Information Sciences 179 (2009) 1870–1880

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

How valuable is medical social media data? Content analysis of the medical web

Kerstin Denecke *, Wolfgang Nejdl

L3S Research Center, University of Hannover, Germany

ARTICLE INFO

ABSTRACT

It is still an open question where to search for complying a specific information need due to the large amount and diversity of information available. In this paper, a content analysis of health-related information provided in the Web is performed to get an overview on the medical content available. In particular, the content of medical Question & Answer Portals, medical weblogs, medical reviews and Wikis is compared. For this purpose, medical concepts are extracted from the text material with existing extraction technology. Based on these concepts, the content of the different knowledge resources is compared. Since medical weblogs describe experiences as well as information, it is of large interest to be able to distinguish between informative and affective posts. For this reason, a method to classify blogs based on their information content is presented, which exploits high-level features describing the medical and affective content of blog posts. The results show that there are substantial differences in the content of various health-related Web resources. Weblogs and answer portals mainly deal with diseases and medications. The Wiki and the encyclopedia provide more information on anatomy and procedures. While patients and nurses describe personal aspects of their life, doctors aim to present health-related information in their blog posts. The knowledge on content differences and information content can be exploited by search engines to improve ranking, search and to direct users to appropriate knowledge sources.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

Electronic media are increasingly used to obtain medical information and advice. Health information on the Internet ranges from personal experiences of medical conditions and patient discussion groups to peer reviewed journal articles and clinical decision support tools. A study on how consumers in America search for health-related information[1] shows that the Web is the most widely used resource for health information. Nevertheless, finding the best knowledge source to comply a specific information need is difficult, because relevant information can be either hidden in web pages or encapsulated in social media data such as blogs and Q&A portals. Through content analysis, this paper tries to give an overview on content differences in the various social media resources on health-related topics.

We focus on health-related information provided in the Internet for two reasons. First, health-related experiences and medical histories offer unique data for research purposes, for practitioners, and for patients. Second, it is still an open question whether existing text and content analysis tools are able to process medical social media data and to identify relevant (medical) information out of them.

* Corresponding author.
E-mail addresses: denecke@L3S.de (K. Denecke), nejdl@L3S.de (W. Nejdl).

[1] How America searches: health and wellness, http://www.icrossing.com/research/how-america-searches-health-and-wellness.php.

0020-0255/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2009.01.025

2. Literature review

2.1. Analysis and assessing social media data in medicine

In the last couple of years, research interest in social media analysis increased due to the growing user interest in these tools. Most of the works focused on weblogs. One research aspect is the analysis of social aspects in weblog communities [16,15]. Sekiguchi et al. detect topics of blogs based on interest similarities of users [22]. Approaches to content analysis and topic detection from weblogs work on determining information diffusion through blogspace [11] or analyze the sentiment of blogs [18,20].

However, the medical domain remained almost unconsidered in most of the existing works in this field. Angel et al. work on a system that measures the quality of medical websites [17,23]. Eysenbach analyzed health-related searches on the World Wide Web [9] and reported that in Google 6.75 Million health-related searches per day are conducted. As presented by Mueller et al. [19], the most common queries for media search in particular are for general terms (e.g. heart and lung). Hillan [13] provides a set of applications and benefits for using blogs by physicians in order to support a patient's treatment. Studies like this show that a huge number of health-related searches are globally conducted every day. But, topics that are mostly discussed and described at medical webpages and medical social media tools are still unknown and will therefore be identified by the analysis presented in this paper.

Nicholas et al. [21] studied the usage of content from five health-related webpages by analyzing transaction files. They also compared the content in terms of 37 health categories or five topics, respectively, and found large differences in content between the five topics. Their content analysis relies on page headings only, and misses therefore an in-depth content analysis as it is presented in this paper. In this work, instead of web pages, the content provided by different medical social media tools is analyzed.

Besides weblog analysis, answer portals and forums are of research interest, in particular, the Yahoo! Answer query portal. Agichtein et al. introduce an approach on separating high-quality items in query–answer portals from the rest [1]. Su et al. present work on how to exploit these portals in order to generate a labeled corpus of named entities [25]. Himmel et al. [14] introduce a classification approach for lay requests in medical expert forums regarding topic or user expectations. Their approach is mainly based on a list of medical terms whose occurrences in a text are determined and used to find the nearest neighboring document. However, there is no in-depth study of medical weblogs and Q&A portals even though this clearly is an interesting and challenging domain due to its influence on a large number of people.

2.2. Information content of documents

Distinguishing *informative* from *affective* posts as considered in this paper is similar to the problem of subjectivity analysis. The main difference in our approach and already existing methods to subjectivity analysis is that the proportion of *affective* and *informative* content is effectively exploited for classification purposes, and specifically targeted to medical blogs. Ni et al. presented in [20] a machine-learning algorithm for classifying *informative* and *affective* articles among weblogs related to the approach in this paper. Their approach differs from ours in the features exploited: They use words as features, while our approach exploits medical concepts and the polarity of words. They tested their algorithm on a blog data collection on different topics in Chinese. In this paper, the focus is on medical texts in English, which are different from blogs and texts of other domains by describing medical concepts and complex coherence like the origin of a disease.

