occurrence (COOC) - “privacy information” or “information about privacy”→ “privat_inform”

We chose the single word W1 (bag of words) for the baseline of this study. W2 and COOC differ in that W2 takes into account the word order, while COOC does not.

D. Feature selection

Feature selection has two goals, first reducing the dimension of the document representation and second distinguishing features that help to determine a document’s class. The latter reduced over fitting, making sure that the model is not too general.

We used document frequency for feature selection, as it is simple and effective. It uses only the features whose document frequency exceeds a predefined threshold (after stop word removal).

We used various levels of features reduction, keeping 30%, 50% or 80% of features with a goal of demonstrating how the reduction of the number of features would influence classifier performance.

TABLE II. PRECISION, RECALL AND F1-MEASURE AVERAGE OF 10 RUNS FOR AUTHORITY, COMPLEMENTARITY, PRIVACY PRINCIPLES

Parameters			Criteria								
Algorithm (Alg.)	*Feature Type (FT)*	*% FT Kept*	*Authority*			*Complementarity*			*Privacy*		
			Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*
NB	W1	80	0.50	0.85	0.63	0.84	0.95	0.89	0.69	0.98	0.81
NB	W1	50	0.50	0.86	0.63	0.83	0.95	0.89	0.69	0.98	0.81
NB	W1	30	0.50	0.87	0.63	0.83	0.95	0.89	0.70	0.98	0.82
NB	W2	30	0.52	**0.88**	0.65	0.84	0.96	0.90	0.85	0.98	0.91
NB	COOC	30	0.51	0.84	0.63	0.78	**0.97**	0.86	0.80	**0.99**	0.88
SVM	W1	30	0.70	0.64	0.67	0.89	0.91	0.90	0.96	0.97	0.97
SVM	COOC	30	0.66	0.56	0.60	0.89	0.89	0.89	0.97	0.94	0.95
SVM	W2	30	**0.73**	0.69	**0.71**	**0.92**	0.91	**0.92**	**0.97**	0.97	**0.97**

TABLE III. PRECISION, RECALL AND F1-MEASURE AVERAGE OF 10 RUNS FOR REFERENCES, JUSTIFIABILITY, TRANSPARENCY PRINCIPLES

Parameters			Criteria								
Algorithm (Alg.)	*Feature Type (FT)*	*% FT Kept*	*References*			*Justifiability*			*Transparency*		
			Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*
NB	W1	80	0.43	0.77	0.55	0.49	0.50	0.50	0.86	0.95	0.90
NB	W1	50	0.41	0.79	0.54	0.45	0.59	0.51	0.85	0.96	0.90
NB	W1	30	0.40	**0.81**	0.54	0.40	**0.65**	0.50	0.85	**0.97**	0.90
NB	W2	30	0.43	0.81	0.56	**0.69**	0.58	**0.63**	0.90	0.97	0.93
NB	COOC	30	0.41	0.73	0.53	0.39	0.53	0.45	0.73	0.90	0.81
SVM	W1	30	0.61	0.60	0.61	0.52	0.53	0.52	0.94	0.95	0.95
SVM	COOC	30	0.51	0.41	0.46	0.41	0.33	0.36	0.85	0.81	0.83
SVM	W2	30	**0.65**	0.64	**0.64**	0.61	0.56	0.58	**0.95**	0.96	**0.96**

TABLE IV. PRECISION, RECALL AND F1-MEASURE AVERAGE OF 10 RUNS FOR FINANCIAL, ADVERTISING, DATE PRINCIPLES

Parameters			Criteria								
Algorithm (Alg.)	*Feature Type (FT)*	*% FT Kept*	*Financial*			*Advertising*			*Date*		
			P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
NB	W1	80	0.57	0.92	0.71	0.60	0.85	0.70	0.94	0.94	0.94
NB	W1	50	0.56	0.94	0.70	0.54	0.91	0.68	0.94	0.95	0.94
NB	W1	30	0.56	**0.94**	0.70	0.51	**0.94**	0.66	0.94	0.95	0.95
NB	W2	30	0.57	0.92	0.71	0.66	0.87	0.75	0.95	0.97	0.96
NB	COOC	30	0.54	0.90	0.67	0.50	0.83	0.63	0.79	0.89	0.84
SVM	W1	30	0.79	0.79	0.79	0.74	0.80	0.77	0.95	0.95	0.95
SVM	COOC	30	0.78	0.69	0.73	0.65	0.72	0.68	0.91	0.81	0.86

SVM	W2	30	**0.84**	0.81	**0.82**	**0.77**	0.81	**0.79**	**0.96**	**0.97**	**0.96**

IV. Results

Here are the examples of the extracts from certified websites taken by the experts for two criteria (transformed into lower case):

- Complementarity: “the information that we provide on our web site is designed to support, not replace, the relationship that exists between a patient/site visitor and his/her physician. Please keep in mind that the text provided is for informational purposes only and is not a substitute for professional medical advice, examination diagnosis or treatment. Always seek the advice of your physician or other qualified health professional before starting any new treatment or making any changes to existing treatment.”
- Privacy: “privacy policy this web site does not collect information from any visitor. Cookies are not used at any time. we do not collect email addresses and any communication will not result in the retention of your information in any form. We do not keep a database of visitor information or any other statistics regarding the demographics or other attributes of users of this web site. There are opportunities to become a patient of the norman endocrine surgery clinic, however, this occurs on two specific pages designed for this purpose. These two pages are hosted on a secure server. You will know that you are entering your data and warnings will be given. This is a clear decision that you will make. These two pages are encrypted and secure and are clearly marked as such. These two pages have the logos and secure certificates clearly visible. The information entered on these two secure pages is not accessible to anybody except the medical staff of the norman endocrine surgery clinic. These two secure pages have been approved and meet all current 2004 standards for secure online medical information as established by the american medical association.”

