Textual and Informational Characteristics of Drug Related Content on Three Kinds of Websites: Drug Review Website, Discussion Board and Hospital Information Portal
LORRAINE GOEURIOT, JIN-CHEON NA, WAI YAN MIN KYAING, CHRISTOPHER S.G. KHOO, YIN-LENG THENG, YUN-KE CHANG, & SCHUBERT FOO
Division of Information Studies
Wee Kim Wee School of Communication & Information
Nanyang Technological University
31 Nanyang Link, Singapore 637718
lorraineg@ntu.edu.sg; tjcna@ntu.edu.sg; kymwai@ntu.edu.sg; assghoo@ntu.edu.sg;
yltheng@ntu.edu.sg; ykchang@ntu.edu.sg; assfoo@ntu.edu.sg

Abstract
The overall goal of this project is to develop techniques to mine medical and health-related social media content for information and opinions about diseases and treatments. This would allow users to benefit from the numerous postings of people’s experience scattered all over the Web. This paper reports the results of a textual analysis of user-generated content on drug-related online sites. The aim is to determine what kinds of content can be expected on these sites, from linguistic and information points-of-view. User postings were harvested from two websites carrying different kinds of user-generated content and compared to information on the same drugs from a hospital information portal. The corpus was analyzed to identify what kinds of drugs were often reviewed, the vocabulary and medical concepts used, and the textual characteristics such as length of postings, sentence length, and part-of-speech distribution. Although drug-related user-generated content is very different from editorial content, the results of this project show that it provides useful information on many drugs. From a linguistic point-of-view, user-generated content is simpler in language and more informal in style than editorial content, but still contains useful health information that tends to be patient-oriented.

Keywords: User-generated content, discussion forums, medical and health information, content analysis, social media, drug reviews, opinion mining.

1. Introduction
The quantity of online health and medical information is constantly increasing. Health related websites have become one of the most important public information sources on health, disease and treatment (California Healthcare Foundation, 2008). Web 2.0 technologies have given rise to a host of social media sites, such as discussion forums, blogs, and user-review sites, which allow users to contribute their own content. The emergence of such new data in large quantities has given rise to questions about information searching and quality of information. Finding relevant, useful and trustworthy information on the Web is a real challenge. One of the attractions of social media is the support for sharing information among people having similar experiences and situations.

Previous studies have shown that this user-generated content is useful from different points of view. First, users are often looking for stories from “patients like them” on the Internet, which they cannot always find among their friends and family (California Healthcare Foundation, 2008). Moreover, studies investigating the impact of social media on patients have shown that for some diseases and health problems, online community support can have a positive effect (e.g., Jaloba, 2009; Schraefel et al., 2009). Because of its novelty as well as quality and trustworthiness issues, user-generated content is underexploited. It needs to be further studied, understood and then leveraged in designing new online tools and applications.
The overall goal of our project is to develop techniques to mine medical and health-related social media content for information and opinions about diseases and treatments. In a previous paper (Goeuriot et al., 2011), we reported preliminary results of a textual analysis of user-generated content on drug review websites. In this paper, we extend the analysis to three different drug information sources: drug review websites, discussion boards and hospital portals. We evaluate the quality of the websites and their content in the light of opinion mining on drugs. In this context, we define quality according to three criteria: (i) Websites must cover a wide range of drugs; (ii) They must contain texts that are both informative (i.e. containing medical information) and affective (i.e. containing opinions); (iii) The texts should be syntactically and lexically correct. However, in this study, the validity of social media content is not investigated since it requires in-depth manual analysis with medical domain knowledge.

In Section 2, we describe related works on health-related social media. Then we describe the compilation of our corpus (Section 3). After an analysis of the drug covered in our corpus (Section 4), we evaluate its linguistic characteristics (Section 5), its medical content (Section 6) and its subjective or opinionated content (Section 7). Finally, in Section 8, we discuss our findings in the light of quality and conclude the paper (Section 9).

2. Related works
Health information on the Internet is a highly discussed topic, raising both concern and enthusiasm. Recently, health on social media, so-called Health 2.0, is attracting much interest. According to the California Healthcare Foundation (2008, p. 2), it can be defined as: “the use of social software and its ability to promote collaboration between patients, their caregivers, medical professional and other stakeholders in health”. Social software refers to blogs, discussion forums, podcast, social networks and wikis. We will focus the literature review on three kinds of studies: social media users, user-generated content, and use of this content and its applications.

2.1 Health-related social media users
According to Nettleton et al. (2005), the increasing use of the Web for health matters raises three kinds of reaction: concern for physicians, enthusiasm for sociologists, and a contingent and embedded response.

The California Healthcare Foundation reported that, according to a 2008 survey in the US, 60% to 80% of Americans had used the Internet to find health information, which rivaled physicians as the main source. Moreover, a Pew Research Center report (2010) on the use of social media among teens and young adults stated that 31% of the interviewed “online teenagers” used the Internet to look for health, dieting and physical fitness information, and 17% of them use it for sensitive health information such as sexual health, drug use, or depression. This substantial proportion of users, especially teenagers, facing uncontrolled health information on the Web has raised concern among physicians and health professionals. However, social media benefits from the positive network effect: “Groups are remarkably intelligent and are often smarter than the smartest people in them” (Surowiecki, 2005). On discussion forums, for example, patients’ stories represent a priceless source of information that can benefit both patients and physicians. The high number of users also leads to self-correction of information.

According to Armstrong & Powell (2009), people want to find understanding and empathy through their online search, and also obtain support through shared experiences from “people like them”. This phenomenon has led to the creation of a well-known social network called Patients Like Me (www.patientslikeme.com). Some studies on the impact of health-related social media like discussion boards have found that online support can have a positive effect
on patients. Jaloba (2009) studied users’ behavior on a discussion forum related to breast cancer (“the club no one wants to join”). She observed that even if cancer diseases implied very special forms of help, online support had a positive effect on active users and lurkers (readers that never post).

As for less severe diseases, people are seeking support but also sharing experience and having other patients’ opinion. For example, Schraefel et al. (2009) observed that users looking for diet and weight loss information were gathering information on diet from multiple sources and shared their actual experience on social media.

It is difficult to measure the number of users of social: the number of registered users is not always known, and the number of lurkers cannot be evaluated. However, some popular websites like Patients Like Me have more than 100,000 registered users.

