

# Text Categorization Models for Identifying Unproven Cancer Treatments on the Web

Yin Aphinyanaphongs<sup>a</sup>, Constantin Aliferis<sup>abc</sup>

<sup>a</sup>*Department of Biomedical Informatics, Vanderbilt University, Nashville*

<sup>b</sup>*Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville*

<sup>c</sup>*Vanderbilt Ingram Cancer Center, Vanderbilt University, Nashville.*

## Abstract

The nature of the internet as a non-peer-reviewed (and more generally largely unregulated) publication medium has allowed wide-spread promotion of inaccurate and unproven medical claims in unprecedented scale. Patients with conditions that are not currently fully treatable are particularly susceptible to unproven and dangerous promises about miracle treatments. In extreme cases, fatal adverse outcomes have been documented. Most commonly, the cost is financial, psychological, and delayed application of imperfect but proven scientific modalities. To help protect patients, who may be desperately ill and thus prone to exploitation, we explored the use of machine learning techniques to identify web pages that make unproven claims. This feasibility study shows that the resulting models can identify web pages that make unproven claims in a fully automatic manner, and substantially better than previous web tools and state-of-the-art search engine technology..

Keywords: *Information Storage and Retrieval, Medical Informatics, Internet, Neoplasms, Text Categorization*

## Introduction

“The killing of all parasites and their larval stages together with removal of isopropyl alcohol and carcinogens form the patients' lifestyle results in remarkable recovery (from cancer), generally noticeable in less than one week [1].” This is one example of an unproven treatment claim made on the web. These unproven treatments are known as *quackery* with the *quacks* promoting them defined as “untrained people who pretend to be physicians and dispense medical advice and treatment [2].” The internet allows quacks to advocate inaccurate and unproven treatments with documented fatal, adverse outcomes in some situations [3-6].

In regards to cancer patients, Metz et al. reported that 65% of cancer patients searched unproven treatments and 12% purchased unconventional medical therapies online [7]. In another study, Richardson reported that 83% of cancer patients had used at least one unproven treatment [8]. Compounding the problem are consumers who are ill-equipped to evaluate treatment information [9]. The language and quality of web

pages with unproven treatments is also highly variable [10]. With a growing internet, the ease of publishing unproven claims, and susceptible and often desperately ill patients, the chance for further adverse outcomes, patient and family despair, and sunk costs is inevitable. It is the mandate of the medical profession to protect patients from inaccurate and poor medical information.

Extensive research has developed several manual methods to combat the propagation of unproven claims on the web. The Health-on-the-Net Foundation advocates self-regulation of health related websites [11]. The foundation applies strict criteria to websites and grants them a seal if they pass. However, most consumers ignore the seals [12]. In another approach, experts produced rating tools that consumers are supposed to apply to websites [13, 14]. Another method is manual review of individual websites that are published either in print or electronically.

Each method has limitations. Self-regulation relies on knowledge of the certification and a vigilant public to report failing web sites. Ratings tools are dependent on a knowledgeable public to apply, difficult to validate, time consuming to produce, and do not always produce consistent ratings [15, 16]. Moreover, the rating tools are not appropriate for use on complementary/ alternative medicine sites [17]. Furthermore, manual review suffers from limits in reviewer time and the selection of web sites to review.

Ideally, we would like a solution that is validated, easy to apply by consumers, and works on any webpage. In this paper, we hypothesize that automated approaches to identifying web pages with unproven claims may provide a solution.

## Previous Work On Automatic Webpage Identification

Previous research focused on automated or semi-automated approaches to identifying *high quality* medical web pages.

Price and Hersh [18] evaluated web page content by combining a score measuring quality proxies for each page. Quality proxies included relevance, credibility, bias, content, currency, and the value of its links. The authors evaluated the algorithm on a small test collection of 48 web pages covering nine medical topics labeled as desirable or undesirable by the investigator. In all cases, the score assigned to the desirable

pages was higher than the scores assigned to undesirable pages.