Tables II, III and IV presented in this section give the results obtained for different combinations of parameters (algorithm (Alg.), feature type(FT), percentage of features kept(Kept%)) described above.

We expressed the results using standard measurements: precision (P), recall(R) and F-measure(F). Precision represents the fraction of all documents assigned to given class by the classifier that really belong to that class, while recall represents the fraction of all documents that belong to the given class according to the test set that were correctly assigned by the classifier. The F-measure is the harmonic mean of P and R.

In Tables II, III and IV the best performance in precision recall and F-measure is given in bold.

The cells in grey show a precision up to 73 % for an F1 measure of up to 71% for all the criteria except for the criteria “References” and “Justifiability” where respectively the precision is 65% and 69% for an F1-measure of 64% and 63% respectively. The automatic classification for the Transparency, Complementarity, Privacy, and Date criteria show a precision over 92% with a good recall over 91%. Authority, Funding and Advertising are above of 73% of precision and a recall up to 69%

A. Effect of the algorithm used

Even though the performance difference between these two algorithms in the study can go up to 53% (“References”, W1) higher precision using SVM rather than NB, for certain criteria (Complementarity, Transparency, Date) this difference never exceeds 15%.

B. Effect of feature type

As it can be noticed from the tables, the choice of the feature type impacted on the classifier performance. In certain cases, as for example the criteria Justifiability, changing the feature type from W1 to W2 with the same feature reduction level leads to relative increase in precision by more than 50%. The W2 and COOC seem to be performing similarly in the most of the cases, although for certain criteria such as Transparency or Justifiability the precision with COOC is much smaller. This is probably the result of the COOC unlike W2 takes no context into account.

C. Effect of reducing feature space

The size of the term space can be significantly reduced without significant loss in the performance of the text classifier. This is the main goal of the features selection process. We chose Naive Bayes, with W1 setting to illustrate the impact of this parameter on the classification results.]

The first three lines in Tables 2, 3 and 4 indicate that difference in the precision is rather small between the 80%, 50% and 30% features kept. The relative decrease in precision of 18,7% between the 80% features kept (0.49, taken as a baseline) and 30% (0.40) features kept for the criteria “Justifiability” given in the Table 3 is the only difference that can be seen as important . On the other hand, it is noticeable for this criterion that the recall in the case of 80% features kept is only at 50%, making the classification rather random.

V. Discussion / Conclusion

In our previous studies conducted by HON in the domain of automatic detection of the HONcode principles [10] we conducted a preliminary feasibility study of the design and the evaluation of an automatic system conceived for the categorization of medical and health documents according to the HONcode ethical principles. Based on our first promising results, the research activities presented in this paper is a prospective validation study where the classifiers has been tuned, features has been evaluated, the corpus has changed and the classifier examined previously unexplored websites.

In the previous study the sentences was used as the unit for the classification. Here we use the document as a unit for classification in order to avoid false positive in the collection creation since each single sentence is not necessary conforming to the criteria if the document itself is.

Based on the article [20] the agreement of the manual classification between two persons rarely gives more than 72% of precision. The automatic classification globally largely outperforms what we can expect from a manual classification.

We used Document Frequency as an algorithm for features selection. It has been shown that the feature space can be reduced by 70% without an important loss in performance.

The results obtained also show, that for a certain number of criteria, we can denominate these criteria as "easy" and even the most "simple" parameter setup returns very good results in precision. We can take the criteria Date, Complementarity or Transparency as examples of "easy" criteria. For the NB, W1, 30% parameter combination precision obtained for these criteria are 0.94, 0.83 and 0.85 respectfully.

Ever since the first work carried out to apply the SVM on text categorization [12], [13] it has proven to be more effective than other baselines. The SVM used in our study gives higher results in terms of precision than the Naive Bayes algorithm for most HONcode criteria, with exception of the "Justifiability" were NB seems to perform better in terms of precision.

However if we take into account the time/resource consumption difference between the two algorithms presented in this paper, this proves that, even though being less efficient, the usage of the NB is still justifiable.

It has also been shown that even with a simple feature selection such as Document Frequency and different learning algorithms and features types it is possible to achieve precision of more than 70%, with the exception of the References and Justifiability criteria.

VI. Future Work

The next steps in our study is to implement and investigate other feature selection algorithms, feature weighting (such as tf-idf [15]) and feature types (ex. character or word n-gram) in order to achieve over 85% of precision for all the HONcode principles including those estimated to be "hard "to automatically retrieved in this study.