Thus, social media do not need to be avoided for health matters. Although the quality of medical information may not be the best, the multiplicity of stories creates important knowledge that both patients and physicians can use. Social media is also a good way to find “patients like me”, which may not be easy to locate for some diseases, and gain support from patients in similar situations.

2.2 Health-related user-generated content

While health-related social media has raised some concerns and studies have been carried out to evaluate the quality of its content, many user studies have evaluated the quality of general health Web content, which focused on the design of the website and the credibility of the content. However, the results can hardly be applied to social media. Most of the studies were based on users’ satisfaction with Web content retrieved by a given query. Sillence et al. (2007) investigated content related to menopause and hormone replacement therapy with fifteen women. They observed that they can easily identify reliable content on Web 1.0 websites, and also used social media to get other patients’ stories. Adams (2010) carried out an inter-disciplinary literature and website review on Web 2.0 health information. She concluded that although users need to be cautious and critical about social media information, it can give useful information when it is combined with other sources. They represent new opportunities for physicians to reach or follow-up on patients, and can help to increase communication between these groups. Hughes et al. (2009) studied how a group of junior physicians were using social media to seek medical information. They observed that those junior physicians were using the Web (mainly Wikipedia or Google search), but were not careful enough about the content. They proposed three policy options to manage risks and improve efficiency of their web use: basic training, tools to raise the availability and credibility of the content, and new tools such as wikis to create physician communities.

To address the difficulty of measuring Web 2.0 content quality and credibility, Denecke and Nejdl (2009) proposed a new method. Assuming that informative content quality is better than affective (related to mood and feelings) one, they created a system using subjectivity words and medical ontology to evaluate the informative content of social media. Their experience showed that social media has to be divided in different categories, suited for different purposes. For example, blogs are good resources for disease or drug related experiences, while wiki and encyclopedia give good information on anatomy.

These studies showed that social media content can be a trustworthy resource regarding patients’ experiences of drugs and diseases. But it should be coupled with other sources of information to get information on a disease or its conditions. As they represent a new opportunity for physicians and patients to communicate, effort has to be taken to educate people (patients or physicians) to use them properly.
2.3 Use of health-related user-generated content

Research on user-generated content related to health is quite recent. Himmel et al. (2008) proposed a system of classification of lay requests on a medical expert forum. The classification, based on subject and sender’s expectations, aimed at helping experts in the task of preparing semi-automatic answers based on a similar request, identified using a nearest neighbor technique. Xia et al. (2009) also focused on a medical website allowing users of the British National Service to post reviews. These reviews are then automatically classified according to their topic and their sentiment polarity.

The international workshop on Web Science and Information Exchange in the Medical Web (Denecke et al., 2010) published several papers on user-generated content. For instance, Klenk et al. (2010) proposed a method to calculate similarity between patients having cancer based on their social network profiles. Stewart and Denecke (2010) proposed a method to extract information related to health events from social media with the objective to support epidemic intelligence.

Although data mining projects previously focused on specialized data like clinical reports and research papers, the emergence of Web 2.0 and the constantly growing user-generated content has created a new opportunity for researchers. In other product or service domains, such as electronic products, hotels and restaurants, user-generated content is well exploited. They allow users to check what other users think about a product and companies to check their product popularity and user satisfaction. However, in the medical domain, people tend to be considered as patients rather than consumers. Little voice is given to their opinions and stories. Medical and health documents tend to be written by physicians or journalists. For instance, well-known health information providers such as MayoClinic (www.mayoclinic.com) and WebMD (www.webmd.com), although they provide discussion forums, display only professionals’ writings in the search results. Quality and trust being a major issue for medical information, researchers tend to study only professional writings. Quality is difficult to define and measure in the case of user-generated content. However, social media can no longer be avoided and studies have shown that their content is useful in different ways for patients (Armstrong & Powell, 2009) and physicians (Hughes et al., 2009). Thus, rather than being hung up on the issue of quality, researchers are now trying to find ways of exploiting useful information in user-generated content.

3. Compilation of the corpus

In this section, we present the different kinds of online drug related content. We describe resources chosen to compile our corpus and finally give its characteristics.

3.1 Online drug-related content

Drugs are highly discussed online. Researchers publish papers and reports about creation of new drugs. Practitioners present and describe drugs on hospital websites or medical information portals. Journalists discuss new drugs in online newspapers. Patients often check drugs prescribed by their doctors, share their experience of the treatment, compare drugs and review them.

To characterize drug content, we first need to identify the different agents involved. Bowker and Pearson (2002) described communicative settings in specialized domains as involving 3 main agent categories: experts, semi-experts, and non-experts. In the medical domain, experts can be researchers, clinicians, doctors, etc. Semi-experts can be medical students, journalists or nurses, and non-experts are all patients who do not have extensive medical knowledge. This paper focuses on drug-related information designed for non-experts. Thus, we distinguish two kinds of documents:
• Documents written by experts or semi-experts: recommendations to patients, institution documents, articles in popular science papers, and blogs.
• Documents written by non-experts: discussion forums, patients' blogs, and social platform content.

Medical and health information portals, such as Mayo Clinic (www.mayoclinic.com) or WebMD (www.webmd.com), are the main sources providing recommendations to patients on a large number of drugs, with a description of their use, side effects, and precautions. However, content related to a given drug cannot be systematically found in popular science journals or blogs (unless the drug is new or highly discussed).

As for patients’ writings, discussion forums and social platforms cover a wide range of drugs. In particular, drug review websites, having a similar framework as discussion forums, offer user reviews on many drugs.

As we are studying drug-related user-generated content, we will be focusing on discussion forums and drug review websites. In order to evaluate their content, we use data from a health information portal as a baseline for comparison. Data in discussion forums are divided in threads, i.e. a set of posts (each post corresponding to a review) composed of a first element, usually a subject or a question, and answers or comments. Threads are also clustered in topics, over different drugs. The topics themselves can be grouped into categories on websites, e.g., analgesic, anti-inflammatory, contraceptive, and diabetes management.