Even though the algorithm perfectly discriminated between desirable and undesirable webpages, several limitations exist. First, the test sample was small and not representative of the scale for a web classification task. Second, the algorithm does not measure content quality directly, but used proxies for quality to compile a score for a web page. Third, the usefulness of some of the explicit criteria may not correlate with content quality [19], and may not be valid or good features to include for scoring.

As a leading search engine, Google has become a de facto standard for identifying and ranking web pages. Pages that rank highly in Google are assumed better quality than those at lower rank. Several researchers have explored this assumption for health pages. Fricke and Fallis [20] evaluated PageRank score as one indicator of quality for 116 web sites about carpal tunnel syndrome. Their results show that PageRank score is not inherently useful for discrimination or helping users to avoid inaccurate or poor information. Of the 70 web sites with high PageRank, 29 of them had inaccurate information.

Griffiths [21] evaluated PageRank scores with evidence based quality scores for depression websites. The authors obtained Google PageRank scores for 24 depression websites from the DMOZ Open Directory Project website. Two health professional raters assigned an evidence based quality score to each site. PageRank scores correlated weakly ( $r = 0.61$ ,  $P=0.002$ ) with the evidence based quality scores.

Tang, Craswell, and Hawking [22] compared Google results with a domain-specific search engine for depression. They found that of a 101 selected queries, Google returned more relevant results, but at the expense of quality. Of the 50 treatment related queries, Google returned 70 pages of which 19 strongly disagreed with the scientific evidence.

## Hypothesis

Our fundamental hypothesis for this feasibility study is that we can model expert opinion and build machine learning models that identify web pages that make unproven claims for the treatment of cancer.

To the best of our knowledge, there is no research on automated techniques for identifying web pages that make unproven claims. In prior work, we showed that text categorization methods identified high quality content specific articles in internal medicine [23]. Extending this work into the web space, we reverse the hypothesis of the previous studies. Rather than identifying high quality pages, we explore automated identification of low quality pages, specifically pages that make unproven claims for cancer treatment.

## Materials and Methods

### Definitions

Our gold standard relied on selected unproven cancer treatments identified by experts at <http://www.quackwatch.org>.

The website is maintained by a 36 year old nonprofit organization whose mission is to “combat health related frauds, myths, fads, fallacies, and misconduct.” The group employs a 152 person scientific and technical advisory board composed of academic and private physicians, dentists, mental health advisors, registered dietitians, podiatrists, veterinarians, and other experts whom review health related claims. By using unproven treatments identified by an oversight organization, we capitalized on an existing high quality review.

### Corpus Construction

For this feasibility study, we randomly chose 8 unproven treatments from 120 dubious cancer treatments listed by [quackwatch.org](http://quackwatch.org) [24]. The randomly selected treatments were “Cure for all Cancers”, “Mistletoe”, “Krebiozen”, “Metabolic Therapy”, “Cellular Health”, “ICTH”, “Macrobiotic Diet”, and “Insulin Potentiation Therapy.” We then identified web pages that have these treatments by appending the words “cancer” and “treatment” and querying Google. We retrieved the top 30 results for each unproven treatment. We used a python script to download and store each result as raw html for further labeling.

### Corpus Labels

We applied a set of criteria for identifying web pages with unproven treatment claims. First, of the initial 240 pages, we excluded not found (404 response code) error pages, no content pages, non-English pages, password-protected pages, pdf pages, redirect pages, and pages where the actual treatment text does not appear in the document<sup>1</sup>. Of the remaining 191 html pages, both authors independently asked the following question of each web page: does the web page make unproven claims about the proposed treatment and its efficacy. We labeled web pages with unproven claims as positive and the others as negative.

A web page that is purely informational in nature but does not make any unproven claims about the cancer treatment and its efficacy were labeled as negative. A web page selling a book with user comments that has unproven claims were labeled as positive. Portal pages that do not make any claim were labeled as negative. Web pages that present a balanced viewpoint of the treatment were carefully reviewed for any unproven claims, and, if so, were labeled positive. Additionally web pages that sell the unproven treatment but do not make claims were labeled negative.

