EXPLORING MACHINE LEARNING TECHNIQUES USING PATIENT INTERACTIONS IN ONLINE HEALTH FORUMS TO CLASSIFY DRUG SAFETY

BY

BRANT WAH KWONG CHEE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Library and Information Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Leslie Gasser, Chair
Professor Bruce Schatz, Director of Research
Assistant Professor Karrie Karahalios
Associate Professor Catherine Blake

Abstract

This dissertation explores the use of personal health messages collected from online message forums to predict drug safety using natural language processing and machine learning techniques. Drug safety is defined as any drug with an active safety alert from the US Food and Drug Administration (FDA). It is believed that this is the first exploration of patient derived data of this type for pharmacovigilance – the study of drugs once released to market for safety. It is believed that this is the first application of machine learning and natural language processing techniques to be used for pharmicovigilance on patient derived data.

We present results demonstrating the identification of drugs withdrawn from market as well as predictions of other potential safety alert drugs. One example includes Meridia, a weight loss drug linked with death for those with cardiovascular disease. The drug is identified based on data presented two years before FDA and European Union (EU) advisory panels were formed and the subsequent withdrawal of the drug from market within the EU and United States.

Acknowledgements

Thank you first and foremost to my family. Your unconditional love and support has gotten me through many years of schooling. Mom and Dad, your belief in education provided the foundation necessary to start and complete my graduate career. Melissa, thank you for being the "buffer" during all the tense situations at home.

Les Gasser, thank you for sparking my interest in research and providing me the opportunity to continue on to postgraduate education. Bruce Schatz, thank you for the support through the years and your fatherly lectures on academia and life. I appreciate the countless hours of help with projects, publications and research. You have helped me to navigate the sometimes-tumultuous waters of Higher Education. Karrie, I appreciate you making me a part of your group and taking me in as one of your "ducklings." Your faith and enthusiasm for my work is highly esteemed. Your perspective has been always helpful and welcomed when dealing with academic issues. Cathy, thank you for your belief in my work and willingness to take me on at such a late stage in my doctoral program.

Thank you to all my friends for the fun times and putting up with me throughout everything. I do not think I would have made it without you. Tim Hogan, your faith and belief in me is amazing; I try to be a better person because of you. Jana Masley, you have been there time and again for me, you are my rock and I'm fortunate to have you in my life. Drew Strellis, I try to follow your lead and look at the world in wonder and optimism. Emma Berdan and Hani James Ebeid, I know I've put you through a lot, thank you for sticking around and being supportive. Kiran Lakkaraju, Eric Gilbert, Tony Bergstrom and Samarth Swarup, thank you all for letting me bounce ideas off you and for

all of your encouragement and intellectual insights. To everyone else, thank you for being there for support and all the times we shared.

Table of Contents

1 Introduction

Chronic illness affects the lives of approximately half of all Americans. It accounted for 70% of the deaths in the US in 2005 and was responsible for more than 75% of the nation's $2 trillion healthcare cost[1]. This dissertation explores the possibility of utilizing patient derived data from online message groups for pharmacovigilance and identification of adverse effect causing drugs within the context of chronic illness treatment.

Chronic illness is defined as any illness lasting more than three months; these illnesses affect people of all races, classes and geographic locations and are wide ranging illnesses from heart disease to Schizophrenia. A troubling aspect of chronic illness is that it is not usually preventable by vaccine or curable by medication, and the number of affected people continues to grow steadily:

> Advances in medical science and technology – new diagnostic testing, new medical procedures, and new pharmaceuticals – are being used to treat acute illness and maintain a level of health and functionality that results in increased numbers of people surviving with chronic conditions. We are also successfully screening and diagnosing chronic conditions with greater frequency and success. Earlier detection means people can live with chronic conditions that used to grow to acute care stages before diagnosis (Anderson et al., 2002).

The first baby boomers will reach age 65 around 2011. This is particularly worrisome because the likelihood for having a chronic condition increases with age. Half of all people with chronic conditions have multiple conditions or co-morbidities and those with five or more co-morbidities are responsible for two-thirds of Medicare spending (Anderson et al., 2002).

Those with chronic illness are on a myriad of drugs ostensibly for the rest of their life. Many drugs lack the rigorous long term testing necessary for lifelong treatment

[1] http://www.cdc.gov/chronicdisease/overview/index.htm

plans. Pharmacovigilance is the study of drugs once released to market and is an important component of public health. While clinical trials aim to prove a drug's safety before it is brought to market, many drugs have been marketed that are unsafe or responsible for numerous deaths or deformities; however, as the World Health Organization states, "safety is not absolute, and it can be judged only in relation to efficacy, requiring judgment on the part of the regulators in deciding on acceptable limits of safety" (World Health Organization [WHO], 2002).

Inherent problems exist with clinical trials due to their controlled environment and small numbers of participants. Clinical trials may not observe rare life threatening effects; for example, fatal blood dyscrasia might occur in 1 in 5,000 patients treated with a drug that will likely be recognized after 15,000 patients have been treated provided that a causal association with the drug is clear and there is zero background incidence of it occurring (WHO, 2002). These studies are in a controlled environment not taking into account the differing ways people take medications along with over the counter drugs, co-morbidities, and long-term effects. There are many drug-drug interactions, genetic variations that are not present in sample populations, or even interactions with food that will not surface for some time after drugs are marketed. Grapefruit juice is an excellent example of a seemingly innocuous and healthy drink that people might want to take in conjunction with medication; however, it can, "markedly augment oral drug bioavailability...the duration of effect of grapefruit juice can last 24h...At least 20 other drugs have been assessed for an interaction with grapefruit juice" (Baily et al., 1998).

The Untied States Food and Drug Adminstration (FDA) which is in charge of drug safety in the United States, continues to monitor drugs for safety once they are

marketed in Phase IV trials. The FDA constructs panels to advise on specific drugs' safety. While the FDA usually follows its' own panel's advice, it is not obligated to do so. Drug safety is ultimately a political process. It is not always clear whether or not to pull a drug from market even though it has dangerous side effects if it also has the potential to do good, especially for life-threatening conditions. The public at large has influence over drugs that concern a subset of the population (for example those with rare diseases) and yet mass media and other influences can dictate the withdrawal of drugs from market.

The Adverse Effect Reporting System (AERS) consists of two components, the first of which is mandatory reports from drug manufacturers, distributors, and healthcare facilities. The second component is voluntary reports from consumers or patients as well as from care providers or physicians. A diagram of the FDA drug review process once a drug is released to market is depicted in Figure 1.

The reports in AERS are evaluated by the Center for Drug Evaluation and Research (CDER) and the Center for Biologics Evaluation and Research (CBER) groups within the FDA. CDER regulates over-the-counter and prescription drugs, including biological therapeutics and generic drugs. This work covers more than just medicines. For example, fluoride toothpaste, antiperspirants, dandruff shampoos and sunscreens are all considered drugs. If a potential safety concern is identified in AERS, further evaluation might include epidemiological studies. Based on an evaluation of the potential safety concern, FDA may take actions as described in the figure below.

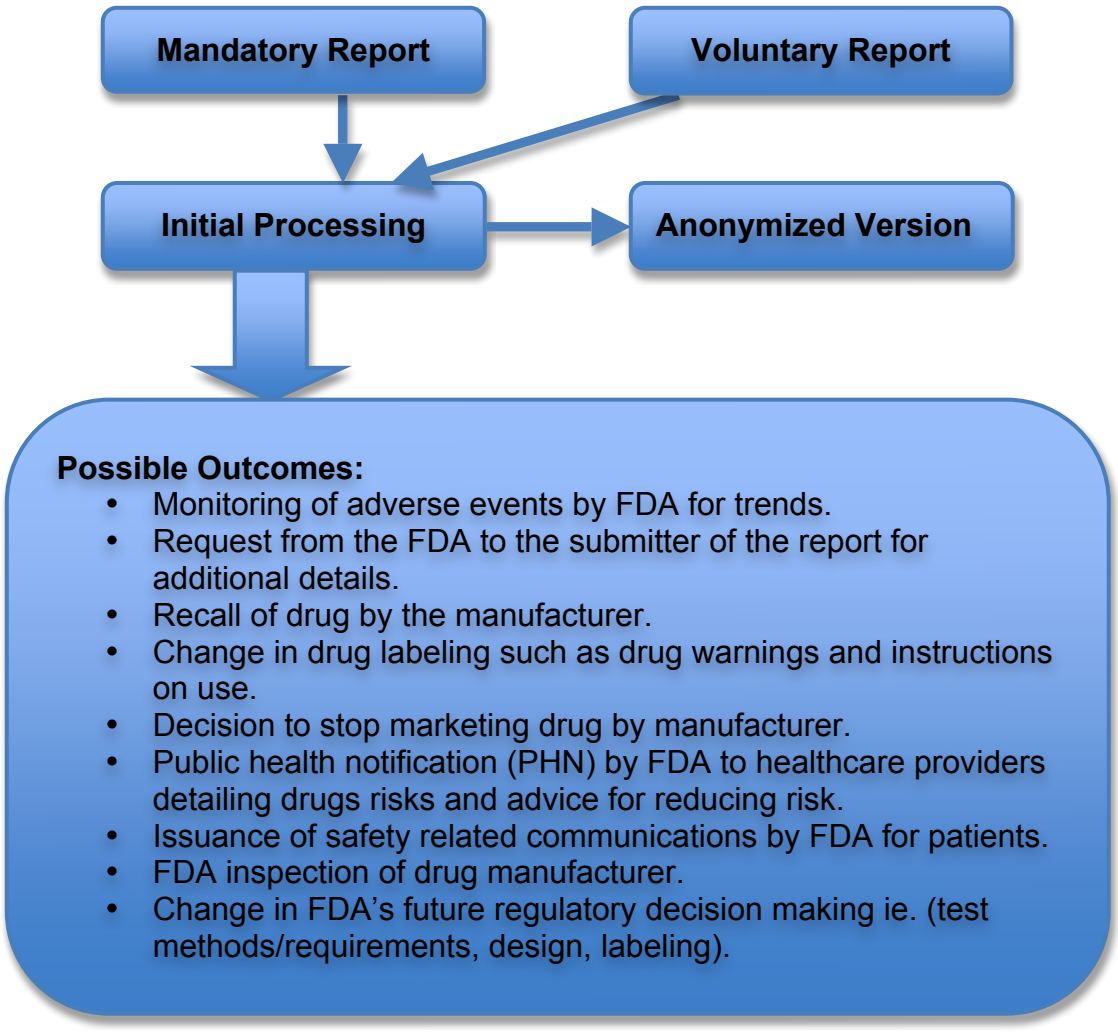


Figure 1: Flow chart of FDA AERS system. Medwatch 3500a is the mandatory form submitted by facilities, distributors, and manufacturers of drugs or devices. Form 3500 is voluntarily submission by healthcare professionals and consumers for spontaneous adverse events surfacing in the course of clinical care. The resulting forms are processed and anonymized, released then further action is pursued.

We hope to enrich the existing AERS system using online health forum messages. We believe that this source of data is currently untapped and has the potential to provide much data. Spontaneous reporting systems such as AERS have been known to have under report rates as high as 98% (Fletcher, 1991). Numerous high-profile safety problems have resulted, not from the lack of the FDA's authority to regulate drugs, but

lack of post marketing surveillance information about drugs (McClellan, 2007). It has been demonstrated that the quality of patient reporting is similar to that of healthcare professionals and that patients are more likely to report adverse drug reactions when they believe that their healthcare professional has not paid attention to their adverse reaction (Leamon et al., 2010).

Patients look to physicians and other healthcare professions as their main source of health information. Yet many physicians do not have a clear picture of their patient's health. It is commonly understood that patients usually spend a very limited amount of time per year directly interacting with a medical doctor, nurse, or health-care system. A typical face-to-face patient-physician encounter is between 15 and 20 minutes long (Travaline, 2005). This is often an insufficient amount of time for a physician to review all aspects of a patient's medical history. This problem is further compounded by the fact that a physician may not have sufficient information through interactions and a patient's medical record to make informed decisions about care. Much of this data is not available in electronic medical records or discussed with health care practitioners; patients are not always honest with their physicians or compliant with treatment plans. Patients construct their own drug regimens consisting of both prescription and non-prescription medication utilized in ways in which they are not intended. Our previous work shows that patients are more honest with peers, "I can't take potassium and only take Lasix as needed," and share patient derived regimens, "I found that 2 Flexiril and a benydryl help me relax and go to sleep."[2] As a primary source of information, physicians need this information to

[2] These quotes were taken from messages on Yahoo! Health groups.

accurately assess a patient's condition and to help a patient arrive at a useful treatment plan.

Many times patients share such information in online health forums or chat groups, "60% of e-patients, or one third of adults access social media related to health" (Fox and Jones, 2009). These social media forums include social networking sites such as MySpace and Facebook. Social media has become entrenched in our daily lives; for example: Facebook has an estimated 500 million active users. Social media forums also include smaller health specific sites such as CHFpatients.com, a site specifically targeted at people with or caring for those with Congestive Heart Failure, "41% of e-patients have read someone else's commentary or experience about health or medical issues on an online news group, website, or blog" (Fox and Jones, 2009). While the topic of health as a social networking site theme is nascent, there are numerous other groups, message forums, and other online communities where people may gather. As of May 25, 2008, there were 162,742 health-related groups on Yahoo Health. The number of community health related sites has grown steadily and will increase as more people turn to online sources for information, support, and decision making help.

Patients seek others like them to share information, support and advice within these forums establishing interpersonal relationships. For rare conditions there may be a limited number of individuals within a close geographic space and for those with serious conditions, they maybe limited in the people they can physically see. People turn to online communities to supplement and augment real world community, information and support system. These interpersonal relationships between patients within a forum or support group form a community and social network. Within these networks patients are

among peers and can share common experiences about their situation. It is known that people do what their peers do, such as friends, family members, or those people with whom they regularly interact (McPherson, 2001).

People want to know what others think and these opinions are important in the decision making process (Pang and Lee, 2008). People often ask family, friends or other peers for recommendations on goods and services or have consulted Consumer Reports for product reviews. Outside of physicians and other healthcare providers, family and friends are the highest rated source of information for online health seekers. In such studies, however, it is not known if the information derived from health forums were considered online sources, friends, or both. The Internet enables the discovery of opinions and experiences of all different types of people from professional critics to grandmothers. This information is used with increasing frequency for many products; in two separate studies of more than 2,000 American adults the following was found (Pang and Lee, 2008):

- 81% of Internet users (or 60% of Americans) have done online research on a product at least once.
- 20% (15% of all Americans) do so on a typical day.
- Among readers of online reviews of restaurants, hotels, and various services (e.g., travel agencies or doctors), between 73% and 87% report that reviews had a significant influence on their purchase.
- 32% have provided a rating on a product, service, or person via an online ratings system, and 30% (including 18% of online senior citizens) have posted an online comment or review regarding a product or service.

Sentiment or opinions expressed in online forums towards a particular treatment is an implicit review for that treatment option. Often these sentiments are causal in nature; for example, "I don't like drug X because it gives me headaches." The causality of these opinions or inclusion of adverse drug events is similar in nature to spontaneous drug reporting systems like ones which the FDA has in place such as MedWatch. The MedWatch voluntary report by consumers asks a patient to, "describe event, problem, or product use error,"[3] and includes sections about "other relevant history, including preexisting medical conditions" which occur in discussions on online health forums. The combination of opinion and latent history is equivalent to reports entered into a spontaneous, even reporting system like MedWatch. It is widely understood that these systems are useful in the area of pharmacovigilance – the study of drugs in populations once introduced into the market (Edwards and Aronson, 2000).

Much of what occurs on in these online health forums is information and experience sharing. Information and experiences about drugs can inform the general public about potential problems not found during clinical trials and should be utilized as an information gathering point for pharmacovigilance. Patients often seek people with first hand experience with treatment options much like consumers seeking product reviews. People do what their peers do; the social network a person resides in informs them and generates information about them. Is this information regarding drugs accurate? Can it be leveraged to form the basis of pharmacovigilance techniques, or are the social network's aggregate sentiment toward a drug the results of mass effects? Does what peers think correlate to FDA watch list? The goals of this dissertation are to

[3] From the MedWatch form available at: http://www.fda.gov/downloads/Safety/MedWatch/HowToReport/DownloadForms/UCM082725.pdf

explore these questions and determine if sentiment within such networks is a useful indicator of FDA watch list potential.

Not only is it important to understand whom a patient interacts with, but also how patients interact with one another and utilize the information shared within their networks. I aim to address the following research questions in exploring the idea of utilizing patient derived data within the context of online message forums:

Q1: Can we develop machine learning classification methods utilizing textual features derived from online health forum messages to differentiate between FDA watch list drugs and non-watch list drugs.

Q2: Are sentiment (positive and negative) and drug effect entities useful features in differentiating FDA watch list drugs and non-watch list drugs?

1.1 Contribution of Work

The main contribution of this work is the exploration of Yahoo! public health message forums as a potential data source for adverse drug event mining and use in pharmacovigilance techniques. Through this exploration I will develop protocol to test the feasibility of utilizing this source, which is then generalizable to other potential sources such as private message groups, blogs or other social media sources.

Currently it has not been shown that machine learning techniques such as classification have been successfully used in predicting drugs correlated with adverse events or FDA watch list drugs. This dissertation is also an exploration on the application of classification techniques to assess the ability to discern FDA watch list drugs from those that are not utilizing textual features.

2 Literature Review

2.1 Adverse Drug Events

Any drug or treatment option that causes a positive effect is also capable of producing an unwanted or adverse effect. These effects are far ranging from minimal (near zero) to high, such as in the case of immunosuppressive or highly toxic chemotherapy treatments. The World Health Organization (WHO) (1972) defines an adverse drug reaction as "a response to a drug that is noxious and unintended and occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease, or for modification of physiological function." However, this definition excludes error as a source of adverse effects and reactions due to contaminants or supposedly inactive excipients in a formulation. Edwards and Aronson (2000) suggest that an adverse drug reaction is "an appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen, or withdrawal of the product."

Though adverse events may occur when taking a drug it is hard to attribute causality that the drug caused the adverse event. If causality is established then the adverse event is an adverse reaction. A report about a possible adverse reaction is listed as an adverse event until causality is established. Edwards and Aronson (2000) describe differing types of adverse drug events such as dose-related, time-related, withdrawal and developed types of mnemonics to describe them. The table below is taken from that publication and succinctly describes the various types of drug reactions.

Table 1: Table discussing various types of adverse drug events from Edwards and Aronson (2000).

Type of reaction	Mnemonic	Features	Examples	Management
A: Dose-related	Augmented	• Common • Related to a pharmacological action of the drug • Predictable • Low mortality	• Toxic effects: Digoxin toxicity; serotonin syndrome with SSRIs • Side effects: Anticholinergic effects of tricyclic antidepressants	• Reduce dose or withhold • Consider effects of concomitant therapy
B: Non-dose-related	Bizarre	• Uncommon • Not related to a pharmacological action of the drug • Unpredictable • High mortality	• Immunological reactions: Penicillin hypersensitivity • Idiosyncratic reactions: Acute porphyria Malignant hyperthermia Pseudoallergy (eg, ampicillin rash)	• Withhold and avoid in future
C: Dose-related and time-related	Chronic	• Uncommon • Related to the cumulative dose	• Hypothalamic-pituitary-adrenal axis suppression by corticosteroids	• Reduce dose or withhold; withdrawal may have to be prolonged
D: Time-related	Delayed	• Uncommon • Usually dose-related • Occurs or becomes apparent some time after the use of the drug	• Teratogenesis (eg, vaginal adenocarcinoma with diethylstilbestrol) • Carcinogenesis • Tardive dyskinesia	• Often intractable
E: Withdrawal	End of use	• Uncommon • Occurs soon after withdrawal of the drug	• Opiate withdrawal syndrome • Myocardial ischaemia (β-blocker withdrawal)	• Reintroduce and withdraw slowly
F: Unexpected failure of therapy	Failure	• Common • Dose-related • Often caused by drug interactions	• Inadequate dosage of an oral contraceptive, particularly when used with specific enzyme inducers	• Increase dosage • Consider effects of concomitant therapy

SSRIs=serotonin-selective reuptake inhibitors.

Currently within the United States the governing organization that oversees drug safety is the U.S. Food and Drug Administration (FDA). As part of programs for post-marketing and safety surveillance the FDA has created computerized information databases, the Adverse Event Reporting System (AERS) and Vaccine Adverse Event Reporting System (VAERS).

Both consumers – including patients and family members and health care professionals such as physicians, pharmacists and nurses may voluntarily report adverse events to AERS or to drug manufacturers. If a manufacturer receives an adverse event report, it is required to send the report to FDA as specified by government regulations. The reports in AERS are evaluated by the Center for Drug Evaluation and Research (CDER) and the Center for Biologics Evaluation and Research (CBER). Based upon these evaluations the FDA may take regulatory action such as updating a product's

labeling information, restricting the use of the drug, communicating new safety information to the public, or, in rare cases, removing a product from the market.[4]

VAERS is similar to AERS; however, it is specifically targeted at vaccines. It was created as an outgrowth of the National Childhood Vaccine Injury Act of 1986 (NCVIA) and is administered by the FDA and Centers for Disease Control and Prevention (CDC).[5]

The goal of medical regulators such as the FDA is to make decisions based on the best evidence available with the aims of promoting and protecting public health (Waller, 2001). Regulators are charged with the task of making prompt decisions such that delays of new medications with positive effects are minimized. The FDA and other governing bodies rely on clinical trials to demonstrate drug safety and efficacy. However such studies exclude certain types of patients and situations, for example, women and the elderly are often underrepresented in pre-marketing clinical trails leading to significant differences (as much as 63%) from drug studies to real world observations of effectiveness and safety (Martin et al., 2001).

These factors highlight the need for post-marketing drug surveillance. This type of post-marketing surveillance falls within the realm of pharmacovigilance, "a principle concern of pharmacovigilance is the timely detection of adverse drug reactions that are novel by virtue of their clinical nature, severity, and/or frequency" (Hauben et al., 2005). The table below from Edwards and Aronson (2000) highlights different methods of post-marketing surveillance indicating their advantages and disadvantages.

---

[4] http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm

[5] http://www.fda.gov/BiologicsBloodVaccines/SafetyAvailability/ReportaProblem/VaccineAdverseEvents/Overview/default.htm

Table 2: Various types of post marketing surveillance from Edwards and Aronson (2000).

Method	Advantages	Disadvantages
Anecdotal reporting (eg, in journals)	Simple; cheap	Relies on individual vigilance and astuteness; may only detect relatively common effects
Voluntary organized reporting* (doctors, pharmacists, pharmaceutical companies)	Simple	Under-reporting; reporting bias by "bandwagon" effect
Intensive event monitoring	Easily organised	Selected population studied for a short time
Cohort studies	Can be prospective; good at detecting effects	Very large numbers required; very expensive
Case-control studies	Excellent for validation and assessment	Will not detect new effects; expensive
Case-cohort studies	Good for studying rare effects with high power	As for cohort and case-control studies; complex calculations
Population statistics	Large numbers can be studied	Difficult to coordinate; quality of information may be poor; too coarse
Record linkage	Excellent if comprehensive	Time-consuming; expensive; retrospective; relies on accurate records
Meta-analysis	Uses data that have already been obtained	Need to obtain unpublished data; heterogeneity of different studies

*Including computerised systems involving, for example, WHO's monitoring programme,[18] the Committee on Safety of Medicines' yellow card system,[19] and the Food and Drug Administration's Adverse Drug Reaction file system.[20]

There is much interest in developing automated surveillance and pharmacovigilance techniques. Databases such as AERS and VAERS provide early warnings of possible safety problems that are difficult to detect during clinical drug development and trials due to power limitations, constricted range of demographics, exclusion of patients with extensive co-morbid illness and co-medications, and limited duration of follow up. However, while databases such as these provide information, they contain much 'noise' where reports are associated with treatment indications, co-morbid illness, protopathic bias, channeling bias, and/or other reporting bias (Hauben et al., 2005). These present two challenges in utilizing this data: the necessity of waiting until enough data accumulates to differentiate a 'signal' from 'noise' and also developing data mining algorithms capable of differentiating 'signal' from 'noise'.

Hauben et al. (2005) present an overview of key algorithms, databases used and their findings. The authors discuss common approaches including contingency tables and dis-proportionality measures, employing a statistical measure of "interestingness" comparing a drug's likelihood to cause an adverse event compared to other drugs. Regression modeling is also employed to address some of the shortcoming associated

with previous two methods, namely leading to false conclusions due to incorrectly modeling the underlying causal models generating the data. However, regression modeling fails to address dependencies between adverse effects and fails to take into account unmeasured or unrecorded measures and often requires assumptions that are not necessarily appropriate.

Other studies have utilized information available to hospitals such as hospital discharge records, clinical notes, laboratory, pharmacy and administrative data. See Bates et al. (2003) for a review article on detecting adverse events using information technology for numerous studies. This dissertation suggests another source of data, online health forums which contain colloquial information through patient-patient and patient-health care professional interactions. This is most similar to studies utilizing clinical notes or text of outpatient encounters like those in Hongiman et al. (2001). However, this study did not use natural language processing, instead it utilized simple keyword search yet still obtained encouraging results (Hongiman et al., 2001). While natural language processing has been utilized to a wide extent within the health domain it has been used within the context of electronic medical records or other clinical setting data (Bates et al., 2003).

2.2 Online Life

The Internet has become increasingly pervasive in all aspects of people's lives satisfying varieties of informational, communicative and entertainment needs (Cotton and Gupta, 2004). It allows people to keep in touch and expand social networks; keep abreast of news; purchase goods and services; and search for product information and reviews (Pang and Lee, 2008)(Cotton and Gupta, 2004).

Social activities such as social media and networking sites have become an increasingly important part of people's use of the web. Social media consists of all forms of media including blogs, podcasts, video whose dissemination is through social means. The number of blogs has increased dramatically in the last decade (Gilbert, Bergstrom and Karahalios, 2009). Similarly the number of people on social networking sites such as Facebook has increased as well; currently there is an estimated 750 million users on Facebook of which 50% log on each day.[6] Social networking sites such as Facebook and online communities like those formed within CHFpatients.com allow people to maintain friendships and relationships formed in the physical world (Kendall, 1999). However, such sites also allow people to find people like themselves in a variety of ways including but not limited to hobbies or interests, sexual orientation, religious affiliation, and chronic illness or disability (Kraut et al., 2002)(Warschauer, 2003). Such sites allow people to cross traditional geographic boundaries enabling those with low mobility or rare illness to find others similar to them to interact with.

Of the various ways people interact socially, this dissertation focuses primarily on online health forums or support groups in which people can communicate with one another within a group setting. Various types of sites fulfill this requirement, for example a blog may provide a venue where others can comment to an initial post, however the overarching goal of a blog is to voice the opinion of the owner. Social media and networking sites blend in various ways taking components of one another.

The increase in popularity of such sites and the connectivity it creates between people has created an area of research on computer-supported social networks (CSSN).

[6] http://www.facebook.com/press/info.php?statistics

Garton et al. (1997) consider a computer network that connects people or organizations a social network. Research in CSSN shifts from individualism in social sciences towards structural analysis of the relationships between people looking for patterns explaining the behavior and attitudes of network members (Garton, Haythornthwaite and Wellman, 1997).

2.3 Support Groups for Chronic Illness

It is well understood that participation in support groups is an important component in the social support network for those with chronic illness; they ideally allow the fusion of ideas, answering practical questions while also allowing the exchange of common concerns and experiences (Silver, 2002). Thus people no longer feel alone and helpless; support groups afford the opportunity to share feelings with people of similar diagnosis, and who therefore understand them in ways that others cannot. In a study of patients with diabetes, it was indicated that support groups were as important as support from family or friends (Landis, 1996). Silver (2002) considers family and friends a type of support group who help patients deal with numerous drug regimens, doctor appointments and dietary restrictions.

Support groups help patients deal with the bitterness and anger often associated with chronic illness that often inhibits successful treatment of the illness. Depression often goes hand in hand with chronic illness, nearly 40% of patients with heart failure in the US have depression and even in the Chinese medical literature, one of the symptoms of heart failure is depression (Silver, 2002).

Physical and face-to-face interactions provide additional cues such as pausing, intonation and facial expression that are not available in purely text-based forums (Walter

and D'Addario, 2001). Human contact – touch, also lacks in these encounters. The impact of a touch in the form of hugs for example, correlates positively in the treatment of fibromyalgia, autism, and general health; touching helps to reduce stress and pain, leading to improvements in the ability to cope and in overall general health measurements (Denison, 2004)(Grandin, 1992)(Weze et al., 2005).

2.4 Online Support Groups

Online support groups are gaining traction and are seen as a way to augment existing social support networks and in some cases possibly supplant them (Houston, Cooper and Ford, 2002). Work looking at the role of online support groups spans a variety of chronic illnesses from depression to cancer (Houston, Cooper and Ford, 2002)(Eysenbach et al., 2004). Online support groups are similar in many ways to traditional support groups and provide avenues for people to meet others that have similar conditions, learning from their experiences as well as socializing them to their illness and providing information and support (Mittem and Cain, 2001)(Lieberman and Goldstein, 2006).

Pennebaker (1997) demonstrates the positive effects – reduction of stress and promotion of physical and mental well-being-of writing about traumatic experiences. Talking and writing about emotional experiences provide comparable biological, mood, and cognitive effects such as t-helper cell growth, antibody response to Epstein-Barr virus and to hepatitis B vaccinations, various short term changes in autonomic activity such as lowered heart rate and electro-dermal activity. Further studies show that communicating and sharing experiences with peers reduces depression, stress, emotional

distress and feelings of social isolation (Gustafason et al., 2001)(Lindsay et al., 2007)(Winzelberg et al., 2003).

2.5 Privacy

It is well understood that health is an extremely private and personal part of one's life. The US government has created provisions for privacy regarding health information within the Health Insurance Portability and Accountability Act (HIPAA) of 1996 (P.L.104-191). The HIPAA Privacy Rule regulates the use and disclosure of Protected Health Information (PHI) held by "covered entities" such as health insurance companies or medical service providers when engaging in certain types of transactions (Terry, 2009). PHI consists of personally identifiably information held by a covered entity concerning health status, payment for healthcare or provision of healthcare (45 C.F.R. 164.501).

HIPAA regulations are currently being augmented through Subtitle D of the Health Information Technology for Economic and Clinical Health Act (HITECH Act), enacted as part of the American Recovery and Reinvestment Act of 2009. The HITECH Act aims to addresses the privacy and security concerns associated with the electronic transmission of health information. It extends the policies for disclosure requirements to information that is used to carry out treatment, payment and health care operations when an organization is using an electronic health record (EHR). Along with this new policy covering disclosure requirements are extensions of HIPPA privacy provisions to business associates of covered entities and notification requirements if a breach of unsecured PHI occurs.

However, while there are privacy safeguards concerning EHRs there is no legislation involving privacy or use of patient volunteered health information on health message boards, forums or websites such as webmd.com or mayoclinic.com. In a study of 1,009 US adults released in January 2000 by the California HealthCare Foundation and the Internet Healthcare Coalition, it was found that:

75% of people are concerned about health web sites sharing information without their permission.

40% would not allow physicians online access to their medical records.

25% would not buy or refill prescriptions online.

17% of people don't go online to seek health information due to concerns over privacy.

A significant percentage of people would not engage in certain health related activities due to concerns over privacy and security.

16% of people would not register on websites.

Internet security and privacy regarding health information is significant. While a patient volunteers this information there are key differences in the information provided by a patient regarding health and other types of information on similar message boards or websites. Some types of data are more sensitive than others; a study of attitudes toward online information practices found that while 82% of respondents were comfortable providing information of their favorite television show about themselves and 52% for a child, only 18% were comfortable providing medical information about themselves and 4% for a child (Cranor, Reagle, and Ackerman, 1999).