We distinguish three types of discussion forums on which drug reviews can be found:
• Drug review websites: websites dedicated to reviewing drugs, giving users information about drugs and the ratings from other users (e.g., www.druglib.com, and www.drugs-expert.com);
• Product review websites: general product rating websites often having a category for health or drug products (e.g., www.rateitall.com);
• Medical forums or health sections on general forums: users participate on those websites to share experiences about a disease or a drug and give their opinion (e.g., WebMD communities, and Yahoo Answers Health section).

Among users of these websites, we distinguish two categories: active users, who read and write on the websites; and passive users, also called lurkers, who only read. They can have any level of knowledge and be either patients or physicians (although they may not have the same use of the websites). However, patients and lay people seem to represent the majority of the users, except for particular discussion forums where doctors respond to patients.

We can distinguish three types of reviews on the websites:
• Structured reviews: to rate a drug, the user has to fill a form asking to rank or comment on different aspects or the drug: patient’s condition, dosage, effectiveness, or side effects. Example (from www.druglib.com):

  **Adderall review by 45 year old female patient**

  | Overall rating: | 9/10 |
  | Effectiveness: | Considerably Effective |
  | Side effects: | Mild Side Effects |

  **Treatment Info**
  - Condition / reason: Major Clinical Depression
  - Dosage & duration: 5mg taken 2x/day for the period of one month
  - Other conditions: none
  - Other drugs taken: Wellbutrin, Paxil

  **Reported Results**
  - Benefits: Reduced fatigue, increased concentration, focus, and
motivation.

**Side effects:** Occasionally an increased heartrate.

**Comments:** Adderall was prescribed for use while antidepressants were taking effect, to give immediate relief from an episode of major depression.

- Semi-structured reviews: each review is composed of a text and other information which can be directly identified (asked separately in the rating form). Usually, the information contains the name of the drug, the condition, and a score. Example (from www.drugs-expert.com)
  
  Zoloft review written by User3
  Rating: 1/10
  Not for Insomnia Sleep doctor thought Zoloft would help with the fatigue I was experiencing due to insomnia. Took it for a month – sleep was more disturbed than ever. Plus I really didn’t feel I was depressed.

- Non-structured reviews: the user gives his rating in text form. Most of the time drug names can be easily identified, but other aspects are addressed in the text. Example (from www.onlinemedsreview.com):
  
  Drug Name: Bellaspas
  I took Bellaspas for night sweats and it was wonderful. I also have several friends who took it also and they wish they would start selling it again. Nothing else seems to work, and I have severe night sweats, soaking sheets, just nasty!

The personal health platform Patients Like Me provides highly structured reviews, and also a summarized view of all the patients’ ratings. Discussion forums always provide non-structured reviews. Drug or product rating websites can provide all types of reviews.

### 3.2 Resources for the corpus

We carried out a content analysis of drug-related content in social media. We chose to compare:

- drug review website: Drug Lib ([http://www.druglib.com](http://www.druglib.com)),
- classical discussion forum: Yahoo Answers ([answers.yahoo.com/dir/index?sid=396545018](answers.yahoo.com/dir/index?sid=396545018))
- a hospital information portal, providing expert content: MayoClinic ([www.mayoclinic.com/health/drug-information/DrugHerbIndex](www.mayoclinic.com/health/drug-information/DrugHerbIndex)).

DrugLib, maintained by two PhD holders in medical sciences and biochemistry, provides drug information and allows anyone to review a drug through a form. Yahoo Answers Health section is very popular, and we can find discussions on many drugs. The discussion process is different from DrugLib: a drug is discussed on this website only if someone decided to create a thread about it or a related disease or drug. A characteristic of this website is that people ask questions by creating a thread and expecting answers on the posts. MayoClinic website provides a list of drugs with a precise description through the following categories: brand names & description, before using, proper use, precautions and side effects.

Since we wanted our corpus to represent a wide range of drugs, we harvested as many threads as possible until we reached our fixed threshold of 400,000 words. Druglib has an organized collection of postings: each drug corresponds to one thread, identified by the name of the drug. We collected all the threads from this website. MayoClinic provides organized data too: each drug is described in detail. As the total number of drugs is higher than for the other websites and they use their generic name (instead of brand name for the two other websites), we randomly picked drugs (among a list of drug names) until we reached our threshold.
However, Yahoo Answers does not provide any drug index. We used the Druglib list of drugs on Yahoo Answers search tool and collected for each drug the first 5 threads. The characteristics of our corpus are discussed in following section.

3.3 Corpus characteristics
Table 1 gives the main characteristics of our corpus. We collected in total more than 3,700 threads or pages. Druglib (DL) and Yahoo Answers (YA) provide threads with various number of posts (thus various lengths), while each drug on Mayoclinic (MC) is described on four pages exactly. Each part of the corpus is composed of approximately 400,000 words. DL has the longest threads, with an average of 700 words per thread. YA and MC contain shorter pages or threads, with an average of 300 words. However, MC has more balanced threads (always 4 pages) in comparison to YA. This explains the unbalanced number of threads in each part of the corpus: 536 threads for DL, while more than a thousand are needed for the other websites to reach the threshold.

<table>
<thead>
<tr>
<th>Website</th>
<th>No. of threads / pages</th>
<th>No. of words</th>
<th>Avg. words / thread</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL</td>
<td>536</td>
<td>375,370</td>
<td>700</td>
</tr>
<tr>
<td>YA</td>
<td>1,306</td>
<td>398,051</td>
<td>305</td>
</tr>
<tr>
<td>MC</td>
<td>1,929</td>
<td>575,426</td>
<td>298</td>
</tr>
<tr>
<td>Total</td>
<td>3,771</td>
<td>1,348,847</td>
<td>358</td>
</tr>
</tbody>
</table>

Table 1: Corpus characteristics

Our corpus is composed of reviews taken individually. We added information to each review as follows:

- External information, general characteristics of the review: origin (website URL), drug concerned, and rating score (if applicable)
- Internal annotations, concerning linguistic features of the review: structure of the review (structured or semi-structured reviews), medical terms tagged with UMLS metathesaurus using Metamap system (Aronson & Lang, 2010), and part-of-speech tags using MedPost (Smith et al., 2004)

4 Drugs covered in the corpus
Although health information portals tend to display information on a large range of drugs, health social media drug coverage relies on users’ interests. In this section, we explore categories and topics in our corpus. This gives us a better idea of the variety of drugs discussed.

