N-gram as an alternative to stemming in the automated HONcode detection for English and French

C. Boyer[1] L. Dolamic[1]

[1]Heath On the Net Foundation

Chemin du Petit-Bel-Air 2, 1225 Chêne-Bourg, Switzerland
celia.boyer@healthonnet.org

Abstract

The HONcode of conduct is composed of eight ethical and quality criteria (www.hon.ch/Conduct.html) This study evaluates supervised automatic classification algorithms capability to determinate whether a health related web page is in compliance with any of those criteria. Various length character n-gram vectors were used to represent health web page documents. Classification performance of the 5-grams was compared to that obtained by words or stems. The study attempts to determine whether the language-independent approach might result in similar classification performance as word-based classification for both English and French languages. The training/testing collection for both languages were created from web page fragments extracted by HONcode experts during the manual certification process as the basis for individual HONcode compliance. Naive Bayes classifier and DF (document frequency) dimensionality reduction metrics were used. The overall results of this study indicate that the n-gram tokenization provides a potentially viable alternative to document word stemming.

Keywords:

Supervised machine learning, Character n-gram, Naive Bayes, SVM, quality criteria, HONcode

1 Introduction

The health related information provided by Internet, to millions of people worldwide includes at the same time reliable, useful information and potentially harmful information [1]. Discerning trustworthy web-based health information from manipulated or biased content is difficult task for end-users [2], since no obligatory regulatory standards exist for ethical and quality content of health-related websites. The Health On the Net Foundation was created in 1996. Its mission is the promotion of transparency and quality for online medical and health-related information. The HON Foundation's Code of Conduct [3], consists of eight procedural guidelines that help to indicate the credibility of online health information. Currently, HONcode expert reviewers manually assess, re-assess, and certify health websites for compliance with the HONcode conduct principles.

The websites respecting the HONcode display the unique HONcode seal. In this manner the certification gives the possibility to the individual user to easily identify whether website respects the HONcode. The presence of such a seal indicates that the website provides the answers to questions such as: "when was the site last updated" or "who is the author of the content", however it does not validate the website's content per se. The detailed HONcode guidelines can be found at http://www.healthonnet.org/HONcode/Conduct.html.

HONcode is the most widely utilized healthcare website Code of Conduct. To date, HON has certified 8'300 health websites worldwide. HONcode certification is multilingual, since identification of quality web-based information in their native languages is crucial to end-users worldwide. The largest number of certified websites is written in English (34%), followed by French (28%) and Spanish (10%). However, the manual initial evaluation and subsequent audits in order to assess HONcode conformity are limited due to the time-and-personnel intensive processes they involve. Furthermore, many high-quality health websites may not have HONcode certification because the review process is voluntary – conducted at the request of a health site's webmaster. Using current methods, large-scale HON reviews for certification of hundreds of thousands of health websites is not feasible.

In this study we are evaluating the possibility to address the HONcode certification as well as its multilingual aspect automatically. In this study we are reporting to what extent the machine learning based automated classification system, in combination to language independent n-gram classification can be of assistance in the process of HONcode certification. This type of tokenization has proven to be effective for different, particularly for the morphologically complex languages [4], especially when linguistic tools such as stemmers [5] were lacking for the given language. The n-gram approach also showed greater robustness in the setting of frequent typographical errors in the source text [6]. This approach is tested on English and French document collections.

This research project is funded by the European Kconnect project (2015-2017, project No. 644753).

Related work

HONcode certification process must become at least partially automated in order to remain relevant, taking into account the quantitative growth of Internet-based health information websites. The research has already been initiated by HON in the direction of exploiting Machine Learning (ML) algorithms for automated text classification [7, 8]. To be used by a machine learning based classifier the content of healthcare document is represented as vectors of weighted tokens [9], since it cannot be directly interpreted by them. Word stems or lemmas is typical approaches for this purpose, with a goal of increasing the system's recall [10, 11]. Linguistic treatment of such stemming has proven to be a powerful tool especially with morphologically complex languages [12]. For the word stem or