Both authors applied the criteria independently. We calculated the inter-observer agreement (Cohen’s Kappa [25]) at 0.76<sup>2</sup>. Of the 20 sites with discrepant labelings, the reviewers discussed the labels until consensus was reached. The final corpus was composed of 191 web pages with 93 labeled as positive and 98 as negative.

---

<sup>1</sup> The Google ranking algorithm relies on anchor text to identify web page content. Anchor text may point to a web page that does not use the text in the web page itself.

<sup>2</sup> We set a threshold of 0.70 for Cohen’s Kappa. If kappa was below 0.70, we would refine the labeling criteria until the threshold was reached.

## Webpage Preparation

For this feasibility study, we chose the simplest web page representation. We converted web pages to a “bag of words” suitable for the machine learning algorithm [23]. First, for each web page, we removed all content between style and script tags. Second, all tags (including the style and script tags) were removed. Third, we replaced all punctuation with spaces. We split the remaining string on the spaces to obtain individual words. Finally, we stemmed each word [23], applied a stop word list [23], removed any words that appear in less than 3 web pages, and encoded as weighted features using a log frequency with redundancy scheme [23].

## Learning Model (Support Vector Machines)

We employed Support Vector Machine (SVM) classification algorithms. The SVMs calculate maximal margin hyperplane(s) separating two or more classes of the data. SVMs have had superior text classification performance compared to other methods [23], and this motivated our use of them. We used an SVM classifier implemented in libSVM v2.8 [26] with a polynomial kernel. We optimized the SVM penalty parameter C over the range {0.1, 1, 2, 5, 10} with imbalanced costs applied to each class proportional to the priors in the data [23], and degree d of the polynomial kernel over the range {1, 2, 5}. The ranges of costs and degrees for optimization were chosen based on previous empirical studies [23]. Different combinations of costs and degrees were exhaustively evaluated by cross-validation.

## Model Performance Estimation

We used 10-fold cross-validation that provide unbiased performance estimates of the learning algorithms [23]. This choice for n provided sufficient high-quality positive samples for training in each category and provided sufficient article samples for the classifiers to learn the models. The cross-validation procedure first divided the data randomly into 10 non-overlapping subsets of documents where the proportion of positive and negative documents in the full dataset is preserved for each subset. Next, the following was repeated 10 times: we used one subset of documents for testing (the “original testing set”) and the remaining nine subsets for training (the “original training set”) of the classifier. The average performance over 10 original testing sets is reported.

In order to optimize parameters of the SVM algorithms, we used another “nested” loop of cross-validation by further splitting each of the 10 original training sets into smaller training sets and validation sets. For each combination of learner parameters, we obtained cross-validation performance and selected the best performing parameters inside this inner loop of cross-validation. We next built a model with the best parameters on the original training set and applied this model to the original testing set. Details about the “nested cross-validation” procedure can be found in [23]. Notice that the final performance estimate obtained by this procedure will be unbiased because each original testing set is used only once to estimate performance of a single model that was built by using training data exclusively.

## Quackometer

We compared our algorithm to a heuristic, unvalidated, and unpublished quack detection tool available at <http://www.quackometer.net>. The exact details of the detection tool are proprietary. In general, the algorithm counts words in web pages that quacks use, and sorted the words into at least 5 dictionaries [27]. It looks for altmed terms such as homeopathic and herbal, pseudoscientific words such as toxins and superfoods, domain specific words such as energy and vibration, skeptical words such as placebo and flawed, and commerce terms such as products and shipping. The algorithm counts the frequency of terms, applies a user-defined frequency threshold, and generates a corresponding score from 0 to 10. The tool is available at [28].

We compared our models to the Quackometer by calculating the corresponding area under the curve (AUC) for each 10 fold-split and reporting the mean and standard deviation.