4.1 Drug categories
In the previous section, we said that forum threads are most of the time classified; their topics are assigned to categories (and sometimes sub-categories). Topics are specified by thread authors in discussion forums such as YA. Topics can also be provided by the website; it is chosen among a list of drugs for DL. Category can be fixed by the website (for fixed topics) or chosen by the user. As categories can be fixed by the website, they differ from one website to another. We can find four types of classification (note that a website can provide several classifications): drugs classified by families, by conditions, by side effects, or no classification (index of drug name). MC only displays drugs by their name. However, it provides disease and symptom browsers, which are linked to the related drugs. Another well-known information portal, WebMD, also provides classification of drugs by condition. Table 2 lists the categories that can be found on DL. It facilitates user’s research by providing two ways to find drugs (by condition or drug family).
No. of categories | Examples
---|---
150 drug families | Anesthetics, anticoagulants, contraceptives, diuretics, and migraine
706 conditions | Attention deficit disorder, epilepsy, glaucoma, sinusitis, and vasculitis

Table 2: DL drug classifications

These two ways to search for drugs correspond to two different processes: searching with the name or the family of the drug implies that the patient already possesses the drug or know of it; whereas searching with a condition or symptoms means that the user is looking for a way to treat a disease or wants to compare products. This is also the case with MC.

The YA Health section provides a classification which does not include drug or condition names: alternative medicine, dental, diet & fitness, diseases & conditions, general health care, men's health, mental health, optical, and women's health. As users assign labels by themselves to threads they create, a drug or condition index cannot be updated. However, the search tool allows finding information on a lot of drugs and the disease & condition section contains also many threads related to drugs.

Even if all the websites do not provide categories, they reflect the way people look for drugs: either to get a “diagnostic” or to check on drugs they know or possess. The different sites support different author’s needs. In review websites, authors, who can be patients or practitioners, already know a drug and can use those websites to share their experience, for example: “This drug alleviates my allergic responses--itchy eyes and sneezing. It's effective because if I don't use it, my symptoms return”. On discussion forums, thread creators post questions and get information, for example: “What’s the difference between Allegra D Pills and the normal Allegra? I have to buy the Allegra D, but i don’t know if it heals the skin rash by food poisoning too”. Other authors then respond and provide information or ask new questions, for example: “D stands for decongestant, like if you have a stopped up nose, usually pseudophed or something similar”.

### 4.2 Drugs discussed

As we fixed a target size for our corpus, the number of drugs covered by the corpus is quite limited. Table 3 presents the number of drugs for each part of the corpus and the corresponding number of threads and reviews. The corpus contains information related to 828 drugs.

<table>
<thead>
<tr>
<th>Website</th>
<th>No of Drug</th>
<th>No of Thread</th>
<th>No of Reviews</th>
<th>Avg review/drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL</td>
<td>536</td>
<td>536</td>
<td>2939</td>
<td>5.4</td>
</tr>
<tr>
<td>YA</td>
<td>477</td>
<td>1306</td>
<td>5408</td>
<td>11.3</td>
</tr>
<tr>
<td>MC</td>
<td>489</td>
<td>1929</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>All</td>
<td>828</td>
<td>3771</td>
<td>10276</td>
<td>8.2</td>
</tr>
</tbody>
</table>

Table 3: Drugs in the corpus

We firstly collected all the threads, of 536 drugs from DL. Then we used the DL drug index to gather data from YA. The quantity of drugs is quite similar. MC contains generic drug names, while DL and YA contains generally brand names. Several drugs on DL or YA may map to only one on MC. More than 300 drugs are common to both websites, and we had to randomly pick other pages to have enough data. Drugs are more discussed on YA, since each drug has on average 11.3 reviews, against 5.4 in DL. However the reviews in a thread on YA may not always focus on a drug. For instance, a thread related to a condition may concern several drugs. The most reviewed drugs of the corpus are given on Table 4.
The most discussed drugs on DL are antidepressant, acne treatment and weight control drugs. The most discussed drugs in YA should be considered differently, as it is only a small sample of the website’s drug discussions. Even if we cannot measure which are the most discussed drugs of the whole website, we can see on this subset of drugs that pain relief, antidepressant and birth control drugs are interesting to users. We notice that all these drugs, except pain relievers, are prescription drugs.

<table>
<thead>
<tr>
<th>Website</th>
<th>Drug</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL</td>
<td>Lexapro</td>
<td>Antidepressant</td>
</tr>
<tr>
<td></td>
<td>Accutane</td>
<td>Acne treatment</td>
</tr>
<tr>
<td></td>
<td>Lipitor</td>
<td>Weight control</td>
</tr>
<tr>
<td></td>
<td>Paxil</td>
<td>Antidepressant</td>
</tr>
<tr>
<td></td>
<td>Retin-A</td>
<td>Acne treatment</td>
</tr>
<tr>
<td>YA</td>
<td>Tylenol</td>
<td>Pain relief</td>
</tr>
<tr>
<td></td>
<td>Advil</td>
<td>Pain relief</td>
</tr>
<tr>
<td></td>
<td>Xanax</td>
<td>Antidepressant</td>
</tr>
<tr>
<td></td>
<td>Yaz</td>
<td>Birth control</td>
</tr>
<tr>
<td></td>
<td>Zoloft</td>
<td>Antidepressant</td>
</tr>
</tbody>
</table>

Table 4: Mostly discussed drugs in the corpus

As DL provides authors a detailed input form, all the reviews are similar; the only difference is the length of the reviews:

**Lexapro review by 39 year old female patient**

**Overall rating:** 8/10  
**Effectiveness:** Considerably Effective  
**Side effects:** Moderate Side Effects

**Treatment Info**

**Condition / reason:** Depression  
**Dosage & duration:** 20 mg taken daily/at bedtime for the period of 2+ years  
**Other conditions:** None  
**Other drugs taken:** Trazodone

**Reported Results**

**Benefits:** I have clinical depression with agitation and anger. This medication seemed to take the edge off the anger as well as decrease my depressive symptoms. I started at 10mg, then shortly increased to 20mg, which left me feeling too drugged. It was only after 12+ months of taking 10 mg that it was increased due to increasing depressive symptoms.

**Side effects:** Initially the medication made me very drowsy. I began taking it at bedtime, which seemed to elevate that particular concern. I think there may have been some weight gain, but I am not certain that it was specifically related to the Lexapro.