lemma to be created the linguistic tools such as stemmers or morphological analyzers are required. Unfortunately, those tools are not always available. Not for all languages, nor for specific language subdomains, such as health. Thus, a language-independent approach to this issue is needed. Linguistic treatment in English does not imply enormous difference in performance, while for the French the performance differences are somewhat more important [12]. Showing the interchangeability between language dependent (stemming, lemmatization) and language independent approach for these two languages, would however give a credible baseline for other languages, for both simple or more complex from the morphological point of view. Employing character n-grams in this purpose is widely present in text classification [13]. The machine learning system represented each document as a vector of mutually independent weighted "terms" (tokens, features). A term itself could be a word, a lemma, or a stem, etc. The lemma or stems used as "terms" have for the goal matching of different words derived from the same root (e.g. derive, derived, derivation). However, accomplishing this goal requires language dependent tools. The current study explores the usability of language independent character n-gram opposing words and stems. Various approaches to creation of character n-grams exist. Some of them take into account the word boundaries where word "privacy" would yield following 5-grams "_priv", "priva", "rivac", "ivacy" and "vacy_" with "_" representing the word boundary. In other cases the word proximity is taken into account, in which case the phrase "privacy policy" would yield 5-grams such as "acy_p". In our approach we consider every word of the document independently and we do not take the word boundaries into consideration as described in [4]. Thus, for the case of n=5, the document phrase "privacy policy" would yield the following n-grams: "priva", "rivac", "ivacy", "polic" and "olicy".

When it comes to dealing with the issue whether information available on the internet can be trusted or not, studies conducted to date have mostly focused on the e-commerce domain. Nevertheless, despite the volume of research in this domain, basic consensus about the meaning of trust remains elusive [14]. The same lack of consensus applies regarding trustworthiness of online health information. Most reported studies have tested one specific aspect of trust as reported in the following articles [15, 16]. The lack of de facto standards regarding online health information makes comparisons among research studies difficult. In the scope of the CLEF initiative eHealth related workshops have been proposed (http://clefehealth2014.dcu.ie/). However, none of the tasks within these workshops focused on the particular issue of the quality of health information. Thus, no collection is available which would allow to have common data in order to appropriately compare results amongst research approaches.

2 Methods

The test system used in this study is based on the machine learning framework described in [17]. Different Machine Learning algorithms such as SVM, Naive Bayes (NB) or KNN have been implemented and tested. Experience in an earlier study [18] in addition to experimental results obtained for this work led authors to retain a Naive Bayes machine learning algorithm (with tf-idf weighting instead of word frequency, due to better precision), as the most appropriate for automated determination whether a webpage complied with each HONcode principles [19]. In addition, the comparison to SVM for certain criteria is performed in order to give support to such a decision. The document frequency algorithm was used to perform the feature selection in this study [10].

There are eight HONcode principles: *Authority, Complementarity, Privacy, Attribution, Justifiability, Authorship, Sponsorship* and *Advertising*. The principle *Attribution* was divided into two separate criteria, namely *Reference* and *Date* due to disparate fulfilment requirements for these two elements. Authors developed separate classifiers for each of the HONcode criteria, because a document's belonging to one HONcode compliance class is independent, under all circumstances, from its conformance with one or more other HONcode compliance classes (*any-of classification*) [10]. All experiments performed used ten-fold cross validation.

The collections used for training of the machine learning algorithms were created from actual webpage excerpts. As the HONcode team previously reviewed candidate websites for HONcode certification, they stored the part of the site's webpage that indicated compliance with one of the HONcode principles. Another challenge the authors faced with was defining the document (classification unit) within these collections. One recent study using machine learning techniques used the sentence as a classification unit [8]. Closer inspection of such an approach revealed that it generats numerous errors, as individual sentences in a document may not conform to criteria even though the document as a whole may [8]. Each document within training/test collection consists of one previously described extract from one website for one specific HONcode criterion. HONcode certification process includes websites in any of a large set of languages. However, the current study was limited to pages written in English and French languages. These languages have been chosen for two main reasons: 1. the collected sets of compliance justification extracts in English and French were the most exhaustive and 2. these languages represent two different languages types from the morphological complexity point of view. Table 1 gives the number of documents (extracts) available in these two collections. This collection is not for the moment publicly available.

