A Machine Learning Approach for Rating the Quality of Depression Treatment Web Pages

Yanjun Zhang1, Hong Cui2, Jacquelyn Burkell1 and Robert E. Mercer3

1 Faculty of Information & Media Studies, The University of Western Ontario
2 School of Information Resources & Library Science, The University of Arizona
3 Department of Computer Science, The University of Western Ontario

Abstract
As health care information proliferates on the web, the content quality is varied and difficult to assess, partially due to the large volume and the dynamicity. This paper reports an automated approach in which the quality of depression treatment web pages is assessed according to evidence-based depression treatment guidelines. A supervised machine learning technique, specifically Naive Bayes classification, is used to identify the sentences that are consistent with the guidelines. The quality score of a depression treatment web page is the number of unique evidence-based guidelines covered in this page. Significant Pearson correlation (p<.001) was found between the quality rating results by the machine learning approach and the results by human raters on 31 depression treatment web pages in this case study. The semantic-based, machine learning quality rating method is promising and it may lead to an efficient and effective quality assessment mechanism for health care information on the Web.

Keywords: machine learning, Naive Bayes classification, quality assessment, web health care information, evidence-based health care guidelines


Copyright: Copyright is held by the authors.

Acknowledgements: We thank Maggie Chu and Chunli Men who provided evidence-based quality rating on the training and testing web pages. We would also like to thank the anonymous reviewers for their valuable comments.

Contact: yzhan266@uwo.ca, hongcui@email.arizona.edu, jburkell@uwo.ca, mercer@csd.uwo.ca

1 Introduction
The last decade has witnessed a dramatic expansion in the amount of publicly available health care information on the World Wide Web, and the use of web health care information has become popular among both health care professionals and patients. Pew Institutes national surveys since 2003 show that over 80% of online users in the United States look for advice or information about health or health care (Pew Internet and American Life Project, 2003-2011) despite 33% increase in Web users (Internet News, 2003; Internet World Stats, 2012).

Although online health care information is widely accessed, the quality of health care information on the web is varied in terms of accuracy, coverage and currency (Eysenbach et al. 2002; Kunset et al. 2002; Griffiths & Christensen, 2005). Surprisingly, information consumers themselves make little effort to verify the information (Eysenbach & Köhler, 2004, Pew Internet and American Life Project, 2006). Misinformation on the web could cause and indeed has caused life-threatening accidents (Crocco et al, 2002; Kiley, 2002). Because of the potential harm that may be caused by inaccurate information, the quality assessment of health care information on the web stays a common interest of various health care information stakeholders, including e-health policy makers, information providers/consumers, and information search service providers.

The large amount of health care information on the web can easily overwhelm the capacity of any manual evaluation system. Automated quality assessment mechanism is therefore in need. In particular, we are interested in quality assessment based on evidence-based health care guidelines.
Evidence-based medicine has been advocated in health care since the original model was presented in the Journal of the American Medical Association (EBMWG, 1992). Many evidence-based clinical practice guidelines have been established under the sponsorship of governmental agencies such as the Agency for Healthcare Research and Quality (AHRQ, 2011) in the United States. The guidelines are established based on the systematic review of scientific evidence in health care and medical literature by multi-disciplinary panel including methodologists, medical experts and scientific reviewers.

In this paper, we propose and evaluate a supervised machine learning approach to rate the information quality of health care web pages based on their content. In our approach, quality of health care web pages is assessed through semantically comparing the text content with evidence-based health care practice guidelines. Content accuracy of a web page is assessed through looking for positive matching between the web page content and any of the health care guidelines; the content coverage is assessed by identifying the number of guidelines covered by this web page. According to a systematic review (Eysenbach et al., 2002) based on 79 distinct health information quality evaluation studies, accuracy and coverage are two most commonly used measures for assessing information quality from the content perspective. We have developed two semantic-based quality rating approaches: one is a knowledge engineering based approach, and the other is a supervised learning method. The former is reported elsewhere (Zhang et al., 2013). We report the latter in this paper. As a case study to proof the concept, the health care guidelines on depression treatment (Appendix C) are used in this research.

This paper is structured as follows. After describing the method, data, and experimental design, we report evaluation results and discuss the results and limitations of the research. A comparison of this work with other web information quality assessment research is then presented. We conclude the paper with plans for future work.

2 Method

We cast the quality rating problem as a sentence classification problem. If a sentence in a web page is consistent with a depression treatment guideline, it is considered as a match (i.e. “positive”) and the quality score of the web page is increased by 1. A Naïve Bayes (NB) classifier is trained to perform the classification on a semantic representation of the sentences.