**Comments:** I started on the medication over 2 years ago at 10 mg a day. The dosage was increased to 20 mg after about 12 months. The side effects were limited, and mostly not particularly bothersome.

YA threads are less precise than DL. It can be related directly to the drug (“Are tylenol 3 tablets the ones with codine available OTC, or are any the tylenol w/ codine available
“OTC???”), to other drugs (“I’m allergic to Motrin and tylenol, can I still take amoxicillin?”) or to a related disease (“How can I make my symptoms of the common cold seem minimal?”). Frequently discussed drugs seem to be related to personal issues, like mental health, weight problems, acne, and sexuality. Users may be more comfortable to discuss those problems anonymously online rather than with their doctors. Teenagers turn to forum users instead of their relatives or a doctor, for example (question from a teenager on YA):

“How do I tell my mom I want to go on a weight loss system?
I’m 14, and I’m very unhappy with my body. [...] I’ve been seeing commercials for various types of weight loss systems (e.g. nutrisystem, alli, weight watchers etc.) and I really wanna try something, I want to be happy with the way I look”

User behavior is not in the scope of our paper, so we will not analyze this. Many drugs seem to be discussed online. Despite the very different functions of YA and DL, they still cover the same drugs (for the same drugs, we collected only a sample of threads on YA). The categories revealed that authors and readers on each website can be looking for information during their treatment, and they can be followed or not by a doctor. The most discussed drugs are either “popular” drugs (i.e. drugs that a lot of people use, like pain relievers) or drugs related to personal issues (i.e. issues dealing with personal troubles, which people may not be comfortable discussing in person, like weight problems or depression). We did not take into account the period of postings. However, it could be useful to observe reactions across time on certain type of drugs and even find applications in epidemic intelligence (Stewart & Denecke, 2010).

5 Linguistic characteristics

In this section, we describe the basic linguistic features of our corpus. Before evaluating our corpus with respect to medical or opinion related-features, we first investigate the kind of text we are studying.

5.1 Structural and syntactic characteristics

The Lingpipe (alias-i.com/lingpipe/) sentence chunker was used to split the text into sentences. Table 5 shows the results obtained using this tool. Reviews in the corpus contain on average 131 words and 9 sentences. Sentences are composed of almost 14 words. Reviews tend to be shorter in user-generated content than in MC, both by number of sentences or number of words. YA has the shortest reviews, containing on average 4.6 sentences and 73.6 words (which is approximately half of DL reviews). In comparison, each drug description on MC contains on average 18 sentences and almost 300 words. The variation calls for comment: MC has a similar pattern for each drug, which includes a similar length, while DL and YA reviews rely totally on users (going from a single word to a whole text).

<table>
<thead>
<tr>
<th>Website</th>
<th>Avg. sentence / review</th>
<th>Avg. word / review</th>
<th>Avg. word / sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL</td>
<td>12.5</td>
<td>127.7</td>
<td>10.2</td>
</tr>
<tr>
<td>YA</td>
<td>4.6</td>
<td>73.6</td>
<td>16.1</td>
</tr>
<tr>
<td>MC</td>
<td>18.3</td>
<td>298.3</td>
<td>16.3</td>
</tr>
<tr>
<td>Corpus</td>
<td>9.4</td>
<td>131.3</td>
<td>13.9</td>
</tr>
</tbody>
</table>

Table 5: Structural characteristics
Surprisingly, YA has the longest sentences, with 16 words, which is similar to MC. However, MC sentence length is deceptive as the texts contain a large number of item lists, which yield the longest sentences. The following examples illustrate sentence length in each website:

**DL:** I take 3 Adderall per day. 1 1/2 pill in the morning, and 1 1/2 in the afternoon. Each pill is 15 mg. (sentences length: 6 / 13 / 5)

**YA:** Take it from me personally. Today I took Adderall for the first time. I took it at around 2 PM. I haven’t eaten all day long. (sentences length: 5 / 8 / 6 / 6)

**MC:** Amphetamine and dextroamphetamine combination increases attention and decreases restlessness in patients who are overactive, cannot concentrate for very long or are easily distracted, and have unstable emotions. It is also used as part of a total treatment program that also includes social, educational, and psychological treatment. (Sentences length: 27 / 19)

Netzel et al. (2003) showed that the MEDLINE database which carries abstracts of English biomedical papers had an average of 27 words per sentence. In comparison, our corpus contains relatively short documents. Moreover, sentences in these documents tend to be short too.

Figure 4 shows the proportion of each part-of-speech (POS) tag in the corpus. We used MedPost (Smith et al., 2004) POS tagger, created for biomedical texts. First of all, the three parts of the corpus have quite similar distributions. This distribution of POS is shared with common POS distribution in general English text (for instance, Shih et al. (2000) gives the BNC corpus distribution). The high proportion of nouns (more than 25% in each part of the corpus) is quite typical of English writing.

![Pronoun usage in corpus](image)

**Figure 1:** Detailed part-of-speech distribution

Pronouns are quite frequent in the corpus, rather than in general English text. Two reasons can explain this phenomenon. First, pronouns play an important role in scientific discourse. Although it can be seen as a very impersonal discourse, involving many passive sentences, the number of pronouns increases for example in research papers (Kuo, 1999). Secondly, it seems quite obvious that users use many pronouns as they describe their experience. Nevertheless, some differences highlight interesting characteristics. We observe that DL has the highest range of nouns, and the lowest range of verbs and modals. This can be explained by the structure of the reviews; users sometimes fill some fields with a list of words instead of complete sentences, for example: “Benefits: no migraines, even when exposed to outside elements (the sun) and winter reflection, brightness”. In YA, a thread can be considered as a dialogue, for example:

**User 1:** I have to take tylenol with codeine, and it tastes horrible. Is there any way to hide the taste, or at least make it taste better?
Most of the time, the user asks about his or her health issues and expects answers from other users. As a result, YA has the highest range of pronouns in the corpus.

This analysis has shown that user-generated reviews are relatively short (in terms of number of sentences, words and sentence length). Compared to editorial content, they contain few sentences and words. The analysis of POS tags showed that drug reviews are closer to editorial content than discussion forums, although some of its content is not structured into sentences. On the other hand, YA answers seem to be closer to spoken language, as it contains more pronouns and fewer determinants.