Previous studies have focused on the privacy policy of health sites themselves, scrutinizing them for compliance, readability, and protecting visitors' privacy (Goldman, Hudson and Smith, 2000). However, few studies consider how the personal health information is utilized on the site, and if it changes, how and what information people are willing to disclose. How information is used also dictates people's willingness to share information. Internet users are more likely to provide information if they remain anonymous and if people understand the purpose for which the information is being collected and used (Cranor, Reagle and Ackerman, 1999).

While this dissertation does not elicit information directly, it depends on patient volunteered information from online sources such as message boards and websites. As such, exploring the mechanisms for sharing information and understanding people's aversions to contributing information is important future work to enable the induction of as much information as possible.

2.6 Crowdsourcing and Collective Intelligence

Crowdsourcing is a term coined by Jeff Howe in 2006 describing the use of the distributed labor networks – akin to "outsourcing" instead of sending jobs to other places or countries, using populations of laborers found online due to the diminishing gap between professionals and amateurs. One example, Mechanical Turk, Amazon's online marketplace allows people to outsource work, often very small tasks for small amounts of money. Examples might include labeling pictures, giving feedback on short videos or evaluating real and hypothetical products[7]. These tasks are considered human intelligence tasks, tasks that humans are good at where computers often falter.

[7] https://www.mturk.com/mturk/welcome

Crowdsourcing enables the participation of many people for short amounts of time with relatively little recruitment or administrative effort and cost. However such systems including Mechanical Turk, have drawbacks as Downs et al. (2010) highlight, the anonymity and lack of accountability – one of the draws of such systems attract people to game the system by participating in many tasks without fully engaging in them. However, Crowdsourcing has been successfully used in academic situations such as evaluating information retrieval results. There have also been many positive examples of Crowdsourcing with larger payouts and accountability, such as InnoCentive[8], where payouts from companies such as Colgate-Palmolive have reached between $10,000 and $100,000 for innovative ideas for challenging problems.

We can pose the pharmacovigilance task as a crowdsourcing task with caveats. Within this paradigm, there is currently no financial incentive for individuals and instead compensations are derived through social engagement or individual feelings of accomplishment. Secondly, the input is not directly usable and in our case requires natural language processing work to extract useful information.

In many ways the pharmacovigilance task as posed is similar to collective intelligence or information cascades. Surowiecki (2004) explored the idea that the aggregation of information from groups results in decisions that are often better than those that could be made by a single member of the group. This differs slightly from information cascades where people follow the actions of others, making similar choices. People rely on the choices others have made, making a rational decision based upon limited information (Easley and Kleinberg, 2010). One such example of an information

[8] http://www2.innocentive.com/

cascade is deciding on a restaurant based on the number of people already eating there. Cascades can be wrong or derived from little information and cause short-lived spikes such as fads, fashion trends, or runs on banks.

One can imagine people basing drug decisions based on information cascade effects, doing what their group of peers does. Similarly people within a group can come to a consensus of the safety of a drug and that notion, wrong or not, can stick or spread within online health communities. It is worthwhile to look at such trends and sources of information about drug safety within online health forums.

2.7 Research Tools and Methodologies

2.7.1 Language Analysis

The words people use correlate with their physical and mental health (Pennebaker and Campbell, 2000). Content analysis introduced in the 1960's detects a person's affective or immediate feeling state based solely on variations in the content of verbal communications (Gottschalk and Gleser, 1969). This technique utilized human interpretation and coding of transcripts of people who were interviewed. The same technique was used in the late 1970's to differentiate schizophrenics from non-schizophrenics (Rosenberg and Tucker, 1979). However, Rosenberg and Tucker's (1979) work utilized computational analysis of transcripts between patients and doctors. Later work focused on written text, finding variations in language usage between depressed and depression-vulnerable students (Rude, Gortner and Pennebaker, 2004). In the Rude et al. experiment, essays were written by college students, then analyzed computationally removing the interviewer from the process.

Words are not the only elements of analysis that provide necessary emotional insights. Face to face communication has many elements providing backchannel communication to speakers including intonation, pauses, or body language. Sarcasm is often indicated by a different intonation that is not easily conveyed in written language, for instance, "you look great" can be interpreted in multiple ways but unless there is context or backchannel cues it is difficult to understand what the writer meant. People have augmented computer-mediated communication to provide the same richness of face-to-face interactions through the use of nonverbal elements (Walter and D'Addario, 2001). Emoticons are nonverbal expressions and are often textual representations of writer's facial expressions (Gajadhar and Green, 2005). For example :) or :-) would correspond to a smile indicating happiness. These cues indicate to the reader the author's intentions that can be hard to determine in informal written communication. Other backchannel cues include spacing within instant messaging to indicate pauses like those in normal conversation.

Chung and Pennebaker (2005) discuss the possibility of assessing people's quality of life through computerized analysis of their responses to open ended questions. They look at function words such as positive/negative emotion words, negations (not, nor, never), prepositions, articles, first person singular pronouns (I, me, my), and other pronouns (she, they we) to determine how these correlated with the Satisfaction with Life Scale. The way people speak reliably related to depression, self-esteem, feelings of isolation and togetherness, self- and other-deception, cognitive complexity and hormone levels. Most importantly they suggest that much information about a person's

satisfaction with life lies in analyzing the way people talk about their lives and not only what they say.

2.7.2 Sentiment Analysis and Opinion Mining

Looking at what a person says can inform others on what they think. Determining what other people think is an important task in information gathering behavior (Pang and Lee, 2008). This can be evidenced by the large number of product review sites online or the feedback found on shopping websites such as Amazon.com. The Internet enables the discovery of opinions and experiences of all different types of people from professional critics to grandmothers. This information is used with increasingly frequency for many products; in two separate studies of more than 2,000 American adults the following was found (Pang and Lee, 2008):

- 81% of Internet users (or 60% of Americans) have done online research on a product at least once.
- 20% (15% of all Americans) do so on a typical day.
- Among readers of online reviews of restaurants, hotels, and various services (e.g., travel agencies or doctors), between 73% and 87% report that reviews had a significant influence on their purchase.
- 32% have provided a rating on a product, service, or person via an online ratings system, and 30% (including 18% of online senior citizens) have posted an online comment or review regarding a product or service.

Sentiment analysis involves performing some or all of the following processes on text: analyzing, extracting, aggregating and representing information. This information is subjective in nature as it is derived from opinion. Since opinions are subjective, opinion

mining is inherently difficult. If humans cannot agree on a sentiment score reliably it is even more difficult for machines to learn to score text accurately. In fact, machine learning based methods of sentiment classification do not perform as well as machine learning topic-based binary classification ones (Pang, Lee, and Vaithyanathan, 2002). There are multiple methods of assigning sentiment scores to a given piece of text; often these are in the form of machine learning based methods employing classifiers utilizing word features, ngrams (multiple words or phrases), part of speech information, and syntax, to classify a piece of text as positive or negative (Pang and Lee, 2008). However, this list glosses over the discussion of level of specificity; sometimes merely labeling a piece of text as positive, neutral or negative is not enough and there needs to be degrees of positivity or negativity. There is also the problem of document segmentation; documents are composed of multiple ideas with multiple degrees of sentiment applying to some or all of those ideas; attributing sentiment to a particular idea or salient thought is difficult. The reader is directed to Pang and Lee, (2008) for an in-depth discussion of many of these ideas.

Sentiment analysis tools are highly domain dependent like many other NLP tools, as the language people use to express emotion differs between domains; for example in the area of movies, unexpected is often thought of as a positive trait, whereas in describing the handling of a car it is often considered a negative feature (Turney, 2002). While sentiment is often used to aggregate product reviews on the web, there is a multitude of non-product evaluation centric based uses of it with regard to applications in the medical domain as previously discussed in the language analysis section. As discussed previously many such uses of sentiment and language analysis within the

context of medicine is determining mental state and cognitive models. However, this dissertation envisions applying sentiment directly to treatment outcomes and evaluation of them. One technique used to measure the effectiveness of a drug is quantifying the side effects that it produces. Sentiment analysis is another way in which we can determine drug satisfaction. A drug may have many serious side effects, yet people may still have a positive attitude towards it, especially if they believe it helps them in some way (Silver, 2002).

On the surface, it appears that sentiment containing words are domain agnostic for words like good, happy, and sad. However, if one looks deeper into sentiment, one can see many non-obvious words refer to someone's emotional alignment. For example, if someone mentions an actor's name whom you detest, your reaction could be very negative; for you this is a negative sentiment containing term. Similarly explosions in general could be negative sentiment words but for an action movie aficionado this could be a positive sentiment term. In this way, if training examples for both the domain as well as population one is trying to track, better performance often results. A common method employs Bayes rule to calculate the sentiment of individual words if a training corpus exists where P(word | positive or negative class). Here we calculate the probability of a word given a class – positive or negative sentiment.

However, in cases where little domain knowledge or training data is available, sentiment words that are predominately positive or negative will still provide useful information. These words are often found in existing hand crafted lexical resources such as Linguistic Inquiry Word Count (LIWC) or SentiWordNet, which ascribe a score to

particular words. These scores incorporate world knowledge as well as corpus statistics to weight the prevalence of a particular sentiment containing word.

One could use the LIWC scores or SentiWordNet scores (or even the Bayes probabilities) to label a piece of text as predominately positive or negative. All that is needed is a scoring metric and a decision rule. If one has probabilities, if one assumes word independence then the probabilities of the words are multiplied and the most likely class is assumed given the data (Naïve Bayes classification). Arbitrary scoring metrics are more heuristic in nature but often work in practice. Whereas Naïve Bayesian classification has strong mathematical foundations, it does not accurately reflect real world data (the independence between word probabilities) however, it also works in practice.

Other approaches are also employed such as Support Vector Machines (SVM) where individual word scores are not necessarily employed but a non-linear mapping of features (words) is used. Conditional Random Fields (CRF) could also be employed which take into account the idea that word ordering matters instead of more traditional bag of word (BOW) approaches where ordering is ignored. SVMs and CRFs may leave users wondering how a particular classification choice was arrived at due to the opaqueness of the algorithms, whereas if individual words are labeled with scores it is relatively easy to determine how a particular class label was arrived at.

Regardless of the method used to determine the polarity of individual words or pieces of text, language evolves and word meanings shift over time. This is apparent even within sentiment containing words. The meaning of "bad" which is a strong sentient containing word associated negatively, can and does mean "good" or have a

positive connotation depending on the domain as well as the time period. Such polar opposites are likely few and far between but do occur. I believe that they are also less likely to occur with "obvious" or traditional sentiment words like good, hate, death, etc. Thus in the same way traditional sentiment words are more stable over time and are less likely to change or drift than other domain specific sentiment terms.

2.7.3 Named Entity Recognition

Named entity recognition (NER) is the area of natural language processing and information extraction that deals with the identification of specific entity types or information units commonly including names of people, locations, organizations, and numeric expressions including time, date, money and percent expressions (Nadeau and Sekine, 2007). As Nadeau and Sekine (2007) point out that named entity recognition is a task often restricted to entities with one or many rigid designators such as proper names, biological species and substances. However, they point out that this designation has been relaxed to include numerical expressions such as dates or months such as June since the month itself does not refer to a specific date.

This dissertation similarly relaxes the constraint on named entities and refers to useful information units such as drug names and noun phrase outcome caused by drug like side effects. Earlier work by Rindfleisch, Tanabe and Weinstein (2000) identified drugs and relations from biomedical literature in Medline. We extend the idea of drug identification to also include the identification of effects from drugs from colloquial literature. Effects include but are not limited to harmful side effects such as headaches or vomiting.

Many NER techniques focus on the identification of information units from unstructured text, the text is generally formal, edited for grammatical and spelling correctness as in the case of news wire or biomedical text (Nadeau and Sekine, 2007). The use of text available on the web is appealing due to the vast quantities and ease of obtaining it. Previous NLP tasks have utilized web data for tasks in machine translation, prepositional phrase attachment, NER, and other-anaphora resolution (Brin, 1998)(Liu and Curran, 2006). However in a comparison of a billion tokens from newswire text to general web text there were significant differences in types of tokens. Web data contains significantly larger numbers of misspellings and typographical errors and much fewer cases of title case – tokens that start with an uppercase letter followed by lowercase letters like "London" (Liu and Curran, 2006).

Identification of named entities within corrected text is easier than general web data due to consistent capitalization and spelling for proper nouns; grammatically correct sentence constructions enabling other supporting NLP tools such as sentence boundary identifiers, part of speech taggers, or syntactic parsers with increased accuracy since tools utilizing machine learning are often trained on similarly corrected text.

However, Brin's (1998) work focused on utilizing lexical features with regular expressions to identify book titles and authors from web data. Later work by Etzioni et al. (2005) focused on general named entity identification from web data. Brin's (1998) work still relies on book titles and author names to follow capitalization conventions across websites, whereas Etzioni et al. (2005) rely on web scale data to provide many examples to learn from. The Yahoo dataset often is composed of such web quality data

but also includes SPAM messages aimed specifically at confounding traditional natural language processing techniques to avoid detection.

Many techniques are utilized for NER, often hard coded heuristics employing patterns such as "city of *X*" to identify cities, for example "city of London"; dates in the form of *MM-DD-YYYY*; or dictionaries for non ambiguous terms like locations or biological names. Machine learning, both supervised and unsupervised methods are also employed utilizing a variety of lexical (*X* is a city), morphological ("ist" endings as in journalist or cyclist), honorifics (Mr. or Dr. before name), capitalization, or part of speech (noun or proper noun) features (Nadeau and Sekine, 2007).

2.7.4 Relationship Extraction

Hearst (1992) suggested a method for automatic acquisition of the hypononymy lexical relation from free text. The goal of this work was to avoid the need for pre-encoded knowledge and apply to a wide variety of text. In the work several examples of lexico-syntactic patterns for identifying and extracting hyponyms from free text are presented. These patterns are in the form:

NP *such as* {NP, NP, …, (*and* | *or*} NP

such NP *as* {NP, }* {(*or* | *and)*} NP

NP {,} *especially* {NP, }* {*or* | *and*} NP

Here NP indicates a noun phrase and italicized words are lexical markers used to identify the pattern. An example might include, "the *bow lute, such as the Bambara ndang*, is plucked and has an individual curved neck for each string," here the pattern "NP such as NP" extracted is illustrated in italics (Hearst, 1992). Later work built upon these hand crafted patterns for hyponym extraction utilizing machine learning techniques to learn

and induce new patterns via the extracted examples from the original patterns (Riloff, 1996)(Etzioni et al., 2004).

Inducing new lexico-syntactic patterns automatically resembles the wrapper induction problem in the area of information extraction. The task of automatic wrapper induction for information extraction involves building systems that learn patterns to extract salient facts from sources providing human-centered interfaces, i.e. web portals (Kushmerick et al., 1997). An example of information extracted utilizing wrappers includes extracting name, telephone number pairs from online telephone books or movie name, rating, location and time tuples from a movie site. Whereas lexico-syntactic patterns rely only on linguistic features, wrapper induction relies both linguistic and other cues such as html or xml tags surrounding text, or placement on a page.

This work is also related to discourse analysis, specifically the area of speech acts involving identification of dialogue acts. Dialogue tags like continuers, assessments, yes-answers, agreements, incipient-speakership, or acknowledgement tokens are attributed to surface level communicative acts in a dialogue or conversation. Much work in the area of dialogue acts involves the classification of act types previously mentioned (Jurafsky et al., 1998). This work focuses on identifying particular types of dialogue acts such as statements and opinions with regard to particular subjects (treatments) and intent (recommendation).

2.7.5 Classification

Classification is a supervised learning task within the area of Machine learning. Its goal is to utilize hand labeled instances (in a binary case positive or negative instances) to train a machine learning classifier to classify or categorize previously

unseen instances as a particular class. Classification is a common task within natural language processing (NLP). Machine-learning tasks within NLP fall into two categories: unsupervised – clustering, or supervised, classification tasks. Clustering involves grouping items based upon features that automatically represent them. Classification tasks include part of speech tagging (POS) that labels a word within a sentence with a part of speech such as noun, verb, preposition; named entity recognition (NER) assigning word(s) to a particular entity class such as person, location, or date; sentiment analysis attributing a score to a segment of text based upon how positive or negative the perceived effect of the text is; or assigning portions of text as particular classes such as SPAM or not. This dissertation looks particularly at NER, sentiment analysis, and assigning class labels to portions of text commonly referred to as text classification.

Many classification algorithms exist; however of these, two are popular for text classification applications. This dissertation focuses on these two, Naïve Bayes and Support Vector Machines (SVM). Naïve Bayesian classifiers are widely used in machine learning NLP approaches due to their efficiency and ability to combine evidence from large numbers of features; however, these classifiers make certain assumptions – ignoring structure and linear ordering inherent in text taking the "bag of words" approach and assuming that words are independent of each other – one intuitively expects the likelihood of "president" to be greater given the observation "Obama" (Manning and Schutze, 1999). Bayesian statistics are also widely used within the field of pharmacovigilance adding support for this method within the area of ADE detection.

SVMs work by non-linearly mapping input (feature) vectors into a high-dimensional feature space using a kernel function (Cortes and Vapnik, 1995). In this

space a linear decision surface (hyperplane utilizing n-1 features) is constructed that maximizes a margin between classes within the higher dimensionality space. Unseen or new instances are mapped into the high dimensional space and classified based upon the side of the hyperplane they fall. Forman (2003) found empirically that SVM was the best performing algorithm in a corpus of 229 text classification problems. Joachims (1998) indicates that text classification is an ideal candidate for SVM due to the fact that text categorization problems are linearly separable; text contains few irrelevant features – all words are important; and that text results in sparse input features - not all words occur in every document.

Whatever type of classifier used, feature reduction is usually an important step to reduce the computational time and memory resources needed to train a classifier as well as label unseen instances. Forman (2003) advises that in order to reduce the size of a dataset "without adversely impacting classification performance to set word cutoff rates low (in their experiment words occurring less than 3 times within the dataset were removed) and perform aggressive feature selection using a metric with linear running time if recall is the sole goal," then more low frequency words should be eliminated.

Feature selection utilizes metrics to rate and rank features by their importance in making a classification decision. Forman (2003) found that Bi-Normal Separation (BNS) outperformed other methods by a substantial margin in most situations and is the top single choice except where the goal is to maximize precision where Information Gain (IG) yielded the best results. IG also proved the best results when the number of features was reduced drastically in situations of 20-50 features. IG measures the decrease when a feature exists versus when absent. BNS is a features selection metric developed by

Forman(2003). It is defined as $F^{-1}(tpr) - F^{-1}(fpr)$, where F^{-1} is the standard Normal distribution's inverse cumulative probability function, tpr sample true positive rate, fpr is the sample false positive rate, and 0.00005 when tpr or fpr = 0. BNS worked best when using between 500 and 1000 features, previous works showed little effect of more than 2000 features (Forman, 2003).

2.7.5.1 Support Vector Machines

Support Vector Machines (SVM) are well founded in terms of computational learning theory and have demonstrated empirical performance in the area of text categorization. SVMs can deal with high dimensional feature space and sparse feature vectors often found when dealing with text classification.

For the SVM experiments LibSVM is utilized with a radial basis function (RBF) kernel which is non-linear in nature. Though linear kernels have a faster running time than RBF kernels and the effectiveness of linear kernels with regard to text classification are quite good (Joachims, 1999), some added nonlinearities help obtain finer improvements in accuracy which the RBF kernel can provide (Keerthi and Lin, 2003). Furthermore in certain cases the RBF kernel behaves like a linear kernel SVM. While kernel selection is one parameter when dealing SVMs it is not the only one. A RBF kernel requires two further parameters C and γ, which are not known apriori.

Given a training set of instance-label pairs (x_i , y_i), $i=1,\ldots, l$ where $x_i \in R^n$ and $y \in \{1, -1\}^l$, the SVM requires the solution of the following optimization problem:

Equation 1: The series of equations whose solution is the linear hyperplane with the maximal margin in a higher dimensional space.

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi i$$

$$y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

Training vectors x_i are mapped into a higher dimensional space by the function ϕ. SVM finds a linear separating hyperplane with the maximal margin in the higher dimensional space. Where $C > 0$ is the penalty parameter of the error term. Furthermore, $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ from the above equation, in the case of the RBF kernel $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$.

A grid search method utilizing cross-validation is employed to look for appropriate values of C and γ due to the relatively few numbers of instances, all of them are utilized to find the best combinations of C and γ. While this is somewhat of a brute force approach it is effective and trivial to parallelize since each combination of C and γ is independent of other runs.

2.7.5.2 Naïve Bayesian Classifier

Naïve Bayesian (NB) classifiers are a probabilistic classifier based on applying Bayes' theorem with independence assumptions (naïve) between features. The assumption is that the presence or absence of features is independent of other features. This assumption is not true in real life, for example given the word "wine" it is clear that the likelihood of seeing "red" or "white" is greater than other colors like "yellow" or "purple". Though it is easy to see that this assumption is not true in real life, in practice this assumption can still lead to optimal decisions even if probability estimates are

inaccurate due to feature dependence (Manning, 1999). A side effect of this model is faster running times due to lower computational complexity. In the simplest form of NB classifiers there are no parameters to tweak and performance is relatively good. Due to the lack of parameters that need tuning and many successful applications for text classification, for example in the area of SPAM detection (Sahami et al., 1998), NB classifiers make sense as a first step when using statistical modeling for classification.

The occurrence of terms is used to generate a probability distribution from which the likelihood of an instance is calculated based upon the combinations of terms in the instance. The use of word-grams helps to alleviate some of the independence problems associated with NB classifiers. Looking at the classification problem presented we can classify drugs based upon the way people talk about them.

Given a set of word-grams w in the aggregate messages mentioning a particular drug, the probability that the i-th word occurs from class C (the classes are watchlist or non-watchlist drug) is written as $p(w_i|C)$. Then the probability of a given drug D given a class C is $p(D|C) = \prod p(w_i|C)$. Because we have a binary class watchlist or not we can denote the class as W or $\neg W$ then the probability of $p(D|W) = \prod p(w_i|W)$. Due to Bayes' rule this can be written as $p(W|D) = p(W)/P(D) \prod p(w_i|W)$ and $p(\neg W|D) = p(\neg W)/P(D) \prod p(w_i|\neg W)$. In this way we build two probability models. These two probability models combined with a decision rule builds a classifier, we utilize maximum a posteriori (MAP) decision rule, which picks the most likely hypothesis.

2.8 Evaluation Metrics

This dissertation focuses on building classification models, and mapping instances of drugs into two classes/groups-in this case "watchlist drugs or non-watchlist drugs".

Because there are two classes the classification is binary in nature and we treat watchlist drugs as positive examples and non-watchlist drugs as negative ones. The prediction of instances based on these models results in labeling input instances as positive or negative. If the prediction is positive and it is a watchlist drug then this is a true positive; however, if it is a non-watchlist drug then it is a false positive. Similarly, if the prediction is negative and it is a non-watchlist drug then the result is a true negative and it is a false negative if the instance is a watchlist drug. True positives, true negatives, false positives and false negatives are the basis for many evaluation criterions.

Table 3: Table depicting the confusion matrix of classification outcomes.

		Actual Value		
		p	n	total
Prediction Outcome	p'	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	Total	P	N	P

For the experiments I perform, separate training and test data sets are derived from the pool of instances mentioned above. Within the separate training data, instances are withheld when training the model and the same data to train was not utilized to generate the results. These instances, though not used to build the final model are utilized for searching the parameter space for model building, and can lead to over-fitting.

When building machine learning classifiers it is assumed that an end state is reached where the classifier is able to predict the correct class for examples not presented during training due to inductive bias. However, over-fitting occurs when learning was performed too long over the instances, for example when searching for optimal

parameters in building models and there are few training examples, which is also the case for our data. The models are biased towards specific random features that have no causal relation to expected outcome.

After the parameter space (for SVMs) is searched and a final model is built the model is tested against the previously unseen testing data. Here we can derive separate scores for the withheld portion of training data and for the completely unseen testing data.

Given a 90/10 split where 90% of instances are one class and 10% are another, it is difficult to outperform a naïve classifier that marks all drugs as non-watchlist will achieve an accuracy rate of 90% if we compare the total number of positives (watchlist drugs) versus negative (non-watchlist) examples. While the accuracy ([True Positives + True Negatives]/[Positives + Negatives)]) rate appears high this is problematic due to the fact that the classifier will never find any watchlist drugs, thereby defeating the purpose of classification and predicting watchlist drugs. The number of true positives of a classifier is zero and the associated true positive rate, the number of true positives divided by the number of classified positives, is undefined.

The assumptions built into many machine learning classifiers is that the goal is to maximize accuracy and that the classifier will operate on data from the same distribution as the training data. If 90% of instances are of one class, the learning algorithm will be hard pressed to perform better than selecting the majority class. Looking at the data, this is the logical thing to do. However, this idea of merely labeling instances the same as the majority class leads to problems as discussed previously. Similarly it is logical to assume that real-world data will behave similarly to the test data insofar as the training data is

real-world data. Generally one relies on past behavior and experience to predict future behavior but this is not always accurate. However, accuracy is intuitive and it is easy to determine how accuracy scores changes based on an additional incorrectly classified example.

There are other methods of model evaluation besides accuracy. Receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC) have been employed previously to evaluation machine learning classification models and are employed in the evaluation of models in this dissertation. Receiver operating characteristic (ROC) curve and the associated area under the curve (AUC) are graphical plots and numerical measures, respectively. These metrics help to differentiate between and select optimal models. ROC curves have been used in the analysis of signal detection such as radar signals, evaluation of diagnostic tests, evaluation of radiology techniques, and in epidemiology and medical research (Green and Swets, 1966)(Zweig and Campbell, 1993)(Pepe, 2003).

ROC curves are plotted over the true positive and false positive rates lying within the space (0,0) to (1,1). It is the plot of sensitivity, the fraction of true positives on the y-axis versus the fraction of false positives or (1-specificity) on the x-axis. Here the True Positive Rate (TPR) = True Positives / Positives and the False Positive Rate (FPR) = False Positives / Negatives. The point (0,1) is considered perfect classification and the line from (0,0) to (0,1) is the line of no discrimination.

AUC is equal to the probability that a classifier will rank a positive instance higher than a negative one (Fawcett, 2006). AUC summarizes the ROC curve into a

single number. However, in doing so information about tradeoffs utilizing a particular model is lost. Yet it provides a convenient way to easily compare various models.

2.9 Health Related Quality of Life

Growing numbers of health measures include social and emotional aspects of health; these are generally considered quality of life measures, though there is no clear distinction between them and other general health status measures. Interest in quality of life is stimulated by the idea that merely surviving is not enough. Health based quality of life indicators were originally derived from those in the social sciences, focusing on people's feeling about their circumstances; many resemble indexes of emotion well-being and life satisfaction (McDowell, 2006).

The Centers for Disease Control (CDC) believes that health-related quality of life (HRQOL) is important in the measurement of effects of chronic illness on patients' lives. It is used with increasing frequency within the US as well as abroad. HRQOL is important in tracking patient's perceived physical and mental health over time and tracking the effects of multiple diseases and disabilities within patient populations. It is also important to weigh the advantages versus adverse effects of treatment of serious chronic illness (deHaes and van Knippenberg, 1985). The U.S. Food and Drug Administration recognized this and included quality of life as a primary criteria in the approval of anti-cancer therapies (Johnson and Temple, 1985).

HRQOL is a broad construct consisting of both objective and subjective measures (Chung and Pennebaker, 2005). Currently many self-reported metrics are used including: The Arthritis Impact Measurement Scales, The Physical and Mental Impairment-of-Function Evaluation, The Functional Assessment Inventory, The

Functional Living Index – Cancer, The Functional Assessment of Cancer Therapy, The EORTC Quality of Life Questionnaire (European Organization for Research and Treatment of Cancer), The Quality-Adjusted Time Without Symptoms and Toxicity Method, The COOP Charts for Primary Care Practices, The Functional Status Questionnaire, The Duke Health Profile, The Older Americans Resources and Services Multidimensional Functional Assessment Questionnaire, and The World Health Organization Quality of Life Scale (McDowell, 2006).

While many different HRQOL metrics exist and the appeal of quality of life is intuitive, the term still remains relatively undefined, which seems to more often reflect the personal values or academic orientation of the research rather than an objective attempt to define the concept (McDowell, 2006). Since most measures are self reported, the subjectivity is influenced by factors such as gender, age, social class, and culture; it becomes increasingly difficult to draw comparisons between different people's HRQOL rating (Stennar, Cooper, and Skevington, 1993). Jia et al. (2007) also notice differences between groups with differing sociodemographic variables and clinical factors; however, these differences were attributed to population health and not individuals. This work also enabled the comparison of two HRQOL metrics, which initially seem quite different but show similar patterns of population health (Jia et al., 2007); it is still unclear, however, if individuals would rank similarly on differing scales. Part of the problem stems from the endemic problem of how to measure health; measurements are influenced by the way we define and think about health (McDowell, 2006). What does the concept of health include and how does it relate to the quality of life and our well-being?

In this chapter we have discussed the foundational topics necessary to frame this dissertation. It is necessary to talk about pharmacovigilance – the study of drugs once released to market because this dissertation hopes to explore a new source of data and techniques for this field. However, one is not able to discuss drugs in relation to health without also discussing quality of life and the subjectivity which people rate it and rate drugs they take. These drugs or other treatments are talked about emotionally, and sentiment analysis enables us to quantitatively analyze the qualitative or descriptive review of drugs and treatments by people. Machine learning enables one to look for patterns in these qualitative reviews and compare them to other drugs.

3 Description of Data

This chapter describes the data this dissertation will utilize. Its goal is to inform the decision making process of choosing language processing tools and techniques. While this data is publically available it is realized that health data is extremely personal in nature. It is possible to personally identify people, from their messages, especially for rare and serious conditions where people have included information on gender, age or geographic locale. However, people choose to share as much or little information as they want with the understanding that the more they share the greater the potential for others to help them or for their information to help others. Companies such as PatientsLikeMe have built businesses around this model of open health sharing. The data utilized within this dissertation is used in aggregate with no particular person's messages analyzed in more detail than others. No attempt is made to identify the party, or linking messages with a profile or email address.

This chapter aims to look at the tokens within messages utilizing automated approaches to decrease the manual annotation time as well as bias and error accompanying it. Here I utilize techniques similar to Liu and Curran, (2006).

As stated previously the Yahoo! corpus contains approximately 12.5 million messages from various groups. Some meta-data about the messages are available such as the name of the group, inception date, if it is moderated, type of group and the language of the group. However much of this information is incorrect or has changed since the group was started. For example moderators may leave groups, the language of the group can change over time or the language utilized within the groups is not the one specified when the group was created.

The actual content of the messages is unknown as well as statistics about the messages. When developing tools to process the messages it is important to understand the data contained and the way in which it is represented. Many machine-learning algorithms are trained on newswire text, which is grammatically correct, with relatively few spelling errors or colloquial terms. Information such as average message length is also important, for example, to understand the amount of time it will take to process a message as well as how much memory each will take.

I developed a variety of lexicon to interpret the data and to give statistics about the composition of the messages. If tokens within messages do not occur within the lexicon the token is manually inspected. The inspection process looks at each token and evaluates them as a spelling error – differentiating between medical terms (including drug names) and general terms, compound words, slang or colloquial terms, abbreviations, garbage (long random nonsensical strings of characters), names, foreign language terms, web terms (artifacts of html stripping), or numbers (such as combinations of numbers and letters indicating numbers like 16lbs).

3.1 Yahoo Data Description

The Yahoo! health message corpus was randomly sampled for 500 messages. The messages were sampled utilizing a random number generator to choose message ids from the corpus. The text from the chosen messages were extracted and processed; if the message contained plain text and html, the html is preferred as html often contained formatting hints and other useful markup data. For this analysis, the html tags were removed. The extracted text was tokenized on white spaces with trailing punctuation,

common numerical tokens (money - $50, percent 10%, dates – 1/2/09), and tokens consisting of only punctuation such as emoticons: :) ;-) were removed.