We exploit SentiWordNet [8] as lexical resource for mining affective content. SentiWordNet provides for each synset of WordNet (http://wordnet.princeton.edu/) a triple of polarity scores (positivity, negativity and objectivity) whose values sum up to 1. For example, the triple 0, 1, 0 (positivity, negativity, objectivity) is assigned to the synset of the term *bad*. SentiWordNet has been created automatically by means of a combination of linguistic and statistic classifiers. Like WordNet 2.0 from which it has been derived, SentiWordNet consists of around 207 000 word-sense pairs or 117 660 synsets. It provides entries for nouns (71%), verbs (12%), adjectives (14%) and adverbs (3%). In contrast to the existing approaches that exploit SentiWordNet for quantifying positive and negative sentiments [4,7] or for extracting subjective adjectives [27], we exploit SentiWordNet for determining polarity scores of single words in order to assess the affective content of a post.

2.3. Content analysis

The base method for content analysis is to analyze high-frequency words and draw conclusions based on this information [6,24]. Therefore, a content analysis starts with word frequencies, keyword frequencies, etc. In order to restrict the content analysis to the actual content, stop words, i.e., content free words, are excluded from the analysis. Problems are evoked by word variations (know, knew, knows) and synonyms. To avoid these problems in our approach and to perform a semantically meaningful analysis, natural language is mapped to a standard medical terminology. We are counting frequencies of concepts instead of terms. Synonyms and lexical variations of words are mapped to the same semantic concept and became comparable in this way.

Herring et al. [12] performed a content analysis on weblogs to identify and quantify structural and functional properties of blogs (e.g., number of words, sentences, images, links, and comments) as well as blogger characteristics (e.g., gender, age, and occupation). But so far, content analysis from a semantic point of view, in particular in the field of medicine, remained unconsidered.

Text analysis of medical documents has been the focus of some previous research work (e.g., the medical information extraction systems MedLEE [10], MedIE [28]), but was centered around the processing of clinical narratives (i.e., texts generated in daily practice in a hospital) or research papers. These documents differ significantly from medical blogs and other social media data, in particular, in language and style used by the authors, but also in content (see Section 4.1). Nevertheless, we will build on an existing medical information extraction system and will test its applicability for the classification of health-related content from different Web resources. The extraction results are exploited for a comprehensive content analysis of medical social media data.

3. Research questions

Weblogs and other social media data gain influence, and for this reason, more sophisticated access to this data needs to be provided. Since different user groups have different requirements on the type of information requested, a search engine should enable patients and health care professionals to find experiences or information on diagnoses, treatments or medications, and to restrict search results to texts written by a particular author class (e.g., by a physician, a nurse, and a patient) or to texts of a particular information type (informative/affective) and/or polarity (positive/negative). Existing medical search engines such as Healia,[2] MedWorm[3] or Medlogs[4] do not provide these facilities. Medlogs allows in restricting search results to specific author groups, but search results are always presented as a flat list without hints to the expressed polarity or information content. In addition, there is no search engine available that directs users to the sources with the largest amount of information on a particular topic (e.g. Where can I find information on diabetes to the largest extent?). In order to do this, an overview on the overall Web content available is required. For this reason, we perform a content analysis for different medical social media platforms to make the first steps towards more sophisticated search possibilities.

The research questions we therefore address in this paper are

- Which topics do the different health-related web resources focus on?
- What similarities and differences in content exist between different medical social media data resources?
- To what extent do medical blogs contain information or experiences?

To answer these questions, we first assess the medical content of the data material presented in Section 4.1. For this purpose, methods to extract the medical content of social media data are exploited (see Section 4.2.1), enabling a semantic analysis of the content of the data material in terms of medical categories. According to our hypotheses, Wikis and encyclopedias offer a larger diversity of medical content while blogs are rather centered around specific disorders.

Second, since weblog data contain experiences or may discuss topics unrelated to health, the information type of this particular data source needs to be determined with a method introduced in Section 4.2.2. We assume that weblogs provide rather experiences or affective statements than information related to health or medicine.

4. Research design

In the Internet, different sources of health-related information can be found. Our work focuses mainly on social media tools, in particular, on answer portals, Wikis, Reviews and weblogs that are well known or that are provided by famous communities or institutes (e.g., Mayo Clinic, National Library of Medicine). In Section 4.1, the data collection is described that has been crawled from the indicated web pages. For the analysis whose results are described in Section 6, methods that identify the medical content in texts and those for studying the information type are necessary. These are introduced in Section 4.2.

4.1. Data collection

4.1.1. Query & Answer Forums

The Question & Answer (Q&A) dataset consists of around 9600 questions crawled from the webpages of Netdoctor,[5] Mayoclinic,[6] Yedda[7] and NHS.[8]

Yedda is a query forum where people ask questions and can find answers on any topic (e.g. computers, home and garden). We collected all the queries and related answers that belong to the category 'health + care'. Answers to queries posted at Yedda can be provided by any person, lay man or specialist. In contrast, answers posted on the portals of Mayo Clinic,

[2] http://www.healia.com.

[3] http://www.medworm.com.

[4] http://www.medlogs.com.

[5] http://www.netdoctor.co.uk, 5457 queries.

[6] http://www.mayoclinic.com, 913 queries.

[7] http://yedda.com, 797 queries.