5.2 Lexical features
The vocabulary size of the corpus is evaluated through the number of unique words. To calculate it, we applied the Lancaster stemmer (Paice, 1990) to the corpus and counted the number of stems. Despite the large number of words in the corpus, the number of stems is quite limited. As words are stemmed, it does not contain inflections.

The corpus contains more than 18,000 unique words. This amount corresponds to the size of the intersection of each part vocabulary. Taken individually, each part of the corpus has a relatively small number of unique words. User-generated content has the largest vocabularies, with about 9,500 and 12,400 words for DL and YA respectively. MC vocabulary is much smaller than user-generated ones, with about 5,600 words.

<table>
<thead>
<tr>
<th>Website</th>
<th>No. of unique words</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL</td>
<td>9,448</td>
</tr>
<tr>
<td>YA</td>
<td>12,376</td>
</tr>
<tr>
<td>MC</td>
<td>5,565</td>
</tr>
<tr>
<td>Total</td>
<td>18,802</td>
</tr>
</tbody>
</table>

Table 6: Vocabulary size in the corpus

As our corpus is balanced across the three websites, and each type of site has the same quantity of data, the vocabulary size can be compared. To explain the variation, we calculate for each part of the corpus the proportion of words belonging to a general language lexicon. We used as a general language lexicon the word lists from the 12dict project, compiled by Alan Beale (wordlist.sourceforge.net/12dicts-readme.html). As the general language lexicon does not contain specialized terms, we also calculated the proportion of words in the vocabulary that map to concepts from the Unified Medical Language System (UMLS, www.nlm.nih.gov/research/umls/) medical metathesaurus. Created by the U.S. National Library of Medicine, this metathesaurus is a collection of 1.7 million medical concepts, grouped in more than 130 semantic types. Each semantic type belongs to one of the 15 semantic groups. For example, the concept Aspirin belongs to the semantic type Pharmacologic substance, which is under the Chemical and Drugs semantic group.

Firstly, we observe that more than 50% of the vocabulary word (unique words) for each type of website belongs to the general lexicon. Although this proportion is quite homogeneous in the corpus, the proportion of words belonging to UMLS metathesaurus is not. As expected, MC vocabulary contains many medical terms, up to 68%. DL and YA vocabularies contain fewer medical terms, 47% and 52% respectively, which is still a high range. The third part of
the figure shows the proportion of words that do not belong to the general lexicon nor to the medical one (some words can belong to the general lexicon and to the medical one). This third part is higher in DL and YA (29% and 31% respectively) than in MC (14%). This difference can be first explained by the fact that MC contains texts that are written by professionals and reviewed before being published. As for DL and YA, there are many possible reasons why such a large part of their vocabularies are not in these lexicons:

- Proper nouns: drugs brand name (e.g., Adderall, or Xanax.), person names (e.g., Alice, or Jeff);
- Abbreviations (e.g., H2O, HIV, or DHS);
- Web neologisms or expressions (e.g., lol, www, thx, or wtf);
- Spelling errors (e.g., bewteen, experiance, or substitute);
- System errors: html extraction errors, numerical entities (e.g., 15mg, or 190lb), lemmatizer errors (e.g., accustomed, or anecdotaly).

![Figure 2: proportion of words from vocabularies that belong to 12dict, UMLS metathesaurus or none](image)

We can hardly detail these sets of words, because it would require manual analysis. We can assume that the biggest the set is, the highest is the number of spelling mistakes, abbreviations, etc.

### 5.3 Most distinguishing terms

In this section, we investigate which are the most distinguishing words in our corpus. To do so, we used the Information Gain (IG) measure, which is known to be an effective term goodness measure in the framework of opinion mining (Thet et al., 2010). It is obtained with the following formula:

$$IG(w) = \Pr(w) \cdot \sum_{i=1}^{K} \Pr(c_i | w) \log \frac{\Pr(c_i | w)}{\Pr(c_i)} + \Pr(w) \cdot \sum_{i=1}^{K} \Pr(c_i | \overline{w}) \log \frac{\Pr(c_i | \overline{w})}{\Pr(c_i)}$$

where \( w \) is the word, \( K \) the number of classes (2 in our case: user-generated content vs. expert content), \( c_i \) the classes, \( \Pr(c_i) \) the percentage of training documents in \( c_i \), \( \Pr(w) \) the percentage of documents containing \( w \) and \( \Pr(\overline{w}) \) the percentage of documents in which \( w \) is absent. \( \Pr(c_i | w) \) is the conditional probability of category \( c_i \) given word \( w \) and \( \Pr(c_i | \overline{w}) \) the conditional probability of category \( c_i \) given that \( w \) is absent.
This method allows comparing two sets of documents and determining which the most distinguishing terms of each set are. We first compared user-generated content (DL and YA) with expert content (MC). Table 7 shows the terms with the highest information gain for each set. We observe that the user-generated content list is highly user-centered: the pronoun I, the determiner my, the verb am and the noun patient. This means that these terms are seldom in MC, which seems quite obvious as authors do not intend to talk about themselves contrary to YA and DL. The most distinguishing terms in MC are more related to general concepts of the medical field (healthcare, medicine, or medical) and also to advice related to drug use (may, or carefully). This list also highlights the language difference between both media. On one hand, user-generated content uses simple terms and personal pronouns, whereas expert content contains more specialized terms.

We carried out the same analysis to compare the two cases of user-generated content. Table 8 shows the most distinguishing terms in DL and YA. We can see that the lists contain some words from the website forms and not users’ comments, for example review and patient for DL and question, answer and member for YA. Disregarding these terms, the result reflects different uses of those sites. On one hand, DL aims at reviewing drugs, so patients describe precisely their treatment, giving details on the dosage, and the length of treatment.

