A Classifier to Evaluate Language Specificity of Medical Documents

Trudi Miller, Gondy Leroy, Samir Chatterjee, Jie Fan, Brian Thoms
School of Information Systems & Technology, Claremont Graduate University
Trudi.Miller@cgu.edu

Abstract

Consumer health information written by health care professionals is often inaccessible to the consumers it is written for. Traditional readability formulas examine syntactic features like sentence length and number of syllables, ignoring the target audience's grasp of the words themselves. The use of specialized vocabulary disrupts the understanding of patients with low reading skills, causing a decrease in comprehension. A naïve Bayes classifier for three levels of increasing medical terminology specificity (consumer/patient, novice health learner, medical professional) was created with a lexicon generated from a representative medical corpus. Ninety-six percent accuracy in classification was attained. The classifier was then applied to existing consumer health web pages. We found that only 4% of pages were classified at a layperson level, regardless of the Flesch reading ease scores, while the remaining pages were at the level of medical professionals. This indicates that consumer health web pages are not using appropriate language for their target audience.

1. Introduction

1.1 Readability of Consumer Health Information

Professionals regularly write documents to assist laypeople understand unfamiliar technologies. Websites like WebMD (www.webmd.com) offer accurate health information targeted to consumers, but it is difficult for those well-versed in the jargon of their profession to eliminate technical terms from their writing. While well-written, easy to understand documentation can augment the layperson's understanding, misunderstood health information can cause harm to its readers [1]. Those with the lowest health literacy report poorer health [2] and have less understanding about the medical care they receive [3]. Informed patients are more likely to engage in positive health behaviors to maintain or improve their health [4].

There is disparity between the readability of available online health information and the reading abilities of the average consumer. Almost half of American adults have difficulty understanding health information [5]. Berland *et al.* [6] found online information to be accurate, but concurred that it requires high reading levels to comprehend. Ownby [7] evaluated 60 sites with the topic of depression in seniors and found them to be well above the average reading level.

1.2 Assessing Readability

There are two methods commonly used to measure readability: Fry's formula and Flesch's Reading Ease. Fry's formula is calculated by selecting three 100-word passages from the text and calculating the average number of sentences and syllables across all three passages [8, 9]. These two values are then plotted on the Fry graph for estimating readability, giving the approximate grade level. The second method is Flesch's Reading Ease, which calculates a percentage between 1 and 100 for documents, based upon the average sentence length and the number of syllables per word. A score between 0 and 60 is difficult, 60 to 70 is standard, and greater than 70 is easy. Both Fry's and Flesch's Reading Ease have been used extensively in the literature to evaluate the readability of consumer health information online [6, 7, 10] and in printed form [2, 11].

There are several criticisms of these traditional readability formulas. Chapman *et al.* [12] noted that readability measures are limited in evaluating comprehensibility due to their focus on sentence and word length. Moreover, authors who use readability statistics in their research note the differences among formulas. For example, D'Alessandro *et al.* [10] found that the calculated Flesch-Kincaid reading

levels were 4 to 5 grade levels lower than Fry for the same documents. Schriver [14] noted the inherent subjectivity of readability scores because they rely on comparison with a standard text. The subjectivity is exemplified by an increase in the estimated reading level of documents that contain bullets without periods at the end of each item. The formulas treat these lists as long sentences, ignoring the mental processing benefits such lists provide. Duffy [13], in his seminal article, points out that sentence length and other commonly used variables are not those most important in determining document comprehensibility. He advocates the use of the formulas as a relative metric for selecting between alternative texts, not as an absolute metric to be measured against one's educational level.

Substantial changes in grade level can be achieved with superficial changes. For example, by transforming lists with short items and no terminal periods into comma-separated lists and by replacing colons with periods, as shown in *Figure 1*. Such substitutions have no effect on readability, but instead exploit the algorithms used by traditional readability metrics. Even though texts are presented at lower grade levels, this does not necessarily improve understanding.

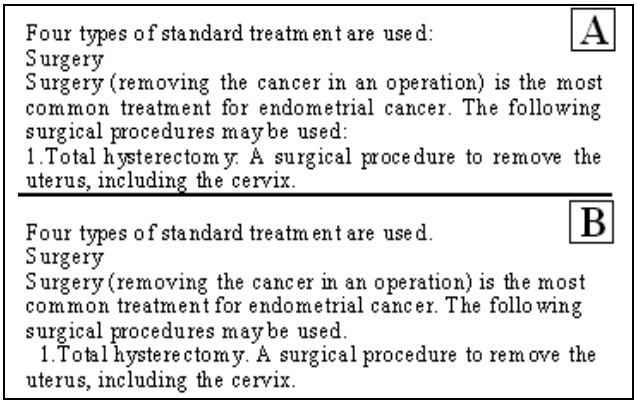

Figure 1. Section from a health information document at 12th grade reading level (A) and at 10th grade level (B).

