Using Machine Learning for Automatic Identification of Evidence-Based Health Information on the Web

Majed M. Al-Jefri
School of Computing, Engineering and Mathematics
University of Brighton, Moulsecoomb
Brighton, UK BN2 4GJ
M.Al-Jefri@brighton.ac.uk

Roger Evans
School of Computing, Engineering and Mathematics
University of Brighton, Moulsecoomb
Brighton, UK BN2 4GJ
R.P.Evans@brighton.ac.uk

Pietro Ghezzi
Brighton & Sussex Medical School
University of Sussex, Falmer
Brighton, UK BN1 9RY
P.Ghezzi@bsms.ac.uk

Gulden Uchyigit
School of Computing, Engineering and Mathematics
University of Brighton, Moulsecoomb
Brighton, UK BN2 4GJ
G.Uchyigit@brighton.ac.uk

ABSTRACT

Automatic assessment of the quality of online health information is a need especially with the massive growth of online content. In this paper, we present an approach to assessing the quality of health webpages based on their content rather than on purely technical features, by applying machine learning techniques to the automatic identification of evidence-based health information. Several machine learning approaches were applied to learn classifiers using different combinations of features. Three datasets were used in this study for three different diseases, namely shingles, flu and migraine. The results obtained using the classifiers were promising in terms of precision and recall especially with diseases with few different pathogenic mechanisms.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; **Natural language processing**; **Machine learning**; **Supervised learning**;

KEYWORDS

Online health information; machine learning; assessing health web pages; evidence-based medicine; text classification

ACM Reference format:
Majed M. Al-Jefri, Roger Evans, Pietro Ghezzi, and Gulden Uchyigit. 2017. Using Machine Learning for Automatic Identification of Evidence-Based Health Information on the Web. In *Proceedings of DH '17, London, United Kingdom, July 02-05, 2017,* 8 pages.
https://doi.org/10.1145/3079452.3079470

DH '17, July 02-05, 2017, London, United Kingdom

ACM ISBN 978-1-4503-5249-9/17/07…$15.00
https://doi.org/10.1145/3079452.3079470

1 INTRODUCTION

The quality of online health information is important in healthcare. Nowadays, people refer to the Internet to ask about everything in their daily life including health information [2, 19]. More people post health questions daily than refer to their doctors [5]. The Pew Research Centre reported that 72% of Internet users sought health information online in 2012 [8]. Furthermore, the emergence of the Web 2.0 technology has transmuted the way that Internet users seek online information including online health information. However, due to the open nature of the World Wide Web and website creation tools, any person can easily create a website and produce any content. This includes health websites or health information content on blogs or forums which could be distributed without being carefully verified and that could have a severe influence on people's health. Accordingly, assessing the quality of online health information automatically is very important.

There is no agreed-upon definition of health information quality (HIQ). Different researchers use different definitions and evaluate it using various indicators and criteria [26]. Many existing instruments that use different criteria have been designed for measuring health information quality. Among those, the most common ones are the Health On the Net (HON) foundation code [11], the Journal of the American Medical Association (JAMA) quality criteria [1] and DISCERN [3]. The HON code is a certification of quality obtained from the HON foundation. A website is accredited when it satisfies eight quality principles, namely authorship, attribution, privacy, complementarity, transparency, justifiability, financial disclosure, and advertising policy. This certification is valid for one year and can be renewed. For the JAMA quality criteria, a website should satisfy four criteria to be considered of good quality, these criteria are authorship, source attribution, site ownership disclosure, and currency (the date of the information contained in the website). DISCERN is an instrument designed in the form of a questionnaire to help health information consumers to judge the quality of written health information about treatment choices [4].

Zhang et al. [26] carried out a survey that examined 165 articles in which researchers evaluated health information quality on the web against predefined criteria. Most of these studies used pre-existing instruments with predefined criteria (such as the JAMA benchmarks, HON code, and DISCERN) to evaluate the quality, while

others used combined criteria based on previous studies and/or criteria from different pre-existing instruments. The evaluation of the results revealed that health information quality varied across medical domains and across websites as well. They stated that the overall quality is still problematic.

In the research considered in Zhang et al.'s study, the assessments of quality were carried out manually, which is not a practical way of addressing HIQ across the web as a whole. It is clear that there is a need to developing better methods that can automatically assess the quality of health documents. This will make it possible to analyse larger numbers of websites, and thereby provide important information to users, professionals and policy-makers about the quality of health information available across the web.