Table 1: Number of extracts per criteria (English and French HONcode compliance extracts collections)

		No. extracts	
Criterion		**English**	**French**
Authority	HC1	2812	2338
Complementarity	HC2	2835	2005
Privacy policy	HC3	2683	2055
Reference (*Attribution*)	HC4	2349	1888
Justifiability	HC5	872	827
Contact details	HC6	2861	2349
Financial disclosure	HC7	2700	2098
Advertising policy	HC8	1412	627
Date (*Attribution*)	HC9	2794	2158

In the current study the words (W1) used as terms is taken as the "baseline". The results obtained by the baseline are compared to those of 5-grams (C5) and stems (W1s). For the English the porter stemmer was used, while for the French we use the snowball stemmer[1]. Various length of character n-grams have been tested in the scope of this evaluation. The 5-gram was retained since it has shown the behaviour closest to the behaviour of word or stems for both languages tested. The goal was to determine the extent to which words might be replaced by 5-grams or stem tokens while not sacrificing system classification performance.

In conducting the study, prior to the tokenization, authors removed stop words such as «le», «la». «du» etc.. from the stud-

1 http://snowball.tartarus.org/algorithms/french/stemmer.html

ied documents. These lists contain 174 words for English and 126 for French.

We have chosen the precision (P), recall (R) and F_1-measure to present the quality of the classification for each of the HONcode criteria. The F_α-measure [19] combines the precision and recall measures, allowing us to give relative importance to each of them. In this study α is set to 1. In this way this measure gives equal importance to both recall and precision.

3 Results

The values for precision (P), recall (R) and F_1-measure for each criterion with NB classifier are given in the Table 2. Those values represent the averages of each respective measure over 10 runs.

Table 2: Precision (P), recall(R) and F_1-measure, NB

HON code		Tokenization					
		English			French		
		W1	W1s	C5	W1	W1s	C5
HC1	P	**0.64**	0.63	0.60	**0.69**	0.66	0.63
	R	**0.75**	0.71	0.65	**0.72**	0.72	0.71
	F_1	**0.69**	0.67	0.62	**0.70**	0.69	0.67
HC2	P	**0.83**	0.82	0.77	**0.95**	0.94	0.91
	R	0.96	0.96	**0.96**	0.88	**0.92**	0.90
	F_1	**0.89**	0.88	0.85	0.92	**0.93**	0.91
HC3	P	**0.91**	0.90	0.89	**0.94**	0.92	0.92
	R	**0.98**	0.98	0.91	0.98	0.98	**0.99**
	F_1	**0.94**	0.94	0.94	**0.96**	0.95	0.95
HC4	P	0.56	0.58	**0.60**	**0.82**	0.77	0.76
	R	**0.61**	0.58	0.53	0.41	0.43	**0.44**
	F_1	**0.59**	0.58	0.56	0.55	0.55	**0.56**
HC5	P	**0.74**	0.69	0.64	**0.97**	0.94	0.76
	R	**0.31**	0.27	0.25	0.38	**0.42**	**0.42**
	F_1	**0.43**	0.37	0.36	0.55	**0.58**	0.54
HC6	P	0.92	**0.93**	0.92	**0.96**	0.94	0.91
	R	**0.94**	0.93	0.86	**0.86**	0.86	0.83
	F_1	**0.93**	0.93	0.89	**0.91**	0.90	0.87
HC7	P	**0.77**	0.76	0.74	**0.92**	0.82	0.75
	R	**0.79**	0.76	0.71	0.43	**0.85**	0.82
	F_1	**0.78**	0.76	0.72	0.59	**0.83**	0.78
HC8	P	**0.77**	0.76	0.72	**1.00**	0.95	0.83
	R	0.73	0.70	**0.79**	0.37	0.35	**0.50**
	F_1	**0.75**	0.73	0.75	0.54	0.51	**0.63**
HC9	P	0.97	**0.97**	0.95	**0.99**	0.99	0.96
	R	**0.95**	0.94	0.93	**0.92**	**0.92**	0.89
	F_1	**0.96**	0.93	0.94	**0.96**	0.95	0.93

Three tokenization schema namely word (W1), stems (W1s) and 5-gram (C5) for both English and French were used, with 30% of most frequent terms being retained for each class. In this table, the highest value for each measure, tokenization and criteria combination is marked in bold.