<table>
<thead>
<tr>
<th>YA + DL</th>
<th>MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Healthcare</td>
</tr>
<tr>
<td>My</td>
<td>Following</td>
</tr>
<tr>
<td>Was</td>
<td>Professional</td>
</tr>
<tr>
<td>Old</td>
<td>Medicine</td>
</tr>
<tr>
<td>None</td>
<td>Medical</td>
</tr>
<tr>
<td>Patient</td>
<td>Label</td>
</tr>
<tr>
<td>Drug</td>
<td>Occur</td>
</tr>
<tr>
<td>Am</td>
<td>May</td>
</tr>
<tr>
<td>Just</td>
<td>Carefully</td>
</tr>
<tr>
<td>How</td>
<td>Adults</td>
</tr>
</tbody>
</table>

Table 7: Most distinguishing terms between user-generated content and expert content

<table>
<thead>
<tr>
<th>DL</th>
<th>YA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review</td>
<td>You</td>
</tr>
<tr>
<td>Patient</td>
<td>Your</td>
</tr>
<tr>
<td>Period</td>
<td>Answer</td>
</tr>
<tr>
<td>Year</td>
<td>Question</td>
</tr>
<tr>
<td>None</td>
<td>Thanks</td>
</tr>
<tr>
<td>Taken</td>
<td>Luck</td>
</tr>
<tr>
<td>Frequency</td>
<td>Wondering</td>
</tr>
<tr>
<td>Dosage</td>
<td>Chosen</td>
</tr>
<tr>
<td>Times</td>
<td>Goodbye</td>
</tr>
<tr>
<td>Once</td>
<td>Member</td>
</tr>
</tbody>
</table>

Table 8: Most distinguishing terms between DL and YA

This linguistic analysis has shown that user-generated documents are quite short, in terms of number of words and sentences, and their sentences are short too, compared to scientific writings. Their part-of-speech distribution is quite similar to general English one. It indicates that discussion forums, with their dialogue pattern, are closer to spoken language, while drug reviews are closer to editorial content. The results of the lexical analysis indicate that user-generated content has a larger vocabulary than editorial content. Nevertheless, they tend to
contain fewer medical terms and more spelling mistakes. Our analysis of the most distinguishing terms indicates that expert content contains more specialized terms than user-generated content. As expected, it reveals that user-generated content is patient-centered while expert content is treatment-centered. It also highlights the difference of use between the two user-generated contents: DL describes in detail a treatment while YA contains dialogues between users on drugs.

6 Medical lexicon
In this section, we evaluate the medical content in our corpus. We used the UMLS metathesaurus and the MetaMap system (Aronson & Lang, 2010) to tag our corpus with UMLS concepts. The Metamap process allocates one or several concepts to each term. Figure 6 gives an example sentence and the concepts tagged by the Metamap system.

This result shows that 38% of the words in the corpus are medical concepts. The high number of medical terms is possibly due to the large vocabulary in UMLS. Concepts are related to very specialized terms (for example, calcified granuloma or mediastinum) as well as general concepts like temporal ones (for example, day or year). All the terms related to drugs and their description belongs to UMLS, except brand names of drugs.

MC has the highest proportion of medical terms, 40%, while DL and YA have 37% and 33% respectively. The following examples illustrate what are considered medical concepts in UMLS (medical concepts are underlined):

YA: Alcohol very occasionally, and then only one or two drinks. On your prescription bottle of Lexapro there should be warnings about "do not drink alcohol with this medication". Alcohol is a depressant - and nullifies the effects of the Lexapro.

MC: Make sure your doctor knows about all the other medicines you are using. Escitalopram may cause serious conditions such as serotonin syndrome or neuroleptic malignant syndrome (NMS)-like reactions when taken with certain medicines.

Table 9 shows the most frequent UMLS terms in the corpus and their semantic types. The first observation is that there is no highly technical term in this list. On the contrary, the terms are linked to treatment (patient, dosage, take, use, day, doctor, clinic, side effects, etc.). The only disease appearing in this list is depression, which results from antidepressant which is one of the most discussed drugs in the corpus. While DL and YA terms are linked to a precise description of the treatment experience, MC seems to provide broader information about it. For instance, no term related to dosage appears in the list.
This analysis shows that the quantity of medical terms in our corpus varies from 40% for editorial content to 33% for discussion forums. This means that in both cases a lot of medical terms are employed in the texts, whoever the author is (expert or non-expert). In user-generated content, drug reviews contain more medical terms than discussion forums. The top-ten terms in each part are strongly related to the drug lexical field: usage, dosage, and side effects.

7 **Opinion aspects**

In this section, we evaluate how subjective are the documents in our corpus. We assume that informative content is less subjective than affective content. We compare user-generated content to MC, which is considered informative content. We used the Subjectivity Lexicon (Wilson et al., 2005) to identify subjective terms. This lexicon contains about 8,000 subjective terms, classified as strongly or weakly subjective. Each term also has a polarity that can be positive (33%), negative (60%), both or neutral (7%).

We first observed in our corpus that 7% of the words used are subjective. The proportion of subjective terms in DL is the highest, with almost 7%, while it is around 5% in YA and MC. Figure 7 shows the distribution of positive, negative and strongly subjective terms in the corpus. Generally speaking, there are more negative terms than positive in our corpus. YA and MC share the same proportion for each category, and DL has a slightly higher range for each. This can be explained by the purpose of DL, which is expressing opinions on drugs.
We collected from our corpus the most frequent positive, negative and neutral terms of this lexicon (Table 10). At first sight, a lot of terms seem to be related to the medical domain: pain, depression, risk or anxiety. We can also see that more than a half of these medical terms are considered negative in the lexicon, which explains why we counted more negative terms than positive. This may be due to the fact that talking about a disease is rarely positive in a general framework. Positive terms are related to the wellness improvement lexical field: well, able, effective, and better. We saw in Section 4 that the most frequently rated drugs are related to depression, this is confirmed by the frequency of the negative words depression and anxiety in DL. This analysis shows that opinions are expressed in our corpus on different aspects of the drugs.