## Google PageRank

The Pagerank algorithm [29] is used by Google to identify higher quality pages on the web. The basic tenet is that a web page will rank highly if the web page has more and higher quality links pointing to it. For example, if a web page has a link from Yahoo (a highly linked page), it would rank higher than a link from a less linked to web page. In detecting web pages with unproven claims, our assumption is that web pages with poor quality information should get fewer and lower quality links than web pages with better quality.

We use Google as a proxy for PageRank<sup>3</sup>. We make the comparison to our algorithms within each topic rather than within each 10 fold split. We compared within each topic to avoid bias in ranking situations where one topic has uniformly higher Google rank than another topic. We invert the labels<sup>4</sup> in the 8 randomly selected topics, calculate the AUC, and report the mean AUC and standard deviation.

## Results

Table 1 shows the AUC performance between our machine learning filter models, Quackometer, and Google. The machine learning method identified web pages that make unproven claims with an AUC of 0.93 with a standard deviation of 0.05 across the 10 folds. Quackometer does worse with an AUC of 0.67 and a standard deviation of 0.10 across the same 10 folds. Finally Google performs least effectively in discriminating web pages with an AUC of 0.63 and a standard deviation of 0.17 across the 8 selected topics. Figure 1 shows the corresponding receiver operating curves for each method.

Table 1 – Area Under Curve for Each Discrimination Method

Model	Mean Area Under the Curve
Support Vector Machine	0.93 (std. 0.05)
Quackometer	0.67 (std. 0.10)

<sup>3</sup> Google uses a proprietary version of Pagerank for ranking.

<sup>4</sup> We test the assumption that Pagerank will rank web pages with *proven* claims higher than web pages with *unproven* claims.

Google	0.63 (std. 0.17) <sup>5</sup>
--------	-------------------------------

---

<sup>5</sup> The mean and standard deviation are calculated across the 8 topics rather than across the test sets of the 10 folds.

Table 2 – Web page excerpts where previous tools fail to detect unproven claims. For a page that makes unproven claims, we want a small support vector machine rank, a large quackometer score, and a large Google rank. SVM rank is calculated over 10 fold cross validation test set composed of 9 positives and 9 negatives. Google rank is out of the top 30 results returned. Quackometer score provides ranks from 0 to 10.

Failure Analysis Excerpts	Support Vector Machine Rank	Quackometer score	Google rank
I am convinced that our mind and emotions are the deciding factor in the cure of cancer.	1	1	16
The hundreds of clinical studies conducted by many competent physicians around the world, including those directed by Dr. Ernesto Contreras Rodriguez at the Oasis of Hope Hospital hospital in Mexico, give us complete confidence that there is no danger.	3	0	9
The cure shows results almost immediately and lasts three weeks only. It is cheap and affordable for everybody and proved with 138 case studies.	3	8	3
Many advanced cancer patients are petrified of their tumor. This knee-jerk reaction is caused by orthodox medicine's focus on the highly profitable (and generally worthless) process of shrinking tumors.	1	1	18
IPT (Insulin Potentation Therapy) has an outstanding 135 doctor-year track record (115 years for cancer) over 72 years, and is ready for clinical trials and widespread use.	1	0	1
We are proud of these findings, which confirm that cellular medicine offers solutions for the most critical process in cancer development, the invasion of cancer cells to other organs in the body. Conventional medicine is powerless in this.	2	1	8

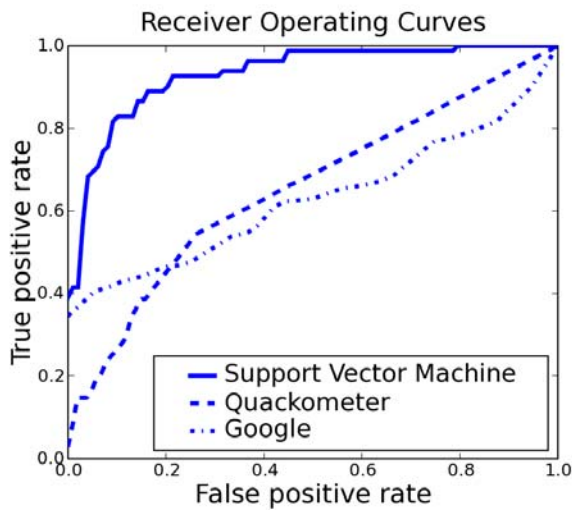


Figure 1: Receiver operating curves for each method.