Naïve Bayes Classifier: A Naïve Bayes classifier is a supervised classification algorithm. The instance to be classified (i.e., test instances) are represented as a vector of features. The training instances are the vectors with class labels, which in our case are “positive” and “negative”. Through a training process, a Naïve Bayes classifier constructs a probabilistic model that can be used to classify new input instances. The training and test instances are sentences from web pages, and each sentence is represented as a set of features (also called “attributes”). When constructing the probability model, a Naïve Bayes classifier assumes that all features are independent of each other given the context of the class (i.e., “conditional independence”). Because of the independence assumption, the parameters for each feature can be learned separately, and this greatly simplifies learning, making the algorithm efficient on large feature spaces. The Naïve Bayes model has many variations. In this study, we used the multivariate Bernoulli event model that is implemented in WEKA (Witten et al., 2011). The same model has been used for text classification in numerous studies (McCallum & Nigam, 1998; Billsus & Pazzani, 1999; Schneider, 2003; Chen et. al., 2009).

Feature Space Construction: Like many other supervised classification applications, selecting the features to represent the instances for Naïve Bayes classifier is arguably the most critical aspect of the algorithm design. Our feature space construction module carries out the following procedure to construct the feature space:
Semantic tagging of original sentences. This step transforms original sentences to vectors of semantic tags. These semantic tags are candidate features.

Feature space reduction.

- Remove semantic tags representing numerical values or web URLs (i.e., tags containing “www”), as they are not relevant to quality assessment (See Table 1 for some examples).
- Further prune the feature dimensions in a vector space.

<table>
<thead>
<tr>
<th>Original Text (underlined part)</th>
<th>Semantic Tag Result (enclosed by square bracket)</th>
<th>POS Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>telephone: 805-967-7636</td>
<td>[805], [967], [7636]</td>
<td>number</td>
</tr>
<tr>
<td>by the 10th postnatal day</td>
<td>[10th]</td>
<td>number</td>
</tr>
</tbody>
</table>

Table 1: Examples of noisy tag to be removed

Although the semantic representation of a sentence includes three types of data: semantic tags, POS tags, and term positions, this study only uses semantic tags to form the feature space for the Naïve Bayes classifier. Other data will be utilized in future studies to improve classification performance. Next, we describe the semantic tagging and feature dimension reduction in vector space in detail.

**Semantic Tagging of Sentences:** A naïve application of NB is often applied on text represented as “bag of words”. In this approach, instead of using the original sentences, a semantic representation of a sentence is created by using three tools, MetaMap API, TaggerClient (v2.4.c) and Lexical Variant Generator (LVG) (McCray et al. 1994; National Library of Medicine, 2009). Semantic tagging is carried out after a sentence cleaning process where citation notations and other extra text and symbols are removed.

MetaMap API tags nouns and noun phrases with their matching semantic concepts in UMLS Metathesaurus (UMLS, 2009). Each successful match comes with a confidence score. For example the nouns “depression” and “depressive illness” are both mapped to a semantic concept “depressive disorder” with a score of 1000. TaggerClient is a part of speech tagger for biomedical domain. LVG is a lemmatization tool. Both are used to control lexical variations and map a word to a preferred synonym. For instance, “ceases” and “cease” both have the preferred synonym “stop”.

As MetaMap API focuses on medical concepts expressed as nouns and noun phrases, TaggerClient and LVG are used in addition to control the lexical variations of verbs, adjective, adverbs, and the nouns that are not tagged by MetaMap API. LVG tags of other word groups such as articles, prepositions, etc. are relatively less useful for shallow semantic analysis purposes and hence are ignored in this study. In addition, when a noun or noun phrase has both a MetaMap tag and a LVG tag, the MetaMap tag is taken only when the confidence score is greater than 850 (set empirically); otherwise, the LVG tag is taken as the semantic tag. After semantic tagging, a sentence can be transformed from its original textual representation to a vector of semantic tags.

Feature Dimension Reduction in a Vector Space: We observe that a depression treatment web page may discuss other aspects of depression, for example, causes, self-diagnosis, research groups, or useful resources. Features associated with these aspects are less useful for identifying positive sentences. Further, due to the content variations across web pages, these features also more likely to have a value of 0 in many sentence vectors, making these features also less useful for identifying negative sentences. Removing these dimensions would address the sparse feature space problem, improve parameter estimation in Naïve Bayes classifiers (Kim et. al., 2005), and reduce computational cost.
The goal of feature dimension reduction is to identify and remove these less useful features. This is achieved through the following measures:

a) Project vectors of training instances onto a multi-dimensional vector space. The number of dimensions of the vector space is determined by the number of unique semantic tags in all the sentence vectors after the removal of numerical and web URL tags.

b) For each sentence vector, measure its cosine similarity with all positive training instances, and take the maximal similarity as the score for this vector.

c) Sentence vectors with a similarity score greater than 0.5 (threshold set empirically) are selected to form a reference group. Features with a value of 0 across all vectors in the reference group are removed.

Feature space reduction is effective. For depression treatment guideline #1, the dimension size was reduced from 3963 to 635. A manual review of these removed dimensions confirmed that their corresponding semantic concepts are semantically unrelated to depression treatment. Some typical examples of these concepts are cancer, clinic, university, etc.

