Content-based Assessment of the Credibility of Online Healthcare Information

Meeyoung Park, Hariprasad Sampathkumar, Bo Luo
Electrical Engineering and Computer science
University of Kansas
Lawrence, U.S.A.
{*mpark, hariprsd, bluo*}*@ku.edu*

Xue-wen Chen
Computer Science
Wayne State University
Detroit, U.S.A.
xwchen@wayne.edu

***Abstract*—Currently, a large amount of data is produced in healthcare informatics due to the growth of web technologies like social networks, wikis, blogs and RSS feeds. However, not all health information provided online is trustworthy. Even though many experts are involved in publishing trusted information, it is difficult for the general population to determine the credibility of the information. Therefore, a reliable mechanism to automatically determine the trustworthiness of online healthcare information is highly desired. In this paper, we propose two novel approaches based on Topic Modeling and Hidden Markov Models (HMMs), that can be applied over a large volume of online healthcare data to assess its trustworthiness. Traditional Topic Modeling is solely based on the "bag-of-words" model, however, we also consider the semantics of the content to identify the underlying topics in a sentence. For the HMM approach, we built our trustworthy and suspicious models after analyzing the characteristics of sentences from such websites. Both methods perform well to assess the trustworthiness, however HMM is less sophisticated to capture the semantics of sentences. We evaluated our method on randomly chosen real dataset and are able to achieve about 90% accuracy in identifying the trustworthiness of the content.**

***Keywords*-Healthcare Informatics; Big Data; Hidden Markov Model; Topic Discovery;**

I. Introduction

Robinson et al. defined Interactive Health Communication (IHC) as "the interaction of an individual- consumer, patient, caregiver, or professional - with or through an electronic device or communication technology to access or transmit health information or to receive guidance and support on a health-related issue." [1] In the US, 66% of healthy adults and 51% of patients were found to be looking for the health information online as of 2010. [2] As IHC has become essential in daily life, medical researchers are paying more attention to the importance of the quality of healthcare information that is available online. [3], [4], [5]

With the rapid growth of Web 2.0 and social networks like Facebook and Twitter, the amount of healthcare information available online has amplified tremendously. In order to seek health information, consumers make use of portal sites or search engines like Google. The importance of search engines in accessing healthcare information has been discussed well in the past studies. [6] Berland et al. showed that of the results displayed in the first page, only 20% included relevant content, less than 50% of the clinical elements were accurate, while 24% of the content of the search engine results were not covered. [7] Peterson et al. also showed that most users preferred the first page results returned by the search engines. [8]

However, most search engines which return pages based on the user's search keywords make use of link analysis algorithms like PageRank or HITS which only use the link structure and number of incoming and outgoing links to score the pages. [9], [10] With no way to score based on the trustworthiness of the content, the veracity of the healthcare information returned by such search engines becomes highly questionable. Furthermore, it is possible to create credible-looking scientific websites with spurious information and have them be highly ranked in search engine results. There have been several studies conducted by medical experts in order to assess the quality of healthcare information available on the internet, which have shown the information to be not trustworthy. [11], [12], [13], [14] Meric et al. used the keywords "breast cancer" on Google and examined the first 200 websites of over 10,000 English sites manually. [12] They evaluated the characteristics of the websites and found that only 57% of the web pages had author information, with the rest of them having partial or no information. In addition, the study also confirmed that the quality of the information had no correlation with link popularity. Jonathan et al. conducted a user study to identify the reliability of the search engine results and asked experts to judge the trustworthiness of the top 20 results returned from the commercial search engine, Google, for 10 healthcare related queries. [15]. The experts' evaluation identified 51% of the sites to be "credible", 45% of the sites to be "suspicious" and the remaining 4% to be irrelevant. The study clearly highlights that search engines do not use trustworthiness of content in their scoring mechanism. Such a mixture of results leads to misinformation and confusion among users.

Even though credible websites like the Center for Disease Control and Prevention (CDC) or the National Institute of Health (NIH) guarantee the trustworthiness of healthcare information they provide, their use of medical terminology can sometimes be hard for users to understand. Users may

978-1-4799-1293-3/13/$31.00 ©2013 IEEE

tend to prefer more explanatory sites such as Wikipedia, a free editable online encyclopedia. Furthermore, Google often returns Wikipedia pages as the first page in its search results. Recently, a small-scale study conducted by Leithner et al. examined the quality of the Wikipedia articles relating to Osteosarcoma [16] in comparison to the ones available in National Cancer Institute (NCI). They observed the quality of the Wikipedia articles to be good and more accessible than the NCI articles, but found them to lack scientific citations.