1.3 Consumer Health Information Vocabulary

If the syntactic structure of a text is not enough to measure readability, one must explore additional characteristics. Neither Flesch's nor Fry's take into account the vocabulary used; use of a short word like "cyst" will lower the reading level assessed by both formulas, but may be too complex for those without sufficient medical knowledge. Gemoets *et al.* [11] evaluated traditional readability formulas and found that those documents with the lowest readability scores also had the lowest "lexical density". Lexical density is the number of unique number of words within a given unit (e.g. sentence, document). Solving the problem of lexical density alone is not enough to bridge the gap for average readers.

Medical professionals use technical words that may be unfamiliar to many patients. Without consumer friendly terms, consumers can misinterpret medical information by filling in the gaps on their own [15].

McCray *et al.* [16] identified three levels of difference between consumers and clinicians: lexical, syntactic, and semantic. Readability formulas address only the syntactic dimension, ignoring the semantic component that is vital to comprehension. Kogan *et al.* [17] described that patients encounter difficulty in understanding the medical jargon found in information retrieval query results. This was borne out by Zeng *et. al.* [18], who found that patients tend to prefer terms related to diseases, syndromes, or body parts over the occupational terms that medical professionals prefer. Slaughter *et al*. [19] noted that, to be applicable to consumer health information research, clinically based resources like the Unified Medical Language System (UMLS) need to expand their vocabulary to include terms used by patients to express their conditions.

1.4 Consumer Focused Vocabulary Initiatives

Research into consumer focused vocabulary has received much attention in the recent past. Zeng & Tse [15] discuss the development of consumer health vocabularies (CHVs), which represent terms commonly used by a given consumer group to express health related topics. They argue that research requires such CHVs to be able to facilitate consumers' understanding of health information. Initial research in this area was done through collection and examination of health-related consumer queries, with the goal of finding a single, unambiguous label for each medical term [20]. Consumer's limited domain knowledge of the health field leads to the construction of simplistic queries observed in Zeng *et al.* [21].

Research into the mapping between clinician and consumer language has begun. Soergel *et al*. [22] advocated the use of an intermediate layer between patients and clinicians, including such resources as a thesaurus that would provide translations. Leroy *et al.* [23] further outline the benefits of an interpretive layer using modification of sentence structure and the words used. Tse & Soergel [24] found that consumers have an understanding that is different from clinicians, and that it is important to understand the mapping between the two.

2. Research Questions

Until a consumer/clinician mapping of medical vocabulary is complete and everyone adheres to it, those who provide consumer health information need to be able to evaluate whether the documents they provide will be comprehensible to their target audience. A metric is needed that takes into account the vocabulary used instead of just treating words like part-of-speech tagged black boxes. If nouns used within a document are unintelligible, those with low reading skills skip over them [9]. Since these unknown nouns can represent crucial medical concepts (e.g. diseases, symptoms, side effects), it is imperative that patients understand them. This raises our first question: Can we automatically discover the vocabulary level of a document?

Different medical documents are targeted towards different audiences. Medical professionals need to communicate with each other and with patients. Too often, documents written by medical professionals for medical professionals are distributed to laypeople with little consideration to their needs. Three audience categories are prevalent within consumer health information: consumers/patients, novice health learners, and medical professionals. Patients are people whose familiarity with medical text is minimal, and whose language is least formal. Novice health learners have no medical training, but the desire to learn appropriate medical terminology from educational materials like websites and brochures. Medical professionals are those who have training in and work in the medical field (e.g. doctors, nurses).

A classifier categorizes documents as being appropriate for a specific audience. Such a measure of language specificity would assist the authors of medical documents in ensuring that their vocabulary that they are using is appropriate for their target group. For example, a public health agency could use the classifier to evaluate a press release to ensure that it would be comprehensible by laypeople. Doctors could use the classifier to ensure that the language used in post-operative instructions would be understood. This leads to our second research question: If we can discover the vocabulary level automatically, at what levels are common "consumer health" documents available today?