The results presented in Table 2 show that the tokenization resulting in the highest values in terms of precision varies between HONcode criteria but also between languages. For the English, the W1 tokenization results in highest value of precision and F1 for all the principles except for "Reference (*Attribution*) HC4" and "Contact details HC6" respectively obtained with 5-gram tokenization C5 (60%) and W1s (93%). The same tendency is noticeable for the French between word and 5-gram tokenization. However the differences between the highest and other values never exceed 10%, using highest value as a baseline.

Even thought the NB classifiers tend to be outperformed by more sophisticated algorithms such as SVM, this is not the case for the collection on-hand. Tables 3 and 4 give the comparison of the classification performance for "Reference (HC4)" and "Privacy policy (HC3)" criteria respectively, between NB and SVM algorithms. These algorithms were compared for both stemming and 5-gram tokenization.

Table 3: NB vs SVM, for stems (W1s) and 5-grams (C5), Reference criterion (HC4)

HC4	English				French			
	W1s		C5		W1s		C5	
	NB	SVM	NB	SVM	NB	SVM	NB	SVM
P	0.58	**0.62**	**0.60**	0.59	**0.77**	0.57	**0.76**	0.59
R	0.58	**0.76**	0.53	**0.77**	0.43	**0.68**	0.44	**0.66**
F_1	0.58	**0.68**	0.56	**0.67**	0.55	**0.62**	0.56	**0.62**

The results presented in Tables 3 and 4 show that, besides or the stem tokenization for the Reference criteria in English, the NB algorithm results in higher precision when compared to SVM for this collection. The most important difference can be noticed for the HC3 criterion (Table 4), where for the French language and W1s tokenization NB achieves precision of 0.92 compared to 0.46 achieved by SVM. In return, SVM results in higher recall, although the relative difference is not as important.

Table 4: NB vs SVM, for stems (W1s) and 5-grams (C5), Privacy policy criterion (HC3)

HC3	English				French			
	W1s		C5		W1s		C5	
	NB	SVM	NB	SVM	NB	SVM	NB	SVM
P	**0.90**	0.70	**0.89**	0.63	**0.92**	0.46	**0.92**	0.48
R	0.98	**0.99**	0.99	**0.99**	0.98	**0.99**	0.99	**0.99**
F_1	**0.94**	0.82	**0.94**	0.77	**0.95**	0.62	**0.95**	0.65

To demonstrate the similar behaviour of different tokenization in relation with dimensionality reduction, we have compared in Tables 5 and 6 the relative precision loss and relative recall gain respectively, between the cases where 80% and 30% of all features are kept.

Table 5: Average precision loss for English and French

	Token.	English			French		
		Kept %		Diff %	Kept %		Diff %
		80	30		80	30	
P	W1	0.89	0.79	-11.08	0.93	0.91	-1.76
	W1s	0.87	0.78	-10.28	0.92	0.88	-4.44
	C5	0.85	0.76	-10.41	0.90	0.83	-7.92

The loss in precision for the English language, shown by the results in Table 5 is the smallest in the case of W1s tokenization (10.28%). The gain in the recall is however the most important in the case of C5 being used (38.86%, Table 6) for this language. For the French the smallest loss in precision is noticed for W1 tokenization as well as the highest gain in the recall.

Table 6: Average recall gain for English and French

	Token.	English			French		
		Kept %		Diff %	Kept %		Diff %
		80	30		80	30	
	W1	0.59	0.78	31.14	0.48	0.71	47.53
R	W1s	0.57	0.76	33.13	0.51	0.72	41.37
	C5	0.53	0.74	38.86	0.52	0.72	40.16

4 Conclusion

The character n-gram tokenization has been evaluated in this publication with a goal to determine whether it could be used as an alternative to bag of words or stems. Based on the results presented in the Table 2, we can conclude that even though the word tokenization results in best performance in terms of precision. When the recall is the measure one desires to augment both the n-gram and stems impose themselves as the better solutions. This tendency is more pronounced for the French than English language, which could be explained by the fact that the French is a morphologically richer language, thus benefits more from linguistic treatment.