The resulting tokens were compared to various lexicons to determine the language of the token, if it was spelled correctly and the type of token. Various lexicon were created, for example general English, foreign language, web/slang, medical terminology, disease, drug names and other names (people, locations, etc).

General language lexicons for English, Portuguese, Italian, Indonesian, Turkish (Kurdish), Filipino (Tagalog), Galacian, Spanish, Afrikaans, Malay, Polish, and Indonesian were utilized from the spell checking portion of the OpenOffice project[9].

Drug names were extracted from the Food and Drug Administration's (FDA) Drugs@FDA website[10], which "contains prescription and over-the-counter human drugs and therapeutic biologicals currently approved for sale in the United States. Drugs@FDA includes discontinued drugs and 'Chemical Type 6' approvals."[11] However, not all therapeutic biological products are in Drugs@FDA. Foreign names of drugs or drugs not approved for market in the US are also not included. Multi-word names were tokenized on white spaces resulting in 5,637 unique drug tokens.

A medical lexicon was generated from medical terms on MedicineNet[12], Inc which is owned and operated by WebMD[13] the authors of whom also wrote content for Webster's New World™ Medical Dictionary. Similarly, the terms from MedicineNet were tokenized and resulted in 11,132 unique tokens.

---

9 http://extensions.services.openoffice.org/dictionary

10 http://www.accessdata.fda.gov/Scripts/cder/DrugsatFDA/

11 http://www.fda.gov/Drugs/InformationOnDrugs/ucm075234.htm#contains

12 http://www.medicinenet.com

13 http://www.webmd.com/

Often forum posts contain names of people – real names as well as profile - or pseudonyms as well as group names; for example names are frequently found in salutations such as "Hi angelheart". A list of names was compiled from the from and to headers of the forum messages from the corpus. These are often in the form of <email address> or Name followed by <email address> as depicted below:

> FROM: <angelheart@domain.com>, John Smith <john.smith@domain.com>
>
> TO:<HigherGround@domain.com>

The names before email addresses were extracted as well as the non domain name portion (the text before the @ mark in an email address) were added to the names lexicon. In the example above angelheart, john, smith, john.smith, and higherground are added to the lexicon. In addition to these names two other sources were utilized: 1,000 most popular male and female (2,000 total) baby names from 2000 to 2008 from the social security website[14]; and the most frequently occurring first names (4,275 female and 1,219 male) from the 1990 census[15]. A total of 1,308,619 names was added to the lexicon from the three sources.

After the lexicons were created, the tokens from each message were tagged utilizing the provided lexicons. The tokens first were filtered by the English lexicons, web lexicon, followed by the medical terminology/disease ones, then drugs and if the token was not tagged as one of those it was tagged as an error for manual inspection.

Upon manual inspection of many numbers of what appeared as foreign language words, the use of foreign language lexicon were utilized to classify many of the terms as a foreign language. If a word was not a foreign language token then it was inspected and

[14] http://www.ssa.gov/OACT/babynames/decades/names2000s.html
[15] http://www.census.gov/genealogy/names/

classified as a specific type: abbreviation, spelling error – general and medical, compound word, slang, number, garbage, web, name or unknown.

The tokens were evaluated in several rounds some of which were easier than others. For example if a token looked like a number or quantity but was not removed by the regular expression, for example 1cc, 16wks, 65k, 5'3, 10px, the token was classified as a number. Similarly if the token was unusually long in length with no vowels, many consonants in a row, or combinations of numbers with letters that looked nonsensical it was classified as garbage; examples include: qv1fxvptv1jbqvhxqhrmufpcw1jdxuz erh9rxfk and jqwywptfo02381217r0.

Artifacts of html conversion and stripping identified and classified as web examples include malformed html tags where the text remains like colspan, href, or nbsp. Compound words were identified as two terms that were put together. They include typical spelling errors like alot or namebrand; artifact of message splitting and html formatting like saiddavid; and medical terms people might inadvertently combine: anticardiolipins, invitro, or overmedicated.

A slang classification was also created to indicate common colloquialism and online abbreviations such as: heh, suks, pushin, otoh, kn0w. Here intentional misspelling of a term, for example sucks to sux or suks, was classified as slang as well as intentional use of numbers or other characters to replace letters $ for s or 0 for o.

To determine the type of other errors such as, names and abbreviations, web searches were utilized on Google. Searches were performed using the word and then the word in conjunction with the search term “health”. If the predominant sense in both was a name – personal, object, location, etc even if it was in a foreign language - then it was

labeled as a name. Common types of names are listed below with examples from the messages.

Table 4: Tokens labeled as names broken down by categories.

People	Oprah
Websites	Xcitefun, Rocketmail
Screen name	Divine_mercy0901, Fixitman
Company	Makita, Glaxosmithkline, Toshiba
Location	Butang, Waltham
Software	PalTalkScene
Group/Organization	Ostomyland – ostomy support community
Drug	Capitata – from Buchenavia capitata
Medical Term/Procedure	Hepatocytes, Transpalatal

A Google search was performed and if the word looked like a common abbreviation (the top hits contained sites referring to it or it was contained in an abbreviation lookup site) it was classified as an abbreviation. Some of the abbreviations found in the messages include: afib - Atrial Fibrillation/Flutter, tnb - Trinitrobenzene Sulphonic acid, pws - Prader-Willi syndrome

If the token could not be classified as one of the other categories or was too ambiguous to determine what it was, then it was classified as unknown.

3.2 Pre-Processing

The statistics for the messages are shown in Table 5 below. Each message has an average of 172 tokens (words). This number is somewhat misleading due to the stripping of all punctuation (emoticons, for example) tokens, and numeric tokens and message lengths may be artificially inflated due to the signatures found in the bottom of many messages. Most noticeable is the relatively few number of errors for each message, errors only account for ~4% of the total tokens. The number of spelling mistakes is even smaller due to the fact that foreign language terms were counted as "errors" leading to

.8% spelling error rate, if only spelling errors, slang, compound and unknown words are taken into account.

Table 5: Total and average number of tokens and types of tokens found in analyzed messages.

	Tokens	Drugs	Errors	Names	Medical
Total	86,105	145	3,546	2,672	404
Average	172.21	.29	7.09	5.34	.81

The distribution of error types is listed below. The classification of error types was done manually except for foreign language tokens which utilized freely available lexicon from spelling checkers. After it was noted that many tokens appeared in different languages, the use of these lexicons was utilized to reduce classification time while also increasing accuracy due to the lack of knowledge of several languages.

Table 6: Tokens not in English, medical, drug, disease, or slang lexicons that were manually classified by type.

Error Type	Total	Unique
Foreign Language	1925	1213
Names	592	386
Spelling Errors	376	314
Compound Words	140	117
Slang	109	80
Abbreviations	107	47
Web	107	41
Unknown Words	92	76
Numbers	68	50
Garbage	11	11

Not surprisingly the second most abundant type of "error" was names because dictionaries cannot capture names of all people, places, or products. The total of column two from Table 2 does not equal the total number of errors listed in table one because the dictionary utilized was not comprehensive and missed some common English terms that

should have been in it or common variations such as British spelling on terms like favour or colour.

The incidence of medical terminology misspellings seems somewhat higher than that of normal words. Medical terminology makes up approximately .637% of the total number of terms within the text, yet makes up approximately 14% of the total spelling mistakes (not including compound words). This seems somewhat intuitive since medical terminology is often used less often than other terms and people are less familiar with it. Similarly, spell checkers on computers often lack medical dictionaries containing terminology or drugs.

Table 7: Spelling error breakdown, medical vs. non-medical terms.

	Total	Unique
Medical Terms	53	48
Non-Medical Terms	323	267

Upon inspection several messages are in a foreign language, including Indonesian and Spanish. It is not known whether the message board started in English and progressed towards the other language or if the board was mis-labeled as English speaking. Terms in other languages are used as slang or as a term of endearment, "my beautiful lochie….your sweet little lochie," and often are interspersed within English messages. Some messages are in English but include foreign language signature blocks at the end or the message text is written in a foreign language with a news release or other form of information in English in lined within the message. The distribution of languages is listed below. However, this may not be an accurate representation. Many words exist in multiple languages, and a single term within another language may indicate that the term was not in the true language's dictionary but denoted as another

language, ie. a single word in French, Norwegian or Pampanga. For example a message contained the word "mahu", which exists in Hawaiian but was not included in the Indonesian dictionary, but upon further inspection the entire message where mahu was found was in Indonesian. Foreign language lexicons, like English, are not comprehensive and do not contain new terminology or slang which makes the identification of language difficult.

Table 8: Breakdown of foreign language tokens by language.

Language	Total	Unique
Afrikaans	3	3
Arabic	3	2
Croatian	4	4
Czech	4	3
Dutch	48	38
Finnish	21	20
French	1	1
Galician	12	9
German	4	4
Icelandic	2	1
Indonesian	1085	604
Italian	50	48
Macedonian	3	3
Malay	54	40
Norwegian	1	1
Pampanga	1	1
Persian	4	1
Polish	56	47
Portuguese	142	114
Russian	1	1
Serbian	6	5
Spanish	179	138
Swedish	34	30
Tagalog	49	30
Turkish (Kurdish)	59	47
Unknown	99	18

The results of this study demonstrate that spelling errors are not as problematic as previously thought, this presumably in part due to the prevalence of spell checking software in email clients and increased amount of medical terminology misspelled. Foreign language is by far more problematic and a worthwhile goal is to look at automated foreign language classifiers. Not surprisingly new names or previously unseen tokens are a challenging problem.

The task of Named Entity Recognition relies on identifying tokens as certain entities. For this dissertation I focus on drugs and drug effects. Drugs are a relatively closed class of nominals; however, as evidenced by this experiment many foreign and non-FDA approved as well as non-drug labeled biologicals such as herbs or other chemicals are not available in a comprehensive list. It remains to be seen if dictionary based approaches are good for the identification of drug outcomes as this is a more open bounded problem and people may use slang terms such as one labeled in our data, "itchers." However, the implications of this work demonstrate that spelling problems are not as problematic as I previously thought, comprising a relatively small amount of total tokens, indicating that dictionary identification approaches are feasible leading to high precision with lower recall. While this is not ideal, high precision approaches are useful for certain tasks.

I have demonstrated corpus statistics of a small sample of the Yahoo! health corpus that will be used in the rest of the dissertation. Here I have developed semi-automated techniques for manual annotation of statistics regarding the prevalence of names, spelling errors and foreign language tokens. While errors are not as prevalent as I had previously believed this does highlight the open ended nature of NER and the

difficulty in creating a pre-calculated dictionary of proper nouns including people and companies.

3.3 General Vocabulary

While all words in messages concerning drugs can be utilized in feature vector generation, this is not necessary and can lead to poorer results due to added noise as well as slower running times due to additional computational overhead. For instance the running time to train linear SVMs is dependent on the number of training instances (n) times the number of features k leading to a running time of O(kn). Many text classification tasks utilize feature vectors of single words taking the "bag of words" approach where the ordering of words does not matter. However, it is intuitive that the ordering of words is important; for instance the series of three words - good, bad, not - have very different meanings depending on their orderings "good not bad" is different from "bad not good." Additionally, certain drugs or medical procedures are composed of multiple words where the meaning of the phrase is different from its constituent components, for example "vitamin a" is different from "vitamin" and "a" occurring within a message. Here word-grams can capture the importance of certain word orderings. For our text classification task we utilize unigrams, bi-grams and tri-grams.

A common approach utilizes the top k-most frequently occurring words. However if no initial stop word removal was performed many of the most frequently occurring words are ones that contain little information such as the, and, a, or. Luhn's (1957) work proposed that the most informative words in a work are the mid-frequently occurring ones. In this dissertation I utilize this feature selection approach. While other feature selection approaches exist such as Information Gain (IG) or Bi-Normal

Separation (BNS) these utilize binary counts, whether features exist or not and if they exist within one class versus another. These feature selection criterions would penalize terms occurring within both classes, including many of the specialized medical terminology, drugs names, diseases, or sentiment words. Instead we would like to see these words included but look at the differences in number of occurrences in vocabulary between the two groups. We take the approach of taking the top k-n most frequently occurring terms where k is the top number of terms minus n to account for function words like a, or, and the.

3.4 Specialized Lexicon

An alternative to using specialized lexicon is taking a subset of the general vocabulary and performing feature selection on it. In this case the feature selection criterion is world knowledge preferencing words that are more related to the specific health domain. I aim to utilize combinations of the set of lexicon since using a small subset of the entire lexicon can lead to over-fitting. For example the drug lexicon is relatively small and each example will have specific drug mentions over represented (the instance's drug). The presence of the watchlist drug will then map to a drug being classified as watchlist or not. While it can have a high degree of accuracy for the training set instances it is not generalizable to unseen drugs. As mentioned before the specialized lexicons used are: drug lists from drugs.com, medical terminology from MedicineNet, Sentiment Lexicon from SentiWordNet and LIWC, MedDRA lexicon from AERS reports and disease lists from Wikipedia. From the five different lexicons we can generate different feature sets combining various combinations of the lexicon. This leads to 29 different datasets with each lexicon a separate dataset, pairs of lexicons, and so on and

the last including all of the combined lexicon terms. The feature vector is then the subset of the generalized lexicon terms contained within the combinations of specialized lexicon; for example if a two combination consisting of sentiment and drug lexicon, the resulting feature vector dimensions would consist of all drug and sentiment terms.

4 Data Preprocessing

4.1 Language Identification

Due to the significant amounts of messages in foreign languages, it is necessary to identify and remove these messages from the corpus. This chapter focuses on techniques for processing the Yahoo! data to reduce the noise and improve classification performance. A major focus of this preprocessing is foreign language identification and removal of those messages. While the text processing techniques are language agnostic, messages in foreign languages would increase the size of feature vectors for each drug in the machine learning experiments. Larger feature vectors increase the amount of storage space and processing requirements and possibly negatively affect classification if sufficient numbers of messages for each language do not occur (data sparsity problems).

Foreign language terms are also problematic for visualizations where the projected audience is English speaking. Drug comparisons depicting features that indicate similarity in languages other than English are difficult for English-only speakers to interpret and evaluate.

An initial first pass over the data is performed utilizing Unicode language detection to remove messages that contain non-romanized text. For example messages written in Japanese alphabets or Kanji, Chinese characters, Arabic abjad, or Cyrillic were eliminated.

Differentiating between English and non-English messages written in Romanized alphabets is more difficult. A variety of methods have been used to identify different languages; among them are the "common words" approach, character n-grams and dictionary based approaches. The common words approach utilizes function or stop

words in a language; for example in English some words include: a, and, be, but, than, the, you (Ingle, 1976). These words carry little information but make up large portions of the content in written text. Later work by Dunning (1994) utilized character n-grams or series of characters (20 characters in one example: e pruebas bioquimica; man immunodeficiency; faits se sont produi) based upon the observation that humans need little amounts of text to correctly identify a language. Dunning's approach requires the use of training data (albeit small amounts – several kilobytes worth) for statistical classifiers. Later work by Rehurek and Kolkus (2009) utilizes dictionaries derived from nine European Wikipedia datasets. Dunning (1994) dismisses the use of dictionaries, however the dictionaries discussed only consist of function word lists. Rehurek and Kolkus (2009) determined 97.16% accuracy on 1,000 documents and 49,943 words.

Rehurek and Kolkus (2009) the authors go on to discuss the advantages of dictionary word based language modeling versus character n-grams. Character n-grams are not reliable when building statistics of web corpus documents that include the character distribution is skewed towards repetitious words or phrases such as "In reply to:" in the case of our Yahoo! Health message corpus. Many web pages (and our input data as well) consist of multiple languages due to logical structure (Yahoo! Signatures at the end of a message) but also the nature of the body text itself.

> dear all,
> semalam 'terlepak' sama ustaz abdul rahman (ABU). dia ada cerita sikit2 about his 'struggle', hmm, i don't actually know where to start! hahaha... anyway, dia ada plans to establish an orphanage in tawau - he went there a few weeks ago, on his friend's invitation, according to his friend, in tawau tak ada orphanage. dia ada plans to go there lagi sekali, this time nak survey tempat, etc. risma, if you're free, maybe you can meet-up with this ustaz or mintak tolong your brother, jusri, hehehe (kawan baik & housemate john masa study kat upm!). maybe can bring him around, etc.

Figure 2: Example of a multilingual message from the Yahoo! corpus.

Character n-grams also have difficulty in dealing with similar languages. For example Slavic languages such as Slovenian, Slovakian, Czech and Polish have significant grammatical and lexical overlap.

Due to the web nature of our input data, I used dictionary-based methods of language identification; the goal of this identification is a simpler problem than traditional language identification. We have a binary problem, determining if a message is predominantly English or not. While the identification task is simpler in some regard, in others it is more difficult. As observed previously the messages are a mix of English and foreign words, a lot of names, medical terminology, and other erroneous tokens such as ascii are often found in signatures, emoticons or punctuation based delimiters. Luckily spelling errors are not as frequent as previously thought; however the unseen word problem is significant. Rehurek and Kolkus (2009) method relies on building language models utilizing large amounts of text (several gigabytes for each language). Such training data does not exist for foreign language (and English) health message board text.

Instead I used dictionaries of commonly found words for various languages available through the OpenOffice project, an open source freely distributable office software suite (www.openoffice.org). In addition medical dictionaries, drug and disease lexicon, as well as lists of names as described previously are utilized. The foreign language dictionaries contain many more words than just function words. While they are not as extensive as language modeling based approaches containing domain specific or esoteric words, they capture many words found in everyday use. Instead of using a statistical scoring function, an inequality of two linear combinations of word counts was utilized:

Equation 2: Linear inequality for scoring a message as English or not.

$$4*\text{foreign} + \text{unknown} + \text{ignore} > \text{english} + \text{drugs} + \text{medical}$$

The constant in front of the number of foreign tokens is greater than one, increasing the amount a foreign token counts for, indicating that fewer foreign tokens are needed than English ones to count the message as non-English – an English message should contain approximately < 25% foreign words. Each token or word in the message is compared to a set of dictionaries to derive word type counts. The word counts are calculated in the following way:

```
If ((word in Ignore List) OR (word length < 2) OR (word contains "@"))
        Ignore Count++
Else if (word in English List)
        English Count++
Else if (word in Drug List)
        Drug Count++
Else if (word in Medical List)
        Medical Count++
Else if (word in Name List)
        Name Count++
Else if (word in Foreign List)
        Foreign Count++
Else
        Unknown Count++
```

Figure 3: Pseudo code for scoring a word token.

Pre-processing of tokens incurs removing web addresses (URLs) and punctuation only tokens (such as emoticons like ;) or :-( ). Short words, email addresses, or tokens that are on an ignore list are filtered out. Then each token is progressively compared to another dictionary. The counting function is conservative erring on the side of English rather than foreign words, tagging a token as English rather than foreign if it occurs in both lists. This is based on the assumption that having a foreign language message is better than removing an English language message. Unknown counts include spelling errors, names, slang, expletives or other words not in any list.

However allowing some foreign language words can lead to problems with short message text in a foreign language followed by a longer English signature block like the message below:

```
Pidu toimub ja Liisa koju ei l?he :)
On , Dmitri Somov <dimsgen@ > wrote:
ei tea mis ?ldse tuleb, aga kolmanda kursuse tudengid on oma ?ppemisega
nii mures, et mingi pidu ei mahu nende plaanidesse:(
---------- Yahoo! Groups Sponsor ----------~-->
Affected by disease? Support health awareness efforts at Network for Good.
----------~->
Yahoo! Groups Links
<*> To visit your group on the web, go to:
<*> To unsubscribe from this group, send an email to:
ehyljuhatus-unsubscribe@
<*> Your use of Yahoo! Groups is subject to:>
```

Figure 4: Example of a non-English message from the Yahoo! corpus containing an English signature block.

Utilizing the sampled original text from the previous section, the linear inequality and coefficients were optimized. This resulted in a smaller corpus consisting of 10,178,710 messages with 20,417,209 unique terms from the original 12,520,438 messages with 31,124,146 terms reducing the number of unique terms per a message from ~2.5 to ~2.

This chapter demonstrates the data preprocessing and cleansing techniques employed on the data used in this dissertation. The number of messages has been substantially reduced although it is still unknown how much noise is still present and to what extent it will affect classification performance; however it is possible to extrapolate these numbers based upon the sampling performed in the last chapter.

5 Experimental Design

This dissertation focuses on generating a classification system for drugs based upon the way people talk about them. The diagram below depicts the architecture of the system. Here a set of cleaned messages from the Yahoo! corpus was organized by the drug mentions in them. For safety alert or watchlist drugs, only messages up to the date the safety alert was released were used. I aim to build a predictive classifier and want to determine if the way people talk about safety alert drugs are different from non safety alert ones.

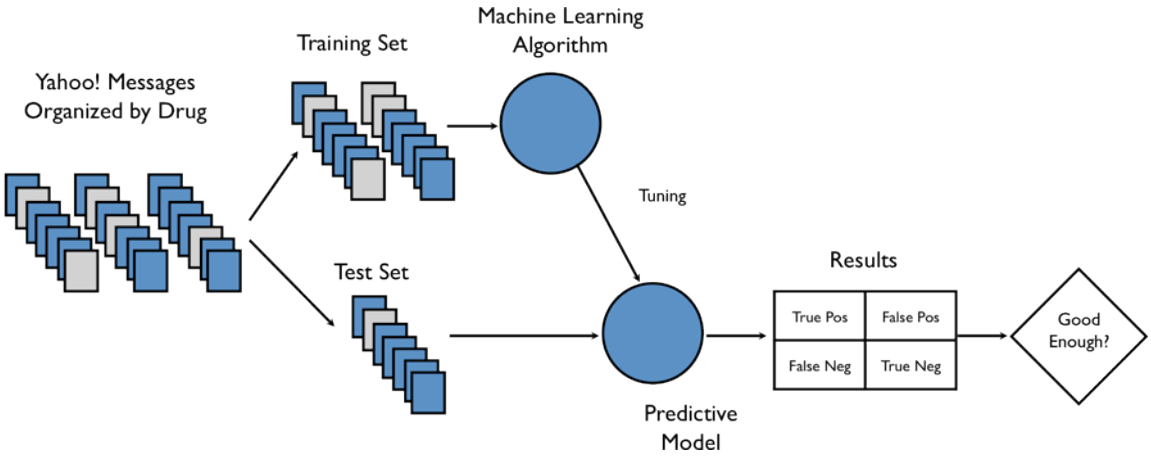
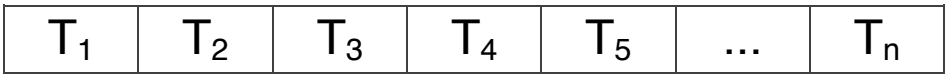


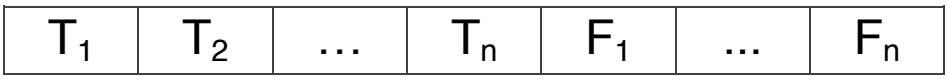
Figure 5: Depiction of the iterative process involved in building a machine learning classifier.

Messages containing drug mentions are divided into two sets, a testing and training set. These messages are then converted into feature vectors, which are run through classification algorithms, in this case Naive Bayesian, and Support Vector Machines. The parameters for each algorithm are tuned and the results against the test set are generated to rate their performance.

This dissertation uses two different feature vector compositions as illustrated in the diagram below:



Word Features



Word and Special Features

Figure 6: Depiction of the feature vectors used by the classification algorithms.

The first type of feature vector consists entirely of words or phrases. Feature vectors do not preserve any of the word ordering that exists in a message. To incorporate some of the ordering word phrases (N-Grams) are utilized. Word ordering is especially important for things like negations; "not good" is very different from "not" and "good". These features consist of words selected from the message corpus themselves that tend to differentiate the classes of messages, for example, preferencing words pairs that exist in watchlist messages but not in non-watchlist ones. Other types of word features include specialty lexicons such as the AERS lexicons, disease names from Wikipedia or drug names from drugs.com. By selecting these lexicon a priori we are imparting our knowledge and intuition that these words are good discriminators between the two classes.

The second type of feature vector combines word features as previously discussed along with "specialty" features. These specialty features consist of counts over the number of instances of specialty lexicon; for example, number of disease mentions, number of drug mentions, number of positive sentiment containing words, and number of negative sentiment containing words.

The identification of drug names and sentiment containing words rely on Named Entity Recognition (NER) of drug names and sentiment analysis tracking. Both approaches are dictionary based but raise validation concerns. Can we reliably extract these entities and are they meaningful? The next sections describe experiments detailing dictionary NER techniques for drug and drug effect detection followed by experiments on tracking sentiment.

However before any machine learning experiments or NER and sentiment tracking experiments are run, one must determine if the two classes are separable by examining the language use of watchlist messages versus ones that are non-watchlist.

5.1 Kullback–Leibler Divergence

As an initial experiment I first looked at the word distributions of the watchlist drug messages and the non-watchlist drug messages. If there is a large difference between the distributions it is indicative that people talk about these two classes of drugs in differing ways. If people talk about the watchlist drugs differently from non-watchlist drugs it demonstrates the tractability of utilizing machine learning classification to separate the two classes of drugs. If there is a metric of distance or difference it should also indicate the ease or difficulty in separating the two classes and building a predictive classifier. In this dissertation I utilize Kullback–Leibler divergence (KL divergence) to quantify the difference in word frequency distributions between watchlist and non-watchlist drug containing messages. KL divergence is a non-symmetric measure of the difference between two probability distributions P and Q. It is often used in the area of information theory and probability theory. While often alluded to as a distance metric, it is not a true distance metric since it does not satisfy the triangle inequality in that the

"distance" from P to Q is different from the "distance" from Q to P. In information theory, KL divergence is the expected number of extra bits needed to encode samples from P when using a code based on Q, rather than using a code based on P. Formally KL divergence is specified as:

Equation 3: Kullback-Leibler Divergence

$$D_{KL}(P \parallel Q) = \sum_{i} P(i) \log \frac{P(i)}{Q(i)}$$

Within the context of this dissertation I create a probability distribution from a frequency distribution over words utilizing smoothing. Smoothing is necessary because when utilizing KL divergence if for some *i* (in this case *i* represents a word or phrase) exists in P but not in Q this leads to an infinite KL divergence indicating the distribution P predicts an event that is possible while Q predicts it is impossible. This is intuitively incorrect, though a word *i* not existing in one distribution does not mean it is impossible in the other, it was simply not present in the data the distribution was generated over and may not have been seen yet. So a small probability is ascribed to each value in one distribution that is not present in the other.

While KL divergence gives us a measure, it is somewhat meaningless without a point of reference. I compared the KL divergence between watchlist and non watchlist drug containing messages to the Google Web 1T 5-gram corpus and the Reuters Corpus. The Google corpus is composed of English word n-grams ranging in size from one to five words. The corpus was generated from approximately 1 trillion words from publicly accessible web pages. However, I utilized only the single words and compared them to the single words in the drug corpus due to memory constraints. The Reuters corpus

consists of 810,000 Reuters English language news stories from August 20, 1997 to August 19, 1997.

The Reuters corpus is more formal in nature leading to fewer spelling mistakes, slang, swear words, etc. The Google corpus, however, is rife with spelling mistakes and swear words and has many more brand names, slang, proper nouns, etc. Due to the colloquial nature of the Yahoo! messages I hypothesize that the KL divergence between them and the Google corpus is smaller than the KL divergence between the Yahoo! messages and the Reuters corpus.

5.2 Dictionary NER Exploration

The next two sections detail special features; drug entities and sentiment scores that will be employed by the machine learning classifiers. These features are different from those found in general text classification tasks. However, it is important to understand these features individually, both how easy they are to identify as well as how accurate they are and what information they contribute.

This section details experiments of dictionary named entity recognition from the Yahoo! corpus. A dictionary approach is employed due to the lack of training data as well as other NLP tools that are necessary if one uses a statistical classification approach.

The drug lexicon used to identify drug named entities is derived from the FDA and the drug taxonomy available at Drugs.com. The taxonomy enables the grouping of name brand and generic drugs by function. The following is a branch from the taxonomy:

Central Nervous System Agents > Antiemetic/antivertigo Agents > 5HT3 Receptor Antagonists

Figure 7: Branch of a drug taxonomy from Drugs.com.

The taxonomic structure enables the ability to group drugs in a semantically meaningful way. Grouping drugs could be used to look for trends within a class, for example generalizing and hypothesizing that a group of drugs might exhibit a particular adverse reaction if it is found that all sub classes of drugs exhibit the same adverse reaction. Taxonomic structure also facilitates the analysis of drugs within the same class and chemical structure if there is not enough data – it can help to alleviate the data poverty problem.

The Adverse Event Reporting System (AERS) uses the Medical Dictionary for Regulatory Activities (MedDRA). It is a lexicon that is applicable to all phases of drug development excluding animal toxicology and includes the effects, operation and malfunction of medical devices. It is used to report adverse event data from clinical trials and for post-marketing reports and pharmacovigilance. MedDRA was developed by the International Conference on Harmonisation (ICH) and is owned by the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA). [16]

A Lucene (an Open Source Java Information Retrieval package[17]) index was created over the approximately 10 million (12.5 million originally with 2.5 million messages removed) messages. The AERS and drug lexicons were then used to generate phrase searches, which were run against the Lucene index. The index is a structure enabling fast retrieval of message text and ease of locating relevant information. There are drawbacks to using an information retrieval (IR) system to look for instances of drugs or medical terminology. I currently use a lowercase filter with stemming; this creates a lower case representation of a term within the inverted index: “Commit”, “commit”,

[16] Taken from: http://www.meddramsso.com/public_about_meddra.asp
[17] Available at: http://lucene.apache.org/java/docs/

"COMMit" are all mapped to the same lexical entry "commit". Similarly "committing" and "commit" are also mapped to the same lexical entry "commit". These two filters have implications in precision and recall for IR.

Precision is the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search, and recall is the number of relevant documents retrieved by a search divided by the total number of existing relevant documents. While precision and recall do not necessarily have to be trade offs, ie optimizing for precision at the expense of recall or vice versa, in practice this is generally the case. Utilizing a lowercase filter will generally retrieve more documents leading to (generally) greater recall, whereas without it will generally lead to greater precision. For example, if someone wanted to retrieve instances of "COMMit" since it is a brand name a lower case filter will return all instances of "commit" leading to many documents not referring to the item but the verb (for example, the act of giving trust or act of consigning someone) leading to a lower precision. On the other hand the lowercase filter would enable the retrieval of all instances of "commit" including instances where the noun is not properly capitalized such as "COMmit" leading to greater recall. This is important because some drug names consist of common words that are capitalized. Some drugs in the lexicon that utilize common words include: Commit, Control, Duration, Perfect Choice, Sleep, Hold, SF, RID, Maternity, Bright Beginnings, and Definity. "Definity" is not a common word and therefore people commonly mistype "definity" as "definitely". This leads to the second problem with using IR for NER, the lack of contextual clues or other information such as part of speech (POS) tags.

Utilizing a dictionary approach does not make use of contextual information,

though it leverages the inverted index nature of an IR system matching index terms to lexicon terms. However, an IR system could be augmented to use context such as fuzzy searches requiring "drug" or "taking" or other keywords indicating a drug along with the drug name. Other contextual information such as POS tags might be useful at differentiating drug names. In the case of the drug "Commit" verb instances could be ignored such as "commit a crime". These rules, however, are not absolute, since the text is colloquial in nature, instead of the fragment, "starting to take Commit" one might instead say, "starting Commit" dropping the verb phrase "to take". Similarly people in colloquial text are not careful about proper capitalization.

For the purpose of this dissertation and the following experiments, drug names with common words were replaced by the generic (chemical) name. For each of the lexicons phrase searches were constructed and run against the index.