**Quality Score of Web Pages:** A Naïve Bayes classifier is trained for each depression treatment guideline using the corresponding training examples. The training instances take the form of vectors of semantic tags. The number of trained classifiers equals the number of the guidelines. The trained classifiers are then applied on each of the test instances one by one. Test instances are in the same format as the training example, except that test instances do not have class labels. The classifiers will classify the test instance as either Positive or Negative. If it is positive, the webpage containing the sentence scores one. The quality score for any webpage ranges from 0 to the number of guidelines. Figure 1 shows the logic flow of the system.
3 Data

The corpus for this study comprised a total of 201 web pages on the topic of depression treatment (Appendix A). The sample data were obtained from multiple sources (Table 2) in May 2009. For search engines, “depression treatment” was used as the query and the first 30 returned web pages from each search engine were collected as candidates. For web portals, candidate pages were collected from depression treatment related sections only. Candidate pages were examined manually to remove duplicate pages and pages that were inappropriate for other reasons (see Appendix B). In the end, 201 web pages were selected to form the corpus.
Table 2: Sources for constructing the corpus

**Training and test data:** Two human raters were hired to identify positive and negative sentences in the 201 web pages. In addition, they were required to highlight in the positive sentences the key phrases that lead them to a positive identification. A five-hour rating workshop was held for the human raters to learn the evidence-based depression treatment guidelines. The intra-class correlation coefficient (ICC) between the webpage quality scores assigned by two raters across all guidelines was .990, with the 95% confidence interval between .979 and .995, as measured by the single measure ICC value i.e. ICC(3,1). The discrepancies were resolved through discussion between the raters.

The quality of each of the 201 pages was rated by the human raters using the health care guidelines. The score of a page is the number of unique guidelines reflected by the page. The scores ranged from 0 to 8. The 201 web pages were divided into 5 bins (i.e., those with a quality score of 0, 1-2, 3-4, 5-6 or 7-8). Stratified random sampling was used to select 31 test web pages (2677 sentences in total). The remaining 170 pages were used as the training set.

**Depression Treatment Guidelines 1, 6 and 12-B:** This research used a subset of the 20-item evidence-based depression treatment guidelines used in Griffiths and Christensen (2005). Appendix C lists these guidelines. When a guideline contains multiple semantic propositions, it is necessary to split it into multiple guidelines. For instance, the guideline #12 “abrupt cessation of antidepressant can cause discontinuation syndrome and that antidepressants should not be stopped suddenly” says

a) Antidepressant should not be stopped suddenly.

b) Abrupt cessation of antidepressant can cause discontinuation syndrome.

Since it is possible for only one of the two points to be mentioned in a sentence, guidelines like this can potentially cause discrepancy among human raters when creating training examples and reviewing test results. To avoid this problem, guideline #12 was split into guideline 12-A and 12-B.

Due to the skewed distribution of the guidelines in the corpus, some guidelines (e.g. #18, 19) end up with too few positive examples (n < 5) for training the Naïve Bayes classifier. However, we were able to select three guidelines, #1, 6, and 12-B, of varied semantic complexity to test the proposed approach. These guidelines have minimum 50 positive training examples found by human raters from different web pages, so the positive data set size is reasonably large for training classifier.

**4 Results**

After the Naïve Bayer classifiers were trained with the 170 web pages, they were used to classify the 2677 sentences in the 31 test pages. In this section, we report the classifiers’ sentence classification performances, as well as the web page quality rating performance.

Table 3 lists the performance of the machine learning approach for each individual guideline as measured by precision, recall, and accuracy. The following equations are used for calculating these measures.
TP stands for the number of true positive (cases) identified by classifier; FP stands for false positives identified by classifier; FN stands for false negatives identified by classifier; and TN stands for true negatives identified by classifier.

**Equation 1:** Precision = the proportion of true positives (TP) over tested positives  
= TP / (TP + FP)

**Equation 2:** Recall = the proportion of true positives (TP) over actually positives  
= TP / (TP + FN)

**Equation 3:** Accuracy = the proportion of correctly identified sentences over all sentences  
= (TP + TN) / (TP + FN + FP + TN)

<table>
<thead>
<tr>
<th>Depression treatment guideline</th>
<th>Human Classification</th>
<th>Machine Learning Classification (Y)</th>
<th>Machine Learning Classification (N)</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Y</td>
<td>42</td>
<td>7</td>
<td>85.7%</td>
<td>13.7%</td>
<td>89.9%</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>263</td>
<td>2365</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#6</td>
<td>Y</td>
<td>16</td>
<td>3</td>
<td>84.2%</td>
<td>76.2%</td>
<td>99.7%</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>5</td>
<td>2653</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#12-B</td>
<td>Y</td>
<td>11</td>
<td>2</td>
<td>84.6%</td>
<td>28.9%</td>
<td>98.9%</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>27</td>
<td>2637</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Performance of sentence classification by machine learning approach

<table>
<thead>
<tr>
<th>Testing Page ID</th>
<th>Quality Score via Human Rating</th>
<th>Quality Score via Machine Learning Rating</th>
<th>Quality Score Difference (machine learning vs. human rating)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>44</strong></td>
<td><strong>55</strong></td>
<td></td>
</tr>
</tbody>
</table>

Note: The quality score was assigned based on guideline #1, #6, and #12-B only.