In this paper, we go beyond these previous studies by showing that incorporating analysis of the textual content of health web pages allows us to distinguish between two classes of advice, namely evidence-based and non-evidence-based ('complementary' or 'alternative') using an automatic procedure. It is now recognized in most countries that criteria for drug approval are based on a hierarchy of evidence, and a high level of evidence that the drug meets the primary efficacy endpoint in randomized clinical trials is required [10, 12]. In contrast, complementary and alternative medicine (supplements) do not need to undergo such stringent criteria as they do not make specific efficacy claims [7]. As such, it is reasonable to argue that our result constitutes a significant contribution to automatic analysis of HIQ.

We used natural language processing (NLP) and machine learning (ML) techniques to assess the quality of online health webpages, combining information from criteria used in previous research with formal and linguistic properties of the text content. Datasets for three different diseases were used, namely shingles, flu and migraine. Different machine learning technique were used to classify health webpages and the obtained results were compared.

The rest of the paper is organized as follows. Section 2 presents the related work in this area. Section 3 presents the datasets that were used. Section 4 explains the methodology. The experimental results are shown in section 5 and discussed on more detail in section 6. Finally, we present our conclusions and future research directions in section 7.

2 RELATED WORK

Many studies have tried to assess health information quality, most of them are manual methods that use some quality guidelines while others are automatic methods that use some metrics to classify health documents (mainly web documents). In the survey of Zhang et al. [26] more than a third of the surveyed studies used instruments based on medical guidelines, text books or literature, while more than a quarter of the surveyed studies needed intervention from medical experts in order to evaluate the content. All of these studies evaluated the quality of information by looking through the content manually which is time consuming and effort-intensive.

Gaudinat et al. [9] present a method for categorization of health documents based on the HON code principles. For the training dataset, they used more than 5,000 HON code accredited websites in four different languages as positive examples. The paragraphs of these documents that demonstrate the HON principles were extracted with a help of human experts. These paragraphs were further segmented into sentences using regular expressions based on HTML tags and punctuation marks. As features, they used word n-grams, word co-occurrences with and without stopwords and stemming. Within each sentence three elements were used, term frequency, inverse term frequency and length normalization. Naive Bayes, Support Vector Machines (SVM), K-Nearest Neighbour (KNN), and Decision Trees were used as machine learning algorithms. This research studied how to categorize manually extracted paragraphs related to the eight HON principles into eight classes. However, there are a number of problems with this approach. First, it requires manual intervention from experts to extract which part of the document relates to which principle. Second, all the examples are positive and it says nothing about whether the principles exist or not. It is only a categorization task and it does not check whether a principle is covered by the document or not.

Wang & Liu [22] developed a tool for detecting health information quality indicators for the purpose of automatically evaluating the quality of health information. Seven criteria categories were used, namely authority, source, currency, content, disclosure, interactivity, and commercialization. Under each category, different criteria were classified, a total of 15 different criteria were used. Then they defined multiple measurable indicators for each criterion. Three datasets were collected on three queries, viz. acne, melanoma, and skin cancer. The reported result reached 93% and 98% for recall and precision, respectively. Wang & Richard [23] further proposed a rule-based method to detecting health information criteria by analysing the structure and the content of health webpages. They defined measurable indicators for each criterion with the indicator value and the expected location within a webpage. Using regular expressions, the expression pattern of each candidate line is identified after being extracted by matching the indicator value with the webpage content. However, not all well-known criteria were explored and the used datasets were very small.

The study by Sondhi et al. [20] is considered one of the latest works that tried to automatically assess the quality of medical web pages. They classified medical webpages as being reliable or not based on the information they contain and the features they have. They used different types of features such page rank, links, and commercial features. They collected a dataset using the HON criteria. A set of webpages that were accredited by the HON foundation were used as the positive sample set, while for the negative set they used general and advertisements webpages that failed the HON reliability criteria. However, no information is available about the rank of the retrieved webpages within the search engine results, especially the negative ones. For classification, SVMs were used to train their system with different combinations of feature sets, and they evaluated it using 5-fold cross-validation. They reported prediction accuracies of over 80%.

3 DATA

One of the most challenging issues in this kind of studies is getting a dataset that is accurately annotated. The migraine dataset collected in [24] and the flu prevention dataset collected in [14] were used in this study. These datasets were collected by searching google.com for the treatments of the two diseases and the first 200 webpages

Table 1: Dataset Statistics

Dataset	No. of positive samples	No. of negative samples	Total samples	Avg. no of words
Shingles	88	23	111	1669
Flu	38	22	60	1171
Migraine	22	53	75	1745
Total	148	98	246	1571

retrieved by the Google search engine were saved and analysed. Webpages were manually classified by the type of intervention and accordingly classified into evidence-based medicine (EBM) and non-EBM approaches.