5.3 Drug Mentions in Messages

I performed an analysis on messages looking at the distance between unique drug messages with more than one drug mention. The goal is to determine the separation between drug mentions in messages. If drugs are close together it is harder to segment the message around a drug to determine the sentiment towards a particular drug as well as attributing a particular effect to drug. Given prior work it is expected that drug mentions would be close together. For example, people often list drugs they are on – their current regimen - or ask for opinions between two drugs of the same class to determine which is better or how to differentiate between the two.

To confirm this hypothesis pairs of unique drug mentions within the same messages were extracted and the distance (number of characters) between the first

mention was calculated. This method has some drawbacks such as messages containing more than two unique drugs; to eliminate some of this I only look at messages with two mentions. Looking at messages where there are frequent co-occurrences between different drugs highlights how the drugs are talked about. The variance in distance illustrates that drugs are usually talked about in the same way. Similarly, a large difference between the median and mean is also indicative of the variability of the data.

```
| for each N pairs of highly overlaping drugs
|   for each message containing drug1 and drug2 mentions:
|   |       find first mention of drug 1
|   |       find first mention of drug 2
|   |       distance = absolute value of (drug 1 first mention – drug 2 first mention)
|   | end
|   calculate mean, standard deviation, median, minimum and maximum distances
| return statistics for each N
```

Figure 8: Pseudo code describing how average distance between drug mentions is calculated.

Mutual information is used to score pairs of unique drugs within the same message to look for highly co-occuring drugs. Mutual information measures the information that two variables X and Y share. It quantifies the extent of knowing one of the variables X or Y reduces uncertainty about the other. If X and Y are independent, then information about X does not provide any information about Y and vice versa, therefore their mutual information is zero. If X and Y are identical, information about X is also applicable to Y.

Equation 4: Formula for calculating Mutual Information between variables X and Y.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \, log \left(\frac{p(x,y)}{p_1(x) \, p_2(y)} \right)$$

Mutual information was used as a scoring function instead of cosine similarity. Cosine similarity, which is the cosine between two vectors A and B, where the vectors represent drugs and the dimensions are messages the drugs occur in. As seen by the formula below, cosine is the intersection of vectors A and B divided by the size of the two vectors multiplied together. Cosine biases towards exact matching of vectors, such that if two drugs only occur in one message each but occur together this would rank higher than drugs that each occur in 100 messages each and occur together 90 times.

Equation 5: Formula for calculating the Cos between vectors A and B.

$$\cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|}$$

5.4 Measuring Sentiment Tracking Drug Outcomes

5.4.1 Introduction

Sentiment is the next type of special feature. Two dimensions are added to the word feature vector, number of positive sentiment containing terms and the number of negative sentiment containing terms. I equate drug opinions within forum messages to product reviews. However, the question arises, if a message contains a drug and also sentiment containing terms, does the sentiment reflect the author's opinion of the drug? Due to the lack of labeled training data, I instead look at aggregate sentiment of messages with drug mentions over time. It is then possible to look at the changes in sentiment to see if they correspond to significant news items such as FDA announcements or manufacturer announcements about the drug. Ideally drug mentions should be statistically different from messages without drug mentions utilizing a sentiment scoring metric.

I believe that sentiment analysis, determining the positive or negative valence, is another way in which we can determine drug satisfaction. A drug may have many serious side effects, yet people may still have a positive attitude towards it, especially if they believe it helps them in some way (Silver, 2002).

Opinion mining and sentiment analysis is a commonly accepted area within natural language processing. It is commonly used to aggregate and evaluate product reviews. In this dissertation I consider drugs analogous to other products. We value others' opinions in our own decision-making processes for purchasing products (Pang and Lee, 2008). Widespread dissatisfaction with a drug is not only alarming to others potentially taking the drug but also to oversight organizations such as the FDA. Widespread unease may be indicative of problems associated with the drug, ranging from adverse effects to concerns over pricing or efficacy; it can indicate to others that further investigation into the drug is warranted.

5.4.2 Methods Using Personal Health Messages

Within the Yahoo! forums, patients seek others' support, advice and information about treatment options. Below is an excerpt from a message from a neurological group that has been anonymized. I interpret these responses that a patient writes as a "review." Traditional sentiment analysis techniques are applied to these texts.

```
…when I gave up coffee and sugar in earnest and stopped the amitrptyline I
was taking I am feeling much better still especially depression wise and the
heavyness and sluggishnes that was in my legs is leaving. I am also getting
back into a more normal sleep pattern of getting sleepy by normal time in
evening and waking up in the morning more normally.  I believe the years of
amitriptyline for muscle relaxant were doing more damage than good and am
doing pretty good controlling my night time bladder spasms by no coffee,
magnesium and the diet…
```

Figure 9: Excerpt of a message from a neurological group from the Yahoo! corpus.

My method utilizes portions of the lexicon in the Linguistic Inquiry and Word Count (LIWC) when calculating sentiment scores for messages (Pennebaker, Francis and Booth, 2007). Previous work utilized LIWC to demonstrate variations in language usage between depressed and depression-vulnerable students (Rude, Gortner and Pennebaker, 2004). Further, it is known that the words people use correlate with their physical and mental health (Pennebaker and Campbell, 2000).

Specifically, I use the words in LIWC corresponding to the following categories: positive emotion, negative emotion, anxiety, anger and sadness. I have augmented the LIWC lexicon to include a wide range of emoticons such as :), :(, :P, ^__^ , LOL ROFL.

The messages from the Yahoo groups were parsed to extract just the textual information and to remove noise such as replies that are often included in messages. While replies help understand the context, a message's emotional context should not be based on what other people write, only on the author's text.

The resulting messages were matched against the LIWC lexicon categories and emoticons discussed previously. Counts containing number of positive emotion words, and negative ones, and total number of words were recorded. It was found that the ratio of negative emotion words to total words was the most helpful in determining negative valence. The following methods and results use a negative ratio of negative emotion words to total words in a message so that when graphing results, the positive y-axis represents less negative messages.

5.5 Machine Learning Experiments

The crux of this dissertation depends on the assumption that people talk about FDA watchlist drugs differently from non-watchlist drugs. While the difference might be

difficult for humans to readily discern, I depend on machine learning algorithms to differentiate the two. The inputs into these algorithms are feature vectors generated over the words people use to talk about these drugs. Feature selection is an important part of all machine-learning tasks. The goal is to utilize sufficient numbers of features to enable an algorithm to differentiate between instances both in the training set as well as unforeseen instances while limiting the amount of noise introduced.

This dissertation focuses on generating features using two approaches. In the first approach the feature vector is generated over general vocabulary terms and "meta features". In this approach all terms within messages are considered and then filtered using heuristics such as frequency cutoffs. These general terms are combined with other "meta features" these meta-features impart world knowledge in the form of counts over specialized lexicons for example the number of drug mentions or positive or negative sentiment words. The second approach focuses on using the specialized lexicons exclusively. An example feature vector might consist of only medical terminology, drugs, diseases, and sentiment containing words or some subset of them. The specialized lexicons include drugs, medical terminology, sentiment, adverse drug event lexicon from MedDRA, and lists of diseases as mentioned previously.

Classification utilizing machine learning algorithms typically require large amounts of training data, and the performance of classifiers are often commensurate with the amount of training data available. However, due to the nature of the data available relatively few positive (watchlist drugs) are available with sufficient amount of data. There are only 435 drugs with more than 500 unique messages mentioning them; of these there are 63 watchlist drugs. Similarly if drugs with more than 250 unique mentions are

used, there are only 575 drugs and 77 watchlist drugs. Approximately 90% of instances are non-watchlist drugs. This is somewhat comforting in terms of health and drug safety; only 10% of drugs are watchlist and are demonstrated to possibly cause adverse effects. However, in terms of machine learning experiments this leads to problems with bias and data scarcity. This is problematic due to the few numbers of examples a classifier can learn and generalize over.

To help alleviate some of the data sparsity problems, cross-fold validation is utilized. Multiple runs utilize the datasets for comparison: watchlist drugs versus non-watchlist drugs. In stratified K-fold cross-validation, the original data is partitioned into K subsections (in this case 10). One of the subsections is withheld as testing data and the other K-1 ones are used for training. The cross validation process is then repeated K times (folds). The folds are selected such that the same proportion of positive and negative classes is chosen. The K results are then averaged to produce a single estimate. All observations are used for training and testing and each one is used for testing once (McLachlan, Do and Ambrose, 2004). Cost weighting is another technique that is used. Two common forms of cost weighting exist: weighting instances or introducing penalties for misclassification. Instance weighting involves weighting instances of a particular class more than others, for example weighting or counting positive instances double what a negative one counts for. Misclassification penalties involve training a model and classifying instances. Instances misclassified by the model depending if they are positive or negative impact the updating of the model in different ways.

Instead of weighting instances one may build balanced datasets, where the number of positive examples is near the number of negative examples using sampling

techniques. Machine learning classification tasks frequently have many training examples available and usually in the binary classification application, approximately equal numbers of positive and negative instances. The dataset consists of only 13.4% positive instances. This leads to an imbalanced dataset where most examples are negative. A variety of sampling techniques are used to generate a balanced dataset, among these are: random undersampling (RUS), random oversampling (ROS), ons-sided selection (OSS), cluster-based oversampling (CBOS), Wilson's editing (WE), SMOTE(SM) and borderline-SMOTE (BSM), see Van Hulse, Khoshgoftaar, and Napolitano (2007) for an overview of various methods. This dissertation focuses on RUS in which instances of the majority class are randomly discarded. RUS resulted in the best performance in empirical tests of 2340 datasets (Van Hulse, Khoshgoftaar and Napolitano, 2007). However in the same work it was demonstrated that sampling does not significantly improve the area under the curve (AUC which is discussed in more detail below) for Naïve Bayesian classifiers and the performance measurement (AUC, true positive rate, accuracy) dictates the performance of the sampling method.

5.5.1 Decision Making

The building of machine learning classifiers depends on making many choices from a broad decision space as discussed previously. Aside from choosing specific algorithms, in this case SVM and Naïve Bayes, there are also associated parameter space decisions for SVM such as the type of kernel and the parameters for each kernel. Furthermore, there are feature selection problems for choosing the types of features as input into the classification algorithms; the features in this case are word grams.

Another consideration when dealing with feature selection and associated feature vectors is to utilize scaling of individual dimensions. One of the advantages of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. One can imagine that common dimensions such as words like "drug" or "pain" might then dominate the classification and overwhelm less frequently occurring but more informative dimensions like "dying". A common approach is to use term frequency-inverse document frequency (tf-idf) to weight a word according to its importance based upon the number of appearances; see Manning (1999) for an in depth discussion on tf-idf weighting. Scaling further helps to avoid numerical difficulties during calculations when using SVMs because kernel values usually depend on the inner products of feature vectors, and large attribute values can cause numerical problems such as number overflows.

Yet another choice is in how to divide the testing and training data. Many times one wants to use as many examples as possible to train a model and 90/10 splits are common where 90% of the total number of instances are used for training a model and 10% are used for testing. However, in this case if only 10% of the instances are utilized for testing, this leaves 58 testing instances or 16 testing instances with a balanced test set.

For these experiments I chose to utilize drugs with more than 250 messages, leading to an overall example set of 575 drugs of which, 77 are watchlist drugs. Separating examples into two different datasets, one with 500 or more messages and between 250 and 500 messages, leads to the question of generalizability. Will a classifier trained on more data (500 messages/drug) work on instances with sparser data (250 messages/drug) or vice versa, and would the results be fair? Instead both types of

instances are mixed and a subset of the total is used for testing while the rest is used for training.

5.5.2 Training and Testing Dataset Size Experiment

An initial experiment aimed at satisfying some of the dataset construction questions was implemented and run. This first experiment was designed to choose between scaling dimensions in feature vectors or not in addition to choosing the ratio of negative to positive training examples for both testing and training datasets. For example a 1:1 ratio for testing would indicate the same number of positive (watchlist) drugs as non-watchlist ones, 2:1 would indicate 2 times the number of non-watchlist drugs to watchlist drugs. There are many more instances of non-watchlist drugs available. I decided on a 80/20 data split due to the fact that if a 90/10 split was used and 1:1 ratio of negative to positive examples were used, the resulting testing set would only have 15 instances, hardly enough for a statistically significant result. The word features utilized for the early test were comprised of the combined special lexicons. The preliminary experiments were performed using SVMs with a RBF kernel and using grid search with 10-fold cross validation for building models. The test dataset was then run against the resulting model and ROC curves and AUC numbers were generated from the resulting output.

5.5.3 Classification Experiments

For the specialty lexicon experiments and the general lexicon experiments, the total set of instances are divided into a test and training set as shown in the Figure 7. Due to the inconclusive nature of the training and testing dataset size experiments presented earlier, the commonly accepted stratified sampling approach is utilized where the test and

training sets are sampled with the same distribution as the original data. The data is divided into a 90/10 split with 90% of examples being used to train and 10% being used for testing and are sampled such that the splits are representative of the original distribution of positive and negative instances. Several types of experiments are run. The first is 10-fold stratified cross validation on the 90% of training data. Several cross validation experiments are run on the 90% of the data, the first a general classification experiment with now cost weighting, another is a cost weighting experiment where the costs of the positive and negative examples are adjusted to make the cost distributions approximately equal. In the case of the cost weighting experiment, a greater penalty is imposed for incorrectly classifying a positive example then a negative one.

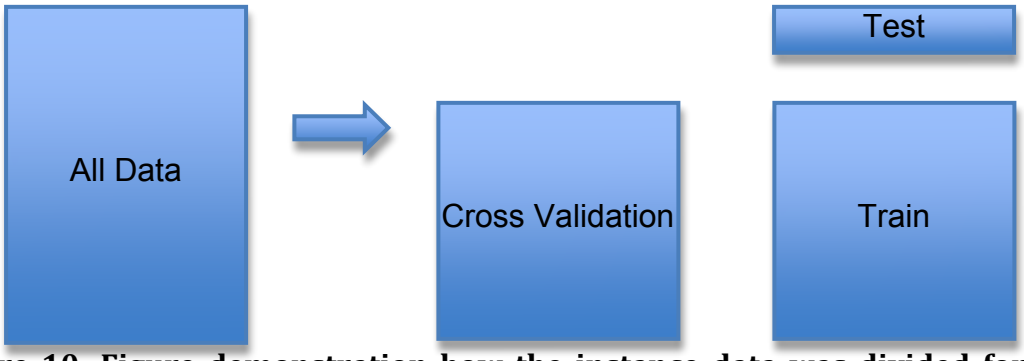


Figure 10: Figure demonstration how the instance data was divided for the various experiments.

The types of classification experiments performed include: Un-normalized Naïve Bayes (UNB), Un-normalized Naïve Bayes with cost weighting (UNBC), Normalized Naïve Bayes (NNB), Normalized Naïve Bayes with cost weighting (NNBC), Un-normalized SVM (SVM), Un-normalized SVM with cost weighting (SVMC), Normalized SVM (NSVM), and Normalized SVM with cost weighting (NSVMC). For the case of cross fold SVM experiments, grid search was performed for each fold.

I believe that the cross fold validation experiments offer a more accurate picture of the classification performance because cross fold validation is the average over multiple runs with multiple divisions of the data. I utilize the smaller subset of data for these initial cross fold validation experiments because similar sized data is used in the feature selection experiments and to make the results comparable over differing experiments. There is no test data for the feature selection experiments, and instead I use the 10% of the full dataset to select word features over utilizing BNS. In this way the feature selection does not preference the watchlist drugs a priori and thus reduces the chances of overfitting.

5.5.3.1 Specialty Lexicon Experiments

An initial experiment of the 5 various combinations of the specialty lexicon (medical, disease names, drug names, sentiment, reaction lexicon) was used to identify the best combination of lexicon that provides the top classification performance. This leads to a total of 30 combinations of lexicons ranging from each individual lexicon to all of them together. Two hundred forty (240) experiments with various parameters were run. Accuracy, F1 (a combination of precision and recall) and area under the ROC curve are chosen as evaluation metrics.

5.5.3.2 BNS Lexicon Experiments

The BNS lexicon experiments as stated previously use the test subset of data to choose the most salient word gram features using Bi-Normal Separation which preferences word-grams that are differentially expressed between watchlist and non-watchlist messages. The top 15,000, 10,000, and 5,000 word grams were chosen from the test subset. Utilizing word grams from the entire message set would artificially

inflate the classification scores since often times classifiers do not have complete knowledge about the entire lexicon apriori.

5.5.3.3 Watchlist Predictions

As with any predictive work it is difficult to quantify the results except basing it upon past performance. This is usually but not necessarily indicative of future performance. I provide predictions of future watchlist or recalled drugs; however these predictions are validated with evidence of past performance cases drugs, which are removed from market.

The previous section's results are used to choose the highest performing classifiers for accuracy, F1 and AUC scores. Utilizing the feature set and classification algorithms, numerous cross validation runs across all of the data are performed multiple times. The output from these classification runs provides insight into future watchlist drug predictions.

Multiple top performing classifiers from each category are combined in an ensemble like approach taking their combination of features and different classification algorithms to produce a meta classifier where the false positives from each category are combined using a linear combination resulting in a score.

I look at the false positives, drugs that are non-watchlist but are classified as watchlist by a classifier. A false positive occurs when a negative instance is incorrectly identified as a positive one. For SVMs this means that the instance falls on the same side of the hyperplane as the positive instances and is usually close to the boundary. For Naïve Bayes, the maximum likelihood estimate is such that the likelihood of the instance being a watchlist drug is greater than a non-watchlist drug.

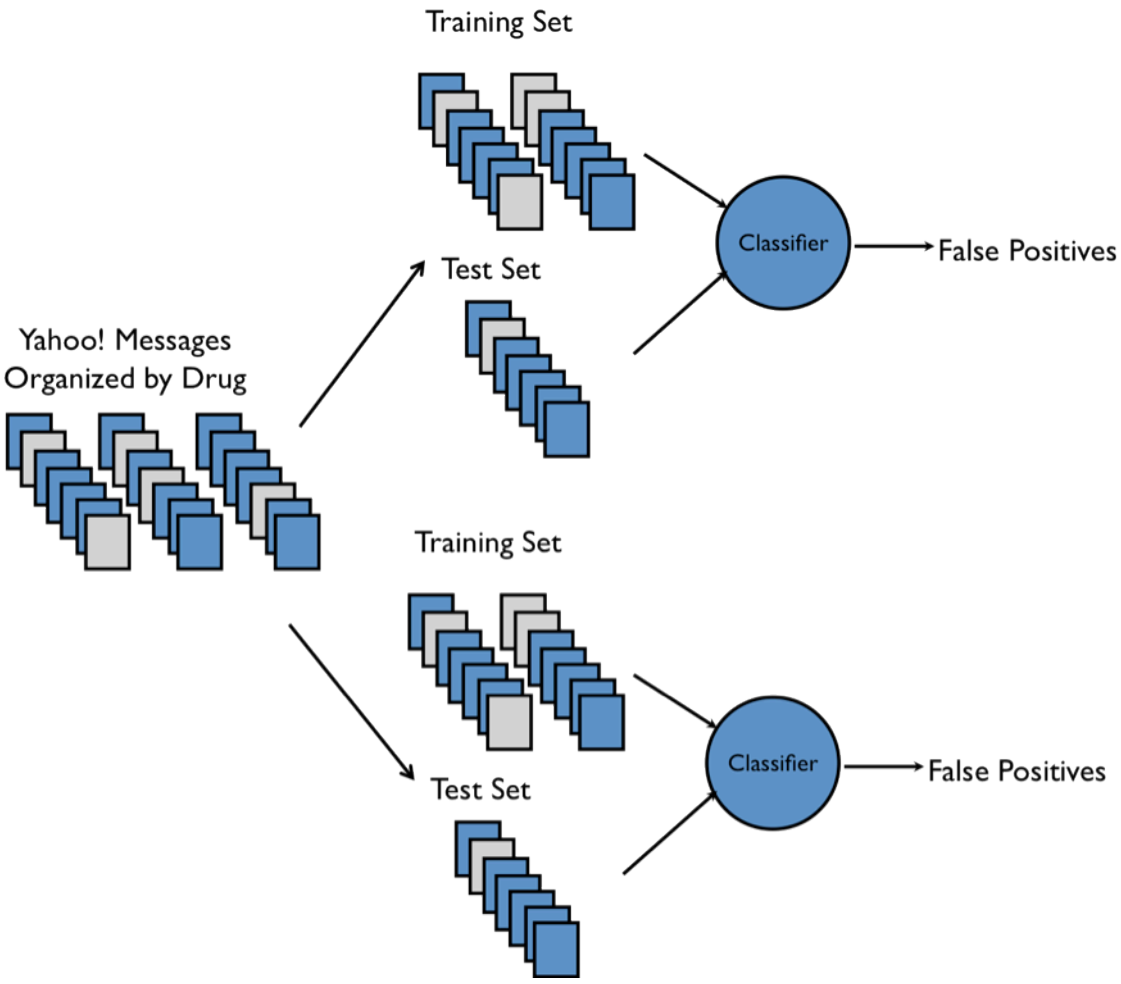


Figure 11: Depiction of ensemble classification method using a divided dataset.

I am interested in drugs that are consistently marked as false positives. I hypothesize that drugs consistently labeled as watchlist are more likely to be "real" or future watchlist or removed from market drugs in the future. In being labeled as a false positive the consistency provides confidence in the prediction. This prediction is based solely on the word features people use to talk about these drugs. Drugs that are false positives could be real watchlist drugs in the future given third party confirmation such as from the FDA.

Given a 90/10 split where 10% of drugs are used to evaluate a classifier, 10 runs should ensure each drug is tested at least once and 50 runs statistically speaking, should allow each drug to be classified 5 times against 5 different classifiers. For these experiments each set of features was used to build a hundred classifiers, test and training sets.

Here outputs from classifiers are utilized because different types of classifiers were found to perform the best and their output cannot be directly compared. It does not make sense to compare a likelihood estimate to a distance to a hyperplane. Multiple rounds of classification with mixed training data increase the confidence in a prediction, as does the use of multiple classifiers.

A weighted ratio is created to score the false positives including the ratio of false positives to number of tests, the number of false positives and the number of classifiers that predicted a false positive: Number of False Positives / Number of occurrences (tests) * Number of False Positives * Number of classifier types. This was done because it is intuitive that a weighted average over the number of false positives is important, a ratio of .5 given 1 false positive to 2 occurrences is different from 100 false positives to 200 occurrences. The number of different classifiers is also important, which classified it as a watchlist drug giving credence to the other classifications.

Two runs were made with drugs withdrawn from the market. Firstly withdrawn drugs were labeled as non-watchlist to determine if the classifiers would accurately identify the withdrawn drugs. This procedure validates this method of watchlist drug identification. Secondly, it demonstrates the robustness of the method for watchlist drug identification. The second run removed the watchlist

drugs and classifies them after the classifier has been built for each fold of the cross-validation run. This second method should more accurately identify the withdrawn drugs with greater confidence because their data is not mixed with the other non-watchlist drugs possibly, reducing the accuracy of the classifiers.

6 Results

6.1 KL Divergence

KL Divergence demonstrates differences in distributions. Here we are interested in the differences in word distributions between messages with watchlist drug mentions compared to those without mentions. Table 9 describes the divergence between the two distributions; Because KL Divergence is non symmetric, both scores (for example Watchlist compared to Non-Watchlist and Non-Watchlist compared to Watchlist distributions) are reported.

Table 9: KL-Divergence scores for the word distributions from watchlist and non-watchlist messages versus the Google 1-Tera gram corpus and the Reuters News corpus.

P	Q	Score
Watchlist	Non-Watchlist	0.1684
Non-Watchlist	Watchlist	0.1778
Watchlist	Google	1.4178
Non-Watchlist	Google	1.1804
Watchlist	Reuters	1.3279
Non-Watchlist	Reuters	1.0815
Reuters	Google	1.2534

Table 9 compares the differences between the watchlist and non-watchlist distributions and between the Google and Reuters distributions. Interestingly the Reuters distribution is closer to the Google distribution than the watchlist drug message. Some of the differences might be attributable to the watchlist distribution generated over less data than the non-watchlist. It was generated from approximately 15.5% of the amount of

data from which the non-watchlist distribution generated. In addition some of the differences in the watchlist and non-watchlist distributions could also be attributed to the drug mentions themselves.

We can look at the terms within each distribution that were attributed to the greatest differences between the distributions. In some sense this shows the "over expressed" terms, ones that proportionately occur more in one distribution than in the other.

Table 10: Terms with the most contribution to the KL-Divergence between the term distributions from the wachlist and non-watchlist message corpora.

D_{KL}(Watchlist||Non)		D_{KL}(Non||Watchlist)	
Term	**Score**	**Term**	**Score**
br	0.015891	the	0.004037
gt	0.011768	nbsp	0.002699
i	0.007875	of	0.002254
lt	0.003781	in	0.00207
my	0.003018	vaccin	0.002002
br&gt	0.0029	oil	0.001032
me	0.002256	fat	0.001032
have	0.001763	by	7.92E-04
you	0.001712	as	7.60E-04
yahoo	0.001395	is	7.56E-04
it	0.001326	or	7.40E-04
take	0.001311	food	7.30E-04
amp	0.001303	use	6.67E-04
sevofluran	0.00123	diet	6.25E-04
u	0.001157	hiv	5.95E-04
med	0.001122	natur	5.75E-04
am	0.001098	autism	5.74E-04
drug	0.001091	product	5.71E-04
was	0.001043	cream	5.53E-04
get	0.001041	ribavirin	5.43E-04
about	0.001026	contain	5.18E-04
i'm	9.64E-04	are	5.12E-04
feel	9.44E-04	acid	4.95E-04
so	8.80E-04	skin	4.87E-04
migrain	8.69E-04	magnesium	4.87E-04

There are marked differences between the two lists of the 25 top terms contributing to differences between the distributions. The first thing that comes to mind are the differences in "garbage" or conversion artifacts of html stripping such as npbsp, br, gt, br&gt, and amp. These exist in both lists indicating that messages in both distributions were written in html though they are more prevalent in the watchlist drug messages.

Ribavirin and Sevofluran are two drugs in the top 25 terms. While drug mentions are not the most prevalent type of term, they do exist. Similarly conditions such as HIV and Autism are mentioned. Drug names and conditions might be mentioned together if drugs to treat serious conditions are more likely to cause side effects or if they occur often with names of watchlist drugs since people are describing the problematic drug.

Furthermore there are differences in the expression of determiners and prepositions such as the, am, you, it, I'm, is, or. This is not surprising as previous work has shown that such features are used to determine authorship between genders. These functional words are also indicative of emotional writing, which one would expect from adverse effect causing drugs. This is also helpful because these types of words are general in nature and likely to occur frequently. General features are preferred since the classifier will not learn on rare words or features that can lead to overfitting, especially with small amounts of data.

6.2 Dictionary NER

The KL Divergence results are encouraging indications that the two word list distributions are separable utilizing machine learning techniques. However, to aid in the classification task identification of drug and effect entities are necessary. This section details results of experiments on identifying and extracting drug and effect entities from

within messages.

For the purpose of this dissertation and the following experiments, drug names with common words were removed and instead the generic (chemical name) was left. Each of the lexicons phrase searches was constructed and run against the index. A total of 13,794,445 AERS instances and 2,228,588 drug instances were found. These are numbers of messages containing a mention. However messages can contain multiple drug and AERS mentions. The differences in the number of instances between numbers of AERS versus drug mentions are clearly delineated in Figure 10. Figure 9 depicts the number of different drug mentions per message. Many messages have relatively few mentions with >96% of messages that contain a drug having 5 or fewer distinct drugs. However, several messages have many different drug mentions with one containing >700 distinct drugs. Messages having many mentions are often lists of drugs people post on groups or SPAM. Messages having more than 5 unique drugs in them are removed. A potential side effect is that this will remove some messages from people with very large drug regimens such as those with CHF or HIV.

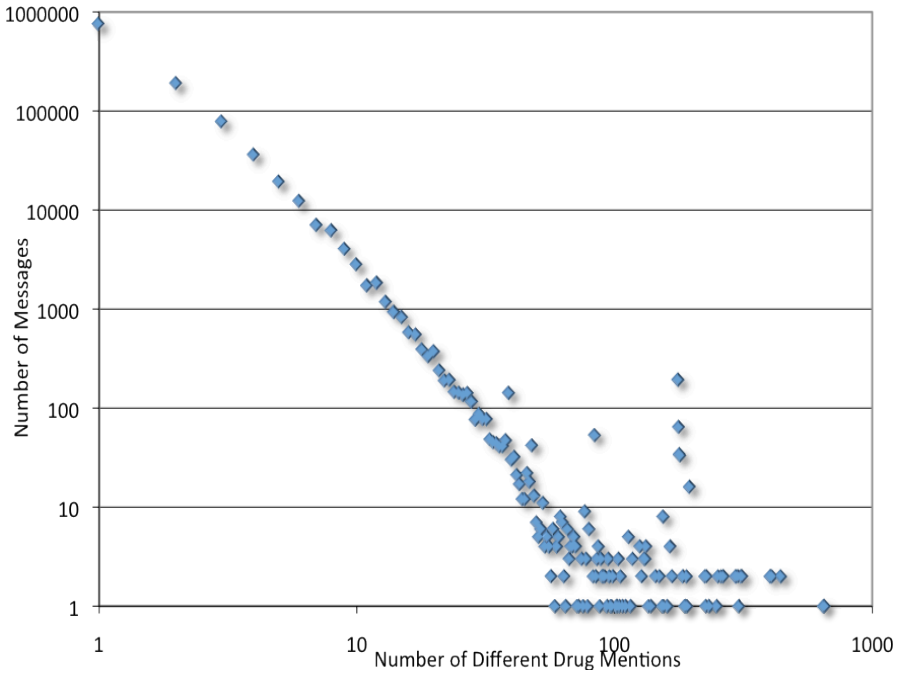


Figure 12: Graph depicting the number of unique drug mentions versus numbers of messages that contain that number of mentions in a log/log scale. The graph follows a Zipf distribution with many messages containing few unique drug mentions.

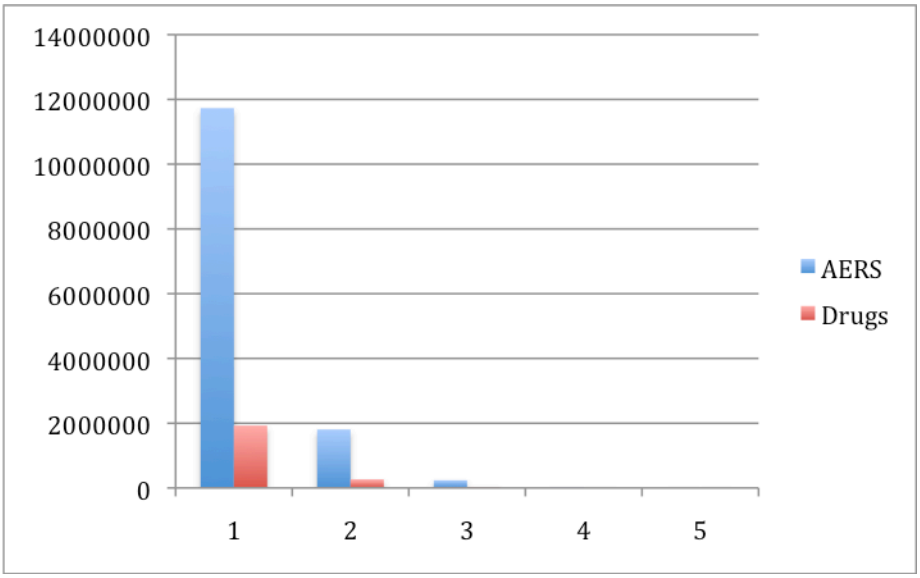


Figure 13: Graph depicting the number of messages with unique AERS or drug lexicon mentions per a message versus term length. For example "Benadryl" would have a term length of one and "abdominal cramps" would have a length of two. The distributions are exponentially decaying.