[8] http://www.nhsdirect.nhs.uk, 2439 queries.

Netdoctor and NHS are provided by physicians only. The Mayo Clinic is a non-profit medical practice dedicated to the diagnosis and treatment of illnesses. On its website, a huge amount of health information and tools is provided (e.g., self-assessments and calculators). NHS Direct is an organization that offers different health-related content including a health encyclopedia, a self-help guide and a set of common health questions that represents our NHS dataset. Netdoctor is a collaboration between physicians, health care professionals and patients. The website delivers information related to health as well as various services such as query facilities. The corresponding queries provide the Netdoctor query set.

The formulation of questions differs significantly between the different portals under consideration. Users posting questions at Netdoctor and Yedda tend to give complete descriptions of symptoms, observations, and experiences, while users of NHS and Mayo post very compact, direct questions. Different types of questions can be distinguished, for example, questions on concrete disorders and symptoms (e.g., *Is pneumonia contagious?*) or queries on the definitions of medical terms (e.g. *What are blood types?*). Depending on the question, the answer length differs significantly from one up to thirty or more sentences throughout all portals. While all queries posted at Netdoctor, Mayo and NHS were answered, some questions at Yedda remained unanswered or got only a comment instead of an answer.

4.1.2. Medical weblogs

The second part of the dataset consists of posts of medical weblogs, i.e., weblogs whose main topics are medicine or health care. Medical weblogs can be differentiated with regard to their author into blogs written by health care professionals and those written by patients. Patients share in their blogs experiences on the disease they are suffering from or on a particular treatment, or exchange other health-related information. Doctors provide health information and insights into their daily clinician life; they report on problems in patient treatment or discuss political decisions. Nurses write on general things (e.g., books and family), about their experiences in taking care of patients, and about conflicts with physicians, but resist on offering information on disease or treatments.

In addition, it is useful to distinguish blog posts based on their medical relevance and information content. The *information content* of a post describes its proportion of *affective* content and *informative* content. Using this measure, we can distinguish primarily *informative* from primarily *affective* posts. Posts with a large information content are considered *medically relevant*.

In an *affective post* the author describes actions he performed during a day; it contains his thoughts on treatments, diseases, medications or his feelings. *Affective* posts can also lack medical content completely and contain, for example, only links to other web sites (so called *filters*). On the other hand, a medical post is considered *informative*, if it contains general or disease- (and/or treatment-) specific information, news on current research results or on the health care system or on general/ transferable experiences regarding a particular treatment or disease.

The weblog dataset consists of 5274 patient-written posts, 7852 physician-written posts (6715 hosted at Blogspot.com, 1137 hosted at WebMD) and 5724 nurse-written posts. The corresponding 95 weblogs have been selected randomly by collecting addresses from the two (medical) weblog search engines Medworm[9] and Medlogs.[10] Most of these weblogs are hosted by Blogspot.com. Physician weblogs were also collected from the webpage WebMD[11] that provides 36 blogs of physicians with different clinical specialities.

From the linguistic point of view, medical weblogs usually consist of syntactically correct sentences (e.g., *The Mayo Clinic will suggest PCR*), but can contain verbless clauses (e.g., *Paperwork, paperwork*, and *paperwork*) or sentences without subject (e.g., *Take out the garbage*). In addition, abbreviations (e.g., *CLL* (*chronic lymphocytic leukemia*)), enumerations and citations of conversations as well as common speech, medical terms and opinion-related words are used frequently in medical blog posts. Patient-written posts are longer in terms of the number of words (approximately 387 words) and sentences (21 sentences in average) than the posts of nurses and physicians, where around 13 sentences and 300 words could be identified in average.

4.1.3. Reviews

The third dataset consists of 3731 drug reviews for 630 drugs that have been collected from Drugratingz.com. This webpage is part of a family of websites dedicated to help consumers in finding the best businesses, places, and services by sharing ratings and reviews. Users can anonymously rate drugs in several categories, including effectiveness, side effects, convenience and value; they can post and read comments. These comments, dealing with symptoms and side effects, provide the data set for our analysis while the ratings remain unconsidered. Their length ranges from one-word reviews (e.g., *Great*) to reviews with up to 15 sentences.

4.1.4. Wikis and encyclopedias

The fourth dataset consists of 725 pages of AskDrWiki[12] and 750 pages crawled from Medline Plus.[13] Dr Wiki is a non-profit educational website made by physicians for physicians, medical students, and healthcare providers. Its purpose is to serve as an

[9] http://www.medworm.com.

[10] http://www.medlogs.com.

[11] http://www.webmd.com.

[12] http://askdrwiki.com.

[13] http://www.nlm.nih.gov/medlineplus/.

online repository of medical information that can be accessed by anyone. The main focus of this Wiki has been on Cardiology and Electrophysiology, but it has been also expanded to other medical specialities.

MedLine Plus is a service provided by the US National Library of Medicine and the National Institutes of Health. Its objective is to direct users to information that helps answering their health questions. From the MedLine Plus webpage, we crawled the 'health topics' which is a collection of 750 topics on conditions, diseases and wellness. The summaries of these webpages provide the MedLine Plus dataset in our analysis.