* The quality score difference = quality score via machine learning - quality score via human rating.

**Table 4: Quality score assigned to testing web pages for guideline #1, #6, and #12-B**

The quality scores generated by the machine learning approach using guidelines #1, 6 and 12-B are listed in Table 4. The quality scores range from 0 to 3 (i.e., a page may match 0 to 3 guidelines), and a total of 55 occurrences of the guidelines were identified from the 31 test pages. Figure 2 shows the overlap between the machine learning results and human rating results.

---

**Figure 2: Identified depression treatment guidelines (#1, #6, and #12-B)**

The linear correlation between machine learning based quality scores and the evidence-based human rating quality scores was high and statistically significant ($r = 0.841$, $r^2 = 0.707$, $p < .001$, $n = 31$, see Figure 3).
Discussion

The sentence classification results (Table 3) show that for all three guidelines, the recalls are above 84%. This suggests that the machine learning approach can effectively identify the sentences reflecting these guidelines, despite the variations in natural language guideline expressions. Fairly complex sentences such as those presented in Figure 4 were correctly identified as positive in different test pages.

Figure 3: Relationship between machine learning quality rating scores and human rating quality scores

5 Discussion

The sentence classification results (Table 3) show that for all three guidelines, the recalls are above 84%. This suggests that the machine learning approach can effectively identify the sentences reflecting these guidelines, despite the variations in natural language guideline expressions. Fairly complex sentences such as those presented in Figure 4 were correctly identified as positive in different test pages.
Guideline #1:
“Antidepressant medication is an effective treatment for major depressive disorder.”

Matched sentence:
1. SSRI s affect mainly serotonin and have been found to be effective in 
treating depression and anxiety without as many side effects as some older 
antidepressants.

Guideline #6:
“The side effect profile varies for different antidepressants.”

Matched sentences:
2. SSRIs and SNRIs are more popular than the older classes of antidepressants, 
such as tricyclics—named for their chemical structure—and monoamine oxidase 
inhibitors (MAOIs) because they tend to have fewer side effects.
3. Side effects may vary depending on the medicine you take, but common ones 
include stomach upset, loss of appetite, diarrhea, feeling anxious or on 
eds, sleep problems, drowsiness, loss of sexual desire, and headaches.
4. However, because TCAs tend to have more numerous and more severe side 
effects, they’re often not used until you’ve tried SSRIs first without an 
improvement in your depression.
5. The side effects vary depending on the type of antidepressant you take.

Figure 4: Examples of correctly classified sentences

These cases demonstrate that the machine learning system is able to successfully map text expressions to 
semantic concepts, including “SSRI” - “antidepressant”, “treating” - “treat”, and “depression” - “major 
depressive disorder”. In addition, these examples include two different ways for expressing the meaning of 
guideline #6. One says directly that side effects “vary” depending on antidepressants; the other one 
indicates variation by a discussion of “fewer/more” side effects between antidepressants. In both cases, the 
machine learning approach successfully identified that the sentences are in concordance with the rating 
guideline #6.

While all the guideline occurrences identified by human raters were successfully identified by the 
machine learning approach, the machine learning approach identified 11 false positive occurrences (Figure 
2). False positive errors may occur when the semantics of a text segment is taken for the entire sentence. 
Because the sentence contains “your response to certain antidepressant”, the classifier mistakenly classified 
the sentence in Figure 5 as a match for guideline no. 1. This partially explains the low precision.

Another limitation of the current implementation is that it uses individual sentences as the 
processing unit. Because of this, guidelines that are expressed across multiple sentences or as part of a bullet 
list in a web page are challenging to the Naïve Bayes classifiers. Although these cases did not occur in this 
experiment we acknowledge that they can lead to false negative results.

Guideline #1:
“Antidepressant medication is an effective treatment for major depressive disorder.”

False Positive Match:
10. The test, called the cytochrome P450, helps pinpoint genetic factors that 
influence your response to certain antidepressants (as well as some other 
medications).

Figure 5: A false positive example
The low precision and imperfect recall of sentence classification did not seem to have greatly affected the page rating performance. This is because a single guideline is commonly paraphrased more than once in a web page. So, suppose the machine identified five sentences as a positive instance of a guideline, as long as one of them is a true criterion, the impact of false positives would not been reflected in the machine-assigned quality rating score. Table 4 shows that for 21 of the 31 pages (67.7%) the machine learning quality scores and the human rating quality scores were identical. In 9 pages (29.0%) the machine learning quality scores were one higher than the human rating quality scores. For one page (3.2%) the difference between the machine learning and human rating quality scores was greater than one (2 in test page #28). The high linear correlation between human and machine ratings suggests that the proposed approach has the potential to evaluate the quality of online health care information (Figure 3).