Apart from the quality of health information available in search engine results and Wikipedia, we also need to consider the quality of information available in social networks. Since more and more people are engaged in their use, their content also plays an important role in the dissemination of health information among general consumers. Weitzman et al. conducted a study to observe the quality and safety of diabetes-related social networks. [17] They reported the quality to be variable, but found security and privacy of user's personal data to be poor. Although the study was conducted on a small scale, it is enough to show that social networks inevitably contain suspicious healthcare information. In order to address these issues, we need new automated approaches for a scientific and objective measurement of trustworthiness of healthcare information.

Natural Language Processing (NLP) and Machine Learning (ML) are now essential techniques to process text data in medical informatics. [18], [19], [20] In this study, we propose a content-based analysis using the above techniques to analyze healthcare text data on the web and assess its credibility. Our proposed method is inspired by an observation: websites whose content are similar to trusted websites are also more likely to be trustworthy. Therefore, our approach tries to identify similarities in website content in comparison to content from known websites. To do this, we first gather healthcare related pages from the internet using a focused crawler. We then use two methods: HMM based sentence models to identify the trustworthiness of healthcare information and a "Bag-of-words" based Topic Discovery method to identify topics within the sentences of those pages. We then perform page-level and site-level classifications based on results from both these methods to identify the trustworthy and suspicious sites. We evaluated our method on randomly chosen real dataset and are able to achieve about 90% accuracy in identifying the trustworthiness of the content.

Our contributions in this paper are primarily three fold: (1) We have proposed two novel approaches for performing content based analysis on healthcare data. (2) We have been able to show that the Topic Modeling approach is able to perform better than the HMM approach due to its ability to effectively capture semantic information and (3) the proposed algorithm for performing content analysis scales linearly making it suitable for handling big data.

II. Preliminaries

We used two approaches for our content-based analysis on healthcare data; Hidden Markov Model [21] and TAGME. [22]

A. Hidden Markov Model

A Hidden Markov model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with hidden states. Primarily, HMMs have been used to model sequence data like speech utterances in speech recognition. [23] They have also been used in Part-of-Speech tagging [24] and Named Entity Recognition [25] tasks. The success of HMMs in identifying patterns in sequential data has motivated us to explore the possibility of using HMM for content-based analysis. In general, a HMM can be defined using the following parameters:

Notation and definition
N: Number of states in the HMM
M: Number of observation symbols in the HMM
$A = [a_{ij}]$: N by N state transition probability matrix
$B = b_j\,(m)$: N by M observation probability matrix
$\Pi = [\pi_i]$: N by 1 initial state probability vector

An HMM is used to model a sequence with hidden states that represent the latent characteristics of the pattern that we are trying to model, which however emit symbols or observations that are visible. The outputs of the hidden states are observable and are represented as probabilistic functions of the state. In case of sentence modeling, the hidden states would represent the characteristics of a sentence, while the words forming the sentence would represent the visible observations. HMM is a supervised learning method where a training set is used to train the model. The Baum Welch algorithm is used for this and it learns the transition and observation probabilities of the HMM. Once trained, the HMM can then be used for computing the probability of a sentence belonging to given model using the Forward-Backward algorithm or can be used to predict the possible hidden state sequence that could have generated a given sequence of observations using the Viterbi algorithm.

B. Short-Text Tagging

Traditional topic modeling methods use probabilistic approaches based on "bag-of-words" model. [26] However, the "bag-of-words" approach is solely based on the frequency of terms in a document; therefore it is hard to capture the semantics of the text. In order to overcome the problem, Latent Semantic Analysis (LSA) [27], Explicit Semantic Analysis (ESA) [28] or Knowledgebase approaches [29] have been proposed. Recently, with the rapid growth of Wikipedia's knowledgebase and its link structure connecting the related concepts efficiently, several ESA based approaches using Wikipedia have been studied. [30], [31], [32] One of the

ESA methods is TAGME, which is a web application tool for identifying underlying topics in short text fragments using Wikipedia and its link structure proposed by Ferragina and Scaiella (http://tagme.di.unipi.it/) [22]. They improved their method based on the studies of Kulkarni et al. and Cucerzan to deal with annotating very short texts or fragments such as tweets or news feed items *on-the-fly*. [33], [32]