3. Methods

3.1 Classifier Corpus Selection

The classifier corpus was populated with documents targeted at each of the three target audiences: patients, novice health learners, and medical professionals. Fifty patient blogs were used to represent the language used by patients. These were collected from different blog sites (e.g. www.blogger.com) through the use of medical keywords like 'treatment' and 'hospital'. Written specifically as educational material, 50 web pages from the City of Hope National Medical Center website (http://www.coh.org/) were added to the corpus as documents representative of the novice health learner level. Medical professionals use clinical terminology to ensure accuracy and brevity, and this part of the corpus was represented by 50 journal articles from the *Journal of the American Medical Association*. *JAMA* was used because it is the most widely circulated medical journal in the world [25], and it is not specialized to one medical specialty or to one type of disease. It provides higher external validity than would a journal like *Radiology* or *Cancer Cell International*.

These three sources provide three distinct levels of readability. Patient blogs had a mean Flesch reading ease of 67.1 and grade level of 7.7, classifying them at a standard reading level. The educational pages had a mean reading ease of 39.8 and grade level of 10.8, a difficult reading level. The journal articles had a mean of 14.5 and grade level of 12.0, also a difficult reading level (*Table 1*).

Table 1. Readability scores for classifier corpus.

N = 150	Flesch Reading Ease		Flesch-Kincaid Grade Level	
	Mean	Std. Dev.	Mean	Std. Dev.
Patient	67.1	7.8	7.7	1.7
Novice Health Learner	39.8	18.0	10.8	1.3
Medical Professional	14.5	9.4	12.0	0.2

3.2 Naïve Bayes Classifier

The relevant pages were downloaded in HTML format and had navigational and extraneous formatting removed, leaving only the content as raw text (*Figure 2*). The text was tokenized using the GATE tokenizer [26] and stored in a database. All numbers were replaced with a placeholder (<literal number>), because the value of the number itself was not as relevant as the fact that a literal number was present. All punctuation marks were removed, leaving only word tokens. After cleaning, the corpus had 196,560 tokens, with 52,111 unique tokens. The frequency for

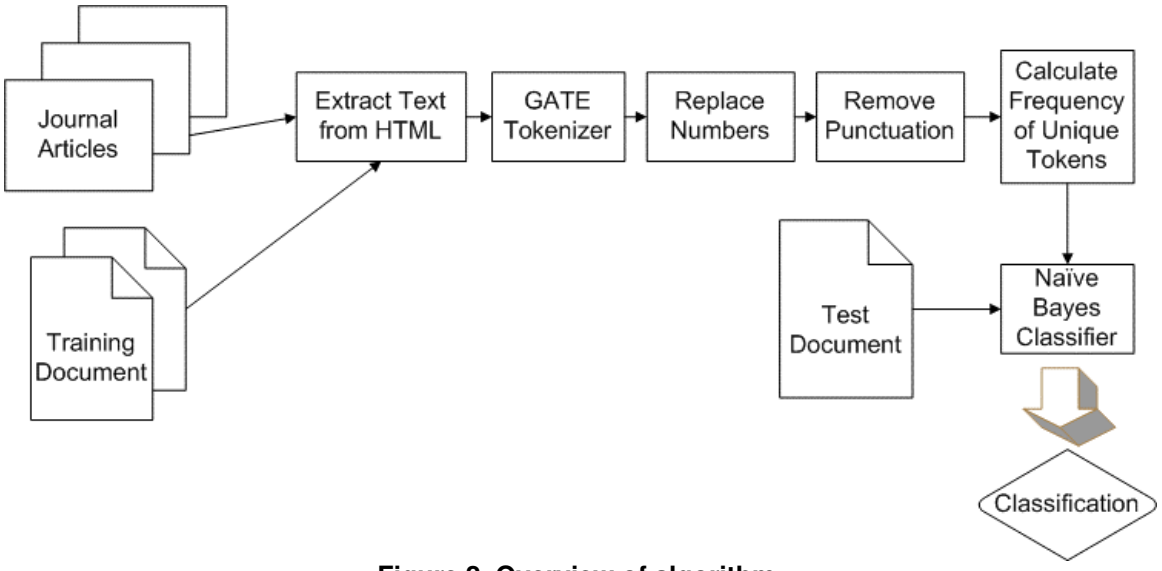

Figure 2. Overview of algorithm.

each unique token was calculated and stored for each document.