## Discussion

This feasibility study showed that machine learning filter models identify web pages that make unproven claims on a select, focused gold standard. The learning filters have superior performance over the Quackometer and Google. We also note that the loose correlation between Google and high quality sites seems comparable to previous work [20-22].

This method has distinct advantages to rating instruments or manual review. First, there is no need to state explicit rating criteria. The model identified patterns in the data that label a page with unproven claims. Second, compared to the limited focus of manual review on select web pages, these models allow application to any web page.

We also highlight a subtle point in this work. We make a distinction between web pages that make unproven claims and

web pages that promote the unproven treatment. Oftentimes, there is no distinction. For this work, we only want to identify pages that make unproven claims. Pages that promote a product but do not make unproven claims are not identified. In future studies, it would be interesting to evaluate models that identify web pages that promote treatments.

In Table 2, we present excerpts from pages where the previous models failed to identify pages with unproven claims. These pages should be identified by the Quackometer and should not appear in the top 30 Google results. Failure to identify or mark these pages may result in patient’s exposure to potentially harmful, unproven treatments. In future work, we will explore potential strategies to fixing these previous models.

In practice, we envision implementing a system that works much like a spam filter works for e-mail. Spam filters identify illegitimate e-mails. In a similar fashion, we envision a system that runs on top of a search engine and flags any web pages that may have unproven health claims.

## Limitations

We tested a small sample comprised of 8 unproven treatments in 240 web pages. We will explore how well the models generalize with an independently collected dataset, more unproven treatments, and more labeled web pages. Collecting an independent dataset would allow for validation of the labeling criteria and the model selection procedures.

For this feasibility study, we purposely limited the topic of this study to cancer treatment. In the future, we will build and evaluate other models identifying web pages that make unproven claims for other conditions such as arthritis, autism, and allergies.

The comparison to Quackometer and Google is limited. Both methods are not designed to identify pages that make unproven claims. The original Pagerank algorithm relies on the link graph between web pages to rank and not the content of the page itself. Quackometer is designed to identify quack pages

and not necessarily pages that make unproven claims (though the distinction is oftentimes the same).

## Conclusions

We present a, first of its kind, feasibility study to build machine learning filter models that exhibit high discriminatory performance for identifying web pages with unproven cancer treatments. This work paves the way for building broadly applicable models involving more health conditions, more pages with unproven claims, and eventually applied systems to protect patients from quackery.

## Acknowledgments

The first author acknowledges support from NLM grant LM007948-02. The second author acknowledges support from grant LM007948-01.

## Address for correspondence

Yin Aphinyanaphongs, [ping.pong@vanderbilt.edu](mailto:ping.pong@vanderbilt.edu)