Given the set of anchor texts $A(X)$ identified from a block of text X, the score for a particular sense p_x for the anchor text x to be associated with the page p is determined through a vote of all other anchor texts y which are in support of the annotation $x \xrightarrow{link} p$. Since the anchor text y can also have many senses, the vote is computed as the average relatedness for each sense p_y of the anchor y in relation to the sense p_x. Since not all senses of y have the same statistical significance, the contribution of p_y is weighted using its *commonness* or prior probability $Pr(p_y \mid y)$. Thus the voting formula is defined as:

$$vote_y(p_x) = \frac{\sum_{p_y \in G(y)} rel(p_y, p_x) Pr(p_y \mid y)}{\mid G(y) \mid} \qquad (1)$$

where $rel(a, b)$ is a measure of relatedness between two pages a and b based on the overlap between their in-linking pages in Wikipedia. The relatedness score makes sure that only the senses p_y that are related to p_x affect the voting measure. The final score that defines the *goodness* of the annotation $x \xrightarrow{link} p$ is obtained by the sum of the votes of all other possible anchors y in the text T. The set of candidate anchors identified from the disambiguation phase are then passed through a pruning phase to discard possibly meaningless anchors. These bad anchors are identified based on the link probability of an anchor and the coherence of its candidate annotation which is computed as the average relatedness between the candidate sense of an anchor and the candidate senses for all other anchors in the given text. Only anchors with high link probability or whose assigned sense is coherent with the senses to other anchors are retained.

III. METHOD

In this section, we propose two novel approaches based on topic modeling and machine learning techniques to assess the trustworthiness of the information provided in healthcare sites by doing content-based analysis automatically. For the topic modeling approach, we make use of TAGME to identify salient topics in the sentences available in the healthcare websites. An analysis of the similarity measures among the topics identified is used to decide if the information from candidate website falls under the suspicious or trustworthy category. For the machine learning approach, we apply Hidden Markov Models to model trustworthy and suspicious sentences using an annotated training set.

A. Data Collection

We gathered our dataset using a special healthcare focused crawler; we used manually curated trustworthy and suspicious pages as the seed to crawl other websites. We used a heuristic based on weighted term frequencies of approximately 150 healthcare related keywords to control the crawling of the websites and stopped crawling if the initial few pages of a website fall below a certain threshold value of this heuristic. Overall, we have collected 316 thousand (316K) web pages from 39,831 domains. Though a lot of healthcare related information is present in social networking sites, we consider such information to be highly inconsistent in quality and hence exclude a total of 187 major active social networking sites like Facebook or Twitter from our list of crawled websites.

B. HMM Analysis

In this approach we create two separate HMMs: one to model the *suspicious sentences* and the other to model *trustworthy sentences*. Once these HMMs are trained, the probability of any new sentence belonging to both the models is determined. The sentence is then classified to belong to a model which has the highest probability value. We make use of the Stanford NLP tagger to identify Part-Of-Speech tags in sentences which then help in identifying the features of trustworthy and suspicious sentences. [34]

1) Trustworthy Sentence Model: We chose a few seed websites from the collection of trustworthy sites and analyzed the sentences manually to identify features that can be used to model a trustworthy sentence. Most of the sentences that provide credible information often tend to be expressed in *passive* voice, which can be detected in general by the presence of the *Noun-Verb-Noun* format where the first noun is the *Object* and the second is the *Subject*. In case of passive voice, the verb forms used tend to be either: a "be" form verb, a "have" form verb or a past tense verb, while the nouns are proper nouns, gerunds, and nouns in singular or plural forms. These features are used to form the states of a trustworthy sentence HMM:

- Proper Noun (PRN): names of places, things, etc.
- Be Form Verb (BFV): be, am, is are, was, were, been, being
- Have Form Verb (HFV): have, has, had, having
- Past Tense Verb (PTV)
- Gerund (GER)
- Participle Verb (PAV): past or present participle verb
- Other Noun (OTN): nouns other than proper nouns
- Other words (OTH): words not in the above categories

2) Suspicious Sentence Model: In general sentences from suspicious websites: (1) try to promote or sell a particular product, (2) provide instructions for the reader to follow, (3) tend to resort to superlatives in order to sell the products or services, (4) contain commercial terms to sell a service or product, (5) contain suspicious words or content that appear to promise a guaranteed solution with no scientific proofs, (6) usually end with an exclamation mark. So the above