In addition to those datasets, we collected and annotated another dataset on shingles treatments using the same methodology used in [24] and [14]. We started by searching on google.com looking for "shingles treatment" (after clearing the history and cache data from the web browser so results were uninfluenced by browser history or any additional filters). The search was done on the 23/5/2016 and the first 204 of the retrieved webpages were annotated by identifying whether the suggested treatments in a webpage are approved (EBM) or not. Inaccessible webpages were excluded, (cases when links were dead or login details were required). To establish whether an approach is EBM-based we looked to see if the treatment was approved by the US Food and Drug Administration, was recommended by the UK NHS or the National Institute of Care Excellence. To follow the same procedure as done in [24] and [14] the pages were also checked for whether they comply with the four JAMA criteria. The annotation was performed by a trained annotator and the data was further partially checked by a medical expert.

The webpages of the three datasets were saved into disk for later to be locally accessed. The pages were first 'cleaned' (extraneous mark-up and non-text information was removed) and labelled. The datasets consist of positive and negative (EBM and Non-EBM) documents that were labelled as *pos* and *neg* respectively.

A total of 111, 60 and 75 webpages were obtained for shingles, flu and migraine corpora, respectively. Table 1 shows the statistics of the collected datasets.

4 METHODOLOGY

Classifying health documents based on their contents is performed using natural language processing and machine learning techniques to create (or 'train') 'classifiers'. After being trained, these classifiers are used to make decisions about whether webpages or documents containing health information are reliable or not based on a set of collected features. A specialized corpus is required in order to achieve this goal. The datasets discussed in the previous section were used to train and test the classifiers. Textual features that identify useful dimensions of health information quality were extracted using NLP techniques. Afterwards, machine learning techniques were used to learn from such features in order to classify health documents and achieve the ultimate goal of providing reliable health information to allow users of a wide range of abilities to make informed decisions about their health needs.

Python programming language was used to implement the prototype: Natural Language Toolkit NLTK [16] and the scikit-learn [18] machine learning libraries were mainly used in this work. The Sketch Engine corpus management tool [13] was used to prepare and label the data for processing.

4.1 Preprocessing

For preprocessing, the first step was to set up the corpus to be saved into separate files with the corresponding labels. For that, the saved webpages were loaded into Sketch Engine to be cleaned of html and web scripts as well as other unrelated text such as advertisements.

Another preprocessing step was to remove the punctuation marks and unigram feature stopwords such as '*a*', '*the*', and '*is*' were also removed from the text before applying any feature extraction method.

4.2 Features

In order to classify webpages as EBM or not using machine learning techniques, suitable features (individual characteristics of the documents for machine learning algorithms to learn from) need to be extracted from a training set and fed to the machine learning algorithm to train a classifier. Different features were extracted in our experiments including text-based features and domain-specific criteria.

Text-based features frequently used in this kind of analysis include word n-grams (sequences of n words occurring in the text), which capture simple, frequently occurring linguistic concepts and relationships, and formal properties, such as capitalisation patterns, punctuation marks, word, sentence and document length. For our experiments we used unigrams (single words), bigrams (word pairs) and trigrams (word triples) that occur sufficiently frequently (in the training set), as well as lemmas (word stems), vocabulary richness ratio, number of capital letters, number of punctuation marks and normalized document length.

Domain specific criteria used as features included some of the well-known instruments for measuring health information quality, such as JAMA scores. The four JAMA criteria were extracted (if encountered) using an approach developed in a separate piece of work (not discussed further here) which performs with accuracies above 80%. The number of approved drugs terminologies were also considered as a feature. (We did not use a list of approved medicines, but counted long words as a proxy measure for domain terminologies.)

A model needs to be trained on these extracted features so that it knows from the given class labels how features differ from the positive class to the negative one. From the training phase, the learned model can then be used to classify unseen documents based on the extracted features from them into the proper class.

4.3 Classification

Machine learning techniques are algorithms that learn from given data and then make predictions or decisions on what they have learnt on new unseen data. They can be supervised, unsupervised, or sometimes semi-supervised algorithms. The supervised algorithms learn from labelled data, usually referred to as training data that has predefined labels or classes. When tested on unseen data

they try to label them, that is to assign each data sample to one of the predefined classes. The unsupervised algorithms have no predefined classes or label, instead they try to predict the classes and cluster data into those classes. In this research, we used supervised learning algorithms to learn classifiers from pre-categorized examples (in our case EBM and non-EBM) so that it assigns unseen samples (the test set) to those categories automatically. The classifiers are validated by testing on a portion of the training data, known as the validation set, and can be adjusted to improve their performance, before final testing on unseen data (the test set).