Once the token frequency was calculated, a naïve Bayes' classifier was used to classify documents (*Figure 2*). Naïve Bayes classifiers' use within classification problems is well-established [27-29]. Each token is examined and the probability that the word occurs in each of the document types is calculated. Summing up the probabilities from all of the tokens, one can obtain numeric estimates representing the likelihood that the document belongs to a given category. The classifier was reinitialized between each document so that no residual knowledge was transferred between sessions. Smoothing was implemented by adding a small non-zero value for each token encountered in the test document that was not present in the classifier corpus. We used our Java-based own implementation of naïve Bayes rather than using a pre-existing tool.

3.3 Consumer Health Information Classification

Once the classifier was validated, it was applied to consumer health information available on the Internet. For this, 30 pages from three different sources were collected. The first was the health section of a non-profit organization (SeniorNet.org), an organization whose purpose is to educate and assist seniors. The second source was a pharmaceutical company (Merck). With increased advertising by drug companies, more information is being made available via their websites. The final source was a government public health website (New York State Department of Health), whose purpose is to communicate with the public about health issues. Together, these three sites comprise a sample of consumer health information from both the private and public sectors.

Readability scores were calculated and are summarized (*Table 2*). Non-profit pages had a mean Flesch reading ease score of 38.5 and a Flesch-Kincaid grade level of 11.4. Government pages had a mean Flesch reading ease score of 45.6 and a Flesch-Kincaid grade level of 10.4. The pharmaceutical manufacturer's pages had a mean Flesch reading ease score of 17.3 and a Flesch-Kincaid grade level of 12.0. All means of Flesch reading ease are in the difficult category.

Using the same algorithm as shown in *Figure 2*, these 90 pages were classified by the naïve Bayes classifier as either patient, novice health learner, or medical professional level language.

Table 2. Readability scores for consumer health information pages.

N = 90	Flesch Reading Ease		Flesch-Kincaid Grade Level	
	Mean	Std. Dev.	Mean	Std. Dev.
Non-Profit	38.5	7.6	11.4	1.0
Government	45.6	10.8	10.4	1.4
Pharmaceutical Manufacturer	17.3	8.1	12.0	0.1

4. Results

4.1 Classifier Validation

The naïve Bayes classifier was evaluated with leave-one-out validation. One hundred forty-nine of the documents were used to train and the remaining document was tested. This was performed 150 times, with each document being "left out" once.

Overall, 96% of the documents (144/150) were correctly classified (*Table 3*). All 50 patient level documents were correctly classified. Forty-five of 50 novice health learner level documents were classified correctly (90%). Forty-nine of 50 medical professional documents were correctly classified (98%).

Table 3. Classifier validation results using leave-one-out validation.

N = 150	% Classified Correctly
Patient	100%
Novice Health Learner	90%
Medical Professional	98%
Overall	96%

4.2 Classifier Application

All 90 consumer health pages were evaluated using the classifier (*Table 4*). Overall, 86 of the 90 (96%) documents were found to use medical professional level vocabulary, with only 4 (4%) documents at the patient level.

The website SeniorNet had only 4 (13%) documents using patient level language. The government website from the New York Department of Health had all of its documents written at the medical professional level, as had the pharmaceutical manufacturer Merck.

5. Discussion

5.1 Classifier Validation

The classifier's accuracy of 96% shows that the difference in clinician and patient language can be automatically detected using a naïve Bayes classifier.

Six documents were not correctly classified during the classifier validation. Four novice health learner level documents were incorrectly classified at medical professional level. The four describe clinical drug trials for menopausal hormone use, lung cancer, lymphoma, and cholesterol reduction. The language in these documents is very technical, including discussions of placebo effects and study methods.

Table 4. Classifier results for consumer health information pages.

	# of Documents	Classifier Output	% of Pages
Non-Profit	4	Patient	13
	0	Novice Health Learner	0
	26	Medical Professional	87
Government	0	Patient	0
	0	Novice Health Learner	0
	30	Medical Professional	100
Pharmaceutical Manufacturer	0	Patient	0
	0	Novice Health Learner	0
	30	Medical Professional	100
Overall	4	Patient	4
	0	Novice Health Learner	0
	86	Medical Professional	96

One novice health learner document was classified at patient level. It contained instructions to follow after an abdominal CT scan, written in a question and answer format. The misclassified medical professional article describes a woman's struggle with ovarian cancer, and is presented in a narrative form that is similar to a newspaper, not typically characteristic of medical professional language.