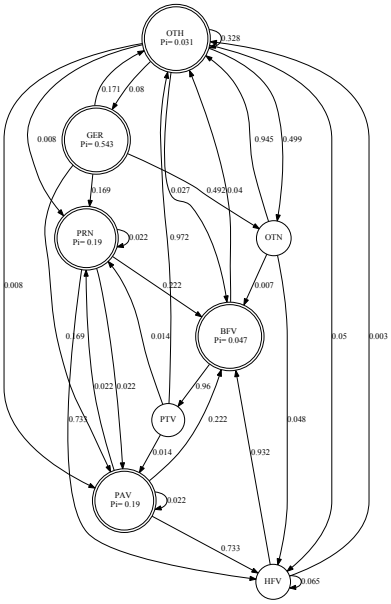

Figure 1: State transitions in a trained HMM for trustworthy sentences

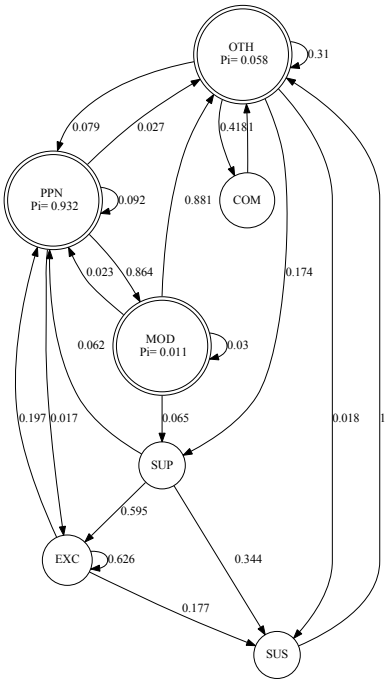

Figure 2: State transitions in a trained HMM for suspicious sentences

features are combined to represent the states of a suspicious sentence HMM:

- Personal Pronoun (PRP): I, You
- Modal verbs (MD): should, must, need to, ought to, have to
- Superlatives (SUP): superlative adjectives or adverbs
- Commercial terms or keywords (COM)
- Suspicious terms or keywords (SUS)
- Exclamation(EXC): exclamation mark at end of sentence
- Other(OTH): words not in the above categories

The commercial and suspicious terms were manually extracted based on term frequency analysis on the extracted content of suspicious websites. Figures 1 and 2 display the states of a trained trustworthy and suspicious sentence models respectively. The double circled nodes are used to denote the possible starting states with a Pi value denoting the starting probability. The state transition probability values displayed on the arcs.

3) Training: A seed data set consisting of manually annotated sentences were used to form the training and testing sets for both the trustworthy and suspicious sentence models.150 sentences, 15 each from 10 sites were used for training the suspicious model, while 120 sentences, 15 each from another 8 suspicious sites were used for testing it. For the trustworthy sentence model, 150 sentences, 50 each from 3 trustworthy sites were used for training, while 150 sentences, 15 each from another 10 trustworthy websites were used for testing.

4) Classification: For each sentence in a web page, the probabilities of it belonging to the trustworthy and suspicious sentence models are calculated. The sentence is then classified to belong to the model that has the highest probability. The probability values of the sentence models are then used in constructing the *page-level features* to classify a web page containing those sentences as trustworthy or suspicious. A Support Vector Machine based classifier is explored to make this prediction. [35] SVMs are supervised machine learning models which are used for classification and regression analysis and have shown to have very high classification accuracy especially for binary classification. For each sentence in a page, we compute the difference between the probability of it being trustworthy to the probability of it being suspicious. This difference is then used to update the counts of a histogram with range -1.0 to +1.0 and intervals of 1.0E-10. The counts are then normalized by the total number of sentences in the page. The normalized counts of the histogram are used to form the page-level features used by the SVM for classification.

C. Topic Analysis

In order to perform topic analysis, web pages from a manually selected set of 20 trustworthy and suspicious sites were gathered to form our reference sites. Plain text sentences were extracted from these pages and TAGME was used to identify semantic topics in these sentences. As described in Section II-B, TAGME provides the goodness value of a topic and its corresponding topic word. We used a certain goodness threshold for choosing meaningful topics among all the topics in a sentence obtained from TAGME.

1) Page-Level Similarity: The similarity between pages is measured after identifying the semantic topics for each

page. Since each page contains representative topics, we can compare the similarities with each other. For calculating page-level similarity, we use the most popular set-similarity method called *Jaccard similarity* and it can be obtained by:

$$PageSim(x_i, y_i) = \frac{Total\ number\ of\ same\ topics}{Total\ number\ of\ topics} \quad (2)$$

where x_i and y_i are pages from each website. Based on this, we can calculate the page-level similarity score between pages.

2) Site-Level Similarity: Next, in order to get the site-level similarity, we adopted *Group Linkage* problem. [36] Group Linkage problem is used to identify the similarity between two groups which have different number of elements. The fundamental idea of the group linkage is based on *Maximum Bipartite Matching(MBM)* problem. [37] The definition of MBM similarity is defined as below:

Let A and B be two sets, $A = a_1, a_2, ..., a_m$ and $B = b_1, b_2, ..., b_n$. The Maximum Bipartite Matching Similarity, *MBM_Sim*, is defined as:

$$MBM\_Sim(A, B) = \frac{\sum_{(a_i,b_j)\in M} sim(a_i, b_j)}{m + n - M} \quad (3)$$

where $sim(a_i, b_j) \geq \rho$ is the similarity of two elements in the two groups A and B. M is the number of maximum weight matching in the bipartite graph. In our approach, $sim(a_i, b_j)$ would be the page-level similarity above ρ. The threshold ρ is to remove the pages having very low similarity scores and can be decided heuristically. Using MBM_Sim, we get the similarity score between two web sites.