4.2. Data analysis

4.2.1. Assessing the medical content

The medical content of a text is determined by extracting medical concepts using an information extraction system, SeReMeD, which was originally developed for the analysis of medical texts describing diagnoses and findings by doctors about a certain patient [5]. The SeReMeD system aims to represent natural language text in a standardized and normalized manner and to extract relevant information from texts. The system's domain knowledge provides the Unified Medical Language System (UMLS, http://www.nlm.nih.gov/research/umls/). Mapping of natural language to UMLS concepts is realized within SeReMeD by MetaMap [2].

The medical terminology UMLS consists of around 1.7 Million biomedical concepts, where each concept is assigned to at least one of the 134 specified semantic types. The semantic types are grouped in turn into 15 semantic groups (e.g., the concept *atrial fibrillation* belongs to the semantic types *Finding* and *Pathologic Function* that in turn belong to the group *Disorders*). Both, semantic groups and semantic types, are exploited by the information extraction system SeReMeD.

For processing a medical text, SeReMeD first analyzes the syntactic structure of each sentence and produces a semantic representation that consists of UMLS concepts, their semantic roles (e.g., NEG, EXCLUSION) and relations derived from linguistic dependencies, such as *accompanied_by*. To extract specific information from a document, the semantic representation is searched for the concepts of a particular semantic group (such as the semantic group *Disorders*), of a specific semantic type (for example, the type *Finding*) or for concepts bearing a particular semantic role. Since semantic groups integrate the concepts of different semantic types, search for concepts belonging to a particular semantic group provides more general results, while search for the concepts of specific semantic types allows a rather specific extraction.

The content analysis described in this paper exploits SeReMeD to extract concepts of the semantic types listed in Table 1. In particular, concepts describing diagnoses, procedures, medications, anatomy and physiology are extracted. We restricted the extraction process to the concepts of these types to avoid an extraction of very general concepts, e.g., the very general Disorder concepts *Problem*, *Medical History*, and *in care*.

For content analysis, the frequency F_{cat_A} of concepts belonging to the main category cat_A, which is one of the five mentioned semantic groups, is calculated. Furthermore, the proportion P_{cat_A} of concepts of a main category cat_A on extracted concepts is determined. The following formula are exploited:

$$(1)\ F_{cat_A} = \frac{n(cat_A)}{doc} \quad \text{and} \quad (2)\ P_{cat_A} = \frac{n(cat_A)}{c}$$

where $n(cat_A)$ is the number of extracted concepts of category cat_A, doc is the number of documents per collection, and c is the number of extracted concepts of all main categories. In addition to compare the content of the different information sources, the frequencies F_{cat_A} for the semantic groups Disorders, Procedures and Chemicals & Drugs are used for post classification described in Section 4.2.2.

4.2.2. Assessing the information type of documents

As mentioned before, weblog posts will be analyzed with regard to their information type. This problem is considered as binary classification problem, i.e., a post is either *affective* or *informative*. The suggested approach is based on two assumptions:

Table 1
Considered semantic types for extraction of medical concepts.

Semantic group	Semantic types
Disorders, e.g., Headache	Pathologic Function, Disease or Syndrome, Sign or Symptom, Injury or Poisoning, Mental or Behavioural Dysfunction, Acquired Abnormality, Neoplastic Process
Procedures, e.g., Hypnosis	Laboratory Procedure, Diagnostic Procedure, Therapeutic or Preventive Procedure
Chemicals and Drugs, e.g., Aspirin	Pharmacologic Substance, Antibiotic
Anatomy, e.g., Back	Anatomical Structure, Body Substance, Body System, Body Location or Region, Body Part, Organ, or Organ Component,Tissue, Body Space or Junction, Cell Component, Cell, Embryonic Structure, Fully Formed Anatomical Structure, Tissue
Physiology, e.g., Thirst	Cell Function, Clinical Attribute, Genetic Function, Mental Process, Molecular Function, Organism Attribute, Organism Function, Organ or Tissue Function, Physiologic Function

1. Extensive use of medical terminology is an indication for informative content.
2. Adjectives are an indication for affective content.

Therefore, our classification algorithm exploits features that are collected by the module for extracting medical content (see Section 4.2.1), and an additional module for assessing affective content, described in this section. The informative content is described by F_{cat_A}-values for the main categories Disorders, Procedures and Chemicals & Drugs. The affective content of a medical text is determined by means of SentiWordNet. For this purpose, all words of the word classes *adjective, noun* or *verb* are stemmed and the corresponding SentiWordNet entries are collected from the word-sense pair entries. Scores of synonyms remain unconsidered, while scores of different synsets are averaged. This results in a polarity score triple for each sentiment-bearing term.

The polarity scores of words belonging to the same part of speech are summed up, and their average is calculated. As resulting features for classification, three document polarity score triples are calculated: one triple for verbs, for adjectives and for nouns. Each triple consists of a positivity, a negativity and an objectivity value. As additional features, the frequency of positive, negative and objective words of a document is determined. For this purpose, we assign to each word a polarity (*positive, negative, objective*) based on its polarity score triple and by applying the Rule 1. The number of *positive*, *negative*, and *objective* words is counted and divided by the number of terms.

```
Rule 1: Determining the polarity of a term
    If $score_{pos}(A) + score_{neg}(A) \leqslant 0.1$ Assign objective
    Else if $score_{pos}(A) > score_{neg}(A)$ Assign positive
    Else if $score_{pos}(A) \leqslant score_{neg}(A)$ Assign negative
    Else Assign objective
where $score_{pos}(A)$ and $score_{neg}(A)$ are the positivity and the negativity scores of the considered term A.
```

The final classification is based on 21 high-level features, i.e., on 4 features that represent the *informative content* of a post (frequency of extracted diagnoses, procedures, medications and concepts), 12 features that describe the *affective content* (frequency of positive, negative, objective words, and the polarity scores), and 5 stylistic features (frequency of adjectives, nouns, verbs, number of sentences and token).