5.2 Classifier Application

Despite recent efforts to improve the readability of health information, it is clear that the vocabulary used plays as important a role as traditional measures like sentence length and syllable count. With only 4 pages out of 90 using language at the patient level, there is still a large discrepancy between the complexity of available consumer health information and the vocabulary of the consumers to whom it is made available. One of the pages classified at the consumer level featured several quotes from a physician. Another was a transcript of a presentation given about cancer therapy. The remaining two consumer level pages were about insomnia, and were written in a question and answer format. The

commonality of appropriate terminology being used by clinicians when speaking leads to hope that health information can be expressed comprehensibly in written form.

The readability scores of the 4 consumer level documents had a mean Flesch's reading ease of 58.5 and a Flesch-Kincaid grade level of 8.5 (*Table 5*). This shows that the documents' vocabulary classification does not mirror that of the traditional readability formulas. If it did, one would expect to see far more documents from the government group also classified at the consumer level, given the government group's greater mean and standard deviation of readability scores (*Table* 2).

Table 5. Readability scores for consumer health web pages classified at consumer level.

N = 4	Flesch Reading Ease		Flesch-Kincaid Grade Level	
	Mean	Std. Dev.	Mean	Std. Dev.
Overall	58.5	6.4	8.5	0.8

Four documents whose readability scores were not extreme were classified at the consumer level, adding credence to the differentiation between traditional readability scores and our classification based on vocabulary. This emphasizes the difference between traditional readability scores and our classifier.

Popular readability measures only address the length of sentences and number of syllables. Simple sentence manipulation can increase the readability level by a grade level or more. Solely through the replacement of semicolons with periods, we were able to increase the readability grade level of some of our documents by half of a grade level. Readability statistics do not take into account the reading behavior of the average patient, who skips words that s/he does not recognize. If the noun subject of a sentence is not understood, the length of that sentence is no longer important. Authors and distributors of consumer health information need to know that ignoring the vocabulary used can undermine other efforts to make their documents readable by patients. Despite the lower average Flesch-Kincaid grade level of the government pages, more pages from the non-profit site were found to use vocabulary suitable to patients. Our classifier allows authors to evaluate the language used within their documents to determine if it is appropriate for their target audience.

6. Conclusions

The research contribution of the classifier is that the algorithm can be performed against any specialized corpus of sufficient size. The model can then be trained for documents aimed at those outside of the specified field. For example, hospitals can calculate the specialization metric for patient educational materials to evaluate whether or not the terminology being used is too complex for the average person to understand. It will assist health care professionals in evaluating their consumer health information. This classifier is not meant to replace traditional readability levels. Measures like the Flesch Readability Ease should still be used to determine if sentence length and density is appropriate for average readers. The classifier should be used to provide an additional dimension: the difficulty of the words within the document.

Improvements to the classifier will be made by expanding to include additional journals, clinical notes, or other sources from across the globe. Increasing the training set to include additional sources of patient, novice health learner materials, and medical professional level documents will further hone the accuracy of the classifier. Future work includes integration with existing readability metrics to provide a single score for both syntax and vocabulary. This will augment our concurrent work visualizing documents with high readability and clinical vocabulary to make them more comprehensible to consumers.

7. Acknowledgments

This work is supported by the NLM grant R21-LM008860-01.

8. References

[1] A. G. Crocco, M. Villasis-Keever, and A. R. Jadad, "Analysis of Cases of Harm Associated with Use of Health Information on the Internet," *Journal of the American Medical Association*, vol. 287, pp. 2869-2871, 2002.

[2] R. E. Rudd, B. A. Moeykens, and T. C. Colton, "Health and Literacy: A Review of Medical and Public Health Literature," in *Annual Review of Adult Learning and Literacy*, J. Comings, B. Garners, and C. Smith, Eds. New York: Jossey-Bass, 1999.

[3] A. M. A. "Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, "Health Literacy: Report of the Council on Scientific Affairs," *Journal of the American Medical Association*, vol. 281, pp. 552-7, 1999.

[4] G. Van Servellen, J. S. Brown, E. Lombardi, and G. Herrera, "Health Literacy in Low-Income Latino Men and Women Receiving Antiretroviral Therapy in Community-Based Treatment Centers," *AIDS Patient Care and STDs*, vol. 17, pp. 283-298, 2003.

[5] Institute of Medicine, "Health literacy: a prescription to end confusion," vol. 2006. Washington DC: National Academy Press, 2004.

[6] G. K. Berland and e. al., "Health Information on the Internet: Accessibility, Quality, and Readability in English and Spanish," *Journal of the American Medical Association*, vol. 285, pp. 2612-2621, 2001.