Finally, in order to evaluate the trustworthiness of unknown site X, we defined the *ContentSim(X)* as below:

$$\begin{aligned} ContentSim(X) = [&Sum\ of\ Top5\ Site - Level- \\ &Similarity\ of\ Trustworthy\ Sites] \\ &- [Sum\ of\ Top5\ Site - Level- \\ &Similarity\ of\ Suspicious\ Sites] \end{aligned} \quad (4)$$

The content similarity of site X, ContentSim(X), is the difference between the summation of top 5 site-level-similarity from trustworthy group and the summation of top 5 site-level-similarity from suspicious sites. If the value of *ContentSim* score is positive, we consider that site X contains trustworthy informatio. If not, it contains suspicious information.

IV. Results and Discussion

Since we have a big data set, approximately 316 thousand web pages from 39,831 domains, it is not possible to evaluate all of them without manually classifying each of them as well. Therefore, we experiment on this big data set by randomly picking a few sites, manually, to evaluate the accuracy of our approach.

A. Results of HMM Analysis

1) Page-level classification: Sentence level classification is carried out for all sentences in all the pages belonging to those selected sites. The page level features as mentioned before are extracted for all the pages. The extracted page level features from the training and testing data set are used for building the SVM page classifier. The average of classification accuracy for the pages in the testing set is over 95 % for suspicious sites and 94% for trustworthy sites.

Website	Total pages	Classification Accuracy of trustworthy pages	Website classification
ChooseMyPlate	200	48.5%	Suspicious
clinicaltrials	200	98.0%	Trustworthy
dana-farber	198	74.24%	Trustworthy
diabetes	193	11.40%	Suspicious
drugabuse	200	48.5%	Suspicious
foodsafety	50	66.0%	Trustworthy
hhs	310	68.71%	Trustworthy
kidshealth	187	44.39%	Suspicious
letsmove	127	46.46%	Suspicious
mayoclinic	294	38.10%	Suspicious
nemours	200	81.0%	Trustworthy
nutrition	69	31.88%	Suspicious
usa	103	48.54%	Suspicious
webmd	198	13.64%	Suspicious
womenshealth	166	53.61%	Trustworthy

Table I: Page Classification on Real Data Set for Trustworthy Sites

Website	Total pages	Classification Accuracy of Suspicious pages	Website classification
blogtalkradio	200	68.0%	Suspicious
comcblog	309	61.49%	Suspicious
devinalexander	168	85.12%	Suspicious
directselling411	203	49.75%	Trustworthy
discovergoodnutrition	192	80.21%	Suspicious
flite	233	55.36%	Suspicious
goodhousekeeping	196	95.92%	Suspicious
losethebellyfatnow	154	72.73%	Suspicious
ocregister	189	48.68%	Trustworthy
planetarynutrition	200	12.0%	Trustworthy
plentyofhealth	201	84.58%	Suspicious
premadeniches	175	88.0%	Suspicious
wholefoodsmarket	199	97.49%	Suspicious
widgetbox	478	85.36%	Suspicious
zendesk	198	87.88%	Suspicious

Table II: Page Classification on Real Data Set for Suspicious Sites

2) Site-level Classification: Table III presents the website classification accuracy which is based of the results obtained from the SVM based page classification approach which are presented in Tables I and II. It can be seen that for the page-level classification on the 15 real dataset for trustworthy sites, the average classification accuracy is only 40%. Since the SVM Page classification is based on the underlying HMM based sentence classification, this lower performance can be attributed to the HMM classifier's inability to identify trustworthy sentences in comparison to suspicious sentences.