A number of different machine learning techniques were used for the experiments: multinomial naive Bayes, K-nearest neighbour (KNN), support vector machines (SVM), stochastic gradient decent (SGD) SVM, logistic regression, and multilayer perceptron (MLP).

The naive Bayes classifier is based on Bayes theorem. It is simple and often outperforms more sophisticated classifiers [17]. It is suited when the dimensionality of the inputs is high, and it is reasonable to assume independence between the features. It is widely used in machine learning applications especially in natural language processing and text classification. The algorithm tries to maximise the probability for a set of features to belong to class k from a set of K different classes.

The K-nearest neighbour (KNN) classifier is considered as one of the simplest machine learning classification algorithms [21]. KNN requires no explicit training, it just uses the training examples to classify the new unseen ones based on their similarity. Each sample is represented by its multidimensional feature space position and the distance between any two instances represents their similarity. This distance can be computed using many different metrics such as Euclidean distance, and Manhattan distance.

Support vector machines (SVMs) are set of related supervised learning methods used for classification and regression. Given a set of training samples S, each to be classified as a binary class 0 or 1, the SVM algorithm builds a model that predicts whether an unseen sample belongs to the 0 class or the 1 class [6]. Intuitively, an SVM model represents the samples as points in space, and tries to cut the space into two halves, so that the samples of the separate classes are divided by as wide a gap as possible. Any new unseen sample can be mapped into that space and can be predicted to be one of the two classes based on the half that this sample falls in. SVMs are widely used in several machine learning applications due to their classification accuracy. Unlike other predictors, SVMs can separate nonlinearly separable data using the concept of hyperplane separation on data. They map predictors onto a higher dimensional space so that they can be separated linearly.

Two other linear classifiers, stochastic gradient decent (SGD) SVMs and logistic regression were also used in this work. Linear classifiers, although simple, are efficient and have been successfully used in many text classification tasks [25]. In addition, a neural network approach was also included. Neural networks are one of the most effective machine learning algorithms [15], but more complex to train. We used the Multilayer Perceptron approach (also known as feedforward neural networks) in this work.

5 EXPERIMENTAL RESULTS

Different experiments using different classifiers trained on different combination of features were performed on the three aforementioned datasets. Because of the small size of the datasets, roughly 15% of each corpus were saved untouched to be used as test sets. The remaining 85% of the datasets were used as training sets.

5.1 Performance Measures

The system performance is evaluated by means of classification recall, precision and F_1 measures. Classification recall is defined as the fraction of correctly classified samples (i.e. the number of correctly classified samples divided by the number of samples that should have been positively labelled). Classification precision is defined as the fraction of predictions that are correct (i.e. the number of correctly predicted samples divided by the number of all positively predicted samples in the test set). F_1, the harmonic mean of precision and recall, is also used as a combined performance measure. It strikes a balance between precision and recall which does not allow either to dominate the overall performance measure. Equations 1 to 3 shows how recall, precision and F_1 are calculated. Here TP is the number of 'true positives' (correctly predicted positive samples), TN is the number of 'true negatives' (correctly predicted negative samples), FP is the number of 'false positives' (negative samples incorrectly predicted as positive) and FN is the number of 'false negatives' (positive samples incorrectly predicted as negative).

$$Recall = TP/(TP + FN) \tag{1}$$

$$Precision = TP/(TP + FP) \tag{2}$$

$$F_1 = 2 \times TP/(2 \times TP + FN + FP) \tag{3}$$

5.2 Cross Validation Results

As mentioned above, six classifiers were used in our experiments, namely multinomial naive Bayes, K-nearest neighbour, logistic regression, SVMs, SGD SVM (referred to as SGD) and multilayer perceptron. These classifiers were trained and validated using 5-fold cross validation.

As a first step, different combinations of features were used in order to assess whether they all contributed positively to performance. Initially we included all the discriminant features in our feature set (frequently occurring word unigrams, bigrams and trigrams, word lemmas, vocabulary richness ratio, number of capital letters, number of punctuation marks and the four JAMA criteria – we excluded document length because we determined that it was not a discriminant feature for our domain). However, when validating the system using all of these features, the classification rate was lower than expected. To investigate this issue we started by looking at the JAMA features extraction results and we discovered that we were not achieving the expected extraction rate (of around 80%). This was because some of the JAMA feature indicators occurred outside the main body of the text of the page, and hence were being removed by the corpus cleaning pre-processing step. The effect of this was that the overall performance was reduced. Further experimentation revealed that the formal text features also contributed negatively to overall performance, and we concluded that the n-gram features alone produced the best results with the