[7] R. L. Ownby, "Influence of Vocabulary and Sentence Complexity and Passive Voice on the Readability of Consumer-Oriented Mental Health Information on the Internet," presented at American Medical Informatics Association (AMIA) 2005, 2005.

[8] E. Fry, "Fry's Readability Graph: Clarification, Validity, and Extension to Level 17," *Journal of Reading*, vol. 21, pp. 242-52, 1977.

[9] C. C. Doak, L. G. Doak, and J. H. Root, *Teaching Patients with Low Literacy Skills*, 2nd ed. Philadelphia: J. B. Lippincott Company, 1996.

[10] D. M. D'Alessandro, P. Kingsley, and J. Johnson-West, "The Readability of Pediatric Patient Education Materials on the World Wide Web," *Archives of Pediatric & Adolescent Medicine*, vol. 155, pp. 807-812, 2001.

[11] D. Gemoets, G. Rosemblat, T. Tse, and R. Logan, "Assessing Readability of Consumer Health Information: An Exploratory Study," presented at MEDINFO 2004, 2004.

[12] W. W. Chapman, D. Aronsky, M. Fiszman, and P. Haug, "Contribution of a speech recognition system to a computerized pneumonia guideline in the emergency department," presented at AMIA Proceedings, 2000.

[13] T. M. Duffy, "Readability Formulas: What's the Use?," in *Designing Usable Texts*, T. Duffy and R. Walker, Eds.: Academic Press, Inc., 1985, pp. 113-143.

[14] K. A. Schriver, "Readability Formulas in the New Millennium: What's the Use?" *ACM Journal of Computer Documentation*, vol. 24, pp. 138-140, 2000.

[15] Q. T. Zeng and T. Tse, "Exploring and Developing Consumer Health Vocabularies," *Journal of the American Medical Informatics Association*, vol. 13, pp. 24-9, 2005.

[16] A. T. McCray, N. C. Ide, R. R. Loane, and T. Tse, "Strategies for Supporting Consumer Health Information Seeking," presented at MEDINFO 2004, 2004.

[17] S. Kogan, Q. Zeng, N. Ash, and R. A. Greenes, "Problems and Challenges in Patient Information Retrieval: A Descriptive Study," presented at American Medical Informatics Association (AMIA) 2001, 2001.

[18] Q. Zeng, S. Kogan, N. Ash, and R. A. Greenes, "Patient and Clinician Vocabulary: How Different Are They?," presented at MEDINFO 2001, 2001.

[19] L. Slaughter, C. Ruland, and A. K. Rotegard, "Mapping Cancer Patients' Symptoms to UMLS Concepts," presented at American Medical Informatics Association (AMIA) 2005, 2005.

[20] Q. T. Zeng, T. Tse, J. Crowell, G. Divita, L. Roth, and A. C. Browne, "Identifying Consumer-Friendly Display (CFD) Names for Health Concepts," presented at American Medical Informatics Association (AMIA) 2005, 2005.

[21] Q. T. Zeng, J. Crowell, R. M. Plovnick, E. Kim, L. Ngo, and E. Dibble, "Assisting Consumer Health Information Retrieval with Query Recommendations," *Journal of the American Medical Informatics Association*, vol. 13, pp. 80-90, 2006.

[22] D. Soergel, T. Tse, and L. Slaughter, "Helping Healthcare Consumers Understand: An "Interpretive Layer" for Finding and Making Sense of Medical Information," presented at MEDINFO 2004, 2004.

[23] G. Leroy, E. Eryilmaz, and B. T. Laroya, "Health Information Text Characteristics," presented at American Medical Informatics Association (AMIA) 2006, 2006.

[24] T. Tse and D. Soergel, "Exploring Medical Expressions Used by Consumers and the Media: An Emerging View of Consumer Health Vocabularies," presented at American Medical Informatics Association (AMIA) 2003, 2003.

[25] Journal of the American Medical Association, "JAMA -- About JAMA," http://jama.ama-assn.org/misc/aboutjama.dtl, 2006.

[26] Sheffield Natural Language Processing Group, "General Architecture for Text Engineering," 3.0 ed. Sheffield, UK: http://gate.ac.uk/, 2005.

[27] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 54, 2002.

[28] J. Langford, "Tutorial on Practical Prediction Theory for Classification," *Journal of Machine Learning Research*, vol. 6, pp. 273-306, 2005.

[29] K. Larsen, "Generalized Naive Bayes Classifiers," *SIGKDD Explorations*, vol. 7, 2005.