B. Results of Topic Analysis

The *Topic Discovery* method needs a reference dataset for the trustworthiness assessment of unknown sites. We

Approach	Real Data Set for	Total number of web sites	Number classified as trustworthy	Number classified as suspicious	Classification Accuracy
SVM Page Classification	Suspicious web sites	15	3	12	80%
SVM Page Classification	Trustworthy web sites	15	6	9	40%

Table III: Website Classification based on SVM Page Classification

manually identified 20 websites to form the reference set for the trustworthy and suspicious classes out of the crawled data. We then ran TAGME on the reference sites to find the semantic topics from the pages. We then identified the page-level similarity between all the pages, followed by the site-level similarity between websites. We first used the reference dataset to evaluate the performance of the Topic Analysis method. Table IV shows the content similarities of trustworthy reference set. Even though trustworthiness is subjective, we choose government sites as our reference set, as we believe them to contain truthful information and hence use them to represent trustworthiness objectively. As seen in Table IV, the content similarities are all positive, showing that the government sites share more similar semantic topics with each other than the suspicious sites. However, unexpectedly NIH shows very low content similarity value. We can probably attribute this to the notion that content available in NIH tends to be more scientific in nature in comparison to other general trustworthy government sites. Table V presents the content similarity results for the reference set of suspicious sites. Except for one site, all other 9 sites show a negative content similarity confirming the suspicious nature of their content.

Website	Sum of Top5 (Trustworthy)	Sum of Top5 (Suspicious)	Content Similarity
health.gov	0.0089333	0.0035985	0.0053348
healthfinder.gov	0.0055555	0.0023102	0.0032452
foodsafety.gov	0.0034835	0.0014873	0.0019962
womenshealth.gov	0.0021668	0.0008630	0.0013037
hhs.gov	0.0012620	0.0004603	0.0008017
cancer.gov	0.0012653	0.0005057	0.0007596
letsmove.gov	0.0010809	0.0006934	0.0003875
cdc.gov	0.0009787	0.0006152	0.0003634
nutrition.gov	0.0015993	0.0014280	0.0001713
nih.gov	0.0005944	0.0004938	0.0001006

Table IV: Content Similarity within Trustworthy Reference Set

C. Comparison of Methods

From the set of all crawled websites excluding the training and the testing sites, we randomly chose some sites to form the real dataset and performed analysis using both the methods to identify the trustworthy and suspicious sites. The first column of the Table VI lists the websites chosen, most of which are suspicious sites followed by columns which present their classifications based on manual verification, Topic Analysis and HMM Analysis, respectively. Based on the results shown in the Table VI, both topic analysis and HMM analysis methods correctly classified 9 out 10 sites.

Website	Sum of Top5 (Trustworthy)	Sum of Top5 (Suspicious)	Content Similarity
dietprescriptions-rx	0.0025721	0.0041871	-0.0016150
fatburningfurnace	0.0008428	0.0014466	-0.0006038
fatvanish	0.0007956	0.0013823	-0.0016150
apple-cider-vinegar-benefits	0.0006026	0.0011556	-0.0005531
eco-diet	0.0006551	0.0011243	-0.0004693
burnthefat	0.0009967	0.0014238	-0.0004271
healthynewage	0.0010024	0.0013259	-0.0003235
amazing-green-tea	0.0006873	0.0009042	-0.0002169
hypnosisnetwork	0.0025721	0.0041871	-0.0016150
carallumaburnreviews	0.0044153	0.0043684	0.0000469

Table V: Content Similarity within Suspicious Reference set

Website	Manually verified as	Topic Analysis	HMM Analysis
premadeniches	suspicious	trustworthy	suspicious
wholefoodsmarket	trustworthy	trustworthy	suspicious
goodhousekeeping	suspicious	suspicious	suspicious
devinalexander	suspicious	suspicious	suspicious
blogtalkradio	suspicious	suspicious	suspicious
comcblog	suspicious	suspicious	suspicious
planetarynutrition	suspicious	suspicious	trustworthy
discovergoodnutrition	suspicious	suspicious	suspicious
plentyofhealth	suspicious	suspicious	suspicious
losethebellyfatnow	suspicious	suspicious	suspicious

Table VI: Comparison of methods with real dataset

D. Discussion

Our proposed content-based analysis approaches are alternative ways for overcoming the traditional link analysis approaches that are used for assessing the semantics of the web page contents. Especially, Topic Analysis based on ESA works well due to the vast information provided by Wikipedia. In our Topic Analysis method, the important thing to consider is a way to scale up the number of seed sets used for the trustworthy and suspicious reference sites. Since the small number of seed sets hardly represent the volume of the big data, we need to scale them up efficiently. We can use automatic expansion methods to include more seed sites without any bias of content. We can use an automated-annotating scheme based on trained HMMs to do this.

The results of HMM Analysis shows that it is difficult to identify content that is truly trustworthy. The HMM is solely based on certain Part-of-Speech patterns and word sequences, which does not take into account the semantics of the content. In addition, our trustworthy sentence model was built primarily based on content from trusted scientific sites. It is difficult to model a generic sentence that contains trustworthy content without also performing some kind of semantic analysis. Therefore, using only some surface level patterns in the sentences is not sufficient for content-based analysis. The Topic Analysis method, though fundamentally based on the *"bag-of-words"* model approach, still seems to perform better than the HMMs due to its ability to identify semantically salient topics for given content.