6 Literature Review

Much research has been done studying quality dimensions of web health care information (e.g. Bopp & Smith, 2000), establishing quality rating codes and indicators (e.g. Eysenbach et al., 2002; Griffiths & Christensen, 2002; URAC, 2007), and exploring automated rating mechanisms (Griffiths et al., 2005; Hawking et al., 2007; Wang & Liu, 2007).

Earlier studies explored using accountability metadata of web pages such as web page authorship, site sponsorship, and disclosure of editorial board of web site as quality indicators (e.g. Chen et al., 2000; Smith, 2002; Barnes et al., 2009, etc.). This type of indicators asks the questions of how the information was presented or what meta-information was provided (Eysenbach et al., 2002). The correlation between information quality and web site/document linkage patterns such as hyperlinks (e.g. in-link counts to a website and Google’s page rank of site home page) has also been explored (e.g. Frické et al., 2005; Griffiths & Christensen, 2005). These quality indicators seem to be domain independent; however, researchers have found that the association between these indicators and the content quality of web health care information is inconsistent in different health care subjects, putting the validity and reliability of these non-content based indicators in question (e.g. Frické & Fallis, 2002; Frické et al., 2005; Griffiths et al., 2005). The major difference between our work from these studies is that our approach assesses the quality of information content directly and not through secondary attributes such as the disclosure of authorship and domains. The content-based quality rating fills the gap in information quality assessment and complements other rating criteria nicely.

The work reported here was inspired by Griffiths and Christensen (2002), where the authors adopted a set of evidence-based depression treatment rules published by the Centre for Evidence-based Mental Health at Oxford (CEBMH, 1998) as the quality rating standard in their study. In their study, human raters used this standard to rate the quality of depression websites. The quality of a website was measured by the number of different treatment rules reflected in the website content. The greater the number, the higher the quality score a site was assigned. Griffiths and Christensen (2002) suggested that the rating scores generated using evidence-based treatment guidelines were highly correlated (r=0.96, p<.001) with the quality scores of subjective rating completed by health professionals (Griffiths & Christensen, 2002). Using the same set of guidelines, Griffiths and Christensen (2005) proposed an automated (website) quality rating approach based on information retrieval techniques.

Our work differs from Griffiths & Christensen (2005) in that 1) our quality rating is implemented based on matching text semantics while theirs is on keyword distribution and 2) our analysis is conducted at sentence level with a finer granularity than web document level. Because of the use of semantics, our approach is able to identify specific health care guidelines as presented in a web page. Consequently, the quality scores assigned to web pages can not only support ranking, but also justify the scoring results in a user understandable manner. In addition to quality score, it is also important that the clear indication of which exact health care guidelines are presented in web pages can be extra assistance for information
consumers. They can select their reading based on not only quality score, but also the extent to which that the web page content fit the users’ individual information need (e.g. focusing on medication treatment instead of psychotherapy) and knowledge background. Looking into the future, the semantic information extracted can contribute to the Linked Data clouds and the Semantic Web.

7 Conclusion and Future Research

We reported a prototype Naive Bayes based machine learning system that rates the quality of health care web pages according to the evidence-based health care guidelines. As a proof of concept, we used depression treatment for a case study.

Our experimental results suggest that the semantics-based quality rating approach can produce quality score results comparable to human rating results. This is achieved by having computer programs to conduct shallow semantic analysis on each sentence in depression treatment web pages, and then use the semantic tags of training sentences to develop classifiers’ capability to identify the sentences that are in concordance with the health care guidelines. The identification of guideline-conforming sentences is treated as a binary classification problem. The classification and page rating performances (Tables 3 and 4, Figure 3) attest to the effectiveness of the automatic quality rating approach.

There are some limitations of this study. First, the automated quality rating system is tested for only 3 of the 20 guidelines for the evidence-based depression treatment. It would be interesting to testify the performance of the semantics-based approach by using more guidelines for quality rating. Secondly, due to the limit of research resources, the size of training and testing data set is relatively small to allow the study on this prototype system controllable. A wider scope of dataset, ideally also including other health care subjects, would further help proving the generalizability of the semantics-based approach. Additionally, the classifier performance can be evaluated with a more robust method such as cross-validation.

In the future, other than overcoming the above limitations, we will further improve the sentence classification performance, especially for the low precision. Some sensitivity analysis can be done for optimizing trade-off between precision and recall. In the health care quality rating context, a practical classifier probably need to be designed to be more conservative as false positives may cause worse consequence than false positives, e.g. leading someone to trust a low quality web page. In future design, the classifiers do not necessarily take the semantic components only into account, but also apply some constraints to screen out negative cases. For example, adding proximity constraints between a pair of semantic tags and taking into consideration of negations in those anti-guideline cases. These may be added to the feature space of the classifiers or used in other ways to reduce false positives. We will also work to identify statements that contradict with approved health care guidelines.