Figure 10 demonstrates the differences in the number of instances of AERS lexicon found versus drug lexicon. Note that many more side effects are mentioned than drugs. Further analysis is warranted to determine if previous mentions of a drug in an earlier post negates the need of mentioning it again and the drug mention becomes implicit, or perhaps people are complaining or listing adverse effects without attributing them to a particular drug, or it is a common symptom of their treatment or illness. For example headaches or migrains are both symptoms of certain illnesses and a side effect of drugs.

The drug lexicon consists of multi-word phrases. However the distribution of phrase decays exponentially as seen in Figures 13 and 14, with most drug and AERS phrases found being of length 1. These results are not surprising as people are more likely to use a brand name of a drug rather than a multi-term generic active ingredient name. Multi-term phrases are often longer than single term ones, increasing the likelihood of a spelling mistake or typo.

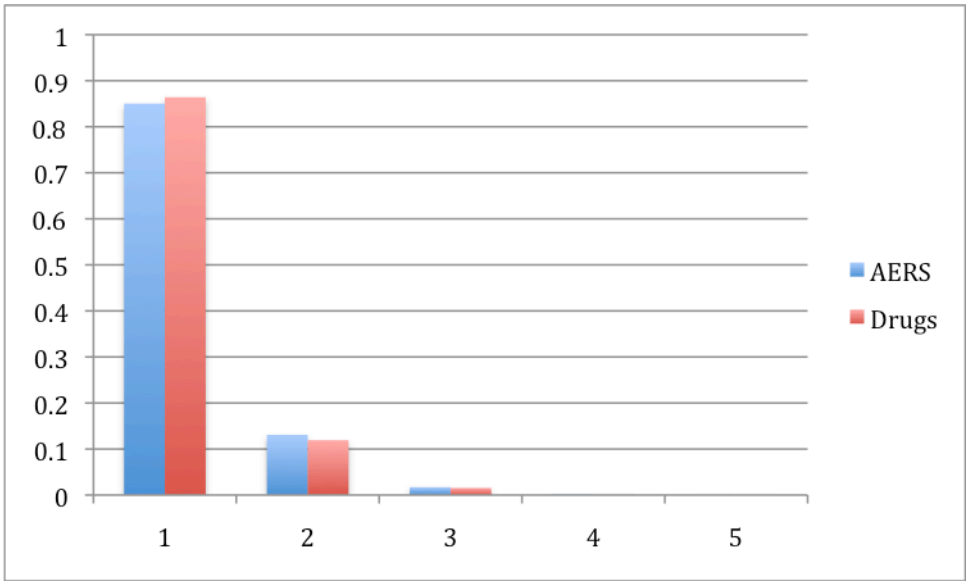


Figure 14: Graph similar to the one in Figure 13 but depicting the percentages of messages with AERS or drug lexicons by term length.

6.3 Drug mentions in messages

Appendix C has a list of the mean, standard deviation, minimum, maximum and median distances for occurences of the top 25 co-occuring drugs. This distance is hard to quantify since English words have an average length of 5.1 characters (7.1 including spaces before and after a word). However many messages contain other text such as signatures, ascii art, emoticons and various amounts of spacing and formatting characters. As seen from the averages of the statistics mentioned for the 25 top scoring drugs, the standard deviation is quite large indicating that people talk about these drugs in very different ways and not only in lists as previously mentioned. Due to the large variance and large differences between the minimum and maximum distances, the median gives a more accurate picture of the data than just the average. The median and average are quite different with the average of the 25 scores ~215 or ~30 words and the median average ~51 or 7 words.

Looking at the number one correlated drug pair, both Viagra and Cialis are used to treat Erectile Dysfunction. The two drugs overlapped in 9,406 messages where they were the only drugs mentioned. Looking at Appendix C, the mean distance between the two drugs is ~90 characters with a standard deviation of 296 characters. The median is 11 and the minimum distance is 7 characters and the maximum distance was 1265. Looking at the distribution of distances between mentions of Viagra and Cialis, Figure 12 demonstrates that the distribution in distance is heavily skewed towards small differences in distances.

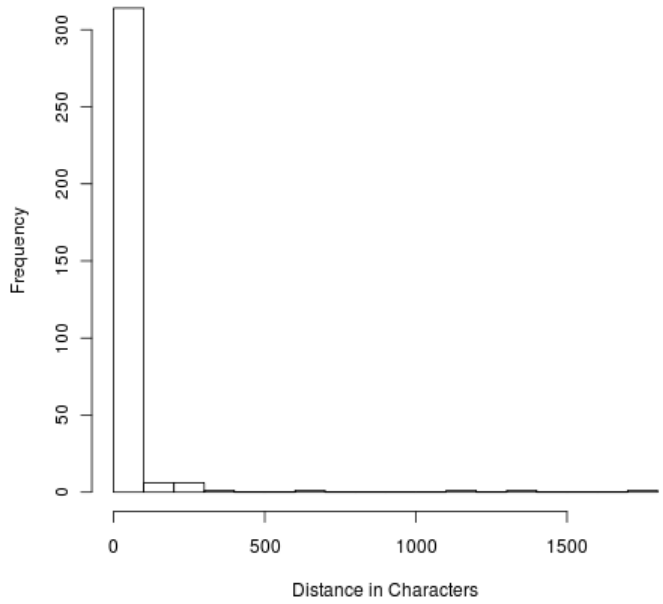


Figure 15: Character distance between mentions of drugs plotted against frequency of occurrence demonstrates a Zipfian distribution.

With this skewed distribution, looking at average distance does not give an accurate picture of what most distances between mentions are like. Upon further inspection the method of calculating distance is biased. The first mention of each drug is not necessarily close to the other. Inspecting the longest distance between messages demonstrated that the first drug (Viagra) was mentioned in the beginning of the message and Cialis was first mentioned near the bottom. However Viagra was re-iterated in the sentence Cialis was first mentioned. Similar results were found in randomly sampled messages with distances that were average, mean, minimum and median cases. In each case both drugs were mentioned either within the same sentence or the following sentence. This makes segmentation of the messages difficult as it becomes hard to discern which adverse effect is attributable to which drug. Therefore for the analysis, attribute effects to all drugs within the messages. The idea is that the varied patterns of

effects and messages combined with the law of large numbers lead to discernable differences by the machine learning classifiers.

This chapter explores dictionary named entity recognition and drug effect mentions and looks at the distribution of drugs within messages. Within the Yahoo! corpus 13,794,445 AERS instances and 2,228,588 drug instances were identified. A SPAM heuristic of >5 different drug mentions in a message was developed. Results also indicate that segmentation of messages is difficult when multiple drugs and effects are mentioned since they are both close in the number characters separating them so it is difficult to attribute an effect to a particular drug.

6.4 Sentiment

The other feature used by the classifiers is sentiment. I believe that sentiment is important in the classification process due to the similarities with product reviews. The sentiment of messages that contain drug mentions are similar to product reviews. People make qualitative and emotional statements about drugs. In order to demonstrate that sentiment can be measured with meaningful results, sentiment of messages with drug mentions are tracked over time. Here we look for a correlation between news articles and changes in aggregate sentiment of drug containing messages.

I hypothesize that sentiment scores increase or decrease following a positive or negative news item. As a control the sentiment scores of drug containing messages are compared to those within the same forums that do not have drug mentions. Two case studies for the drugs Tyysabri and Vioxx are presented.

6.4.1 Tysabri

Tysabri is a recently introduced prescription medication approved for patients with relapsing forms of Multiple Sclerosis (MS). It was originally approved by the FDA in November 2004, and then was subsequently withdrawn by the manufacturer, Biogen-Idec in February 2005. In June 2006 it was then approved for resumed marketing (FDA, 2006). The use of this drug is narrow, specifically for MS.

We demonstrate the ability to track changes in sentiment within a specific group for a limited use drug. Two MS groups that contained more than 500 instances of Tysabri were selected. The messages were evaluated using our augmented LIWC lexicon. A one way ANOVA was run to determine if there was a statistically significant difference between the scores of messages in the following groups: messages containing Tysabri references pre-recall, during recall, and after the recall, and messages not containing Tysabri references pre-recall, during recall, and after the recall. We found that the results were statistically significant with $p < .001$. We plotted the outlier corrected means of the messages containing Tysabri references versus control (messages with no Tysabri references). See Figure 16 for this graph. Data was binned by quarters to improve the number of samples per a data point.

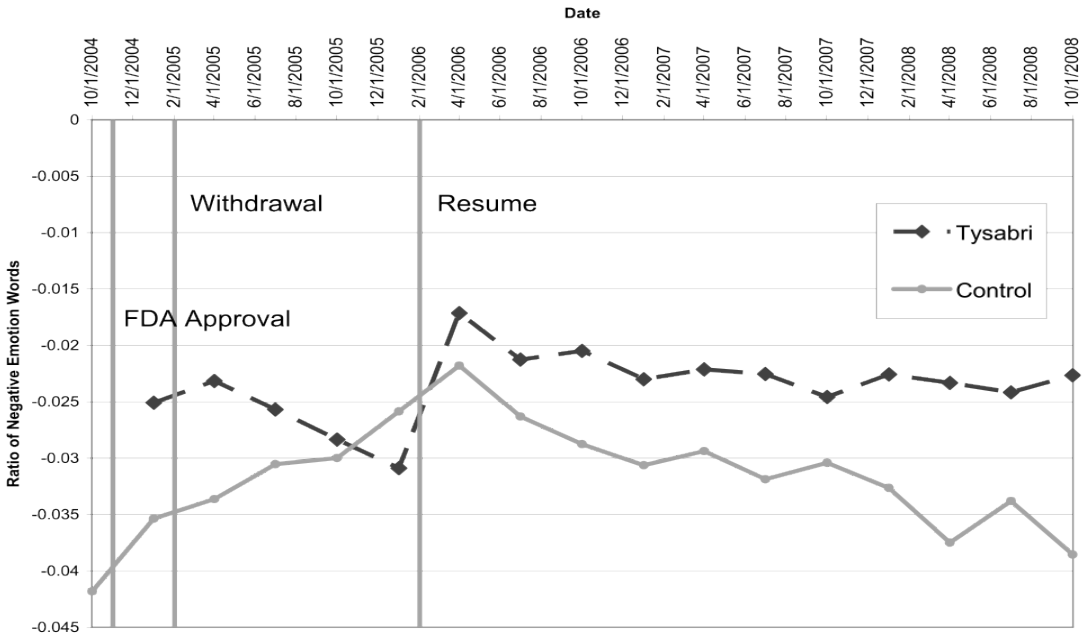


Figure 16: Sentiment of messages mentioning Tysabri versus those that do not for two MS groups. Vertical bars indicate dates for FDA approval of Tysabri, voluntary withdrawal, and remarketing.

Tysabri first appears shortly after it was approved by the FDA in the November/December timeframe as seen by the first data point. After its introduction the sentiment initially improved then got increasingly negative during the time period Biogen-Idec removed it from the market. After the drug's re-introduction to the market the sentiment was extremely positive compared to the time period before and during its' recall. Now that the drug was back on the market, the sentiment has seemed to stabilize at a more slightly negative point than at its reintroduction. We conjecture that the highly positive sentiment was due to people's hope for the drug but the actual effect on the population lagged later and can be seen by the stabilized sentiment.

6.4.2 Vioxx

We looked for groups that contained 100 more instances of Vioxx or Celebrex. Our data was spread over more than 25,000 news groups spanning 7 years. This

selection process was used to weed out news groups that did not contain any references to either of the drugs and ensure there were similar numbers of drug messages as compared to control ones. We only consider the sentiment of control messages of groups that mention either of the drugs. The pruning resulted in 40 groups, 681,516 people (we consider an email address a proxy for a person), containing 867,659 messages of which had an average message length of 1,513 characters. We were interested in determining if our system was able to accurately determine effects for large groups as well as for a broad coverage drug. Vioxx and Celebrex are commonly used pain relievers. Vioxx was first marketed in 1999 and subsequently withdrawn in 2004 (FDA, 2004a). We tracked the sentiment of these two drugs over time. See Figure 10 for data points and dates.

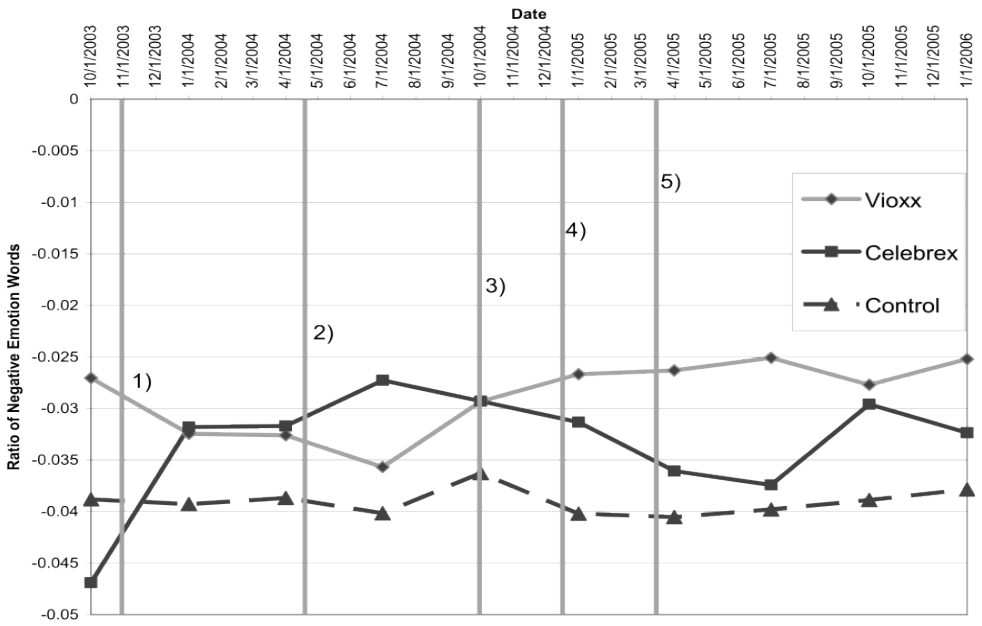


Figure 17: Sentiment of two drugs, Vioxx and Celebrex tracked over time. See Figure 16 for data points and dates.

The messages were evaluated using our augmented LIWC lexicon. A one way ANOVA was performed to determine if there was a statistically significant difference

between the scores of messages in the following groups: messages containing Vioxx pre and post recall; messages containing Celebrex pre and post recall; and messages not Vioxx or Celebrex pre and post recall. The results are statistically significant with $p < .001$. The outlier corrected message means containing Vioxx references, Celebrex and control (messages with no Vioxx or Celebrex references) are plotted in Figure 10. Data is binned by quarters to improve number of samples.

1) Sentiment sharply increases for Celebrex and similarly decreases for Vioxx. On October 30, 2003 the Wall Street Journal published an article from a medical conference that was funded by Merck, the maker of Vioxx, which found "an increased risk of heart attack…compared with patients taking a competing painkiller, Celebrex, from Pfizer Inc" (Burton and Callahan, 2003).

2) The Celebrex sentiment graph increases again and decreases for Vioxx. On April 14, 2004, in study sponsored by Merck, researchers found an elevated risk of acute myocardial infarction associated with Vioxx but not with Celebrex (Solomon et al., 2004). A similar study was published in May 2004 showing that rates of admission for congestive heart failure were higher for patients on Vioxx and non-selective NSAIDs (Mamdani et al., 2004).

3) On September 30, 2004 Merck voluntarily pulled Vioxx from the market (FDA, 2004a). Intuitively one would expect that sentiment towards the drug decreases, however one can imagine that sentiment might increase due to the relief people would have since they are no longer on the drug and that it is no longer available.

4) Sentiment for Celebrex declined, whereas Vioxx sentiment looks somewhat static. On December 17, 2004 the FDA released a statement that the National Cancer

Institute and Pfizer had stopped a clinical trial for Celebrex after determining: "Patients in the clinical trial taking 400 mg. of Celebrex twice daily had a 3.4 times greater risk of CV [cardiovascular] events compared to placebo. For patients taking 200 mg. of Celebrex, the risk was 2.5 times greater" (FDA, 2004b).

5) Again, sentiment for Celebrex declined. On April 7, 2005 the FDA ordered Pfizer, the maker of Celebrex to remove a related drug Bextra (FDA, 2006). However, Celebrex was allowed to remain on the market with a boxed warning about potential cardiovascular events and life-threatening gastrointestinal bleeding (FDA, 2006).

6.5 Lexicon Experiments

6.5.1 Specialty Lexicon Experiments

An initial experiment of the 5 various combinations of the specialty lexicon (medical, disease names, drug names, sentiment, reaction lexicon) were used to identify the best combination of lexicon that provide the top classification performance. This results in a total of 30 combinations of lexicon ranging from each individual lexicon to all of them together. A total of 240 experiments with various parameters were run. Accuracy, F1 (a combination of precision and recall) and area under the ROC curve are chosen as evaluation metrics. Below we sort the experiments into three main categories, Cross Validation, Test, and Test with Cost Weighting. The results are labeled as follows: lexicon + classification type with dis = disease lexicon, react = AERS reaction lexicon, drugs = drug lexicon, sent = sentiment lexicon and med = medical lexicon. Graphs of all of the runs for Accuracy, F1 and Area Under the ROC Curve are in the appendix. Below are the results of the top 10 performing classification algorithm with the corresponding feature set for each of the evaluation metrics.

Table 11: Table depicting the accuracy of the various classification algorithms with different feature sets with the lower and upper bounds with a 95% confidence interval.

Experiment	Accuracy	Lower Bound	Upper Bound
Cross Validation			
dis_react_NSVM	0.903288201	0.793206235	0.957882075
drugs_dis_sent_react_NSVM	0.901353965	0.790801818	0.956684149
drugs_sent_react_NSVM	0.901353965	0.790801818	0.956684149
dis_sent_react_NSVM	0.901353965	0.790801818	0.956684149
react_NSVM	0.901353965	0.790801818	0.956684149
sent_react_NSVM	0.901353965	0.790801818	0.956684149
drugs_dis_sent_NSVM	0.899419729	0.788404626	0.955478999
all_NSVM	0.899419729	0.788404626	0.955478999
drugs_dis_NSVM	0.899419729	0.788404626	0.955478999
drugs_NSVM	0.899419729	0.788404626	0.955478999
Test			
drugs_dis_sent_react_UNB	0.879310345	0.771204077	0.940291098
drugs_sent_react_UNB	0.879310345	0.771204077	0.940291098
dis_sent_react_UNB	0.879310345	0.771204077	0.940291098
drugs_dis_NSVM	0.879310345	0.771204077	0.940291098
sent_react_UNB	0.879310345	0.771204077	0.940291098
dis_SVM	0.877192982	0.768668117	0.938855394
dis_NSVM	0.877192982	0.768668117	0.938855394
dis_NNB	0.877192982	0.768668117	0.938855394
drugs_sent_SVM	0.862068966	0.750738501	0.928415984
drugs_sent_NSVM	0.862068966	0.750738501	0.928415984
Test With Cost Weighting			
drugs_dis_NSVMC	0.879310345	0.771204077	0.940291098
drugs_sent_SVMC	0.862068966	0.750738501	0.928415984
drugs_sent_NSVMC	0.862068966	0.750738501	0.928415984
drugs_dis_sent_react_SVMC	0.862068966	0.750738501	0.928415984
drugs_dis_sent_react_NSVMC	0.862068966	0.750738501	0.928415984
med_dis_sent_SVMC	0.862068966	0.750738501	0.928415984
med_dis_sent_NSVMC	0.862068966	0.750738501	0.928415984
med_drugs_dis_sent_SVMC	0.862068966	0.750738501	0.928415984
med_drugs_dis_sent_NSVMC	0.862068966	0.750738501	0.928415984
drugs_sent_react_SVMC	0.862068966	0.750738501	0.928415984

We initially utilize accuracy to provide some basis of differentiation from the baseline classifier that would label all instances as negative. Labeling all instances as negative would lead to an accuracy of 86.7%. For all three of the experiments, cross fold

validation, simple classification and classification with cost weighting, the top performing system is more accurate than the naïve baseline classifier. However, once we calculate the error bounds utilizing a binomial distribution (since the classifier is binary in nature so each trial is a Bernoulli trial) the lower confidence level is less than the naïve accuracy rate, leading to uncertainty in determining if the classifiers are more accurate than the naïve one.

Table 12: Table depicting the F1 score of the various classification algorithms with different feature sets.

Experiment	F1
Cross Validation	
drugs_dis_react_NSVM	0.476190476
drugs_dis_react_UNB	0.472527473
drugs_UNB	0.469879518
drugs_react_UNB	0.463276836
dis_sent_react_NSVM	0.463157895
sent_react_NSVM	0.463157895
sent_react_UNB	0.460431655
dis_react_NSVM	0.456521739
dis_sent_react_UNB	0.453900709
drugs_sent_UNB	0.452054795
Test	
drugs_UNB	0.470588235
drugs_dis_UNB	0.444444444
drugs_dis_sent_react_UNB	0.363636364
drugs_sent_react_UNB	0.363636364
dis_sent_react_UNB	0.363636364
sent_react_UNB	0.363636364
dis_UNB	0.358974359
med_dis_NNB	0.333333333
drugs_dis_NNB	0.333333333
drugs_NNB	0.333333333
Test With Cost Weighting	
drugs_dis_SVMC	0.444444444
drugs_UNBC	0.444444444
drugs_react_SVMC	0.384615385
drugs_dis_react_SVMC	0.384615385
drugs_dis_UNBC	0.357142857
drugs_SVMC	0.352941176
sent_react_UNBC	0.307692308
dis_sent_UNBC	0.266666667
sent_UNBC	0.266666667
med_dis_sent_NNBC	0.25

The accuracy results of classifiers using cost weighting are less than classifiers not using cost weighting. This makes some sense as the training and test distributions are biased. However what is not apparent from the accuracy score itself is the number of positive (watchlist drugs) each classifier has predicted (either correctly or not). The F1 and AUC (area under the ROC curve) scores provide additional insight into the classifier performance.

The F1 score is the harmonic mean of precision and recall and can be thought of as their weighted average. F1 = 2* precision * recall / (precision + recall). Precision (measure of exactness) and recall (measure of completeness) are common metrics used in information retrieval. An inverse relationship is often seen where optimizing for one leads to a decline in the other. This is particularly apparent in a plot of precision versus recall graph of the classifiers shown in Figure 18 below.

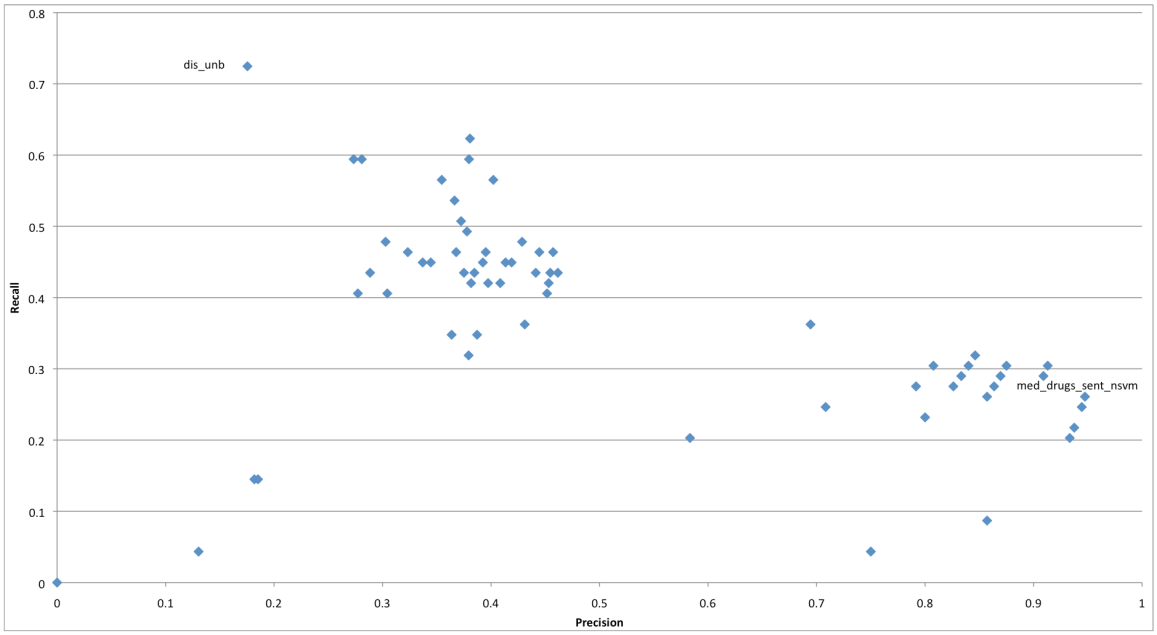


Figure 18: Graph plotting precision versus recall for the various classification algorithms and feature sets for cross validation experiments with classifiers with the highest recall and precision labeled.

Here we see that in the cross validation graph there seem to be two main clusters where the classifiers lay, one cluster with more mass near 1.0 for precision (perfect precision) but relatively low recall (0.3) whereas the other cluster has a mass near 0.45 for recall and 0.4 for precision. Here we see that the classifier with the highest precision uses the disease lexicon using an un-normalized naïve bayes algorithm. The precision is relatively low at approximately 0.18 but has a relatively high recall at 0.73. However, both points are not in the list of top 10 F1 score list because they only optimize one component of the F1 score.

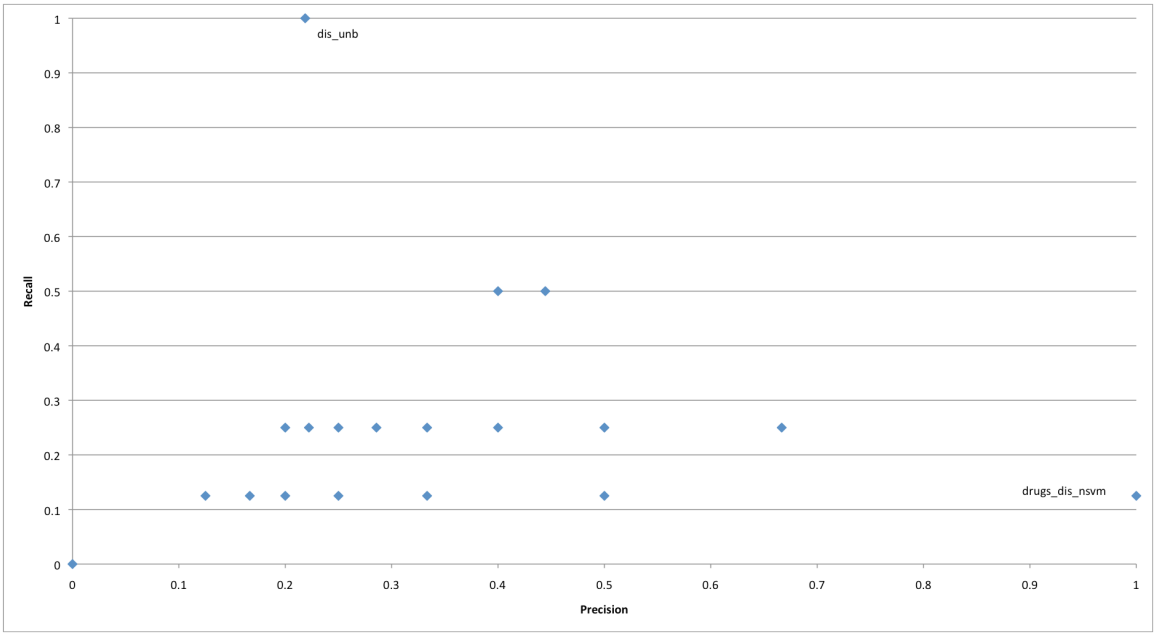


Figure 19: Graph plotting precision versus recall for the various classification algorithms and feature sets for separate test and evaluation instance experiments with classifiers with the highest recall and precision labeled.

Like the accuracy scores, the F1 scores were lower for the experiments with cost weighting and the cross validation scores were the greatest by a small amount. The

patterns of scores for the precision versus recall plots differed significantly from the cross validation ones with overall masses with lower precision and recall with few outliers.

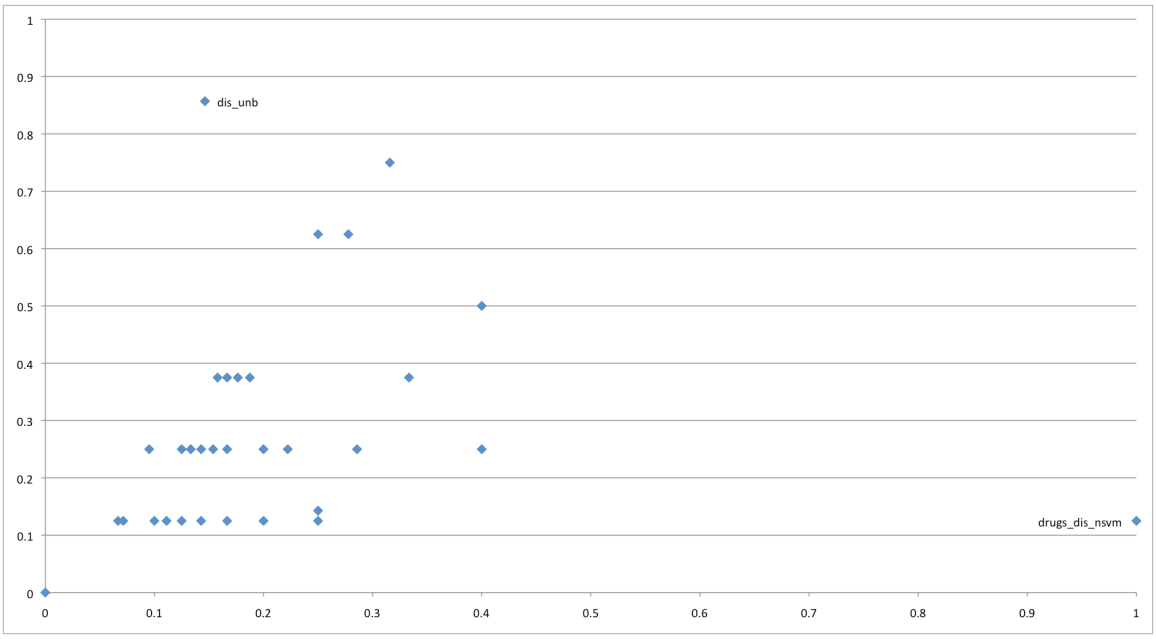


Figure 20: Graph plotting precision versus recall for the various classification algorithms and feature sets for cost weighted classification experiments with classifiers with the highest recall and precision labeled.

Surprisingly the AUC score is the only metric where cost weighting scores higher than non-weighted scores. However, only one score is greater than the non-weighted scores and it is unclear whether or not the difference is statistically significant.

Looking at AUC for non-weighted classifiers, Naïve Bayes dominates. Similar results occur for F1 whereas for accuracy, SVM dominates. The results are more even for the classifiers using cost weighting for F1 and AUC. Looking at the lexicon that occurs in the top performing systems no one set of lexicon dominates. However, drugs and disease lexicon occur most often in the top performing classifiers. The occurrence of each lexicon in the top 10 scoring systems for accuracy, F1 and AUC are shown in Figure 14.

Table 13: Table depicting the F1 score of the various classification algorithms with different feature sets.