For classification purposes, different algorithms implemented in the WEKA library [26] have been tested on our feature set. The best performing algorithm in 10-fold cross-validations (see Section 5) is a SimpleLogistic Classifier that is based on logistic regression models and is a well-known technique for learning scenarios where all attributes are numeric. Compared to a Naive Bayes Classifier, the different attributes depend on each other.

5. Evaluation of the information type classification

5.1. Evaluation methodology

Before we apply the introduced method for blog post classification on the weblog dataset, its performance in a 10-fold cross-validation is tested. For this purpose, some weblogs from all author groups have been randomly selected. The corresponding 1509 posts were classified manually *affective* and *informative*. The evaluation corpus is almost balanced and consists of 771 affective and 738 informative posts. Table 2 shows the distribution on the two different classes per author group.

The purpose of the evaluation is to determine the classification quality of the approach and to compare these results with some baseline results. Furthermore, differences in classification accuracy for the posts of different author groups will be studied. A standard method in text classification is to exploit the frequencies of words as classification features (aka bag-of-word approach or vector space method). Therefore, the baseline results are determined by calculating the frequency of words in all posts of the evaluation material. The final feature set only considers words that occur in at least three posts of the document collection. Stop words were removed from the corpus. This results in 6066 attributes (frequencies of words) for the baseline. The approach presented here exploits 21 features as described above. A last set of features combines the baseline attributes and the high-level features and results in 6087 attributes for classification.

Table 2
Evaluation material.

Author	Informative texts	Affective texts
Patient	303	369
Doctor	271	188
Nurse	164	214
Total	738	771

Table 3
10-Fold cross-validation results of patient- and physician-written posts for different feature sets in % (P = precision, R = recall, F = F-score, Acc = accuracy).

ID	Informative			Affective			Accuracy
	Precision	Recall	F-score	Precision	Recall	F-score	
Bag-of-words approach	70	72.1	71	70.4	68.2	69.3	70
High-level feature	**77.9**	**76.3**	**77.1**	**74.9**	**76.6**	**75.7**	**76**
Combination	77	77.3	77.1	77.7	77.4	77.5	77.3

Precision, recall and F-measure values are determined per class for the different feature subsets. Precision per class is calculated by dividing the number of correct classified texts by the number of texts of one class. Recall is calculated by dividing the number of correct classified texts by the number of all texts. F-measure is the weighted harmonic mean of precision and recall. All results are tested for statistical significance using T-tests.

5.2. Classification results

A 10-fold cross-validation on the complete training set achieves the accuracy values of 71%. These results are not significantly better than the results of the bag-of-words approach on the same material (67% accuracy). The main reason for the low accuracy value is the classification of nurse-written posts. If these posts are excluded from the data material, our approach performs much better than the baseline (see Table 3).

It can be seen that by exploiting high-level features for information content classification, better results can be achieved than by exploiting a simple bag-of-words approach. There is a statistically significant difference between these two accuracy values (99% confidence). The combination of the feature sets results in a slight improvement of classification accuracy. Informative posts are slightly better classified than affective posts for all feature sets.

In Table 4, the performance of the described approach on the posts of different author groups is compared. The results have been determined by means of 10-fold cross-validations for the posts of each author group separately. As mentioned before, nurse posts are classified with the lowest accuracy. The best results are achieved for doctor-written posts (accuracy of 85%). These results differ significantly from the values determined for the other two author groups and it shows that the high-level features are well suited for information content classification. But, it is more difficult to extract these features from patient- and nurse-written posts due to differences in the language used by the different author groups (clinical language vs. common language). Patients tend to avoid their illness' name (e.g., *the beast* instead of *migraine*) or use only abbreviations (e.g., *Type 2* instead of *diabetes type 2*). As mentioned in Section 4.1, nurses have different topics to write about than patients and doctors. Instead of focussing on diseases and particular treatments, nurses write about a huge variety of topics (daily life, wellness, family, etc.). This leads to a more descriptive language and results in more frequent misclassification.

If we extend the feature set by the frequencies of other main categories (e.g., Anatomy and Physiology), the accuracy values remain almost stable. This suggests that these features are already well suited for distinguishing *informative* and *affective* medical posts correctly. Even for our rather simple approach for determining polarity scores and polarity of words, promising classification results can be achieved.

6. Content analysis results

In Section 6.1, we study the medical content of the five resources of our dataset. In Section 6.2, the distribution of the two information types on the weblog dataset is presented. The results are discussed in Section 7.1.

6.1. Medical content of medical social media data

Result 1: *The main topics in medical weblogs are medications, physiology and disorders. Patients are rather concerned by medications, while physicians focus on illnesses. Posts written by physicians differ in the extent of medical information offered.*

Fig. 1 shows the results of a comparison of concept distributions for different categories on blog posts of the three author groups. Obviously, nurses offer in their posts the smallest amount of medical concepts for almost all categories considered.