<table>
<thead>
<tr>
<th>Term</th>
<th>Polarity</th>
<th>Term</th>
<th>Polarity</th>
<th>Term</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>patient</td>
<td>P</td>
<td>abuse</td>
<td>N</td>
<td>problems</td>
<td>N</td>
</tr>
<tr>
<td>pain</td>
<td>N</td>
<td>pain</td>
<td>N</td>
<td>potential</td>
<td>P</td>
</tr>
<tr>
<td>depression</td>
<td>N</td>
<td>block</td>
<td>N</td>
<td>adequate</td>
<td>P</td>
</tr>
<tr>
<td>severe</td>
<td>N</td>
<td>even</td>
<td>P</td>
<td>need</td>
<td>O</td>
</tr>
<tr>
<td>anxiety</td>
<td>N</td>
<td>bad</td>
<td>N</td>
<td>risk</td>
<td>N</td>
</tr>
<tr>
<td>able</td>
<td>P</td>
<td>better</td>
<td>P</td>
<td>even</td>
<td>P</td>
</tr>
<tr>
<td>even</td>
<td>P</td>
<td>luck</td>
<td>P</td>
<td>help</td>
<td>P</td>
</tr>
<tr>
<td>less</td>
<td>N</td>
<td>well</td>
<td>P</td>
<td>specific</td>
<td>O</td>
</tr>
<tr>
<td>effective</td>
<td>P</td>
<td>problem</td>
<td>N</td>
<td>important</td>
<td>P</td>
</tr>
<tr>
<td>felt</td>
<td>O</td>
<td>safe</td>
<td>N</td>
<td>necessary</td>
<td>P</td>
</tr>
</tbody>
</table>

Table 10: Most frequent opinion terms and their polarity (N = negative, P = positive and O = objective / neutral)

Although our method does not measure subjectivity, it indicates the quantity of subjective terms in our corpus. We observed that the quantity of subjective terms is quite low in the corpus (7%). As it contains reviews, DL has the highest proportion, with 7%. There are more negative terms than positive ones in the corpus, and this distribution is observed for each website. YA and MC have similar distributions, in subjective terms and in polarity.
8 Discussion
In this section, we gather our findings and discuss them in the light of quality and trustworthiness.

User-generated content seems to cover a wide range of drugs, which is one of our quality criteria. Among the most popular drugs, we can find over-the-counter drugs such as pain reliever or allergy drugs, wellness-related drugs such as slimming pills or acne treatment, and also mental illness drugs such as for depression or anxiety. In the case of over-the-counter drugs, as they can be purchased without doctor prescription, we can assume that people are able to use them on their own and then trust other people’s advice. As for prescription drugs, considering that patients have met a doctor, their experience with the drugs can be a good support for other patients on certain aspects: side effects, benefits, cost, etc. Nevertheless, advice about a drug being better than another one may not be trusted as they do not have enough medical knowledge. Finally, although mental illness and wellness drugs are usually prescription drugs, their usage and encouragement to use them raise ethical issues. This question is out of the scope of this paper.

Our analysis also shows that user-generated content tends to have its own linguistic characteristics: texts are shorter, contain more spelling errors than scientific writings but they share a quite similar part-of-speech distribution. Discussion forums seem to be closer to spoken language than drug review websites, which can be expected from their use. Both cases of user-generated content seem to reach a process guarantee quality, i.e. even though there are some syntax or lexicon problems, the quality is enough to allow text processing. Moreover, spelling correction techniques (Brill & Moore, 2000, for instance) and studies on Web expressions and acronyms (Tagliamonte & Denis, 2008, for instance) can facilitate the process.

We proposed as a third quality criteria both informative and affective content in the documents. Our analyses on the frequencies of UMLS medical concepts and subjectivity terms show that user-generated content contains a significant proportion of medical terms (although lower than scientific documents) and opinion expressions. Denecke & Nejdl (2009) proposed a more complex method to evaluate it, but we think that our results show that both discussion forums and drug review websites contain informative and affective content. Users’ reviews are texts about people’s experience on a certain drug. Although one needs a good medical knowledge to discuss a priori about drug effectiveness, users can talk about their own health, body and feelings, which does not require any medical knowledge. Moreover, doctors use similar information to treat some patients: they adapt the treatment to the patient’s response and feelings. In that sense, patients’ reviews fit our quality requirements and can be trusted to a certain degree.

Considering our quality requirements, drug review websites seem to be the best choice: many drugs covered and easily accessible (through drug indexes); reasonable linguistic quality of texts for information processing; both informative and affective content. Moreover, they provide an organized structure that facilitates access to opinions (in contrast to discussion forums where opinion and experience are mixed). We intend our opinion mining system to be based on different aspects of the drug: benefits, side effects, conditions, cost and dosage. This will allow us to select the content we want and avoid untrustworthy information such as advice on the treatment or other drugs.
9 Conclusion

We carried out an analysis of three kinds of online drug-related content: drug reviews, discussion board postings and information on a hospital portal. We compiled a corpus composed of threads/documents from three websites. We analyzed the corpus with respect to the drugs covered, their linguistic characteristics, and their medical and subjective content.

We observed that, among 800 drugs covered by the corpus, the most discussed ones were popular drugs (e.g., pain relievers) or drugs related to personal issues (such as depression or weight control). The results of the linguistic analyses show that user-generated content threads tend to be short, in terms of number of words and sentences than expert content. User-generated content tends to be very patient-centered, while expert content tends to focus on usage of drugs. Among user-generated content, we noticed that the discussion board has characteristics close to spoken language and tends to be in dialogue form. Compared to drug reviews and expert content, it has a bigger dictionary, with more “unknown” words from proper nouns, spelling errors, and abbreviations. The medical analysis indicated that user-generated content contains fewer medical terms than expert content. Nevertheless, the medical content in drug reviews is richer, with more frequent terms related to drug use. Finally, our subjectivity analysis showed that subjective terms are not frequent (less than 10%). The frequency is the highest in drug reviews. For the purpose of opinion mining, drug reviews appears to be the most appropriate resource as it contains more subjective words and more medical terms. Drug review websites are also centered on the drug and the patient, contrary to discussion boards which are primarily focused on the patient.

In this study, we have identified a set of characteristics of user-generated content on the Web. This provides a foundation for analyzing the quality of information. Further work can investigate to what extent quality can be analyzed automatically and the criteria to use in the evaluation. Temporal sequence can be incorporated in the analysis, following Stewart and Denecke (2010) to analyze people’s reactions towards medical events (such as H1N1 in 2009). Some ethical issues can also be observed on discussion forums, for example, to study how teenagers use social media to discuss their health and its impact on their health (like Pew Research Center, 2010).

References


Pew Research Center (2010). *Social Media & Mobile Internet Use Among Teens and Young Adults*. Washington, DC: Lenhart, A., Purcell, K., Smith, A. & Zickuhr, K.


---

1 Hughes et al. (2009) published an interesting study on physician’s use of Web 2.0.

2 Although YA seems particular (one question as a topic, answers in the thread), it is still relatively similar to classical discussion forums.