Table 2: Cross Validation Results on the Three Datasets

Dataset	Classifiers	Recall	Precision	F_1
Shingles	MNB	93.14	94.84	93.6
	KNN	84.25	87.86	83.2
	Logistic Regression	96.47	96.25	96.28
	SGD	95.36	96.24	94.52
	SVM	96.47	96.25	96.28
	MLP	94.18	95.37	94.5
Flu	MNB	97.78	98.33	97.84
	KNN	87.11	89.49	86.34
	Logistic Regression	100	100	100
	SGD	97.78	98.33	97.84
	SVM	100	100	100
	MLP	100	100	100
Migraine	MNB	75.45	75.79	75.4
	KNN	68.79	72.95	68.12
	Logistic Regression	84.24	80.97	82.32
	SGD	80.91	82.51	81.16
	SVM	89.55	88.03	88.62
	MLP	84.39	81.06	82.18

validation test set. Consequently, just those features were used in the subsequent experiments.

Table 2 shows the obtained results using 5-fold cross validation on the training datasets. It is clear from the results that all classifiers in general performed well on the flu and shingles datasets. However, this is not the case for the migraine dataset (to be discussed further in the discussion section).

We then ran another experiment in which we combined the three datasets into one set, which we refer to as SFM. The results for the SFM dataset are shown in table 3 and it is clear from the results that the system did not perform well on the combined dataset. We hypothesise that this is because of the bad performance on the migraine dataset. To investigate this further, we ran another experiment in which we only combined the flu and the shingles datasets, which we refer to as SF. The results for this set are also shown in table 3. It is clear from the results that the performance of the classifiers in general is better than when applied on the SFM dataset yet not as good as when applied on each dataset separately.

5.3 Results on the test sets

After obtaining the above results on the training datasets we ran the system on the remaining 15% test sets. The results of classifying each dataset separately are shown in table 4 . It is clear that in general all classifiers performed as expected and the classification rate was high in both the shingles and flu test sets especially in the flu test set. An extra row is added to this table in which we used a voted classifier that uses the decisions from all classifier and takes the common decision amongst them.

Two other experiments were performed, the first one on the SFM test set and the other on the SF test set. The results on the two experiment are shown in table 5 . The voted classifier was also used to classify samples in these test sets. Unsurprisingly, and as

Table 3: Cross Validation Results on the Combined Datasets

Dataset	Classifiers	Recall	Precision	F_1
SFM	MNB	85.93	87.31	86.11
	KNN	83.59	84.24	83.39
	Logistic Regression	92.78	93.42	92.74
	SGD	90.81	91.18	90.84
	SVM	90.35	91.39	90.32
	MLP	90.36	90.86	90.26
SF	MNB	93.74	94.79	93.95
	KNN	89.48	89.84	89.47
	Logistic Regression	98.6	98.66	98.56
	SGD	96.48	96.44	96.43
	SVM	97.88	97.91	97.86
	MLP	98.6	98.73	98.6

when validating the system on the training sets, the system did not perform well when having the three test sets combined together SFM. However, the results are better when applying on the SF set.

6 DISCUSSION

Amongst the different classifiers that were used in our experiments, logistic regression and multilayer perceptron performed well in general in comparison to the other classifiers. Figure 1 and 2 show the comparison between the different classifiers when applied to the three test sets and the combined test sets, respectively. It is clear from the figure that all classifiers performed well in all of the

Table 4: Results on the Three Test Sets

Dataset	Classifiers	Recall	Precision	F_1
Shingles	MNB	88.24	94.12	89.78
	KNN	88.24	94.12	89.78
	Logistic Regression	88.24	94.12	89.78
	SGD	94.12	95.59	94.43
	SVM	82.35	95.59	86.73
	MLP	82.35	95.59	86.73
	Voted Classifier	82.35	95.59	86.73
Flu	MNB	100	100	100
	KNN	88.89	92.59	89.57
	Logistic Regression	100	100	100
	SGD	100	100	100
	SVM	100	100	100
	MLP	100	100	100
	Voted Classifier	100	100	100
Migraine	MNB	66.67	77.08	70.37
	KNN	58.33	57.29	57.41
	Logistic Regression	58.33	62.5	60.08
	SGD	66.67	77.08	70.37
	SVM	58.33	57.29	57.41
	MLP	58.33	80.21	67.54
	Voted Classifier	58.33	62.5	60.08