We collected 316 thousand (316K) webpages from 39,831 domains, however, we only used randomly selected websites for testing our approaches due to the labor intensive process to label all of them. For addressing the needs of *big data*, the critical factor to consider is the speed and efficiency of the algorithms used to implement the proposed approaches. Our Topic Analysis method has a runtime complexity of only *big-O* of (N) where N is the number of web pages. Due to this linear complexity, this approach can be easily scaled to work in real-time even with large volume of data.

Case Study: Among the websites we tested, we chose a suspicious website to look into the trustworthiness of its contents. We chose *www.discovergoodnutrition.com* which provides nutrition and health advice to users. The website as such is well designed, so many users might tend to trust the information they provide. However, if we take a look at their advice pages, we find that they provide information without any scientific citations. For example, one page talks about bacteria which are known to help in digestion, and promotes consumption of *Yogurt* containing such bacteria. However, there are no supporting evidences or links to scientific studies that describe the role of the bacteria and how it is beneficial to our health. Our method was able to predict this site as "suspicious" even though the site itself looks trustworthy at a glance.

V. Conclusion

Healthcare Informatics is a promising field to utilize the blood of healthcare related data through adoption of information technologies. Currently, a large amount of data is generated in this area due to web technologies like social networks, wikis, blogs and RSS feeds. However, not all health information provided online is trustworthy. Even though many experts are involved in publishing trusted information, it is difficult for people to determine the credibility of the information. Most search engines have the ability to control spam pages, but cannot determine the trustworthiness of the page yet and it is not easy for users to distinguish trustworthy sites from the mixed results. Hence, identifying the credible information on the complicated web society is a challenging problem. Therefore a reliable mechanism to automatically determine the trustworthiness of online healthcare information is highly desired.

In this paper, we propose a novel approach over a large volume of online healthcare information collected by a focused crawler. We have tried to measure the trustworthiness of the information based on content-based analysis. Since due to the ambiguity and language specific characteristics of the content analysis, NLP and ML have been used often in medical and healthcare informatics. We applied TAGME and HMM among many methods in NLP and ML to analyze healthcare text data on the web and assess its credibility. Our proposed method is inspired by an observation; websites whose content and opinions are similar to trusted websites are more likely to be trustworthy. Our approach consists of three steps: first, we preprocessed the dataset gathering healthcare related pages from the internet using a focused crawler. We then built HMMs to identify the trustworthiness of healthcare information (HMM Analysis). Finally, we identified topics within the sentences of those pages based on the "bag-of-words" model (Topic Discovery). In HMM analysis, we did page-level classification using SVM and majority voting on site-level to identify the trustworthy and suspicious sites. We also measured the page-level and site-level similarity between credible and suspicious contents for topic discovery.

Both methods perform well to assess the trustworthiness, however HMM is less sophisticated in capturing the semantics of sentences. We evaluated our method on randomly chosen real dataset and are able to achieve about 90% accuracy in assessing the trustworthiness of the content.

References

[1] T. N. Robinson, K. Patrick, T. R. Eng, D. Gustafson *et al.*, "An evidence-based approach to interactive health communication," *JAMA: the journal of the American Medical Association*, vol. 280, no. 14, pp. 1264–1269, 1998.

[2] S. Fox and K. Purcell, *Chronic disease and the Internet*. Pew Internet & American Life Project Washington, DC, 2010.

[3] J. A. Diaz, R. A. Griffith, J. J. Ng, S. E. Reinert, P. D. Friedmann, and A. W. Moulton, "Patients' use of the internet for medical information," *Journal of general internal medicine*, vol. 17, no. 3, pp. 180–185, 2002.

[4] G. Eysenbach and C. Köhler, "How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews," *BMJ: British Medical Journal*, vol. 324, no. 7337, p. 573, 2002.

[5] M. Benigeri and P. Pluye, "Shortcomings of health information on the internet," *Health Promotion International*, vol. 18, no. 4, pp. 381–386, 2003.

[6] L. Greenberg, G. D'Andrea, and D. Lorence, "Setting the public agenda for online health search: a white paper and action agenda," *Journal of medical Internet research*, vol. 6, no. 2, 2004.

[7] G. K. Berland, M. N. Elliott, L. S. Morales, J. I. Algazy, R. L. Kravitz, M. S. Broder, D. E. Kanouse, J. A. Muñoz, J.-A. Puyol, M. Lara *et al.*, "Health information on the internet," *JAMA: the journal of the American Medical Association*, vol. 285, no. 20, pp. 2612–2621, 2001.