Table 4
Classification results for different author groups in % (P = precision, R = recall, F = F-score, Acc = accuracy).

Author	Informative			Affective			Accuracy
	Precision	Recall	F-score	Precision	Recall	F-score	
Patient	69.5	67.9	68.7	72.2	73.7	72.9	71
Doctor	87	86.7	86.8	81	81.4	81.2	85
Nurse	65.9	55.8	60.5	69.9	78	73.7	68

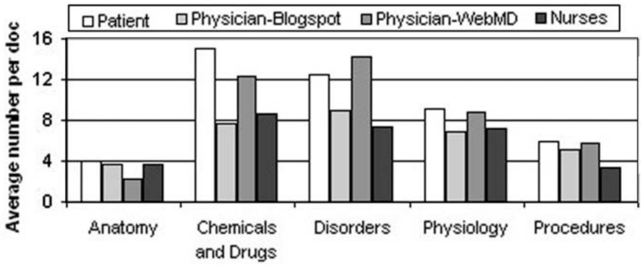

Fig. 1. Content comparison for posts of different author groups.

As mentioned before, nurses mainly do not focus on health-related topics, but they focus on a huge variety of topics. Significantly more drug-related concepts are identified in patient-written posts than in posts of the other author groups. WebMD posts provide the largest number of concepts related to disorders. The other categories seem to be discussed by patients and physicians at WebMD to a similar extent. In contrast to this, physician-written blogs hosted at Blogspot.com contain significantly less medical concepts than patient-written posts for most of the categories. This shows on the one hand that patient-written posts can provide medical information to a larger extent than physician-written posts. On the other hand, the extent of medical information described differs depending on the platform where the blog is hosted. Physician-written blogs at Blogspot.com are more general in content meaning that the authors focus on their life and experiences. For this reason, WebMD posts are probably better suited for information search in weblogs than physician-written posts from blogspot.com due to a larger amount of medical knowledge provided.

Result 2: *In Q&A portals, people mainly post queries related to drugs and disorders. In particular, the Mayo portal provides the largest amount of medical content on these categories. Information on treatment-related issues captures a smaller part only.*

Fig. 2 shows the distribution of main categories for the different query portals. Even though the Mayo dataset was the smallest one, it provides the largest number of concepts on anatomy, medications and disorders. This shows that Mayo answers provide medical content to a larger extent than the other portals. Concepts determined for the Netdoctor corpus are, for example, more general and not always related to healthcare. NHS texts mainly deal with the topics procedures and physiology. These topics are less covered by the other portals that focus mainly on illnesses, medications and physiology.

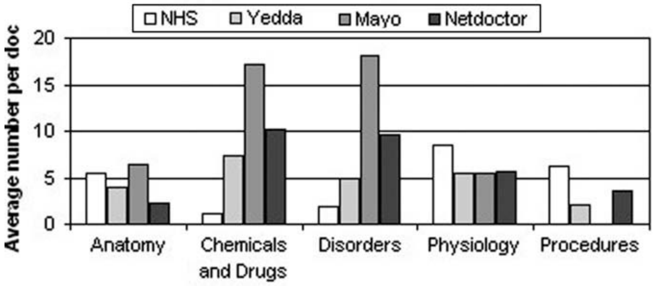

Fig. 2. Comparison of distribution on main categories for different Q&A portals.

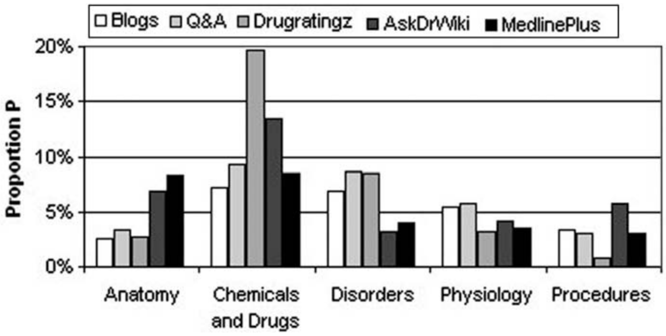

Fig. 3. Proportions of concepts on main categories for different data sources.

Table 5
Distribution of affective and informative posts of the weblog material.

Author	Informative (%)	Affective (%)
Patient	10.4	89.6
Doctor	41.9	58.1
Nurse	29.5	70.5

Result 3: *To find information on disorders and physiology, weblogs and Q&A portals are better suited than the other resources. In particular, information on diseases can be found best in physician-written blogs at WebMD. AskDrWiki and MedlinePlus should be queried for requests on anatomy and procedures. Information on drugs are provided in drug reviews and AskDrWiki best.*

We compared the proportion of extracted concepts of the different main categories for the different data sources (see Fig. 3). Unsurprisingly, drug ratings contain the largest number of concepts related to drugs. But, these comments also provide disease-related concepts to a large extent. This reflects the overall purpose of drug ratings: People write about medications and their influence on a disease or a physical condition.