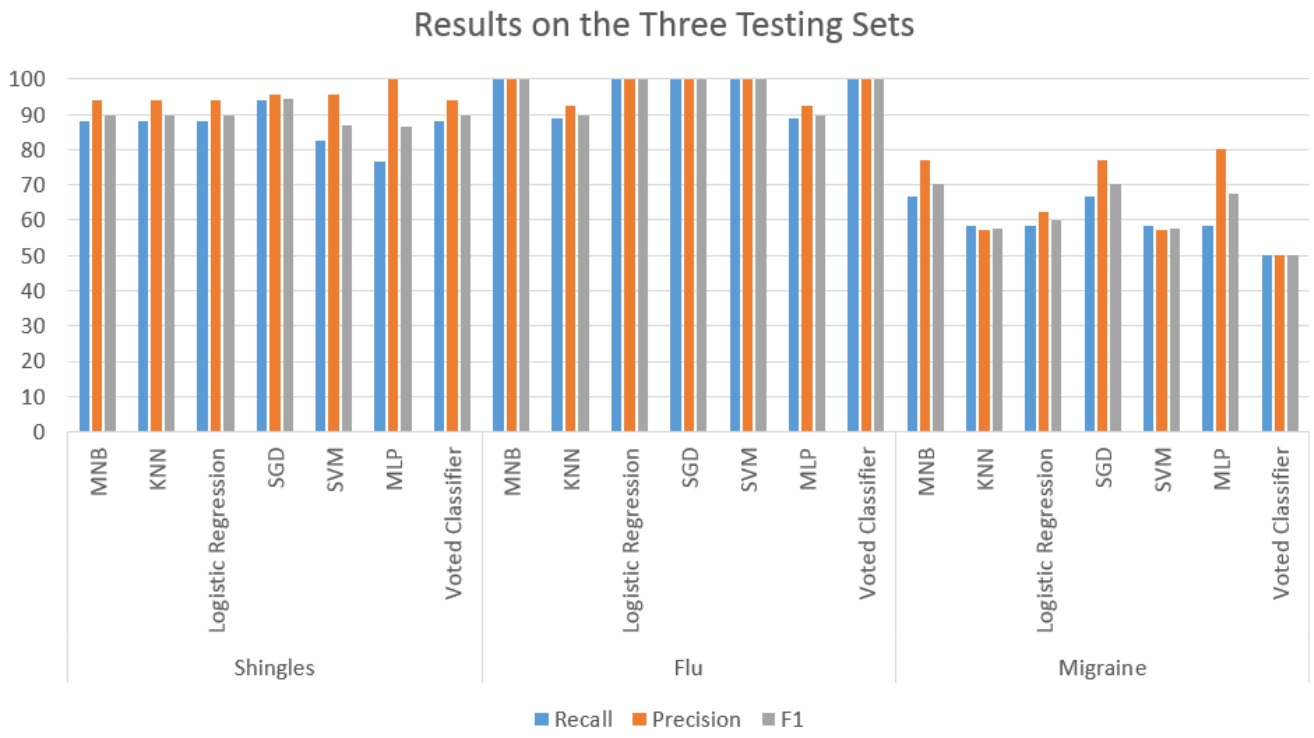

Figure 1: Results on the three test sets

datasets except for KNN which on average performed the worst among them.

It is clear that in general all classifiers performed as expected and the classification rate was high in both the shingles and flu test sets, and especially in the flu test set. However, this is not the case when applied to the migraine dataset. We suggest that this might be due to the fact that migraine is a more complex pathology, arising with many different pathogenic mechanisms, with several types of approved drugs and many different complementary alternative medicine approaches suggested to either cure or prevent it. Additionally, the number of negative (non-EBM) examples in this dataset is more than double the number of positive (EBM) examples and this is exactly the opposite in the other two datasets, which may affect classifier performance. By contrast, influenza (flu) is a disease with a clear pathogenesis (infection with a specific virus), and only two main evidence-based medicine approaches to its prevention (vaccination, hygiene and, in a few cases, antiviral agents).

Table 5: Results on the Combined Test Sets

Dataset	Classifiers	Recall	Precision	F_1
SFM	MNB	73.68	73.65	73.46
	KNN	76.32	85.34	78.36
	Logistic Regression	81.58	81.94	81.7
	SGD	73.68	73.68	73.68
	SVM	78.95	80.02	79.26
	MLP	76.32	76.74	76.47
	Voted Classifier	81.58	81.94	81.7
SF	MNB	84.62	84.62	84.62
	KNN	88.46	93.41	89.61
	Logistic Regression	88.46	93.41	89.61
	SGD	88.46	93.41	89.61
	SVM	88.46	93.41	89.61
	MLP	88.46	93.41	89.61
	Voted Classifier	88.46	93.41	89.61

After combining the three datasets the system did not perform well. We believe this is because of the bad performance on the migraine dataset. It is clear from the results of applying on SF that the performance of the classifiers in general is better than when applied on the SFM dataset yet not as good as when applied on each dataset separately.