[8] G. Peterson, P. Aslani, and K. A. Williams, "How do consumers search for and appraise information on medicines on the internet? a qualitative study using focus groups," *Journal of Medical Internet Research*, vol. 5, no. 4, 2003.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web." 1999.

[10] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[11] G. Eysenbach, J. Powell, O. Kuss, and E.-R. Sa, "Empirical studies assessing the quality of health information for consumers on the world wide web," *JAMA: The Journal of the American Medical Association*, vol. 287, no. 20, pp. 2691–2700, 2002.

[12] F. Meric, E. V. Bernstam, N. Q. Mirza, K. K. Hunt, F. C. Ames, M. I. Ross, H. M. Kuerer, R. E. Pollock, M. A. Musen, and S. E. Singletary, "Breast cancer on the world wide web: cross sectional survey of quality of information and popularity of websites," *Bmj*, vol. 324, no. 7337, pp. 577–581, 2002.

[13] H. Kunst, D. Groot, P. M. Latthe, M. Latthe, and K. S. Khan, "Accuracy of information on apparently credible websites: survey of five common health topics," *BMJ: British Medical Journal*, vol. 324, no. 7337, p. 581, 2002.

[14] P. Z. Stavri, D. J. Freeman, and C. M. Burroughs, "Perception of quality and trustworthiness of internet resources by personal health information seekers," in *AMIA Annual Symposium Proceedings*, vol. 2003. American Medical Informatics Association, 2003, p. 629.

[15] J. Lutes, M. Park, B. Luo, and X.-w. Chen, "Healthcare information networks: Discovery and evaluation," in *Healthcare Informatics, Imaging and Systems Biology (HISB), 2011 First IEEE International Conference on*. IEEE, 2011, pp. 190–197.

[16] A. Leithner, W. Maurer-Ertl, M. Glehr, J. Friesenbichler, K. Leithner, and R. Windhager, "Wikipedia and osteosarcoma: a trustworthy patients' information?" *Journal of the American Medical Informatics Association*, vol. 17, no. 4, pp. 373–374, 2010.

[17] E. R. Weitzman, E. Cole, L. Kaci, and K. D. Mandl, "Social but safe? quality and safety of diabetes-related online social networks," *Journal of the American Medical Informatics Association*, vol. 18, no. 3, pp. 292–297, 2011.

[18] N. Sager, M. Lyman, C. Bucknall, N. Nhan, and L. J. Tick, "Natural language processing and the representation of clinical data," *Journal of the American Medical Informatics Association*, vol. 1, no. 2, pp. 142–160, 1994.

[19] P. Spyns, "Natural language processing," *Methods of information in medicine*, vol. 35, no. 4, pp. 285–301, 1996.

[20] C. Friedman and G. Hripcsak, "Natural language processing and its future in medicine," *Academic Medicine*, vol. 74, no. 8, pp. 890–5, 1999.

[21] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[22] P. Ferragina and U. Scaiella, "Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 1625–1628.

[23] L. Rabiner and B. Juang, "An introduction to hidden markov models," *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4–16, 1986.

[24] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," in *Proceedings of the third conference on Applied natural language processing*. Association for Computational Linguistics, 1992, pp. 133–140.

[25] G. Zhou and J. Su, "Named entity recognition using an hmm-based chunk tagger," in *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 473–480.

[26] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424–440, 2007.

[27] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *JASIS*, vol. 41, no. 6, pp. 391–407, 1990.

[28] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis." in *IJCAI*, vol. 7, 2007, pp. 1606–1611.

[29] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 2012, pp. 481–492.

[30] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer, "Semantic wikipedia," in *Proceedings of the 15th international conference on World Wide Web*. ACM, 2006, pp. 585–594.

[31] M. Strube and S. P. Ponzetto, "Wikirelate! computing semantic relatedness using wikipedia," in *AAAI*, vol. 6, 2006, pp. 1419–1424.

[32] S. Cucerzan, "Large-scale named entity disambiguation based on wikipedia data." in *EMNLP-CoNLL*, vol. 7, 2007, pp. 708–716.

[33] S. Kulkarni, A. Singh, G. Ramakrishnan, and S. Chakrabarti, "Collective annotation of wikipedia entities in web text," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 457–466.

[34] K. Toutanova and C. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*. Association for Computational Linguistics, 2000, pp. 63–70.

[35] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[36] B.-W. On, N. Koudas, D. Lee, and D. Srivastava, "Group linkage," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007, pp. 496–505.

[37] D. B. West *et al.*, *Introduction to graph theory*. Prentice hall Englewood Cliffs, 2001, vol. 2.