A substantial number of concepts describing drugs and procedures could be identified from Wiki pages. MedlinePlus mainly offers information on anatomy. Diseases are obviously less relevant in the Wiki and the encyclopedia compared to other resources. The content of weblogs and Q&A portals seems to be similar, mainly focussing on disorders, medication and physiology. Compared to the Wiki and the encyclopedia, less information on anatomy and procedures could be identified.

6.2. Information type of medical weblogs

The information type classification is only performed for the weblog dataset, since the other resources are considered to be mainly *informative*. The manually classified posts of physicians and patients used for the evaluation in Section 5 are exploited to train the classifier. The percentage of posts that have been classified *informative* and *affective* per author group is determined (see Table 5). Around 42% of the physician-written posts are *informative*. Therefore, physicians offer in their blog posts information to a larger extent than patients and nurses who rather write about experiences or focus on more general topics, sometimes unrelated to health.

Surprisingly, more nurse-written posts are classified *informative* than patient-written posts. The content analysis results presented in Section 6.1 showed that less medical concepts could be extracted from nurse-written posts than from patient-written posts. This result may lead to the assumption that more patient-written posts are *informative* than nurse-written posts. Considering the results of the 10-fold cross-validations presented in Section 5.2, a possible explanation for this contradiction is that nurse-written posts have been classified incorrectly.

7. Conclusions and further research

7.1. Limitations and discussion of the results

Several conclusions can be drawn from the aforementioned results. Our hypotheses proved only to be partly true. Instead of offering a large diversity on topics as hypothesized, a focus on anatomy could be identified in the Wiki and the encyclopedia. We conclude that the latter are best suited to find information on anatomy, while people searching for information on disorders should be directed to weblogs or Q&A portals. We remark that we only considered one Wiki and one encyclopedia and that other sources of these types can completely differ in content. In future, we plan to expand our analysis to other sources in order to further support the results presented in this paper.

The presented results allow to suggest platforms suited best for specific search tasks. Because of the large amount of extracted medical content, people searching for information, for example, on disorders, should be directed to the Mayo Query Portal. It turned out that physician-written posts, in particular, those provided on a health platform like WebMD, are better suited when searching for health-related information. People searching for experiences on drugs or disorders will make a find on patient-written weblogs.

The evaluation of the information type classification shows that the introduced approach depends on the extraction of relevant features. If medical concepts are insufficiently identified in a text, the algorithm fails. In the posts of doctors, these concepts can be identified more easily. This leads to better classification results. But also for the other author groups, good classification results are achieved. In this work, we applied the classification approach only to weblog data. In future work, we plan to test the proposed method to identify informative (or ‘good’) answers to health-related queries in rather general Q&A portals such as Yedda, where answers can be given by any person. This could help to filter out comments and irrelevant answers.

In cases, where a post contains as much affective content as informative content, it is difficult for our algorithm, but also for humans, to categorize correctly (e.g., posts that introduce new aspects on a treatment that are immediately commented). To overcome this limitation, the classification problem could be redefined, such that for each class the problem is considered as a binary classification problem. Another possibility is, instead of deciding for one of the two classes, to present the proportion of informative and affective content as result. A classification per sentence would be useful to determine this proportion.

As an interesting subtopic, aspects on healthcare politics remained unconsidered since they cannot be extracted as medical content. Only diagnoses, procedures and medications are extracted as medical content. Posts that discuss health-politics may be classified incorrectly. An extension by means of a named entity recognition system that recognizes political parties, names of politicians and similar entities is necessary in order to overcome this limitation.

The performance of the SeReMeD system that is exploited to identify medical concepts in social media data remains unconsidered in this paper. As reported in [5], it performs with precision and recall values between 81-96% and 83-98% on clinical narratives. We recognized that false-positive concepts are extracted due to ambiguities (e.g., the string *Stress* is mapped to the concept [*Organic Chemical: Stress bismuth subsalicylate*] instead of [*Finding: Stress*]). Since none of the existing medical information extraction systems has originally been developed to process social media data, it is still an open issue to investigate their robustness and quality regarding processing this particular kind of data. Concepts are sometimes not identified because of misspellings that can occur quite often in social media data. SeReMeD does not currently dispose of a writing error correction or a soundex algorithm. This would be a future extension.

Bath argued in his article that the identification of health information needs is still a challenge in health informatics [3]. Content analysis of social media data like the one described in this paper can be used to identify these information needs: Under the assumption that people tend to write and ask questions about things they are interested in, a content analysis of weblog posts and query portals, in particular, reveals topics of high user interest. In future work, we will focus on these issues in more detail.

7.2. Summary

In this paper, results of a comprehensive content analysis of different health-related Web resources have been presented. It turns out that the analyzed data resources vary significantly in content and amount of presented medical information. This work is unique in a sense that such an analysis was still missing in particular for the domain of medicine. Studies like this provide an overview on the content available in the (medical) Web. They help to identify the best-suited information source in order to comply a specific information need. Search engines can benefit users of these results by ranking search results appropriately or by directing users to the best-suited information source.

Furthermore, this paper introduced a new approach for classifying medical weblogs according to their information type. In contrast to existing approaches, we focused on weblogs in the domain of medicine, which remained largely unexplored until now. We described how the results of entity extraction and sentiment analysis can be exploited for efficient and effective classification purposes. Our approach allows us to distinguish *affective* and *informative* medical posts based on the extraction of *informative* and *affective* content. A potential application of our algorithm is its exploitation for sorting or ranking search results within a blog post search engine. Posts that are mainly *informative* can be ranked higher than mainly *affective* posts, for most users. Search results could also be restricted to *informative* posts, if preferred. *Affective* posts can be sorted by their main opinion (e.g., positive and negative).