8 References


9 Table of Figures

Figure 1: Process flow charts for semantic tagging & Naïve Bayes classification........................................196
Figure 2: Identified depression treatment guidelines (#1, #6, and #12-B)..................................................199
Figure 3: Relationship between machine learning quality rating scores and human rating quality scores 200
Figure 4: Examples of correctly classified sentences ..................................................................................201
Figure 5: A false positive example .............................................................................................................201

10 Table of Tables

Table 1: Examples of noisy tag to be removed ..........................................................................................194
Table 2: Sources for constructing the corpus .........................................................................................197
Table 3: Performance of sentence classification by machine learning approach ....................................198
Table 4: Quality score assigned to testing web pages for guideline #1, #6, and #12-B ..........................199
## Appendix A: The corpus: 201 web pages

<table>
<thead>
<tr>
<th>ID</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://www.mayoclinic.com/health/depression/DS00175/DSECTION=treatments-and-drugs">http://www.mayoclinic.com/health/depression/DS00175/DSECTION=treatments-and-drugs</a></td>
</tr>
<tr>
<td>2</td>
<td><a href="http://www.emedicinehealth.com/depression/page5_em.htm">http://www.emedicinehealth.com/depression/page5_em.htm</a></td>
</tr>
<tr>
<td>3</td>
<td><a href="http://www.emedicinehealth.com/depression/page6_em.htm">http://www.emedicinehealth.com/depression/page6_em.htm</a></td>
</tr>
<tr>
<td>4</td>
<td><a href="http://www.emedicinehealth.com/depression/page7_em.htm">http://www.emedicinehealth.com/depression/page7_em.htm</a></td>
</tr>
<tr>
<td>6</td>
<td><a href="http://www.mentalhealth.com/rx/p23-md01.html">http://www.mentalhealth.com/rx/p23-md01.html</a></td>
</tr>
<tr>
<td>8</td>
<td><a href="http://www.waldenbehavioralcare.com/depression_treatment.asp">http://www.waldenbehavioralcare.com/depression_treatment.asp</a></td>
</tr>
<tr>
<td>9</td>
<td><a href="http://www.healthlinkbc.ca/kbase/topic/special/hw30709/sec1.htm">http://www.healthlinkbc.ca/kbase/topic/special/hw30709/sec1.htm</a></td>
</tr>
<tr>
<td>10</td>
<td><a href="http://www.depressioncenter.org/treatments/cbt.asp">http://www.depressioncenter.org/treatments/cbt.asp</a></td>
</tr>
<tr>
<td>11</td>
<td><a href="http://www.depressioncenter.org/treatments/meds.asp">http://www.depressioncenter.org/treatments/meds.asp</a></td>
</tr>
<tr>
<td>12</td>
<td><a href="http://www.depressioncenter.org/treatments/default.asp">http://www.depressioncenter.org/treatments/default.asp</a></td>
</tr>
<tr>
<td>14</td>
<td><a href="http://www.medpagetoday.com/Psychiatry/Depression/14476">http://www.medpagetoday.com/Psychiatry/Depression/14476</a></td>
</tr>
<tr>
<td>17</td>
<td><a href="http://www.healthlinkbc.ca/kbase/topic/special/hw30709/sec10.htm">http://www.healthlinkbc.ca/kbase/topic/special/hw30709/sec10.htm</a></td>
</tr>
<tr>
<td>18</td>
<td><a href="http://www.webmd.com/anxiety-panic/features/alternative-depression-treatment-risks">http://www.webmd.com/anxiety-panic/features/alternative-depression-treatment-risks</a></td>
</tr>
<tr>
<td>19</td>
<td><a href="http://en.wikipedia.org/wiki/Clinical_depression">http://en.wikipedia.org/wiki/Clinical_depression</a></td>
</tr>
<tr>
<td>20</td>
<td><a href="http://www.mayoclinic.org/depression/treatment.html">http://www.mayoclinic.org/depression/treatment.html</a></td>
</tr>
<tr>
<td>23</td>
<td><a href="http://online.wsj.com/article/SB119128055574245655.html?mod=health_home_stories">http://online.wsj.com/article/SB119128055574245655.html?mod=health_home_stories</a></td>
</tr>
<tr>
<td>24</td>
<td><a href="http://www.medicalnewstoday.com/articles/81578.php">http://www.medicalnewstoday.com/articles/81578.php</a></td>
</tr>
<tr>
<td>26</td>
<td><a href="http://www.depression-guide.com/">http://www.depression-guide.com/</a></td>
</tr>
<tr>
<td>27</td>
<td><a href="http://psychologyinfo.com/depression/treatment.htm">http://psychologyinfo.com/depression/treatment.htm</a></td>
</tr>
<tr>
<td>30</td>
<td><a href="http://depressionandanxietyhelp.com/depression-treatment.html">http://depressionandanxietyhelp.com/depression-treatment.html</a></td>
</tr>
<tr>
<td>31</td>
<td><a href="http://www.helpguide.org/mental/medications_depression.htm">http://www.helpguide.org/mental/medications_depression.htm</a></td>
</tr>
<tr>
<td>32</td>
<td><a href="http://helpguide.org/mental/treatment_strategies_depression.htm">http://helpguide.org/mental/treatment_strategies_depression.htm</a></td>
</tr>
<tr>
<td>33</td>
<td><a href="http://safedepressiontreatment.com/">http://safedepressiontreatment.com/</a></td>
</tr>
<tr>
<td>34</td>
<td><a href="http://www.depressiontreatmentworks.org/">http://www.depressiontreatmentworks.org/</a></td>
</tr>
<tr>
<td>35</td>
<td><a href="http://thedepressiontreatment.com/antidepressants/index.htm">http://thedepressiontreatment.com/antidepressants/index.htm</a></td>
</tr>
<tr>
<td>36</td>
<td><a href="http://www.depression.com/treatment_tips.html">http://www.depression.com/treatment_tips.html</a></td>
</tr>
<tr>
<td>37</td>
<td><a href="http://depression-assistance.com/2006/07/30/depression-treatment/">http://depression-assistance.com/2006/07/30/depression-treatment/</a></td>
</tr>
<tr>
<td>38</td>
<td><a href="http://depressiontreatment.net/">http://depressiontreatment.net/</a></td>
</tr>
<tr>
<td>39</td>
<td><a href="http://www.depression-guide.com/treatment-of-depression.htm">http://www.depression-guide.com/treatment-of-depression.htm</a></td>
</tr>
<tr>
<td>40</td>
<td><a href="http://www.depression-help-treatment.com/depression-medication.html">http://www.depression-help-treatment.com/depression-medication.html</a></td>
</tr>
<tr>
<td>41</td>
<td><a href="http://www.depressiontreatment-help.com/depression-treatment/depression-treatment.htm">http://www.depressiontreatment-help.com/depression-treatment/depression-treatment.htm</a></td>
</tr>
</tbody>
</table>