One of the limitations of this work is that the webpages that contain both type of interventions (EBM and non-EBM) were excluded from the study, this is because such documents contain features from both categories and hence further work needs to be done. Another limitation is that some diseases, such as cancer and AIDS have no EBM treatments, and hence this method cannot be applied on such diseases.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we tried to assess the quality of online health webpages by identifying evidence-based medical advice automatically

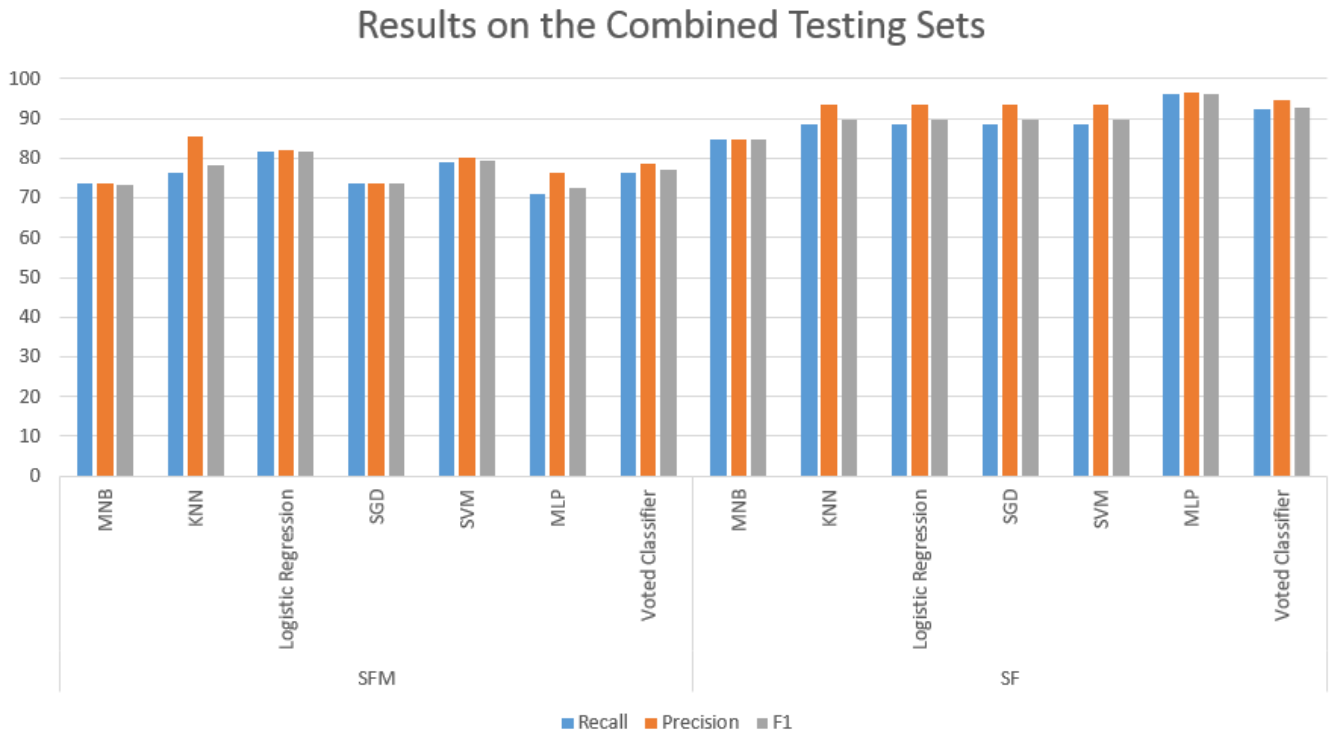

Figure 2: Results on the combined test sets

using natural language processing and machine learning techniques. Three datasets relating to three different diseases were used. A selection of different classifiers were trained and tested on each dataset. Classification recall and precision were high for each individual test set especially on the shingles and flu test sets. However, when the system was trained and tested on all the test sets, combined the recall and precision rates degraded.

In future work we are planning to apply feature selection techniques in order to minimize the number of features and to obtain better classification results. Webpages containing both types of intervention need to be included in future studies as well. As part of the future work, we also plan to increase the size of the existing datasets and also to collect other datasets on other diseases. We will also investigate whether different diseases fall into different classes (easy ones like flu and shingles, and hard ones like migraine) and try to understand the reason behind that. Another direction of research is to assess the quality of online health information using different distinction methods (other than EBM/non-EBM) such as automating the existing quality instruments by rating their criteria mechanically.