References

[1] E. Agichtein, Finding high-quality content in social media, in: WSDM 2008: Proceedings of the International Conference on Web Search and Web Data Mining, 2008, pp. 183–194.

[2] A. Aronson, Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap, in: Program Proceedings of the AMIA Symposium, 2001, pp. 17–21.

[3] P.A. Bath, Health informatics: current issues and challenges, Journal of Information Science 34 (4) (2001) 501–518.

[4] F.R. Chaumartin, UPAR7: a knowledge-based system for headline sentiment tagging, in: Proceedings of the SemEval Workshop, Prague, 2007, pp. 422–435.

[5] K. Denecke, Semantic structuring of and information extraction from medical documents using the UMLS, Methods of Information in Medicine 47 (5) (2008) 425–434.

[6] G.W. Ryan, H.R. Bernard, Data management and analysis methods, in: N.K. Denzin, Y.S. Lincoln (Eds.), Handbook of Qualitative Research, second ed., Sage Publications Inc, 2007, pp. 768–802.

[7] A. Devitt, K. Ahmad, Sentiment polarity identification in financial news: a cohesion-based approach, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, 2007, pp. 984–991.

[8] A. Esuli, F. Sebastiani, SentiWordNet: a publicly available lexical resource for opinion mining, in: Proceedings of the Fifth Conference on Language Resources and Evaluation, LREC, 2006.

[9] G. Eysenbach, C. Kohler, What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the Internet, in: Proceedings of the AMIA Annual Symposium, 2003, pp. 225–229.

[10] C. Friedman, A broad-coverage natural language processing system, in: Proceedings of the AMIA Annual Symposium, 2000, pp. 270–274.

[11] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, Information diffusion through blogspace, in: Proceedings of the 13th International Conference on World Wide Web, 2004, pp. 491–501.

[12] S.C. Herring et al, Longitudinal content analysis of weblogs: 2006–2007, in: M. Tremayne (Ed.), Blogging, Citizenship, and the Future of Media, Routledge, London, 2006.

[13] J. Hillan, Physician use of patient-centered weblogs and online journals, Clinical Medicine and Research 1 (4) (2003) 333–335.

[14] W. Himmel, U. Reincke, H.W. Michelmann, Using text mining to classify lay requests to a medical expert forum and to prepare semiautomatic answers, in: SAS Global Forum, 2008.

[15] W.H. Hsu, A.L. King, M.S.R. Paradesi, T. Pydimarri, T. Weninger, Collaborative and structural recommendation of friends using weblog-based social network analysis, in: Proceedings of Computational Approaches to Analyzing Weblogs – AAAI 2006 Technical Report SS-06-03, Stanford, CA, 2006, pp. 55–60.

[16] R. Kumar, J. Novak, P. Raghavan, A. Tomkins, On the bursty evolution of blogspace, in: WWW'03: Proceedings of the 12th International Conference on World Wide Web, 2003, pp. 568–576.

[17] M. Angel Mayer et al, Quality labelling of medical web content, Health Informatics Journal 12 (1) (2006) 81–87.

[18] Q. Mei, X. Ling, M. Wondra, H. Su, C. Zhai, Topic sentiment mixture: modeling facets and opinions in weblogs, in: WWW'07: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 171–180.

[19] H. Mueller et al., Analyzing web log files of the health on the net HONMedia search engine to define typical image search tasks for image retrieval evaluation, in: K. Kuhn et al. (Eds.), Proceedings of the MEDINFO 2007, 2007, pp. 1319–1323.

[20] X. Ni, G.R. Xue, Y. Yu, Q. Yang, Exploring in the Weblog Space by Detecting Informative and Affective Articles, in: WWW'07: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 181–190.

[21] D. Nicholas, P. Huntington, J. Homewood, Assessing used content across five digital health information services using transaction log files, Journal of Information Science 29 (6) (2003) 499–515.

[22] Y. Sekiguchi, H. Kawahima, H. Okuda, M. Oku, Topic detection from blog documents using users' interests, in: Proceedings of the Seventh International Conference on Mobile Data Management, 2006.

[23] C. Sellitto, S. Burgess, Towards a weighted average framework for evaluating the quality of web-located health information, Journal of Information Science 31 (4) (2005) 260–272.

[24] S. Stemler, An overview of content analysis practical assessment, Research & Evaluation 7 (17) (2001).

[25] Q. Su, Internet-scale collection of human-reviewed data, in: WWW'07: Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 231–240.

[26] I. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, second ed., Morgan Kaufmann Publishers, Amsterdam, 2005.

[27] E. Zhang, Y. Zhang, UCSC on TREC 2006 Blog Opinion Mining, TREC 2006 Blog Track, Opinion Retrieval Task, 2006.

[28] X. Zhou, H. Han, I. Chankai, A. Prestrud, A. Brooks, Approaches to text mining for clinical medical records, in: Proceedings of 2006 ACM Symposium on Applied Computing, Dijon, France, 2006, pp. 235–239.