Experiment	AUC
Cross Validation	
drugs_UNB	0.7592
drugs_dis_UNB	0.7564
med_drugs_NNB	0.7545
med_drugs_dis_NNB	0.7545
med_drugs_dis_react_NNB	0.7536
med_drug_react_NNB	0.7536
sent_UNB	0.7441
drugs_react_UNB	0.7427
all_NNB	0.7406
med_drugs_dis_sent_NNB	0.7405
Test	
dis_UNB	0.7514
drugs_dis_UNB	0.6850
dis_sent_react_UNB	0.6675
drugs_UNB	0.6675
sent_react_UNB	0.6675
drugs_sent_UNB	0.6550
drugs_dis_sent_UNB	0.6300
sent_UNB	0.6300
dis_sent_UNB	0.6250
med_dis_NNB	0.6100
Test With Cost Weighting	
drugs_UNBC	0.7825
drugs_dis_UNBC	0.7075
drugs_dis_SVMC	0.6900
drugs_react_SVMC	0.6825
drugs_dis_react_SVMC	0.6825
dis_UNBC	0.6514
sent_react_UNBC	0.6325
drugs_SVMC	0.6275
dis_sent_react_UNBC	0.6238
dis_NNBC	0.6200

Table 14: Table demonstrating the number of occurrences of each of the five lexicon occurring in the 10 best performing classifiers for the F1, AUC and accuracy scores. The table on the top is for non-weighted classifiers and the one below is for cost weighted classifiers.

Non-Weighted Classifiers	
Drugs	33
Disease	32
Sentiment	28
Reactions	27
Medical	7

Weighted Classifiers	
Drugs	20
Disease	18
Sentiment	15
Reactions	10
Medical	5

Of the 5 lexicons the medical lexicon seems to be the least useful in building top performing classifiers leading one to believe that people do not use medical terminology frequently or it is used similarly within the two groups. It is surprising that the AERS lexicon is the second least performing lexicon indicating also that the reactions are talked about in similar ways or the lexicon does not capture the way people discuss their adverse events that is plausible. It is expected if not somewhat disappointing that the drug and disease lexicons do well. However, as stated previously this leads to overfitting due to classification learning on drug names or the diseases associated with the watchlist drugs.

6.5.2 BNS Lexicon Experiments

The BNS lexicon experiments as stated previously use the test subset of data to choose the most salient word gram features using Bi-Normal Separation which preferences word-grams that are differentially expressed between watchlist and non-

watchlist messages. Here we choose the top 15,000, 10,000, and 5,000 word grams from the test subset. Utilizing word grams from the entire message set would artificially inflate the classification scores since often times classifiers do not have complete knowledge about the entire lexicon a priori.

Listed below are the top terms ranked by the BNS score. There are many unlikely terms that are ranked the highest including "26" and "as to". Many of these are due to the relatively small sample size (58 drugs, 8 of which are watchlist). Terms in the top 10,000 list include names "rita tx_genesis7" and "romero" but also general medical terms that seem related to adverse effects such as "root cause" or "risk[s] are" and include adverse effects "rosacea …"

The BNS lexicon features were combined with the top scoring special lexicon in decreasing scoring order so drugs, diseases, and sentiment were one group followed by drugs and diseases and finally drugs. Along with the lexicon items special features including the counts of numbers of disease mentions, drug mentions, medical terminology, sentiment containing terms and AERS terminology were used. It was hypothesized that the most BNS features along with the specialty lexicon and numerical features utilizing normalized SVM would yield the best results.

Table 15: Table demonstrating the top 25 word grams using BNS as a scoring metric for the various feature cutoff sizes. All feature sizes (15k, 10k and 5k) contain "All" of the word grams. Only the 15,000 and 10,00 feature sizes contain the terms in the "Top 5,000 Terms" and only the 15,000 feature sizes contain the terms in the "Top 10,000 Terms".

All	Top 5,000 Terms	Top 10,000 Terms
they were	culinari	right med to
claim	culprit i also	right now for
26	culprit of	right now in
time the	culprit that	right now this
lab	cumul dose	right reserv republ
not onli	cup oliv	right where
out that	cup oliv oil	risk are
research and	cure rosacea	risk patient
pre	cure was	rita that
find that	current i take	rita tx_genesis7 tx_genesis7
upon	current math	river
as to	current math requir	rn and
brand	current research	road for
day of	current topic	romero
decid to	current topic that	roof then
indic that	current undergo	roof then the
pattern	curti	room temperatur away
that some	cut into	root caus
uniqu	cut it and	rosacea although
assum	cuti	rosacea and other
is my	cvd	rosacea bump
suspect	cycl and then	rosacea can
brown	cytokin that	rosacea can be
correl	cytokin tnf	rosacea dr
infant	d sinc	rosacea flush

Again the various feature combinations were combined with the Naïve Bayes and SVM classification algorithms with and without both normalization, and with and without cost penalties. Due to the numerous combinations of options in classification and feature selection it was necessary to divide the output in some way to demonstrate meaningful results. However, the way data is presented leads to bias such that some features are preferred over others. Here we present results divided into three categories

based upon the most salient feature of this experiment, which is the number of BNS features used.

Table 16 below lists the top ten highest scoring (in accuracy) feature set combinations with their associated classification algorithms. The associated lower and upper bounds for error are also listed. None of the scores is significantly higher than a naïve classifier that labels all instances as non-watch list. It is also disappointing that these accuracy scores are lower than the highest scores for cross validation in the last set of experiments.

Table 16: Table depicting the accuracy of the various classification algorithms with different feature sets with the lower and upper bounds with a 95% confidence interval.

	Accuracy	Lower Bound	Upper Bound
All Features			
bns drugs diseases sentiment NNB	0.8762	0.7602	0.9405
bns drugs diseases sentiment NNBC	0.8762	0.7602	0.9405
bns drugs diseases sentiment NSVM	0.8762	0.7602	0.9405
bns drugs diseases sentiment NSVMC	0.8762	0.7602	0.9405
bns drugs diseases sentiment no features NNB	0.8762	0.7602	0.9405
bns drugs diseases sentiment no features NNBC	0.8762	0.7602	0.9405
bns drugs diseases sentiment no features NSVM	0.8762	0.7602	0.9405
bns drugs diseases sentiment no features NSVMC	0.8762	0.7602	0.9405
bns drugs diseases NNB	0.8762	0.7602	0.9405
bns drugs diseases NNBC	0.8762	0.7602	0.9405
Five Thousand Features			
bns drugs diseases sentiment NNB	0.8762	0.7602	0.9405
bns drugs diseases sentiment NNBC	0.8762	0.7602	0.9405
bns drugs diseases sentiment NSVM	0.8762	0.7602	0.9405
bns drugs diseases sentiment NSVMC	0.8762	0.7602	0.9405
bns drugs diseases sentiment no features NNB	0.8762	0.7602	0.9405
bns drugs diseases sentiment no features NNBC	0.8762	0.7602	0.9405
bns drugs diseases sentiment no features NSVM	0.8762	0.7602	0.9405
bns drugs diseases sentiment no features NSVMC	0.8762	0.7602	0.9405
bns drugs diseases NNB	0.8762	0.7602	0.9405
bns drugs diseases NNBC	0.8762	0.7602	0.9405
Ten Thousand Features			
bns drugs diseases sentiment NNB	0.8762	0.7602	0.9405
bns drugs diseases sentiment NNBC	0.8762	0.7602	0.9405
bns drugs diseases sentiment NSVM	0.8762	0.7602	0.9405
bns drugs diseases sentiment NSVMC	0.8762	0.7602	0.9405
bns drugs diseases sentiment no features NNB	0.8762	0.7602	0.9405
bns drugs diseases sentiment no features NNBC	0.8762	0.7602	0.9405
bns drugs diseases sentiment no features NSVM	0.8762	0.7602	0.9405
bns drugs diseases sentiment no features NSVMC	0.8762	0.7602	0.9405
bns drugs diseases NNB	0.8762	0.7602	0.9405
bns drugs diseases NNBC	0.8762	0.7602	0.9405

The most accurate classifiers for each of the three BNS feature sets are ones that utilize the specialty lexicon; the highest-ranking ones feature all the specialty lexicon, drugs, diseases and sentiment whereas the lower ranking ones contain drugs and diseases. The number of BNS features does not change the overall accuracy indicating that they have little effect on the classifiers. Similarly, it appears that the numerical features – counts over the number of disease mentions, number of drug mentions, etc. also have little effect on the classification performance.

The cross fold validation scores also coalesce since many of the scores tend to be close to one another. This can be seen through the similar accuracy scores as well as the following precision recall graphs and F1 score tables.

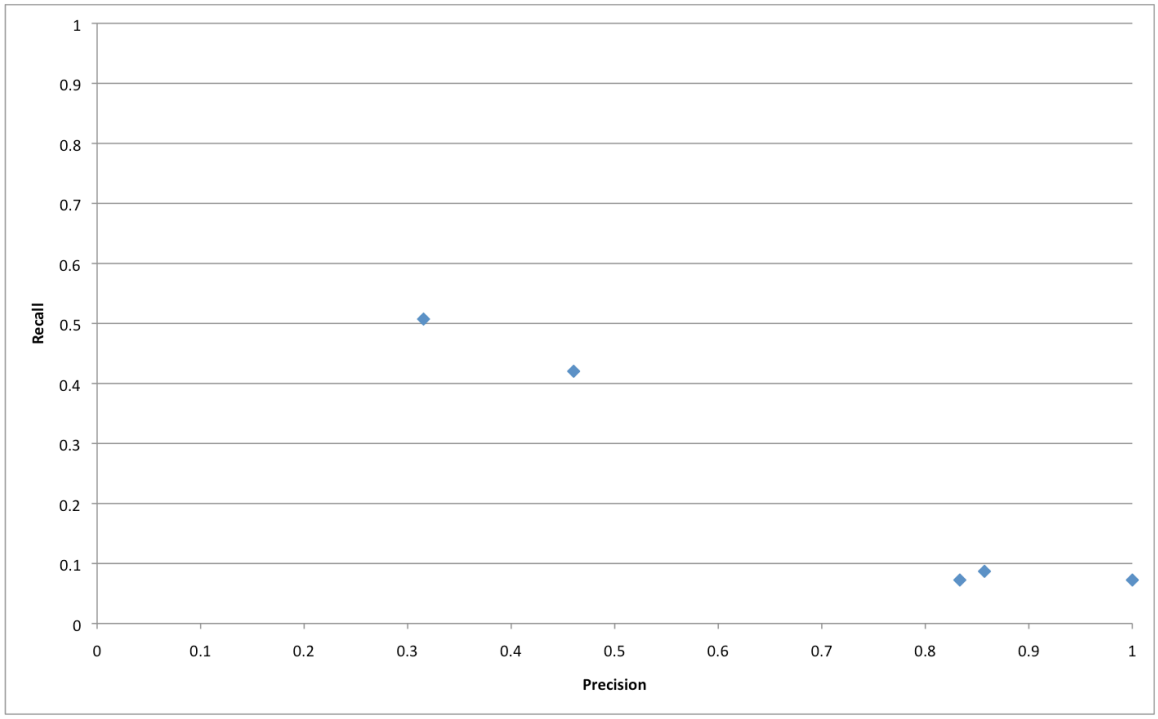


Figure 21: Graph plotting precision versus recall for the various classification algorithms utilizing all 15,000 BNS word features.

The high recall and precision scores were not highlighted because many of the feature set combinations had the same performance. The similarities among the three graphs further emphasize the fact that the differences in BNS feature sizes have little impact on the classification performance.

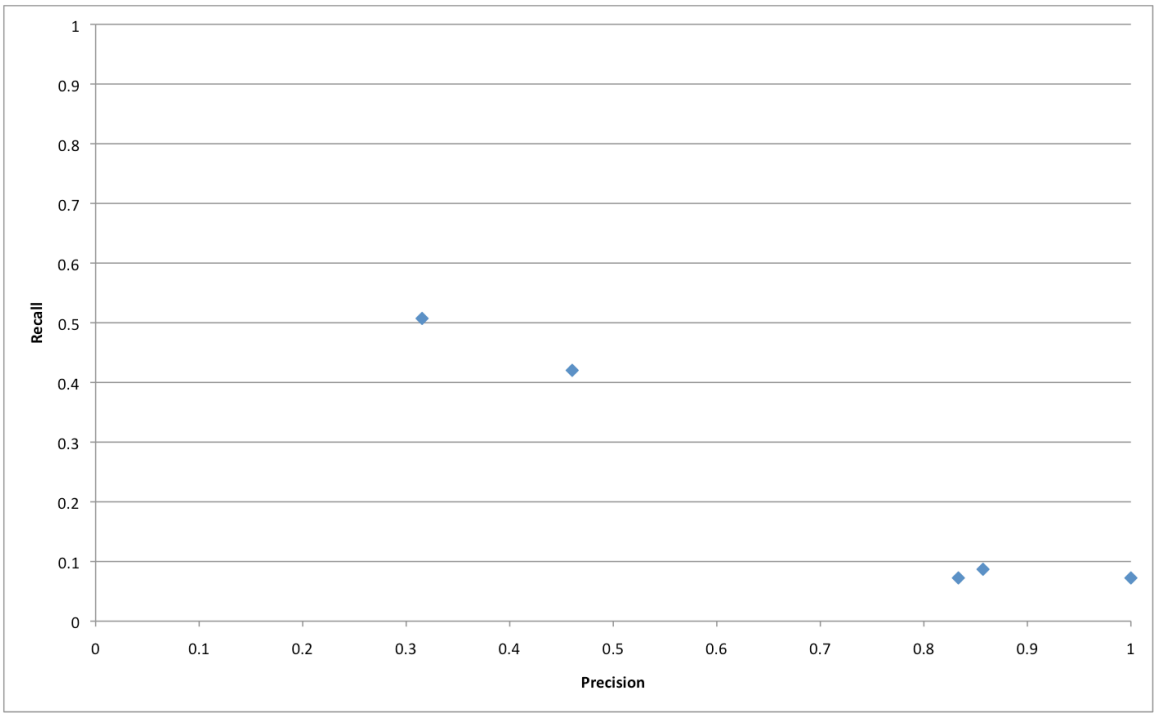


Figure 22: Graph plotting precision versus recall for the various classification algorithms utilizing 10,000 BNS word features.

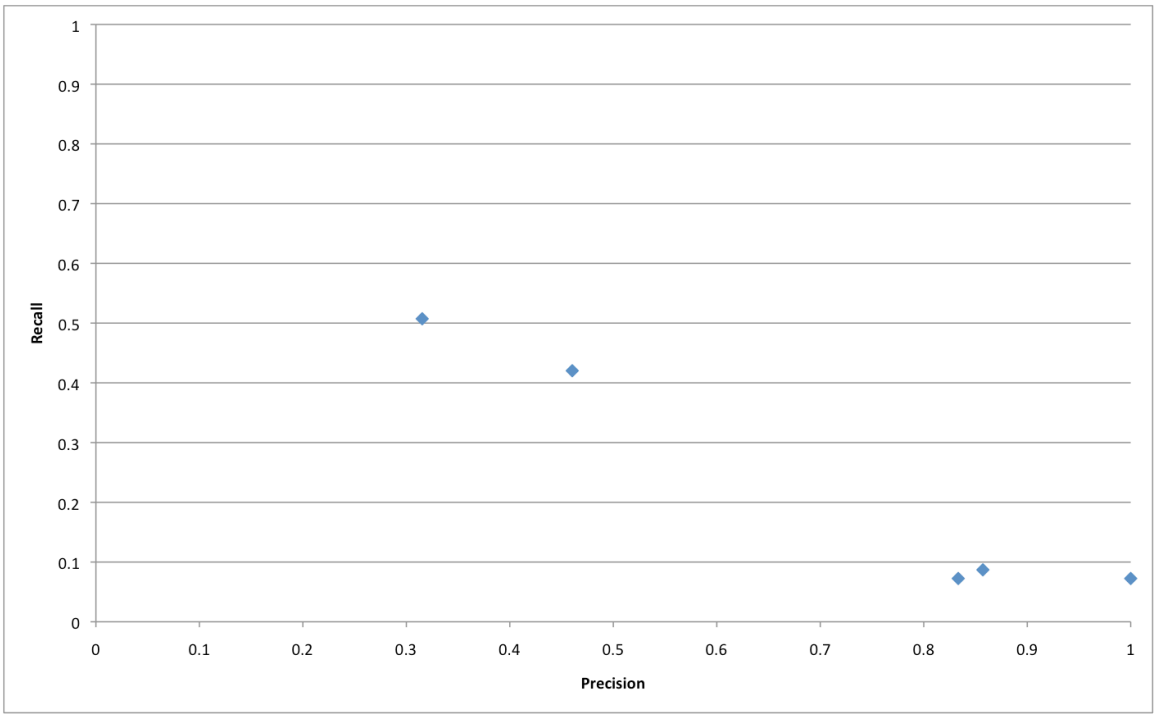


Figure 23: Graph plotting precision versus recall for the various classification algorithms utilizing 5,000 BNS word features.

Like accuracy, the F1 scores are very similar. However there is some differentiation in the top ten. Interestingly, non-normalized Naïve Bayes performs best, outperforming the other classification algorithms despite different feature set combinations. However, there are no bounds on the error for the F1 scores so the differences may not be statistically significant. This is markedly different from the other set of experiments where there was no clear indication that one algorithm performed better than another. In the previous cross validation experiments non-normalized Naïve Bayes and normalized SVM performed best, disappointingly the previous scores greater than the top F1 scores for all feature sets and algorithms in these experiments.

Table 17: Table depicting the F1 score of the various classification algorithms with different feature sets.

Experiment	F1
All Features	
bns drugs diseases sentiment UNB	0.4394
bns drugs diseases sentiment no features UNB	0.4394
bns drugs diseases UNB	0.4394
bns drugs diseases no features UNB	0.4394
bns drugs UNB	0.4394
bns drugs no features UNB	0.4394
bns UNB	0.4394
bns no features UNB	0.4394
bns drugs diseases sentiment UNBC	0.3889
bns drugs diseases sentiment no features UNBC	0.3889
Five Thousand Features	
bns drugs diseases sentiment UNB	0.4394
bns drugs diseases sentiment no features UNB	0.4394
bns drugs diseases UNB	0.4394
bns drugs diseases no features UNB	0.4394
bns drugs UNB	0.4394
bns drugs no features UNB	0.4394
bns UNB	0.4394
bns no features UNB	0.4394
bns drugs diseases sentiment UNBC	0.3889
bns drugs diseases sentiment no features UNBC	0.3889
Ten Thousand Features	
bns drugs diseases sentiment UNB	0.4394
bns drugs diseases sentiment no features UNB	0.4394
bns drugs diseases UNB	0.4394
bns drugs diseases no features UNB	0.4394
bns drugs UNB	0.4394
bns drugs no features UNB	0.4394
bns UNB	0.4394
bns drugs diseases sentiment UNBC	0.3889
bns drugs diseases sentiment no features UNBC	0.3889
bns drugs diseases UNBC	0.3889

The AUC scores provide additional insight into the classification performance. The results of the AUC scores show similar results to the F1 scores in that the best forming algorithm is un-normalized Naïve Bayes. However, un-normalized Naïve Bayes with cost

weighting also is in the top scoring algorithms. Unfortunately, similar to the F1 scores, the performance of the BNS feature experiments in AUC scores are less than those in the previous set of experiments.

Table 18: Table depicting the AUC score of the various classification algorithms with different feature sets.

Exepriment	AUC
All Features	
bns drugs diseases sentiment UNB	0.7028
bns drugs diseases sentiment no features UNB	0.7028
bns drugs diseases UNB	0.7028
bns drugs diseases no features UNB	0.7028
bns drugs UNB	0.7028
bns drugs no features UNB	0.7028
bns UNB	0.7028
bns no features UNB	0.7028
bns drugs diseases sentiment UNBC	0.7006
bns drugs diseases sentiment no features UNBC	0.7006
Five Thousand Features	
bns drugs diseases sentiment UNB	0.7028
bns drugs diseases sentiment no features UNB	0.7028
bns drugs diseases UNB	0.7028
bns drugs diseases no features UNB	0.7028
bns drugs UNB	0.7028
bns drugs no features UNB	0.7028
bns UNB	0.7028
bns no features UNB	0.7028
bns drugs diseases sentiment UNBC	0.7006
bns drugs diseases sentiment no features UNBC	0.7006
Ten Thousand Features	
bns drugs diseases sentiment UNB	0.7028
bns drugs diseases sentiment no features UNB	0.7028
bns drugs diseases UNB	0.7028
bns drugs diseases no features UNB	0.7028
bns drugs UNB	0.7028
bns drugs no features UNB	0.7028
bns UNB	0.7028
bns drugs diseases sentiment UNBC	0.7006
bns drugs diseases sentiment no features UNBC	0.7006
bns drugs diseases UNBC	0.7006

The pattern of similar classification performance between the BNS feature sets indicates that the BNS features do not factor into classification performance or that more than 5,000 features do not help.

The BNS feature performance overall is poorer than the specialty lexicon experiments. This may be partly due to the poor feature selection, which is a side effect of the few numbers of messages from which the BNS features were selected. This task, like many other machine learning and data mining tasks are data driven, requiring large amounts of annotated data to learn and generalize over.

Quantitative analysis of the classifiers is performed on various combinations of feature sets and algorithms in this chapter. The knowledge rich features, diseases and drug names seemed to perform best overall. However, these numbers without some comparison or task to perform, give no idea of how well this task is performed. What is 85% accuracy? What is good enough when predicting drug safety? How accurate does a predictive classifier need to be in order to be useful? If even one dangerous drug is found before it harms, is it worthwhile?

6.5.3 Watchlist Predictions

The previous section's results were used to choose the highest performing classifiers for accuracy, F1 and AUC scores. Examining the top performing classifiers, a normalized SVM using the disease and AERS lexicon yields 90.33% accuracy. A normalized SVM using drugs, disease and AERS lexicon yields a F1 score of 0.4762 and an un-normalized Naïve Bayes classifier yields an AUC score of 0.7592.

We are interested in drugs that are consistently marked as false positives. We hypothesize that drugs that are consistently labeled as watchlist are more likely to be

"real" or future watchlist or removed from market drugs in the future. The consistency in being labeled as a false positive provides confidence in the prediction. The prediction is based solely based on word features. Drugs that are discussed in similar ways will be labeled alike. Drugs that are false positives could be real watchlist drugs in the future given third party confirmation such as from the FDA.

Two runs were made with drugs withdrawn from the market. Firstly withdrawn drugs were labeled as non-watchlist to determine if the classifiers would accurately identify the withdrawn drugs. This procedure validates this method of watchlist drug identification. Secondly it demonstrates the robustness of the method for watchlist drug identification. The second run removes the watchlist drugs and classifies them after the classifier has been built for each fold of the cross-validation run. This second method should more accurately identify the withdrawn drugs with greater confidence because their data is not mixed with the other non-watchlist drugs possibly reducing the accuracy of the classifiers.

For the first run, four drugs that were withdrawn from market were identified, Vioxx, Trovan, Baycol, and hydromorphone. The following demonstrates the top scoring results from the first run. All drugs with a score > 0 are in Appendix J.

Table 19: Top scoring false positive generating drugs with withdrawn drugs mixed in and labeled as non-watchlist. Next to the drug name, the Pos column denotes the number of times a classifier marked the drug as a false positive. Occ indicates the number of occurrences, or number of times the drug was classified. Class indicates the number of different types of classifiers (1-3) that predicted a false positive for a drug. Score indicates the linear combination of Pos, Occ, and Class resulting in a score of the confidence in prediction. Drugs are arranged in descending order.

Drug	Pos	Occ	Class	Score
clozapine OR Clozaril OR FazaClo	31	64	3	45.047
fludarabine OR Fludara OR Oforta	29	61	3	41.361
methylphenidate OR Concerta OR Daytrana OR Metadate CD OR Metadate ER OR Methylin OR Methylin ER OR Ritalin OR Ritalin LA OR Ritalin-SR	25	50	3	37.500
morphine OR Astramorph PF OR Avinza OR Duramorph OR Infumorph OR Kadian OR MS Contin OR MSIR OR Morphine IR OR Oramorph SR OR RMS OR Roxanol	14	38	3	15.474
meloxicam OR Mobic	15	50	3	13.500
Extraneal	10	36	3	8.333
aripiprazole OR Abilify OR Abilify Discmelt	9	30	3	8.100
evening primrose OR Evening Primrose Oil OR Primrose Oil	17	56	1	5.161
quetiapine OR Seroquel OR Seroquel XR	15	52	1	4.327
trazodone OR Desyrel OR Desyrel Dividose OR Oleptro	14	46	1	4.261
(acetaminophen AND diphenhydramine) OR Anacin P.M. Aspirin Free OR Coricidin Night Time Cold Relief OR Excedrin PM OR Headache Relief PM OR Legatrin PM OR Mapap PM OR Midol PM OR Percogesic Extra Strength OR Sominex Pain Relief Formula OR Tylenol PM OR Tylenol Severe Allergy OR Tylenol Sore Throat Nighttime OR Unisom with Pain Relief	12	34	1	4.235
thalidomide OR Thalomid	13	44	1	3.841

Hydromorphone is a narcotic. It is a semi-synthetic opiod derived from morphine. An extended-release version of hydromorphone called Palladone was available in the United States before it was voluntarily withdrawn after a July 2005 FDA advisory warned of a high overdose potential when taken with alcohol (FDA, 2005). However, as of March 2007, it is still available in many European countries.

Cerivastatin (Baycol) is a synthetic statin introduced in 1997 and used to lower cholesterol and prevent cardiovascular disease. Statins work by inhibiting the enzyme

HMG-CoA reductase, which is important in the production of cholesterol. It was voluntarily withdrawn in 2001 due to reports of fatal rhabdomyolysis, which is the breakdown of skeletal muscle that can lead to kidney failure. At the time of withdrawal the FDA had reports of 31 deaths due to rhabdomyolysis (USA Today, 2001).

Trovafloxacin (Trovan) is a broad spectrum antibiotic that was withdrawn from market due to the risk of heptatoxicity - causing liver damage and failure. In 1996, Pfizer violated international law during an epidemic by testing the unproven drug on 100 children and infants with brain infections (Stephens, 2006). Currently the FDA is aware of 14 cases of liver failure linked to Trovan and over 100 cases of liver toxicity (FDA, 1999).

Rofecoxib (Vioxx) is a nonsteroidal anti-inflammatory drug (NSAID) first marketed in 1999 as a safer alternative to drugs such as Tylenol or Aleve. It was subsequently withdrawn in 2004 due to a significant increased risk of acute myocardial infarction (heart attack) (FDA, 2004a). Rofecoxib was one of the most widely used drugs to be withdrawn from market. Merck, the maker of Vioxx, reported revenue of $2.5 billion the year before it was withdrawn (Reuters, 2006).

Thalidomide was also flagged as a high-scoring false positive. Thalidomide was used as an anesthetic but was withdrawn from market in the 1960's after it was found to cause birth defects resulting in babies having no limbs or limbs with finger or toes fused together. However, thalidomide has been remarketed with narrow focus and strong labeling.

Temazepam is an intermediate acting benzodiazepine prescribed as a short term sleeping aid and is sometimes used as an anti-anxiety, anticonvulsant, and muscle

relaxant. Sweden and Norway withdrew the drug from market in 1999 due to diversion, abuse, and high rate of overdose deaths in comparison to other drugs of its group. It is still available in the US with strong warnings for severe anaphylactic and anaphylactoid reactions and cautions about complex behavior such as "sleep driving" - driving while not fully awake and having amnesia about the event (FDA, 2009).

Primrose oil is derived from Oenothera biennis and is sometimes used to treat eczema, rheumatoid arthritis, menopausal symptoms, premenstrual syndrome, cancer and diabetes. This supplement had an unusually high score given that it is a supplement. However, the broad range of uses and associations with other diseases and medications could lead to misclassification especially since one of the classifiers is based solely upon other drug mentions. This misclassification could also apply to acetaminophen, vitamin e, fish oil and Metamucil.

The most striking find was *Sibutramine (Meridia)*, which is an appetite suppressant and is used to treat obesity. It is currently still available in the US but has been removed from market in the United Kingdom and European Union. Currently the drug is under review but is not considered a watchlist drug. A FDA early communication about the drug was posted on 11/20/2009 and a subsequent follow-up on 1/21/2010 indicating an increased risk of heart attack and stroke in patients with a history of cardiovascular disease (FDA, 2010b). This drug was repeatedly marked as a false positive. The latest dated Yahoo! messages mentioning Sibutramine were from 12/11/2008, almost year before FDA advisories and a little over a year before the UK withdrawal (BBC, 2010). Table 20 contains example messages about adverse events attributed to Sibutramine from the Yahoo! corpus. These are two examples of the

numbers of messages that exist and contain serious effects including heart arrhythmia, high blood pressure, confusion, etc.

Table 20: Example messages about Sibutramine highlighting side effects.

Date: 4/20/2003 Subject: I will be leaving the group Hello to All I have been on Meridia for a year and have lost about 45 pounds but now I am going off it due to health problems which my doctor feels was caused by it. I had high , very high blood pressure, the doctor was going to try me on a blood pressure medication but felt I should stop Meridia first to see if it could be the possible trigger. i stopped Meridia two weeks ago and my blood pressure is back to normal. I was also having hair loss and mental confusions, unless I wrote things down I would forget all the time. I am so frustrated as I have lots more weight to lose and am scarred that I will gain all my other weight back but I don't want to die from high blood pressure. I guess I will just have to deal with being over weight and learn to love myself. My finance was very upset by the article below and told me to stop taking the drug also. I also read the following on line. Take care everyone kimberly Date: 2/9/2006 Subject: Re:[Meridia Forum] Hi, I'm new! I was on Meridia and lost over 27 pounds, all of which I have kept off. However, I had alarming heart rhythm problems so I stopped the Meridia after 1 month. Please keep in mind that Meridia WAS listed in Consumer Reports as a potentially unsafe drug. I am living testimony that it was dangerous for me. I have lost the rest of the weight by sheer willpower. The Meridia was a wonder drug but do weigh the risks of obesity with the risks of the drug. Just my opinion -

In the second run, results of which are in Appendix K, a re-ordering of the drugs is seen. Below is a list of the top twelve scoring drugs to compare against the first run. If we take the identification of European Union drugs with higher scores then the second run performs better. Sibutramine is ranked higher, 5th in the list of predictions. However looking at the overall score including the withdrawn drugs, disappointingly Baycol has an overall lower score.

Table 21: Top scoring false positive generating drugs with withdrawn drugs removed from the cross validation data. Next to the drug name, the Pos column denotes the number of times a classifier marked the drug as a false positive. Occ indicates the number of occurrences, or number of times the drug was classified. Class indicates the number of different types of classifiers (1-3) that predicted a false positive for a drug. Score indicates the linear combination of Pos, Occ, and Class resulting in a score of the confidence in prediction. Drugs are arranged in descending order.

Drug	Pos	Occ	Class	Score
methylphenidate OR Concerta OR Daytrana OR Metadate CD OR Metadate ER OR Methylin OR Methylin ER OR Ritalin OR Ritalin LA OR Ritalin-SR	30	34	3	79.412
morphine OR Astramorph PF OR Avinza OR Duramorph OR Infumorph OR Kadian OR MS Contin OR MSIR OR Morphine IR OR Oramorph SR OR RMS OR Roxanol	13	38	3	13.342
quetiapine OR Seroquel OR Seroquel XR	14	31	2	12.645
indomethacin OR Indocin OR Indocin IV OR Indocin SR	19	37	1	9.757
sibutramine OR Meridia	17	34	1	8.500
meloxicam OR Mobic	17	35	1	8.257
vigabatrin OR Sabril	14	31	1	6.323
losartan OR Cozaar	13	28	1	6.036
oxycodone OR ETH-Oxydose OR OxyContin OR OxyIR OR Oxyfast OR Percolone OR Roxicodone OR Roxicodone Intensol	13	30	1	5.633
doxepin OR Adapin OR Prudoxin OR Silenor OR Sinequan OR Zonalon	14	37	1	5.297
aripiprazole OR Abilify OR Abilify Discmelt	13	32	1	5.281
guaifenesin OR Altarussin OR Amibid LA OR Drituss G OR Duratuss G OR GG 200 NR OR Ganidin NR OR Guaifenesin LA OR Guaifenex G OR Guaifenex LA OR Hytuss OR Liquibid OR Mucinex OR Mucinex for Kids OR Muco-Fen 1200 OR Organidin NR OR Q-Bid LA OR Robitussin Chest Congestion OR Scot-Tussin Expectorant OR Tussin	13	34	1	4.971

In both cases we see strong psychiatric drugs such as Ritalin and Clozapine ranked near the top as well as strong painkillers (opiates) such as morphine, Mobic and Oxycodone. This might indicate intuitively that these classes of drugs are more dangerous or likely to cause more serious effects than other types of drugs.