REFERENCES

[1] American Medical Association. 2017. JAMA. (2017). Retrieved January 2, 2017 from http://jama.jamanetwork.com/journal.aspx

[2] David Blumenthal. 2002. Doctors in a wired world: can professionalism survive connectivity? *Milbank Quarterly* 80, 3 (2002), 525–546.

[3] Deborah Charnock and Sasha Shepperd. 2017. DISCERN. (2017). Retrieved February 20, 2017 from http://www.discern.org.uk/

[4] Deborah Charnock, Sasha Shepperd, Gill Needham, and Robert Gann. 1999. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of epidemiology and community health* 53, 2 (1999), 105–111.

[5] Elizabeth Cohen. 2010. Your top health searches, asked and answered. (2010). Retrieved January 2, 2017 from http://edition.cnn.com/2010/HEALTH/10/21/top.health.searches.answered/

[6] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[7] Kathleen C Ellwood, Paula R Trumbo, and Claudine J Kavanaugh. 2010. How the US Food and Drug Administration evaluates the scientific evidence for health claims. *Nutrition reviews* 68, 2 (2010), 114–121.

[8] Susannah Fox. 2013. Pew Research Center. (2013). Retrieved February 18, 2016 from http://www.pewinternet.org/files/old-media/Files/Reports/PIP

[9] Arnaud Gaudinat, Natalia Grabar, Célia Boyer, and others. 2007. Machine learning approach for automatic quality criteria detection of health web pages. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*. IOS Press, 705.

[10] Evidence-Based Medicine Working Group and others. 1992. Evidence-based medicine. A new approach to teaching the practice of medicine. *Jama* 268, 17 (1992), 2420.

[11] Health on the Net Foundation. 2016. Health on the Net. (2016). Retrieved February 20, 2017 from https://www.healthonnet.org/

[12] Jeremy H Howick. 2011. *The philosophy of evidence-based medicine*. John Wiley & Sons.

[13] Lexical Computing CZ s.r.o. 2017. Sketch Engine. (2017). Retrieved February 20, 2017 from http://www.sketchengine.co.uk

[14] Ali Maki, Roger Evans, and Pietro Ghezzi. 2015. Bad news: analysis of the Quality of information on influenza Prevention returned by google in english and italian. *Frontiers in immunology* 6 (2015).

[15] Tom M Mitchell. 1997. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill* 45, 37 (1997), 870–877.

[16] NLTK project. 2015. Natural Language Toolkit -- NLTK 3.0 documentation. (2015). Retrieved February 20, 2017 from http://www.nltk.org/

[17] Irina Rish. 2001. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3. IBM New York, 41–46.

[18] scikit-learn Project. 2016. scikit-learn: machine learning in Python - scikit-learn 0.18.1 documentation. (2016). Retrieved February 20, 2017 from http://scikit-learn.org/stable

[19] Edward H Shortliffe, RB Altman, PF Brennan, B Davie, WM Detmer, V Florance, A Friede, M Frisse, J Glaser, J Huffman, and others. 2000. Networking health: Prescriptions for the Internet. *Computer Science and Telecommunications Board, The National Academies. Washington, DC: The National Academies Press* (2000).

[20] Parikshit Sondhi, VG Vinod Vydiswaran, and ChengXiang Zhai. 2012. Reliability prediction of webpages in the medical domain. In *European Conference on Information Retrieval*. Springer, 219–231.

[21] Songbo Tan. 2005. Neighbor-weighted k-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications* 28, 4 (2005), 667–671.

[22] Yunli Wang and Zhenkai Liu. 2007. Automatic detecting indicators for quality of health information on the Web. *International Journal of Medical Informatics* 76, 8 (2007), 575–582.

[23] Yunli Wang and Rene Richard. 2007. Rule-based automatic criteria detection for assessing quality of online health information. *Journal on Information Technology in Healthcare* 5, 5 (2007), 288–299.

[24] Mubashar Yaqub and Pietro Ghezzi. 2015. Adding dimensions to the analysis of the quality of health information of websites returned by Google: cluster analysis identifies patterns of websites according to their classification and the type of intervention described. *Frontiers in public health* 3 (2015), 204.

[25] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. 2012. Recent advances of large-scale linear classification. *Proc. IEEE* 100, 9 (2012), 2584–2603.

[26] Yan Zhang, Yalin Sun, and Bo Xie. 2015. Quality of health information for consumers on the web: a systematic review of indicators, criteria, tools, and evaluation results. *Journal of the Association for Information Science and Technology* 66, 10 (2015), 2071–2084.