We will need to face the automatic system to the manual one. So the next work will be to then have manual HONcode experts to review the results of the automatic classifier to determine accuracy and precision in practice.

Another important problem concerning health related information on the web is the presence of fake websites addressed in [19]. Current concept of the HONcode doesn't deal directly with the quality of the website content. The question that one might ask is: "What if a website complies to all HONcode principles but with a content of a very bad quality?." We will explore two techniques, based on external knowledge, to enhance the performance of the evaluation algorithm.

A. Accessing scientific information

An alternative or a complementary element to automatic detection of the references in health website could be the usage of the MEDLINE database (via pubmed) for identifying the outcome for a given field and compare it with the content of the given page.

B. Incorporating an algorithm for the automatic detection of fake websites

Some websites copy contents from other website with no added value; their only purpose is to attract advertising. Algorithms have been developed to detect these sites.

We will design a way to incorporate them into our detection algorithm and test if they bring some performance improvement.

Finally the automatic classifier should be implemented and integrated into real system to assess the usage by HONcode reviewers or by Internet searchers.

References

[1] Wong LM, Yan H, and al. 'urologists in cyberspace: A review of the quality of health information from american urologists' websites using three validated tools. Can Urol Assoc J., 7(3-4):100–107, 2013.

[2] Carrion, Fernandez, and Toval. Are personal health records safe? a review of free web- accessible personal health record privacy policies. Jmir, 14(4), 2012.

[3] Pew Internet & American Life Project - 2013 report

[4] P. López-Jornet, F. Camacho-Alonso The quality of Internet sites providing information relating to oral cancer, Oral Oncology, 2009

[5] van Straten, A., Cuijpers, P., & Smits, N. (2008). Effectiveness of a Web-Based Self-Help Intervention for Symptoms of Depression, Anxiety, and Stress: Randomized Controlled Trial. Journal of Medical Internet Research, 10(1), e7

Humphrey, T. (2009, December 23). Internet Users Now Spending an Average of 13 Hours a Week Online at http://news.harrisinteractive.com/profiles/investor/ResLibraryView.asp?BzID=1963; accessed on January 8, 2012.

[6] Artz D., Gil Y (2007) A survey of trust in computer science and the Semantic Web. Web Semantics: Science, Services and Agents on the World, 2007.

[7] Gil Y., Artz D.: Towards content trust of web resources, Web Semantics: Science, Services and Agents on the World Wide WebVolume 5, Issue 4, World Wide Web Conference 2006 Semantic Web Track, December 2007, Pages 227-239.

[8] Boyer C., Baujard V., Scherrer J (1999).: HONcode: a standard to improve the quality of medical/health information on the internet and HON's 5th survey on the use of internet for medical and health purposes. In 6th Internet World Congress for Biomedical Sciences (INABIS 2000), 1999.

[9] Arnaud Gaudinat, Natalia Grabar, and Celia Boyer. Machine learning approach for automatic quality criteria detection of health web pages. In Klaus A. Kuhn, James R. Warren, and Tze-Yun Leong, editors, MedInfo, volume 129 of Studies in Health Technology and Informatics, pages 705–709. IOS Press, 2007.

[10] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. Introduction to in- formation retrieval. Cambridge University Press, 2008.

[11] K. Williams and RA. Calvo. A framework for document categorization. 7th Australian document computing symposium, 2002.

[12] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. 1998. Inductive learning algorithms and representations for text categorization. In Proceedings of the seventh international conference on Information and knowledge management (CIKM '98), K. Makki and L. Bouganim (Eds.). ACM, New York, NY, USA, 148-155. DOI=10.1145/288627.288651 http://doi.acm.org/10.1145/288627.288651

[13] Thorsten Joachims. 1999. Making large-scale support vector machine learning practical. InAdvances in kernel methods, Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola (Eds.). MIT Press, Cambridge, MA, USA 169-184.

[14] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. 1999. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc.,

Boston, MA, USA.[16] Ardion Beldad, Menno de Jong, and Michaël F.How shall I trust the faceless and the intangible? A literature review on the antecedents of online trust. Steehouder. Computers in Human Behavior 26(5):857-869 (2010)

[15] Gunther Eysenbach, John Powell, Oliver Kuss, and Eun-Ryoung Sa. Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web -- A Systematic ReviewJournal of the American Medical Association JAMA 287(20):2691—2700 (May 2002)

[16] Sillence, E., Briggs, P., Harris, P.R. and Fishwick, L. (2007) How do patients evaluate and make use of online health information? Social Science and Medicine, 64 (9). pp. 1853-1862. ISSN 0277-9536

[17] Ahmed Abbasi, Fatemeh "Mariam" Zahedi, and Siddharth Kaza.Detecting Fake Medical Websites using Recursive Trust Labeling, ACM Trans. Inf. Syst. 30(4):22 (2012)

[18] Andrew Mccallum, 1999, Text Classification by bootstrapping with keywords, EM and shrinkage. ACL99 - Workshop for Unsupervised Learning in Natural Language Processing