As stated earlier the prevalence of vitamins or over the counter painkillers that seem innocuous abound on both lists. This might be attributed to their wide spread use among many conditions and used in combination with many different drugs. Both runs demonstrate a relatively high score for acetaminophen and acetaminophen containing products. However, the causality of scores is not established. Like many other machine learning and classification tasks, even if the features used by a classifier are known, there is no established causation, only the correlation between feature and class. It remains to be seen if the outcome is corroborating the recent allegations over the safety of acetaminophen with regard to overdosing and safety of children products (FDA, 2010a) or due to the their widespread use and association with many different drugs and diseases.

Both watchlist and drugs withdrawn from market have been successfully identified using multiple machine learning classifiers. These drugs were correctly identified even when incorrectly labeled. Other non-watchlist drugs were also identified. These drugs are false positives in the sense that they are not currently watchlist or recalled drugs. However, these drugs, at some point in the future could become watchlist or withdrawn drugs. A list of potential watchlist drugs was produced; the most significant of these is a weight loss drug Sibutramine (Meridia). Sibutramine was withdrawn from the US market in October 2010 following the removal from the European Union and United Kingdom drug markets earlier in the year. The Yahoo! data only contained messages three years prior to the removal from US and European markets.

7 Conclusions

This dissertation explores the use health forum data as a possible data source for pharmacovigilance tasks. The major contribution of this work is methods used explore the identification of drugs for further drug safety study. Further contributions include methods for analyzing this dataset for prevalence of drug entities, drug misspellings and other lexical attributes.

The methods presented here are able to accurately identify drugs withdrawn from market as well as drugs that have been put on the FDA's watchlist. The way in which people discussed drugs before the drugs were placed on the watchlist was used to classify them. A drug withdrawn from market was accurately identified based on data prior to its withdrawal.

7.1 KL Divergence

Table 9 compares the differences between the watchlist and non-watchlist distributions and between the Google and Reuter's distributions. The differences in the distributions are smaller than when compared to Google or Reuters, which is expected. However, it is still difficult to quantify the amount of difference between the distributions. This is a limitation of this metric. Future work might include other distributions to compare against or the same distribution with perturbations to demonstrate how these differences impact the KL divergence scores.

Most of the differences between the distributions are due to common words, which is useful for machine learning classifiers. It is more useful to have classifiers trained on common words since they are more likely to exist in more messages leading to

classifiers that work on a greater number of messages. While the "distance" between the two distributions is difficult to quantify it is encouraging to see that the differences are attributable to non-drug or disease specific terminology.

7.2 Sentiment

We demonstrate the ability to track sentiment of messages containing drug mentions over time. There are statistically significant lexical differences between the word distributions of watchlist and non-watchlist messages within the same forum. These differences tend to track news items about drugs showing. While the differences are demonstrated numerically, a numerical result can hide some of the complexity in the language distributions.

A similar numerical result can be obtained from very different word distribution pairs, for example in one case the differences can occur in common words like determiners or pronouns that occur in both word distributions, whereas in another most of the differences are attributable to unique or rare words. It appears that there are differences in general word features that exist in both sets, which are helpful since the classifiers will hopefully be able to discern patterns based upon these common features, and are generalizable to previously unseen examples. These results agree with results in the KL divergence section leading to the conclusion that classifiers can be built over general/common word features.

7.3 Machine Learning Experiments

Quantitative analysis of the classifiers is performed on various combinations of feature sets and algorithms. The knowledge rich features, disease and drug names seemed to perform the best overall. However, these numbers without some comparison

or task to perform, give no idea of how well this task is performed. What is 85% accuracy? What is "good enough" when predicting drug safety? How accurately does a predictive classifier need to be in order to be useful?

Implementations of machine learning classifiers have real world costs and consequences. Costs are involved in building classifiers and annotating data. Similar costs in human capital are involved in evaluating the results of these classifiers.

7.3.1 Watchlist Predictions

Chapter 6 provided a quantitative analysis of classification results. The raw numbers look somewhat disappointing due to the small performance difference between a machine learning classifier and always labeling a drug as non-watchlist. However, we now look for more interesting qualitative results of the data. The low performance scores is an indicator that many drugs are boundary cases, that a classifier has a hard time differentiating between the watchlist and non-watchlist drugs and are close together in n-dimensional space because the features do not adequately separate the instances.

There are several reasons for the low performance scores: the instances are nearly indistinguishable; the algorithm does not efficiently combine or utilize features; the features were poorly chosen and therefore the instances look similar within the feature speace (for example trying to distinguish between a sphere and a conical object from a 2-d representation looking downward at the objects); instances are mislabeled. This third reason is what we are after; a watchlist drug is a non-watchlist drug that has been declared so by some outside entity (the FDA). Drugs can cause potentially serious adverse effects but until the FDA puts them on a watchlist they are not watchlist drug.

Here, we look at boundary drugs to hopefully identify drugs that are not considered watchlist ones but are adverse effect causing relative to other watchlist drugs.

Future withdrawal or watch list drugs are predicted based upon multiple false positives from many different classifiers. These drugs are false positives in the sense that they are not currently watchlist or recalled drugs. However, these drugs at some point in the future could become watchlist or withdrawn drugs. The most significant outcome was the identification of Sibutramine or Meridia, a weight loss drug. Our dataset only has posts up to a year before it was put on a watchlist then subsequently withdrawn from the European market. Further this drug has been recently voluntarily withdrawn from the market by its manufacturer, almost three years beyond the data we currently have. Further work is necessary to determine if predictions are accurate and further refinement of the methods discussed in this dissertation may give further confidence in the other predictions.

Table 22: Table of drugs withdrawn from the market with their associated scores for the two experiments. The first two columns are the scores associated with each experiment. The following columns are the position of each drug within the list of results for each experiment.

Drug	Score Exp 1	Score Exp 2	Rank Exp 1	Rank Exp 2
Palladone	1.929	10.89	33	4
Trovan	1.761	10.89	40	5
Vioxx	1.62	10.24	50	6
Baycol	0.03	0.04	117	107

These scores indicate the classifiers were better at identifying the drugs withdrawn from market in the second run. The raw scores of Palladone, Trovan and Vioxx were almost a magnitude of order higher and similarly ranked almost a magnitude higher in the list. Disappointingly Baycol's score did not improve much and its relative rank while higher

was not significantly higher. However, all drugs withdrawn from market had a score and were identified in both runs leading indicating that this method is promising in detecting adverse event causing drugs.

Drug safety is not black and white, but a gradient. Some of the decisions made in this dissertation were almost arbitrary, for example choosing to differentiate between active safety alert drugs versus non-current safety alert drugs. For instance one might choose all drugs that were not relabeled, or alternatively only drugs recalled from market. These drug divisions give different quantities of positive and negative examples and one must balance the sensitivity of the resulting classifier with available data. If one chooses drugs withdrawn from market, there would be fewer than 15 positive examples, not enough to build a classifier. This dissertation demonstrates that patient derived data is useful in building classifiers for this particular division of data. We cannot predict that online health forum posts would be useful for other classification tasks or that other types of online health forum posts from other sources would similarly be appropriate. Other forums may be narrower in scope and have skewed or otherwise biased information.

Machine learning classifiers make decisions based upon previously seen examples. While the previous decisions might be biased, future decisions using machine learning will be biased in the same way, giving consistency to the decisions made. We present methods using machine learning classifiers in hopes of providing information to people based upon precedent; to forum contributors as well as agencies or groups to guide their individual decision-making processes.

The machine learning classifiers developed perform better than a naive classifier predicting all drugs as non-watchlist ones. A naïve classifier would obtain 86.54%

accuracy; however the F1 score would be 0 and the AUC score would be 0.5. Looking at the top performing classifiers, a normalized SVM using the disease and AERS lexicon yields 90.33% accuracy. A normalized SVM using drugs, disease and AERS lexicon yields a F1 score of 0.4762 and an un-normalized Naïve Bayes classifier yields an AUC score of 0.7592. Contrary to earlier work demonstrating that group sentiment concerning a particular drug can be tracked, the sentiment word features themselves seemed not to improve classification performance regarding identification of watchlist or potential watchlist drugs. However, words associated with negative sentiment such as pain are contained within the AERS lexicon.

The performance of these classifiers is better than a naïve always non-watchlist classifier because the machine learning ones are able to actually predict positive (watchlist) instances. While predicting positives also leads to generating false positives, false positives with regard to the pharmacovigilance task are preferred over false negatives. A positive (even false positive) instance will generate a hypothesis or at least participate in the information gathering and evaluation process for a drug, whereas a false negative will incorrectly add credence about a drug's safety.

Further work is needed in looking into the "false positives" generated by our system. We demonstrate a watchlist prediction model utilizing false positives on current data and have successfully predicted a drug that is currently undergoing FDA review and has been pulled from the EU and UK markets with data that predates these events by a year.

We are able to demonstrate effective identification of these watchlist and potential watchlist drugs based upon people's mentions within online health forum messages. This

dissertation serves as an exploration in the use of health forum messages for datamining tasks. Obvious hurdles we have dealt with include the lack of suitable amounts of training data as well as nascent tools to work with this type of data.

The contributions of this work include the exploration and annotation of health forum data, providing statistics over the incidence of disease, drug, medical and adverse effect terminology mentions. We have developed algorithms to identify and differentiate spelling mistakes from foreign language words. We have also provided statistics over the provenance of foreign language containing messages and spelling errors within a online health forum corpus.

We demonstrate that specialty lexicons such as drug and disease names perform better than word features using BNS indicating that world knowledge imparted through the specialty lexicon is important. While it is disappointing that drug and disease names are the most salient features in identifying possibly harmful drugs it is intuitive that drugs used to treat certain diseases are more likely to cause harmful effects. For example drugs to treat life threatening conditions such as AIDS or cancer are more taxing on the body and therefore are likely to cause more harmful effects. This effect is seen in the prediction of potential watchlist drugs. Psychiatric medications, strong pain relievers as well as broad-spectrum items such including over the counter pain relievers or vitamins and supplements are often flagged as potentially harmful.

We contribute a new methodology and toolset to aid in the pharmacovigilance task. We demonstrate that patient derived data can be used to differentiate and predictively classify drug safety leading to hypothesis generation about potential drug safety for future drugs.

As stated earlier, drug safety is relative and is determined in part by people's choices about drug benefits compared to their drawbacks. With any drug there are potential side effects and society deems if those side effects are sufficiently bad to label a drug or remove it from market. There are many differing views across many societies with drugs that are available in the United States but not available elsewhere and vice-versa. This dissertation makes no ethical or moral consideration of any drug. Who is to determine drug safety if a drug can potentially save a life but also has the potential to kill a small percentage of the persons taking it? Would the same drug be better if it only had that side effect on a smaller percentage of the people taking it? Is there a point in which the likelihood of dying from a drug is small enough it outweighs its potential for good?

This dissertation constructs predictive classifiers for drug safety based upon the societal norms that have judged past drugs – the norms and judgments that created the training data we use. I aim to provide hypothesis and information not only to drug safety organizations or companies, but also to individuals who ultimately have to make the decision to put a drug in their body and want to enable them to make informed decisions.

7.4 Limitations

The method presented identifies 127 candidate drugs with scores ranging from 79.4 to 0.02. The 127 candidates are a relatively small number of drugs compared to the 11,706 prescription drugs, 390 over the counter drugs, and numerous herbal remedies that exist. However, it is a larger percentage of the 575 total drugs we had data for, illustrating that we have data for relatively few drugs compared to the total numbers of existing drugs. This type of system requires large amounts of data, a weakness of many

machine-learning techniques. However, we have demonstrated that this methodology could be useful in identifying drugs for further study.

This technique assesses which drugs are discussed in similar ways; like comparing a drug to its peers. This may not be a fair assessment of a drug and may inaccurately predict a drug for safety warning or withdrawal based upon its perception. Perception of a drug is different from its actual effects. Many people may like their drug despite low efficacy or serious side effects.

A current drawback of this method is that we aggregate all messages across all disease groups. Drugs have different audiences and certain segments of the population are at greater risk for specific diseases. For example women are more likely to suffer from multiple sclerosis than men, therefore the question remains is it valid to group drugs together that are targeted to different segments of the population? It is unknown whether or not it is correct to group all drugs and messages together.

The number of messages for each drug is not evenly distributed, for example the numbers of drugs mentioning nonsteroidal anti-inflammatory drugs (NSAIDs) are much greater than those mentioning Tysabri (a multiple sclerosis drug). Many NSAIDs such as Aleve are available over the counter and have multiple applications. This differs significantly from a narrow purpose drug used by a smaller population. As stated previously the prevalence of vitamins or over the counter painkillers that seem innocuous abound on both lists. This might be attributed to their wide spread use among many conditions and in combination with many different drugs. Both experiments demonstrate a relatively high score for acetaminophen and acetaminophen containing products. However the causality of scores is not established. Like many other machine learning

and classification tasks, even if the features used by a classifier are known, there is no established causation, only the correlation between feature and class. It remains to be seen if the outcome is corroborating the recent allegations over the safety of acetaminophen with regard to overdosing and safety of children products or due to the their widespread use and association with many different drugs and diseases.

The filtering of false positives is outside of the scope of this dissertation. However one could imagine using this approach as a signal detection method and combining the resulting signals with signals generated from other sources. Currently the FDA generates signals over medical claims data from the Mini-Sentinal initiative, spontaneous reporting systems (AERS/VAERS), as well as clinical trials. This data source and technique represents just one tool out of many to generate signals of possible adverse effect correlated drugs.

We rely on words and groups of words as features. Spelling errors introduce problems. Misspelled words are not correctly attributed to their correct word or phrase resulting in lower classification accuracy. Greater numbers of messages help to mitigate this problem but the problem is more pronounced for drugs with fewer mentions or that exist in disease communities with cognitive impairment. Drugs that are discussed in ways that are different from most of the training set will also be misclassified. Drugs causing rare but serious effects, or effects that are different from the other watchlist drugs is a possible cause that Baycol score was so low.

This method will classify drugs that are discussed in similar ways. However there are no controls for volume of messages. Intuitively messages that have greater volumes of negative sentiment or adverse effects than other drugs should have a higher score. An

implementation of reporting ratios which rank drugs with greater proportions of negative sentiment containing messages compared to it's peers will score higher.

The language processing techniques used were coarse grained; no spelling correction was performed or more sophisticated syntactic parsing or anaphora resolution. Threading was also not utilized so "me too" messages did not provide any additional information. Both messages were not counted or used if a message stating a dislike of a drug or containing an adverse effect was followed by another person in the forum saying, "same here" or "me too". Ideally we would want to attach some score or weight to these types of messages.

This method is a black box; the inputs are groups of messages with drug mentions and the output a single score. This output is lacking in explanation. If one used a single Naïve Bayesian classifier the sets of words or word groups that contributed most to the groups classification could be examined. However this approach is unfeasible when using multiple classifiers and multiple methods. A justification or explanation for a prediction might be more satisfying.

A limitation of this work is that crowd intelligence can fail due to emotional factors. People want to belong and succumb to peer pressure. Discussing ones' health is a highly personal and emotional subject. The media or other events could also influence people's perceptions of drugs and cause them to talk about things in similar ways.

7.5 Future Work

This dissertation explores the online health message data from a purely technical standpoint and does not look at the societal connections and impacts of such information and technology.

Furthermore, we ignore much of the sociological data that could provide useful information to improve prediction performance merely because we do not currently understand how to extract and utilize such information. Understanding people's interpersonal relationships within these online forums as well as this implicit social hierarchy could provide useful information on data provenance – providing data quality metrics that might be used to filter SPAM or unhelpful or derogatory messages. We all implicitly evaluate data quality, for example giving more credence to advice given by a good friend with a background in the problem situation than perhaps an acquaintance with a background in a different discipline. Many online message boards try to make such information explicit; for example depicting the number of posts by a user, the number of helpful responses, the length of forum membership, etc. All of these factors need exploration to possibly develop a data provenance or quality metric.

This dissertation did not explore many of the aspects of message data that exist. For example, times beyond pre and post watchlist warnings were not explored. The way people discuss a drug could change over time. For example, an avalanche effect could occur right before a watchlist report is released due to information cascades. A more ideal situation would be to identify a watchlist drug candidate before an avalanche occurred, decreasing the amount of time to detect a watchlist drug, though an avalanche effect could provide more confidence in a prediction.

I also believe that colloquial tools for natural language processing would also help to improve classification performance. To date few tools are trained on colloquial data and medical data. Currently there are no tools trained on colloquial medical text like those found in online health forum data. Further compounding this is the lack of

annotated corpora for this task. As such, regular NLP tools such as part of speech taggers, syntactic parsers and other tools work poorly on such data. Spelling errors, capitalization issues, non-grammatical sentences and html and java script tags within this text make it difficult for even simple part of speech tagging. It is reasonable to believe that such tools would provide helpful information since they have proved useful in many other classification tasks.

However, each of these improvements in NLP is a milestone in and of themselves. At first glance spelling correction would appear to be easy. However without knowing the prior distribution of drug names compared to other common words, it is difficult to use tools such as edit distance or approximate string matching. There is a drug named Drize. Instances where people talk about driving to a hospital would incorrectly be corrected to Drize. There are similar problems for Doctar and doctor, Blistex and blister, etc.

While there are relatively few polysemous drug names, they do exist. Some examples include: Amen, Commit, Compete, Control, Cope, Liberate and Muse. While these drugs did not exist in our corpus, a comprehensive drug safety system would need to account for drug names like these. In this case, word sense disambiguation or some sort of statistical named entity recognizer using part of speech tags and other features is necessary.

Currently we are also using general-purpose lexicon for sentiment analysis. Trained sentiment classifiers or weighted lexicon should similarly improve sentiment scoring for this task. It is apparent that sentiment is domain specific. For example in the

Rotten Tomato lexicon, "explosion" scores high in the sentiment lexicon; however explosions are generally thought of negatively.

Further study into the specifics of disease communities is also necessary. Each community such as Multiple Sclerosis or Diabetes has disease specific terminology as well as colloquial terms, for example dx for diagnosis, TSH for Thyroid-stimulating hormone or Thyrotropin a test to help diagnosis thyroid disorders or to monitor hyperthyroidism. While different terminology exists for each of these disease groups it is unclear whether or not the way in which people discuss drugs differs enough to build different watchlist classifiers for each of the disease groups.

Many of these problems are due to one of two things, lack of sufficient data or lack of hand labeled training data. The development of NLP tools such as part of speech taggers or sentiment classifiers depends on having annotated data.

In conclusion, I have demonstrated a scalable technique that needs little manually annotated training data, which is a limitation of the application of machine learning for many tasks. I believe that this method can be generalized to different types of data sources such as twitter or blogs with little augmentation due to the lack of custom features or advanced natural language processing techniques such as full syntactic parsing. I have demonstrated that our method was able to identify drugs removed from market both when the data was intentionally mislabeled and trained upon as well as when it was used only for testing. I propose that this method could be used as a coarse signal detection technique that can augment existing SRS data and methods using unstructured information directly found in online sources.

References

Angell, M. (2009). Drug companies and doctors: A Story of corruption. *New York Review of Books*, 56(1).

Anderson, E. (1957). A semigraphical method for the analysis of complex problems. *Proc. of the National Academy of Sciences* (PNAS), 13, 923-927.

Baily, D.G., Malcom, J., Arnold, O. and Spence, J.D. (1998). Grapefruit juice-drug interactions. *Br J Clin Pharmacol*., 46(2), 101-110.

Bartlett D.F. (1999). The new health care consumer. *J Health Care Finance*, 25(3), 44-51.

Bates, D.W., Evans, R.S., Murff, H., Stetson, P.D., Pizziferri, L. and Hripcsak, G. (2003). Detecting adverse events using information technology. *JAMIA*, 10, 115-128.

BBC News. (2010, January 22). Top obesity drug sibutramine being suspended. Retrieved from http://news.bbc.co.uk/2/hi/health/8473555.stm

Beamer, B., Bhat, S., Chee, B., Fister, A., Rozovskaya, A. and Girju, R. (2007). UIUC: A knowledge-rich approach to identifying semantic relations between nominals. In Proc. of *The Semantic Evaluation Workshop* (SemEval 2007) in conjunction with ACL, Prague, June 2007.

Bertin, J. (1983). *Semiology of Graphics*. Wisconsin: University of Wisconsin Press.

Bikhchandani, S., Hirshleifer, D. and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5), 992-1026.

Bren, L. (2001). Frances oldham kelsey: FDA medical reviewer leaves her mark on history. FDA Consumer Magazine, March-April.

Brin, S. (1998). Extracting patterns and relations from the world wide web. In Proc. *Conf. of Extending Database Technology. Workshop on the Web and Databases*.

Brownstein, J.S., Sordo, M., Kohane, I.S. and Mandl, K.D. (2007). The tell-tale heart: Population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. *PLoS ONE* 2007, 2(9), e840.

Burton, T.M., Callahan P. (2003). Vioxx study sees heart-attack risk: Merck funded research after concerns were raised about its painkilling drug. *Wall Street Journal* October 30, 2003; B1-2.

Calabretta, N. (2002). Consumer-driven, patient-centered health care in the age of electronic information. *J Med Libr Assoc*., 90(1), 32-37.

Chee, B. (2010). Together in sickness and health: Homophily in online health forums. In Proc. of the *iConference*, Urbana, IL, January 2010.

Chee, B., Berlin, R. and Schatz, B. (2009). Measuring population health using personal health messages. In Proc. of the *American Medical Informatics Association Annual Symposium* (AMIA 2009), San Francisco, November, 2009.

Chee, B., Karahalios, K.G. and Schatz, B. (2009). Social visualization of health messages. In Proc. of the *Hawai'i International Conf. on System Sciences* (HICSS-42), Waikoloa, HI, January 2009.

Chee, B. and Schatz, B. (2007). Document clustering using small world communities. In Proc. of the 7th ACM/IEEE-CS *Joint Conf. on Digital Libraries* (JCDL '07), Vancouver, June, 2007.

Chee, B., Schatz, B. and Berlin, R. (2009). Information visualization of drug regimens from health messages. In Proc. of the *International Conf. on Health Informatics* (HEALTHINF 2009) in conjunction with BIOSTEC 2009, Porto, January 2009.

Chung, C.K. and Pennebaker, J.W. (2005). Assessing quality of life through natural language use: Implications of computerized text analysis. In W.R. Lenderking and D.A. Revicki (eds.), *Advancing health outcomes research methods and clinical applications* (pp 79-94). Washington, DC: Degnon Associates.

Cortes, C. and Vapnik, V. (1995). Support-vector network. *Machine Learning*, 20, 273-297.

Cotton, S.R. and Gupta S.S. (2004). Characteristics of online and offline health information seekers and factors that discriminate between them. *Social Science and Medicine,* 59, 1795-1806.

Cranor, L.F., Reagle, J. and Ackerman, M.S. (1999). Beyond concern: Understanding net users' attitudes about online privacy. Technical Report TR 99.4.3, *AT&T Labs-Research*, April 1999.

Cyber Dialogue and Institue for the Future. (2000). Ethics survey of consumer attitudes about health websites. Retrieved from *California HealthCare Foundation* website: http://www.chcf.org/topics/view.cfm?itemID=12493.

deHaes, J.C.J.M. and van Knippenberg F.C.E. (1985). The quality of life of cancer patients: A review of the literature. *Soc. Sci. Med.*, 20, 809-817.

Denison, B. (2004). Touch the pain away: New research on therapeutic touch and persons with fibromyalgia syndrome. *Holist. Nurs. Pract.* 18(3):142-151.

Downs, J.S., Holbrook, M.B., Sheng, S. and Cranor, L.F. (2010). Are your participants gaming the system? Screening mechanical turk workers. *CHI*, April 10-15, 2010, Atlanta, GA.

Easley, D. and Kleinberg, J. (2010). *Networks, Crowds and Markets: Reasoning about a Highly Connected World*. New York: Cambridge University Press.

Edwards, I.R. and Aronson, J.K. (2000). Adverse drug reactions: Definitions, diagnosis, and management. *The Lancet* 256(9237), 1255-1259.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D.S. and Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence* 165: 91-134, Essex: Elsevier Science Publishers.

Fawcett, T. (2006). A simple generalization of the area under the roc curve to multiple class classification problems. *Machine Learning* 45, 171-186.

Feinberg, S.E. (1979). Graphical methods in statistics. *The American Statistician*, 33(4), 165-178.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Special Issue on Variable and Feature Selection, *J. of Machine Learning Research*, 3, 1289-1305.

Fox, S. and Jones, S. (2009). *The Social Life of Health Information*. Pew Internet & American Life Project.

Gajadhar J, Green J. (2005). An analysis of nonverbal communication in an online chat group. *EDUCAUSE Quarterly* 2005, 24(4), 63-64.

Garton, L., Haythornthwaite, C. and Wellman, B. (1997). Studying online social networks. *J. of Computer-Mediated Comm.*, 3(1).

Gilbert, E. and Karahalios, K. (2007). CodeSaw: A social visualization of distributed software development. In Proc of *INTERACT* 2007.

Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media. In *Proc. CHI* 2009, ACM Press, 211-220.

Goldman, J., Hudson, Z., and Smith, R. (2000). Privacy: Report on the privacy policies and practices of health web sites. Professional Ethics Report, 13(1).

Gottschalk L.A. and Gleser G.C. (1969). The measurement of psychological states through the content analysis of verbal behavior. Berkeley, CA: University of California Press.

Grandin, T. (1992). Calming effects of deep touch pressure in patients with autistic disorder, college students, and animals. *J. Child Adolesc. Psychopharmacol.*, 2(1), 63-72.

Green D.M. and Swets, J.M. (1966). Signal detection theory and psychophysics. New York: John Wiley and Sons, Inc.

Hauben, M., Madigan, D., Gerrits, C.M., Walsh, L. and Van Pijenbroek, E.P. (2005). The role of data mining in pharmacovigilance. *Expert Opin. Drug Saf.*, 4(5), 929-948.

Hongiman, B., Lee, J., Rothschild, J., Light, P., Pulling, R. M., Yu, T. and Bates, D.W. (2001). Using computerized data to identify adverse drug events in outpatients. *JAMIA,* 8, 254-266.

Jia, H., Lubetkin, E.I., Moriarty, D.G. and Zack, M.M. (2007). A comparison of healthy days and euroqol eq-5d measures in two us adult samples. *Appl. Res. in Qual. Of Life*, 2, 209-221.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Proc of the *Tenth European Conf. on Machine Learning* (ECML), Berlin, Germany, 137-142.

Johnson, J.R. and Temple, R. (1985). Food and drug administration requirements for approval of new anticancer drugs. *Cancer Treat. Rep.*, 69, 1155-1157.

Juni, P., Altman, D.G and Egger, M. (2001). Assesing the quality of controlled clinical trials. *BMJ,* 323, 42-46.

Keerthi, S.S. and Lin, C.-J. (2003). A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. Technical Report, Department of Computer Science, *National Taiwan University*.

Landis, B.J. (1996). Uncertainty, spiritual well-being and psychosocial adjustment to chronic illness. *Issues in Mental Health Nursing*, 17(3), 217-231.

Leamon, R., Wojtulewicz, L., et. al. (2010). Towards internet-age pahrmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In Proc of the *2010 Workshop on Biomedical Natural Language Processing*, 117-125.

Liu, V. and Curran, J.R. (2006). Web text corpus for natural language processing. In Proc. of the *11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, 233–240.

Lubkin, M. and Larsen, P.D. (2008). *Chronic Illness: Impact and Intervention*. Jones & Bartlett Pub.

McLachlan, G.J., Do, K.A., and Ambroise, C. (2004). *Analyzing microarray gene expression data*. Wiley.

Mamdani M., Juurlink D.N., Lee D.S., Rochon P.A., Kopp A., Naglie G., et al. (2004). Cyclo-oxygenase-2 inhibitors versus non-selective non-steroidal anti-inflammatory drugs and congestive heart failure outcomes in elderly patients: a population-based cohort study. *The Lancet*, 363(9423), 1751-1756.

Manning, C., and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Martin, K., Begaud, B., Latry, P., Miremont-Salame, G., Fourrier, A., and Moore, N. Differences between clinical trials and postmarketing use. *Br. J. ClinPharmacol.*, 57(1), 86-92.

MacLeod, J.S. and Austin, J.K. (2003). Stigma in the lives of adolescents with epilepsy: A review of the literature. *Epilepsy & Behavior*, 4(2), 112-117.

McClellan, M. (2007). Drug safety reform at the fda-pendulum swing or systemic improvement? *NEJM*, 356(17), 1700-1702.

McDowell, I. (2006). Measuring health: A guide to rating scales and questionnaires. New York: Oxford University Press.

McPherson, M., Smith-Lovin, L. and Cook, J.M. (2001). Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.*, 27, 415-444.

Morin, K., Rakatansky, H., Riddick, F. A., Morse, L.J., O'Bannon, J.M., Goldrich, M.S., Ray, P., Weiss, M., Sade, R.M. and Spillman, M.A. (2002). Manging conflict of interest in the conduct of clinical trials. *JAMA*, 287(1), 78-84.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, (1–2), 1-135.

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proc. of *the Conf. on Empirical Methods in Natural Language Processing* (EMNLP), 79–86.

Pennebaker, J.W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science*, 8(3), 162-166.

Pennebaker, J.W. and Campbell, R.S. (2000). The effects of writing about traumatic experience. *Clinical Quarterly,* 9, 17-21.

Pennebaker, J.W., Francis M.E. and Booth R.J. (2007). Linguistic inquiry and word count: LIWC 2007. Lawrence Erlbaum Assoc, New Jersey: 2007.

Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford.

Reuters. (2006, December 7). Merck sees slightly higher 2007 earnings. Retrieved from:http://www.nytimes.com/2006/12/07/business/07drug.html?ex=1323147600&en=19d27b5814f1c1e8&ei=5088&partner=rssnyt&emc=rss

Rindfleisch, T.C., Tanabe, L. and Weinstein, J.N. (2000). EDGAR: Extraction of drugs, genes and relations from the biomedical literature. In *Proc. Pacific Symposium on Biocomputing*.

Rosenberg, S.D. and Tucker, G.J. (1979). Verbal behavior and Schizophrenia. *Arch Gen Psychiatry*, 36, 1331-1337.

Rude, S.S., Gortner, E.-M. and Pennebaker, J.W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8), 1121-1133.

Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E. (1998). A bayesian approach to filtering junk e-mail. *AAAI'98 Workshop on Learning for Text Categorization*.

Silver, M. (2002). *Success with Heart Failure*. Cambridge, MA: Perseus Publishing.

Solomon, D.H., Schneeweiss S., Glynn, R.J., Kiyota, Y., Levin, R., Mogun, H. and Avorn, J. (2004). Relationship between selective cyclooxygenase-2 inhibitors and acute myocardial infarction in older adults. *Circulation*, 109, 2068-2073.

Stennar, P.H.D., Cooper, D. and Skevington, S.M. (1993). Putting the q into quality of life: The identification of subjective constructions of health-related quality of life using the q methodology. *Soc Sci Med*, 57, 2161-2172.

Stephens, J. (2006, May 7). Panel faults pfizer in '96 clinical trial in nigeria. Washington Post. Retrieved from: http://www.washingtonpost.com/wp-dyn/content/article /2006/05/06/AR2006050601338.html

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York: Doubleday.