206
http://health.yahoo.com/depression-treatment/depression-treatment-overview/healthwise-
aa25747.html
http://www.depressiontreatmenthelp.org/depression_treatment.php
http://www.psychologyinfo.com/depression
www.psychologyinfo.com/depression/treatment.htm
http://mayoclinic.com/health/depression/DS00175/DSECTION=treatments-and-drugs
http://psychcentral.com/health/depression/nil/006/depression-treatment/all/1/
http://www.drweil.com/drw/u/ART00696/depression-treatment
http://www.mayoclinic.com/health/treatment-resistant-depression/DN00016
http://www.healthlinkbc.ca/kbase/dp/topic/ty6745/dp.htm
http://en.wikipedia.org/wiki/Depression_and_natural_therapies
http://www.aboutourkids.org/families/disorders_treatments/az_disorder_guide/depression/treatment
http://www.cancer.gov/cancertopics/pdq/supportivecare/depression/Patient/page4
http://www.cancer.gov/cancertopics/pdq/supportivecare/depression/Patient/91.cdr
http://www.healthlinkbc.ca/kbase/topic/detail/drug/hw29716/detail.htm
http://www.ahrq.gov/clinic/epcsums/deprsumm.htm
http://www.oas.samhsa.gov/2k8/depression/depressionTX.cfm
http://www.mayoclinic.com/health/depression-and-aging/MY00259
http://www.healthline.com/adamcontent/depression-1
http://www.nami.org/Template.cfm?Section=About_Treatments_and_Supports&template=
http://netwellness.org/healthtopics/depression/depressiontreatment.cfm
http://www.med.unic.edu/depression/treatment.htm
http://yourtotalhealth.ivillage.com/depression-treatment-hasnt-worked.html
http://www.realmentalhealth.com/alternatives/pleasant_activities.asp
http://www.dukenews.duke.edu/2000/02/mm_depressiontreatment.html
http://www.wdxcyber.com/psychotherapy.html
http://pibhs.uams.edu/Depression/Depression_treatment.asp
http://www.suicideandmentalhealthassociationinternational.org/depressiontreat.html
http://allcare.net/s2/anxiety.php
http://www.healthlinkbc.ca/kbase/topic/major/ty4640/descrip.htm
http://www.healthlinkbc.ca/kbase/topic/major/ty4640/trtovr.htm
http://www.webmd.com/depression/understanding-depression-treatment
http://www.webmd.com/depression/guide/depression-treatment-options
http://www.webmd.com/depression/postpartum-depression/understanding-postpartum-depression-treatment
http://www.webmd.com/depression/psychotherapy-treat-depression
http://www.webmd.com/depression/treating-depression-medication
http://www.webmd.com/depression/pediatric-prozac
http://www.webmd.com/depression/news/20080221/hope-may-take-time-after-depression
http://www.webmd.com/depression/continuum-care-treatment-resistant-depression
http://www.webmd.com/depression/guide/treatment-resistant-depression-psychotherapy
http://www.webmd.com/depression/experimental-treatments-depression
http://www.healthlinkbc.ca/kbase/dp/topic/ty6886/dp.htm
http://www.healthwise.net/moh/print/PrintTableOfContents.aspx?token=moh&localization=en
http://www.healthwise.net/moh/print/PrintTableOfContents.aspx?token=moh&localization=en
http://www.healthlinkbc.ca/kbase/topic/detail/drug/zp2718/detail.htm
http://www.healthlinkbc.ca/kbase/topic/major/tn9653/descrip.htm
http://www.webmd.com/depression/medication-options
http://www.webmd.com/depression/news/20081201/which-kids-need-antidepressants
http://www.webmd.com/depression/ssris-myths-and-facts-about-antidepressants
http://www.healthlinkbc.ca/kbase/topic/major/tn9653/trtovr.htm
http://www.healthlinkbc.ca/kbase/topic/major/tn9653/drugtrt.htm
http://www.webmd.com/depression/guide/optimizing-depression-medicines
http://www.webmd.com/depression/guide/chronic-illnesses-depression
http://www.healthlinkbc.ca/kbase/topic/major/tn9653/othertrt.htm
http://www.healthlinkbc.ca/kbase/topic/major/tn9653/hometrt.htm
http://www.webmd.com/depression/guide/depression-support
Table A: URL of depression treatment web page samples