## References

- [1] Clark H. *The Cure for All Cancers*: New Century Press; 1993.
- [2] *American Heritage Dictionary*.
- [3] Hainer MI, Tsai N, Komura ST, Chiu CL. Fatal hepatorenal failure associated with hydrazine sulfate. *Ann Intern Med*. 2000 Dec 5;133(11):877-80.
- [4] See KA, Lavercombe PS, Dillon J, Ginsberg R. Accidental death from acute selenium poisoning. *Med J Aust*. 2006 Oct 2;185(7):388-9.
- [5] Bromley J, Hughes BG, Leong DC, Buckley NA. Life-threatening interaction between complementary medicines. *Ann Pharmacother*. 2005 Sep;39(9):1566-9.
- [6] Mularski RA, Grazer RE, Santoni L, Strother JS, Bizovi KE. Treatment advice on the internet leads to a life-threatening adverse reaction: hypotension associated with Niacin overdose. *Clin Toxicol (Phila)*. 2006;44(1):81-4.
- [7] Metz JM, Devine P, DeNittis A, Jones H, Hampshire M, Goldwein J, Whittington R. A multi-institutional study of Internet utilization by radiation oncology patients. *Int J Radiat Oncol Biol Phys*. 2003 Jul 15;56(4):1201-5.
- [8] Richardson MA, Sanders T, Palmer JL, Greisinger A, Singletary SE. Complementary/alternative medicine use in a comprehensive cancer center and the implications for oncology. *J Clin Oncol*. 2000 Jul;18(13):2505-14.
- [9] Sagaram S, Walji M, Bernstam E. Evaluating the prevalence, content and readability of complementary and alternative medicine (CAM) web pages on the internet. *Proc AMIA Symp*. 2002:672-6.
- [10] Ernst E, Schmidt K. 'Alternative' cancer cures via the Internet? *Br J Cancer*. 2002 Aug 27;87(5):479-80.
- [11] Health on the Net. [accessed 11-27-2006]; <http://www.hon.ch/>
- [12] Eysenbach G, Kohler C. How Do Consumers Search For and Appraise Health Information on the WWW? *BMJ*. 2002 March 9;324.
- [13] Bernstam EV, Shelton DM, Walji M, Meric-Bernstam F. Instruments to assess the quality of health information on the World Wide Web. *Int J Med Inform*. 2005 Jan;74(1):13-9.
- [14] Kim P, Eng TR, Deering MJ, Maxfield A. Published criteria for evaluating health related web sites: review. *Bmj*. 1999 Mar 6;318(7184):647-9.
- [15] Bernstam EV, Sagaram S, Walji M, Johnson CW, Meric-Bernstam F. Usability of quality measures for online health information. *Int J Med Inform*. 2005 Aug;74(7-8):675-83.
- [16] Ademiluyi G, Rees CE, Sheard CE. Evaluating the reliability and validity of three tools to assess the quality of health information on the Internet. *Patient Educ Couns*. 2003 Jun;50(2):151-5.
- [17] Walji M, Sagaram S, Sagaram D, Meric-Bernstam F, Johnson C, Mirza NQ, Bernstam EV. Efficacy of quality criteria to identify potentially harmful information. *J Med Internet Res*. 2004 Jun 29;6(2):e21.
- [18] Price SL, Hersh WR. Filtering Web pages for quality indicators: an empirical approach to finding high quality consumer health information on the World Wide Web. *Proc AMIA Symp*. 1999:911-5.
- [19] Fallis D, Fricke M. Indicators of accuracy of consumer health information on the Internet. *J Am Med Inform Assoc*. 2002 Jan-Feb;9(1):73-9.
- [20] Fricke M, Fallis D, Jones M, Luszko GM. Consumer health information on the Internet about carpal tunnel syndrome. *Am J Med*. 2005 Feb;118(2):168-74.
- [21] Griffiths KM, Tang TT, Hawking D, Christensen H. Automated assessment of the quality of depression websites. *J Med Internet Res*. 2005;7(5):e59.
- [22] Tang TT, Craswell N, Hawking D, Griffiths KM, Christensen H. Quality and Relevance of Domain-specific Search: A Case Study in Mental Health. *Info Retr*. 2006;9(2):207-25.
- [23] Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text Categorization Models for High Quality Article Retrieval in Internal Medicine. *J Amer Med Inform Assoc*. 2005;12(2):207-16.
- [24] Cancer Patients Seeking Alternative Treatments. [accessed 11-26-2006]; <http://www.quackwatch.org/00AboutQuackwatch/altseek.html>
- [25] Cohen J. A coefficient of agreement for nominal scales. *Education and Psych Measurement*. 1960;20(1):37-46.
- [26] Chang C, C. L. LIBSVM. 3-13-2006 [accessed; <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [27] Science of Quackometrics. [accessed 11-26-2006; <http://www.quackometer.net/blog/2006/04/science-of-quackometrics.html>
- [28] Quackometer. [accessed 11-26-2006]; <http://www.quackometer.net/?page=quackometer>
- [29] Brin S, Page L. The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems*. 1998;30:107-17.