We have also shown that Naive Bayes algorithm is capable of outperforming the algorithms such as SVM. For the HC3 criterion, using W1s tokenization for the French the SVM results in precision not less than 50% smaller than that of NB, using NB as baseline. On the other hand the SVM results in somewhat higher recall for all language, tokenization, criterion combinations. The relative difference between two algorithms is however less important raising up to 37% (French, W1s, HC4). These results as well as time/resource consumption difference between the two algorithms are in favour of NB as a solution for the task presented.

The results presented in the tables 5 and 6 demonstrate that different tokenization techniques evaluated in this publication, namely word (W1), stem (W1s) or five gram (C5), show the same behaviour in precision/recall evolution towards dimensionality reduction by feature selection. This also indicates interchangeability of these tokenization technique. The baseline established here for the English and confirmed for the French, show that the language independent approach would be a viable alternative to word-based tokenization for wide variety of languages, especially for the morphologically complex ones.

Acknowledgements

The research is conducted in the scope of the European project Kconnect and funded by this project (2015-2018, project No. 644753).

References

[1] Fahy E, Hardikar R, Fox A, Mackay S. Quality of patient health information on the Internet: reviewing a complex and evolving landscape. Australasian Med J. 2014; 7(1) 24–28. PMID: 24567763

[2] White R, Horvitz E. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. http://research.-microsoft.com/pubs/76529/TR-2008-178.pdf Archived at: http://www.webcitation.org/6RoyoFIYT

[3] Boyer C, Baujard V, Scherrer J. 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.

[4] Mc Namee P, Mayfield J. Character n-gram tokenization for european language text retrieval. Information Retrieval, 2004;7(1-2):73–97

[5] Mc Namee P, Mayfield J, Nicholas CK. Don't have a stemmer?: be un+concern+ed. In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, SIGIR, ACM, 2008; 813–814.

[6] Naji N, Savoy J, Dolamic L. Recherche d'information dans un corpus bruité(ocr). In Gabriella Pasiand, Patrice Bellot, editors, CORIA, 6Editions Universitaires d'Avignon, 2011;271–28.

[7] Gaudinat A, Grabar N, Boyer C. Automatic Retrieval of Web Pages with Standards of Ethics and Trustworthiness Within a Medical Portal: What a Page Name Tells Us. Artificial Intelligence in Medicine [Internet]. Springer; 2007;185–189.

[8] Gaudinat A, Grabar N, Boyer C. 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, Studies in Health Technology and Informatics, IOS Press, 2007;129:705–709.

[9] Sebastiani F. Machine Learning in Automated Text Categorization, ACM Computing Surveys,2002; 34: 1-47.

[10] Manning CD, Raghavan P, Schutze H. Introduction to information retrieval. Cambridge University Press, 2008.

[11] Junker M, Hoch R. An experimental evaluation of OCR text representations for learning document classifiers. International Journal on Document Analysis and Recognition, 1998, 1(2):116-122

[12] Tomlinson, S. (2004). Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServerTM at CLEF 2003. In Comparative evaluation of multilingual information access systems. LNCS #3237 (pp. 286–300). Berlin: Springer-Verlag.

[13] Cavnar W.B., Trenkle J.M. N-Gram-Based Text Categorization, In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994, 161—175

[14] Beldad A, de Jong M, Steehouder M. How shall I trust the faceless and the intangible? A literature review on the antecedents of online trust. Computers in Human Behavior 2010;26(5):857-869

[15] Eysenbach G, Powell J, Kuss O, Sa ER. Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web – A Systematic Review Journal of the American Medical Association JAMA 2002;287(20):2691—2700

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

[17] Williams K, Calvo RA. A framework for document categorization. 7th Australasian Document Computing Symposium. December 2002. Sydney, Australia. 13-19.

[18] Boyer C, Dolamic L. Feasibility of automated detection of honcode conformity for health related websites. IJACSA, 2014 Mar; 5(3):69-74, doi: 10.14569/IJACSA.2014.050309

[19] Baeza-Yates RA, Ribeiro-Neto B. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.1999

Address for correspondence

Célia Boyer
Health on The Net Foundation
Chemin du Petit-Bel-Air 2
1225 Chêne-Bourg, Switzerland
celia.boyer@healthonnet.org
+41 (0) 22 37 26 250