12 Appendix B: Criteria for Sampling Depression Treatment Web Pages

Web pages of the following nature are removed
- pages which focus on other diseases instead of depression, or pages that address depression, but discuss only non-treatment topics such as diagnosis; ***determination was based on document heading & sub-heading.***
- pages protected by password.
- pages not in text format (e.g. video/audio clips, PPT slides).
- pages with tables or spread sheets as major part of page content.
- portal pages which do not have their own content, but just hyperlinks referring to other relevant pages. (e.g. URL menus/categories, list of search returns from search engines)
- pages which have article titles or bibliographic information only
- home pages of business or organizations (e.g. medical center or depression clinic)
- pages too long for human rating (e.g. online books or chapters) --- they were filtered out due to the consideration of human rating expense.
- advertisement pages which do not really provide depression treatment content, such as Amazon book advertisement
- academic articles which are targeted for professional audience instead of public online users --- due to academic complexity, some very specific research questions and terminologies can make the articles not quite understandable for most common users and human raters to conduct rating.

13 Appendix C: Evidence-based Depression Treatment Guidelines

Evidence-based Depression Treatment Guidelines Used in (Griffiths & Christensen, 2005)

Evidence-Based Rating Scale for Human Raters (Copied from (Griffiths & Christensen, 2005))

The evidence-based rating scale was developed from statements in the treatment section of A systematic guide for the management of depression in primary care published by the Centre for Evidence-based mental health, Oxford (CEBMH, 1998).

1. Antidepressant medication is an effective treatment for major depressive disorder.
2. Antidepressants are all equally effective.
3. The effectiveness of antidepressants is around 50 to 60%.
4. Full psychosocial recovery can take several months.
5. Drop out rate is same for different antidepressants.
6. The side effect profile varies for different antidepressants.
7. The choice of antidepressant should depend on individual patient factors (e.g. presence of comorbid psychiatric or medical conditions, previous response to a particular drug, patient preference regarding the desirability of specific side-effects, concurrent drug therapy, suicidal risk)
8. Antidepressants are not addictive.
9. A trial of 6 weeks at full dose is needed before a drug can be considered to have failed and another tried.
10. A second-line drug should probably be from a different class of antidepressant.
11. Once improved continuation treatment at the same dose for at least 4-6 months should be considered.
12. Discontinuation syndrome may occur with abrupt cessation of any antidepressant so antidepressants should not be stopped suddenly. Where possible antidepressants should be withdrawn over a 4 week period, unless there are urgent medical reasons to stop the drug more rapidly. [To score 1, need to make general points that abrupt cessation can cause discontinuation syndrome and that antidepressants should not be stopped suddenly]*
13. St John’s Wort appears to be as effective as tricyclic antidepressants and causes fewer side effects, but little is known about any long term adverse effects.**
14. Cognitive therapy can be an effective treatment for depression.
15. Cognitive behaviour therapy is at least as effective as drug treatment in mild-to-moderate depression.
16. Cognitive behaviour therapy may be valuable for people who respond to the concept of Cognitive behaviour therapy, prefer psychological to antidepressant treatment, have not responded to antidepressant therapy. [Score 1 if mention at least one of these]

17. Problem-solving may be effective for depression.

18. [Generic] counselling is probably no more effective than treatment as usual from the GP for depression.

19. Written information (usually based on a cognitive model of depression) can improve mild-to-moderate depression. [Score 1 if cognitive model]

20. Exercise can be effective - alone or as an adjunct to other treatments.

For each item, score 1 if the site information is consistent with the statement. Cumulate item scores across the scale to yield a total evidence-based score for the site.

*, ** Guidelines 12 and 13 each contain multiple “meaning pieces”. They are split into multiple guidelines in this study (see reasons in the article body).

12-A. Antidepressants should not be stopped suddenly.

12-B. Abrupt cessation can cause discontinuation syndrome.

13-A. St John’s Wort appears to be as effective as (tricyclic) antidepressants.

13-B. St John’s Wort causes fewer side effects than (tricyclic) antidepressants.

13-C. Little is known about any long term adverse effects of St John’s Wort.