Terry, K. (2009). Patient privacy, the new threats. *Physicians Practice J*, 19(3).

Travaline, J.M., Ruchinskas, R. and D'Alonzo, G.E. (2005). Patient-physician communication: Why and how. *JAOA*, 105(1), 13-18.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proc. of the Association for Computational Linguistics* (ACL), 417–424.

USA Today. (2001, August 8). Fda statement on Baycol withdrawal. Retrieved from http://www.usatoday.com/money/general/2001-08-08-bayer-fda-statement.htm

U.S. Food and Drug Administration. (2006, June 5). Fda approves resumed marketing of tysabri under a special distribution program. Retrieved from http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2006/ucm108662.htm

U.S. Food and Drug Administration. (2004a, September 30). Fda issues public health advisory on vioxx as its manufacturer voluntarily withdraws the product. Retrieved from http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2004/ucm108361.htm

U.S. Food and Drug Administration. (2004b, December 17). Fda statement on the halting of a clinical trial of the cox-2 Inhibitor celebrex. Retrieved from http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2004/ucm108384.htm

U.S. Food and Drug Administration. (1999, June 9). Food and drug administration 09 june 1999 trovan (trovafloxacin / alatrofloxacin mesylate) interim recommendations. Retrieved from http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/DrugSafetyInformationforHeathcareProfessionals/PublicHealthAdvisories/UCM053103

U.S. Food and Drug Administration. (2005, April 7). Important changes and additional warnings for cox-2 selective and non-selective non-steroidal anti-inflammatory drugs (nsaids). Retrieved from http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm150314.htm

U.S. Food and Drug Administration. (2010a, April 30). McNeil consumer healthcare announces voluntary recall of certain OTC infants' and children's products. Retrieved from http://www.fda.gov/safety/recalls/ucm210443.htm

U.S. Food and Drug Administration. (2010b, January 21). Meridia (sibutramine hydrochloride): Follow-up to an early communication about an ongoing safety review. Retrieved from http://www.fda.gov/Safety/MedWatch/SafetyInformation/SafetyAlertsforHumanMedicalProducts/ucm198221.htm

U.S. Food and Drug Administration. (2005, July 14). Palladone (hydromorphone hydrochloride). Retrieved from: http://www.fda.gov/Safety/MedWatch/SafetyInformation/SafetyAlertsforHumanMedicalProducts/ucm152047.htm

U.S. Food and Drug Administration. (2009, June 19). Restoril (temazepam) capsules. Retrieved from http://www.fda.gov/Safety/MedWatch/SafetyInformation/Safety-RelatedDrugLabelingChanges/ucm113808.htm

Van Hulse, J., Khoshgoftaar, T.M. and Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. In Proc. of 24th International Conf. on Machine Learning, Corvallis, OR.

Viégas, F. and Smith, M. (2004). Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspaces. In Proc. of the *Hawai'i International Conf. on System Sciences* (HICSS-37), Waikoloa, HI, January, Hawaii, HI, January 5-8.

Waller, P. (2001). Pharmacoepidemiology – a tool for public health. *Pharmacoepidemiology and Drug Safety*, 10, 165-172.

Walther, J.B. and D'Addario, K.P. (2001). The impacts of emoticons on message interpretation in computer-mediated communication. *Soc. Sci. Comp. Rev.*, 19, 323–345.

Wise, J.A. (1999). The ecological approach to text visualization. *JASIS,* 40(13), 1224-1233.

Weze, C., Leathard, H.L., Grange, J., Tiplady, P. and Stevens, G. (2005). Evaluation of healing by gentle touch. *Public Health*, 119(1), 3-10.

World Health Organization. (1972). International drug monitoring: The role of national centres. Report of a WHO meeting, *World Health Organ. Tech. Rep. Ser.* 498, 1-25.

World Health Organization. (2002). The importance of pharmacovigilance safety monitoring of medicinal products. *World Health Organ.*, 1-52.

Zweig, M.H. and Campbell, G. (1993). Receiver-operating characteristic (roc) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry,* 39(8), 561-577.

Appendix A: FDA Drug List with Important Information

"Advair Diskus" OR ("fluticasone propionate" AND "salmeterol xinafoate")
"Alosetron hydrochloride"
"Ciprofloxacin Extended Release"
"Fentanyl buccal"
"Fluticasone propionate"
"Levothyroxine sodium"
"Plan B" OR levonorgestrel
"Salmeterol xinafoate"
"Serevent Diskus" OR "salmeterol xinafoate"
"Testosterone gel"
Abilify OR aripiprazole
Accolate OR zafirlukast
Alimta OR pemetrexed
Ambien OR "zolpidem tartrate"
AndroGel
Aranesp OR "darbepoetin alpha"
Aspirin
Avandamet OR rosiglitazone OR Avandaryl OR Avandia OR
Avastin OR bevacizumab
Bupivicaine
Buprenorphine
Butisol sodium
Carbrital OR "pentobarbital and carbromall"
Cerebyx
Cerezyme OR imiglucerase
Chlorprocaine
Cialis OR tadalafil
Clozaril OR clozapine
Coly-Mycin M OR colistimethate
Creon OR pancrelipase
Dalmane OR flurazepam
Dilantin
Doral OR quazepam
Drotrecogin alfa activated
Elidel OR pimecrolimus
Eloxatin OR oxaliplatin
Endrate OR "edetate disodium"
Epogen OR "epoetin alfa"
Eprex OR "epoetin alfa"
Erbitux OR cetuximab
Erythromycin
Exjade OR deferasirox
Foradil OR "formoterol fumarate"
Fosphenytoin
Gadolinium
Gemifloxacin mesylate
Geodon OR ziprasidone
Gleevec OR "imatinib mesylate"
Halcion OR triazolam
Heparin
Ibuprofen
Innohep OR tinzaparin

Invirase OR saquinavir
Iplex
Kepivance OR palifermin
Ketek OR telithromycin
Lenalidomide
Levitra OR vardenafil
Lidocaine
Lindane
Lotronex OR "alosetron hydrochloride"
Lunesta OR eszopiclone
Magnivist OR "gadopentetate dimeglumine"
Maxipime OR cefepime
Mecasermin rinfabate recombinant
Mepiivicaine
Meridia OR "sibutramine hydrochloride"
Methadone OR dolophine
Multihance OR "gadobenate dimeglumine"
Naratriptan
Natalizumab
NeutroSpec OR technetium-99m
Nexium OR esomeprazole
Nimotop OR nimodipine
Norelgestromin
Omniscan OR gadodiamide
Onsolis OR "fentanyl buccal soluble film"
OptimarkM OR gasoversetamide
Ortho Evra OR ("norelgestromin" AND "ethinyl estradiol")
Pallodone OR hydromorphone
Pentobarbital and carbromal
Peramivir
Permax OR pergolide
Phenytek
Phenytoin
Placidyl OR ethchlorvynol
Plavix OR "clopidogrel bisulfate"
Prilosec OR omeprazole
Procaine
Procrit OR "epoetin alfa"
Prohance OR gadoteridol
Propoxyphene
Prosom OR estazolam
Protopic OR tacrolimus
Raptiva OR efalizumab
Regranex OR becaplermin
Relenza OR zanamivir
Restoril OR temazepam
Revatio OR sildenafil
Revlimid OR lenalidomide
Risperdal OR risperidone
Ropivicaine
Rozerem OR ramelteon
Seconal OR secobarbital
Seroquel OR quetiapine
Singulair OR montelukast
Sonata OR zaleplon
Spiriva OR "tiotropium bromide"

Suboxone
Subutex
Tamiflu OR "oseltamivir phosphate"
Testim
Thalomid OR thalidomide
Trasylol OR aprotinin
Unithroid OR "levothyroxine sodium"
Valproate
Velcade OR bortezomib
Viagra OR "sildenafil citrate"
Victoza OR liraglutide
Videx OR didanosine
Vioxx OR rofecoxib
Xigris OR "drotrecogin alfa"
Zarfirlukast
Zelnorm OR "tegaserod maleate"
Zetia OR ezetimibe
Zicam
Zocor OR simvastatin
Zyflo and "Zyflo CR" OR zileuton
Zyprexa OR olanzapine

Appendix B: FDA Watch List Drugs

"Diastat AcuDial"
"Diazepam rectal gel"
"Fentanyl buccal"
"Fentanyl transdermal system"
"Fleet Accu-Prep" OR "sodium phosphate"
"Fleet Phospho-Soda" OR "sodium phosphate"
"Mitoxantrone hydrochloride"
"Proquin XR" OR "ciprofloxacin extended release" OR "Proquin" OR "ciprofloxacin"
"Sodium phosphates"
Accupril OR quinapril
Accutane OR isotretinoin
Aceon OR perindopril
Actimmune OR "interferon gamma-1b"
Actonel OR risedronate
Actoplus OR pioglitazone OR Actos
Adderall OR "amphetamine salts"
Aleve OR "naproxen sodium" OR Anaprox
Alli OR orlistat Or Xenical
Altace OR ramipril
Amerge OR naratriptan
Amiodarone AND Simvastatin
Amnesteem OR isotretinoin or Accutane OR Claravis OR Sotret
Aredia OR pamidronate
Avelox OR moxifloxacin
Axert OR "almotriptan malate"
Bextra OR valdecoxib
Biaxin OR clarithromycin
Boniva OR ibandronate
Botox OR OnabotulinumtoxinA
Botulinium Type A and B
Brovana OR "arformoterol tartrate"
Budesonide
Byetta OR exenatide
Campath OR alemtuzumab
Capoten OR captopril
Carbatrol OR carbamazepine
Celebrex OR celecoxib
Celexa OR "citalopram hydrobromide"
CellCept OR "mycophenolate mofetil"
Chantix OR varenicline
Cimzia OR "certolizumab pergol"
Cipro OR ciprofloxacin
Cipro XR OR "ciprofloxacin extended release"
Citalopram
Codeine
Colcrys OR colchicine
Colistimethate
Compazine OR prochlorperazine
Cordarone OR amiodarone
Crestor OR "rosuvastatin calcium"
Cylert OR pemoline
Cymbalta OR duloxetine

Cytotec OR misoprostol
Depacon OR valproate
Depakene OR valproate OR Depakene
Didronel OR etidronate
Diprivan OR propofol
Duragesic OR "fentanyl"
Dysport OR abobotulinumtoxinA
Effexor OR venlafaxine
Enalapril AND Enalaprilat
Enbrel OR etanercept
Equetro OR carbamazepine
Ezetimibe AND simvastatin
Factive OR gemifloxacin
Felbatol OR felbamate
Fentora OR "fentanyl buccal tablets"
Floxin OR ofloxacin
Fluvoxamine
Fosamax OR alendronate
Frova OR "frovatriptan succinate"
Gabitril OR tiagabine
Haldol OR haloperidol
Humira OR adalimumab
Imitrex OR sumatriptan
Iressa OR gefitinib
Januvia OR "sitagliptin "
Keppra OR levetiracetam
Lamictal OR lamotrigine
Levaquin OR levofloxacin
Lexapro OR escitalopram
Lotensin OR benazepril
Loxitane OR loxapine
Lyrica OR pregabalin
Mavik OR trandolapril
Maxalt
Mellaril OR thioridazine
Mifeprex OR mifepristone
Minirin OR desmopressin OR DDAVP
Moban OR molindone
Monopril OR fosinopril
Myfortic OR "mycophenolate acid"
Myobloc
Naprosyn OR naproxen
Navane OR thiothixene
Neoral OR cyclosporine OR Sandimmune
Neurontin OR gabapentin
Noroxin OR norfloxacin
Novantrone OR "mitoxantrone hydrochloride"
Ofloxacin
Optison OR "perflutren protein-type A microspheres" OR perflutren OR Definity
Orap OR pimozide
Osmoprep OR "sodium phosphate"
Pacerone OR amiodarone
Paxil OR paroxetine
Performist OR "formoterol fumarate"
Phenergan OR promethazine
Prezista OR darunavir

Prinivil OR lisinopril OR Zestril
Prolixin OR fluphenazine
Propylthiouracil
Prozac OR fluoxetine
Rapamune OR sirolimus
Razadyne OR galantamine
Reclast OR "zoledronic acid"
Relpax OR eletriptan
Remeron OR mirtazapine
Remicade OR infliximab
RimabotulinumtoxinB
Rituxan
Rituximab
Rizatriptan
Rocephin OR ceftriaxone
Rosuvastatin
Serzone OR nefazodone
Simponi OR golimumab
Simvastatin OR Amiodarone
Skelid OR tiludronate
Stelazine OR trifluoperazine
Stimate Nasal Spray OR desmopressin
Strattera OR atomoxetine
Symbicort OR ("budensonide" AND "formoterol fumarate")
Symbyax OR (olanzapine AND fluoxetine)
Tegretol OR carbamazepine
Tequin OR gatifloxacin
Thorazine OR chlorpromazine
Topamax OR topiramate
Trilafon OR perphenazine
Trileptal OR oxcarbazepine
Tussionex OR hydrocodone
Tysabri OR natalizumab
Univasc OR moexipril
Vasotec OR (enalapril AND enalaprilat)
Visicol OR "sodium phosphate"
Vivitrol OR naltrexone
Vytorin OR (ezetimibeANDsimvastatin)
Wellbutrin OR bupropion
Xolair OR omalizumab
Ziagen OR "abacavir sulfate"
Zoloft OR sertraline
Zometa OR "zoledronic acid"
Zomig OR zolmitriptan
Zonegran OR zonisamide
Zyban OR "bupropion hydrochloride"
Zyvox OR linezolid

Appendix C: String Distance Between Drug Mentions

The data is in the format:
Drug Name 1
Drug Name 2
Average Standard Deviation Median Minimum Maximum

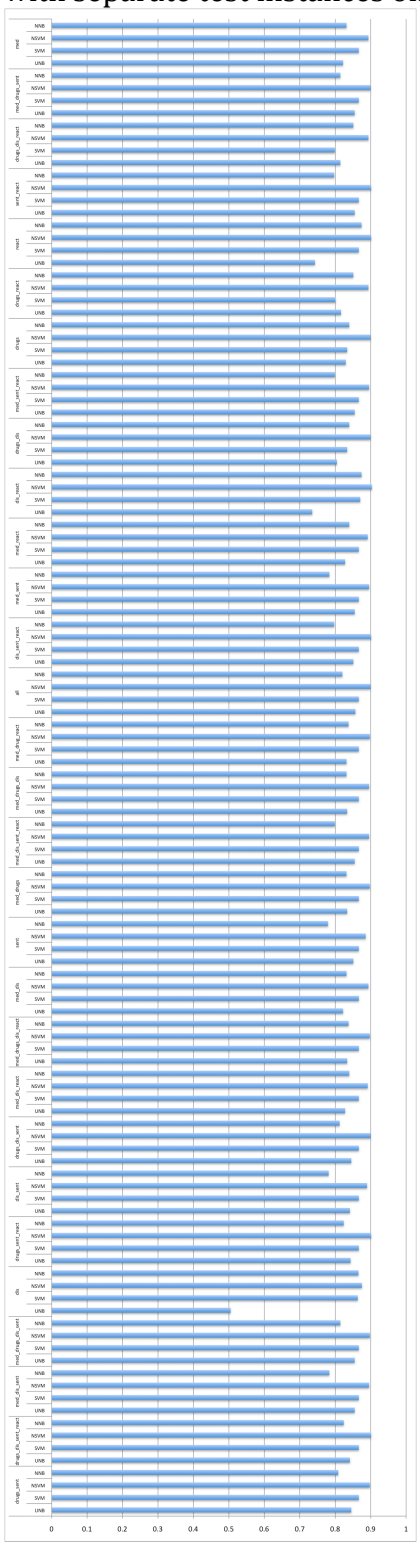
```
message:("sildenafil" OR "Revatio" OR "Viagra")
message:("tadalafil" OR "Adcirca" OR "Cialis")
9.67425      296.0729     11.0   7.0    12165.0

message:("tadalafil" OR "Adcirca" OR "Cialis")
message:("vardenafil" OR "Levitra")
28.476190476190474  85.36018922637123   10.0   8.0    401.0

message:("carisoprodol" OR "Soma" OR "Vanadom")
message:("tramadol" OR "Ryzolt" OR "Ultram" OR "Ultram ER")
106.48780487804878  147.9179032354129   26.0   5.0    484.0

message:("sildenafil" OR "Revatio" OR "Viagra")
message:("vardenafil" OR "Levitra")
76.85082872928177   101.62253064857481  42.0   8.0    699.0

message:("ascorbic acid" OR "Acerola" OR "Ascor L 500" OR "Ascorbic Acid Quick
    Melts" OR "Ascot" OR "Betac" OR "C-Time" OR "C/Rose Hips" OR "Cecon" OR
    "Cemill" OR "Cemill 1000" OR "Cemill 500" OR "Cenolate" OR "Centrum Singles-
    Vitamin C" OR "Cevi-Bid" OR "Ester-C" OR "N Ice with Vitamin C" OR
    "Protexin" OR "Sunkist Vitamin C" OR "Vicks Vitamin C Drops" OR "Vitamin C"
    OR "Vitamin C TR" OR "Vitamin C with Rose Hips")
message:("vitamin e" OR "Alpha E" OR "Amino-Opti-E" OR "Aquasol E" OR
    "Aquavite-E" OR "Centrum Singles-Vitamin E" OR "E Pherol" OR "E-400 Clear"
    OR "Nutr-E-Sol")
217.0563829787234   844.3520141274772   21.0   10.0   19301.0

message:("alprazolam" OR "Alprazolam Intensol" OR "Niravam" OR "Xanax" OR
    "Xanax XR")
message:("diazepam" OR "Diastat" OR "Diastat AcuDial" OR "Diastat Pediatric" OR
    "Diazepam Intensol" OR "Valium" OR "Valrelease")
163.68214285714285  552.1191780555655   10.0   6.0    6112.0

message:("fluoxetine" OR "Prozac" OR "Prozac Weekly" OR "Rapiflux" OR "Sarafem"
    OR "Selfemra")
message:("sertraline" OR "Zoloft")
483.3007246376812   1135.2676128879045  86.0   8.0    10377.0

message:("diazepam" OR "Diastat" OR "Diastat AcuDial" OR "Diastat Pediatric" OR
    "Diazepam Intensol" OR "Valium" OR "Valrelease")
message:("sildenafil" OR "Revatio" OR "Viagra")
23.142857142857142  33.64054712916879   11.0   8.0    206.0

message:("fluoxetine" OR "Prozac" OR "Prozac Weekly" OR "Rapiflux" OR "Sarafem"
    OR "Selfemra")
message:("paroxetine" OR "Paxil" OR "Paxil CR" OR "Pexeva")
595.9365750528541   1122.2559869325905  151.0  7.0    9644.0

message:("acetaminophen" OR "Acephen" OR "Actamin" OR "Adprin B" OR "Anacin
    Aspirin Free" OR "Apra" OR "Atasol" OR "Bromo Seltzer" OR "Children's
    ElixSure" OR "Children's Silapap" OR "Children's Tylenol" OR "Dolono" OR
    "Ed-APAP" OR "Elixsure Fever/Pain" OR "Febrol Solution" OR "Feverall" OR
    "Genapap" OR "Genebs" OR "Infants' Tylenol" OR "Jr. Tylenol" OR "Mapap" OR
```

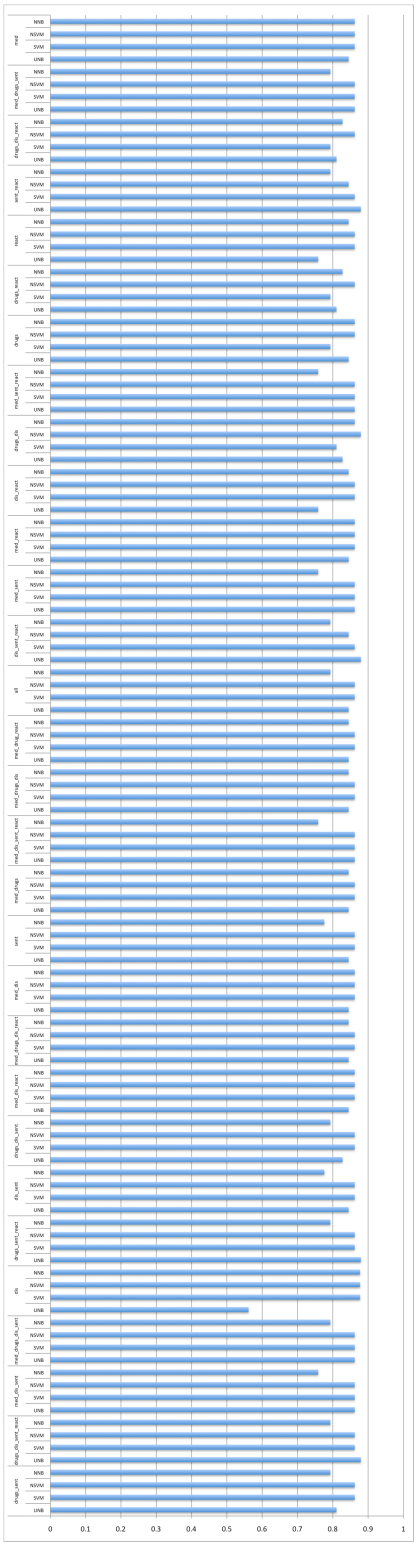
```
    "Mapap Arthritis Pain" OR "Mapap Children's" OR "Mapap Infant Drops" OR
    "Mapap Meltaway" OR "Mapap Rapid Release Gelcaps" OR "Mapap Rapid Tabs" OR
    "Pain-Eze" OR "Q-Pap" OR "Q-Pap Extra Strength" OR "Silapap Childrens" OR
    "Silapap Infants" OR "St. Joseph Aspirin-Free" OR "Tactinal" OR "Tempra" OR
    "Tempra Quicklets" OR "Tycolene" OR "Tylenol" OR "Tylenol 8 Hour" OR
    "Tylenol Arthritis Pain" OR "Tylenol Extra Strength" OR "Tylenol GoTabs" OR
    "Tylenol Sore Throat Daytime" OR "Tylophen" OR "Uniserts" OR "Vitapap")
message:("ibuprofen" OR "Advil" OR "Advil Children's" OR "Advil Junior
    Strength" OR "Advil Liqui-Gels" OR "Advil Migraine" OR "Caldolor" OR
    "Children's Motrin" OR "Childrens Ibuprofen Berry" OR "Genpril" OR "Haltran"
    OR "IBU" OR "IBU-200" OR "Midol IB" OR "Midol Maximum Strength Cramp
    Formula" OR "Motrin" OR "Motrin Childrens" OR "Motrin IB" OR "Motrin Infant
    Drops" OR "Motrin Junior Strength" OR "Motrin Migraine Pain" OR "NeoProfen"
    OR "Nuprin" OR "Q-Profen")
141.33816863100634  431.6383666324089   14.0   6.0    6402.0

message:("ascorbic acid" OR "Acerola" OR "Ascor L 500" OR "Ascorbic Acid Quick
    Melts" OR "Ascot" OR "Betac" OR "C-Time" OR "C/Rose Hips" OR "Cecon" OR
    "Cemill" OR "Cemill 1000" OR "Cemill 500" OR "Cenolate" OR "Centrum Singles-
    Vitamin C" OR "Cevi-Bid" OR "Ester-C" OR "N Ice with Vitamin C" OR
    "Protexin" OR "Sunkist Vitamin C" OR "Vicks Vitamin C Drops" OR "Vitamin C"
    OR "Vitamin C TR" OR "Vitamin C with Rose Hips")
message:("vitamin a" OR "A-25" OR "A/Fish Oil" OR "Aquasol A")
169.63956639566396  722.1454430073155   16.0   10.0   15420.0

message:("paroxetine" OR "Paxil" OR "Paxil CR" OR "Pexeva")
message:("sertraline" OR "Zoloft")
379.21270718232046  825.3031116495735   89.0   7.0    6691.0

message:("carisoprodol" OR "Soma" OR "Vanadom")
message:("sildenafil" OR "Revatio" OR "Viagra")
146.26315789473685  159.7243714095961   94.0   6.0    580.0

message:("diazepam" OR "Diastat" OR "Diastat AcuDial" OR "Diastat Pediatric" OR
    "Diazepam Intensol" OR "Valium" OR "Valrelease")
message:("tadalafil" OR "Adcirca" OR "Cialis")
24.25  14.174507634012077  25.5   7.0    39.0

message:("sibutramine" OR "Meridia")
message:("tadalafil" OR "Adcirca" OR "Cialis")
69.66666666666667   37.18960428220051   67.5   17.0   77.0

message:("vitamin a" OR "A-25" OR "A/Fish Oil" OR "Aquasol A")
message:("vitamin e" OR "Alpha E" OR "Amino-Opti-E" OR "Aquasol E" OR
    "Aquavite-E" OR "Centrum Singles-Vitamin E" OR "E Pherol" OR "E-400 Clear"
    OR "Nutr-E-Sol")
243.67532467532467  389.09838605122957  73.5   10.0   2701.0

message:("diazepam" OR "Diastat" OR "Diastat AcuDial" OR "Diastat Pediatric" OR
    "Diazepam Intensol" OR "Valium" OR "Valrelease")
message:("zolpidem" OR "Ambien" OR "Ambien CR" OR "Edluar" OR "Zolpimist")
713.9166666666666   1212.4585126624072  95.5   8.0    3597.0

message:("carisoprodol" OR "Soma" OR "Vanadom")
message:("vardenafil" OR "Levitra")
25.5   6.363961030678928   25.5   21.0   0.0

message:("alprazolam" OR "Alprazolam Intensol" OR "Niravam" OR "Xanax" OR
    "Xanax XR")
message:("tadalafil" OR "Adcirca" OR "Cialis")
253.57142857142858  334.0248778872901   69.0   8.0    734.0

message:("phentermine" OR "Adipex-P" OR "Ionamin" OR "Obenix" OR "Oby-Cap" OR
    "Pro-Fast SA" OR "Teramine" OR "Zantryl")
message:("sildenafil" OR "Revatio" OR "Viagra")
```

```
239.73809523809524  301.42264788780886  171.0  7.0    1338.0

message:("finasteride" OR "Propecia" OR "Proscar")
message:("sildenafil" OR "Revatio" OR "Viagra")
103.11111111111111  147.36048015363926  28.0   21.0   391.0

message:("ascorbic acid" OR "Acerola" OR "Ascor L 500" OR "Ascorbic Acid Quick
    Melts" OR "Ascot" OR "Betac" OR "C-Time" OR "C/Rose Hips" OR "Cecon" OR
    "Cemill" OR "Cemill 1000" OR "Cemill 500" OR "Cenolate" OR "Centrum Singles-
    Vitamin C" OR "Cevi-Bid" OR "Ester-C" OR "N Ice with Vitamin C" OR
    "Protexin" OR "Sunkist Vitamin C" OR "Vicks Vitamin C Drops" OR "Vitamin C"
    OR "Vitamin C TR" OR "Vitamin C with Rose Hips")
message:("selenium" OR "Selepen")
256.8776758409786   654.7162245462931   44.0   10.0   6588.0

message:("alprazolam" OR "Alprazolam Intensol" OR "Niravam" OR "Xanax" OR
    "Xanax XR")
message:("sildenafil" OR "Revatio" OR "Viagra")
244.88235294117646  313.10219624607817  28.0   7.0    607.0

message:("carisoprodol" OR "Soma" OR "Vanadom")
message:("diazepam" OR "Diastat" OR "Diastat AcuDial" OR "Diastat Pediatric" OR
    "Diazepam Intensol" OR "Valium" OR "Valrelease")
168.12 306.9787343340469   28.0   8.0    1105.0

message:("aspirin" OR "Arthritis Pain" OR "Aspergum" OR "Aspir 81" OR "Aspir-
    Low" OR "Aspirin Lite Coat" OR "Aspirin Low Strength" OR "Aspiritab" OR
    "Bayer Aspirin" OR "Bayer Aspirin Extra Strength Plus" OR "Bayer Aspirin
    Regimen" OR "Bayer Children's Aspirin" OR "Bayer Women's Aspirin With
    Calcium" OR "Buffered Aspirin" OR "Bufferin" OR "Bufferin Arthritis
    Strength" OR "Bufferin Extra Strength" OR "Easprin" OR "Ecotrin" OR "Ecotrin
    Adult Low Strength" OR "Ecotrin Maximum Strength" OR "Empirin" OR "Fasprin"
    OR "Genacote" OR "Halfprin" OR "Litecoat Aspirin" OR "Medi-Seltzer" OR
    "Norwich Aspirin" OR "St. Joseph 81 mg Aspirin Enteric Safety-Coated" OR
    "St. Joseph 81 mg Chewable Aspirin" OR "St. Joseph Aspirin" OR "Stanback
    Analgesic" OR "Tri-Buffered Aspirin" OR "YSP Aspirin" OR "ZORprin")
message:("ibuprofen" OR "Advil" OR "Advil Children's" OR "Advil Junior
    Strength" OR "Advil Liqui-Gels" OR "Advil Migraine" OR "Caldolor" OR
    "Children's Motrin" OR "Childrens Ibuprofen Berry" OR "Genpril" OR "Haltran"
    OR "IBU" OR "IBU-200" OR "Midol IB" OR "Midol Maximum Strength Cramp
    Formula" OR "Motrin" OR "Motrin Childrens" OR "Motrin IB" OR "Motrin Infant
    Drops" OR "Motrin Junior Strength" OR "Motrin Migraine Pain" OR "NeoProfen"
    OR "Nuprin" OR "Q-Profen")
425.72280701754386  1165.0117618667502  32.0   6.0    13985.0
```

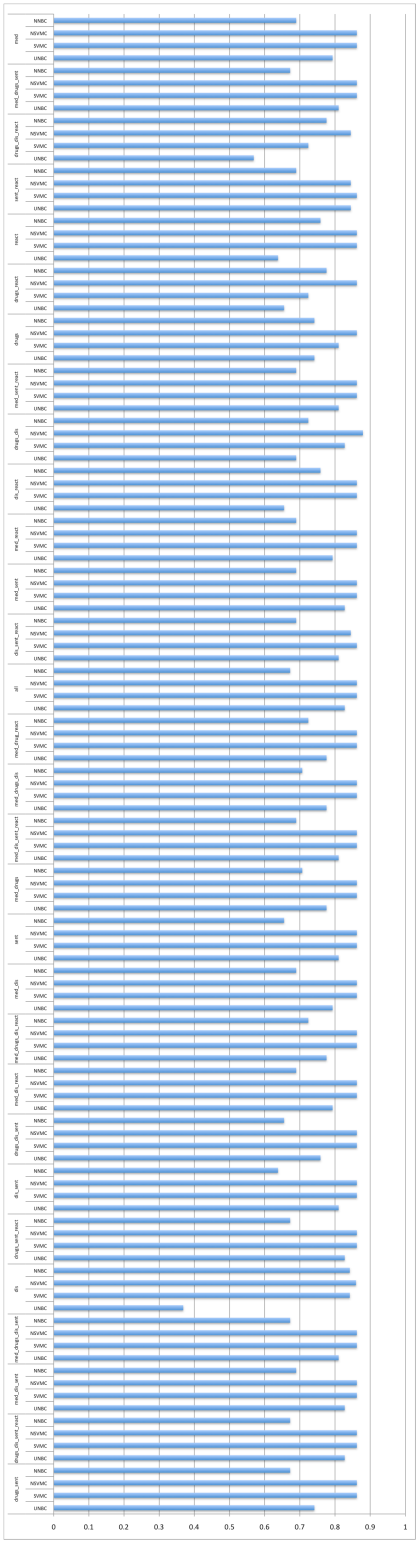
Appendix D: Accuracy Graphs

The graph on the left depicts the cross validation experiments and classification with separate test instances on the right.



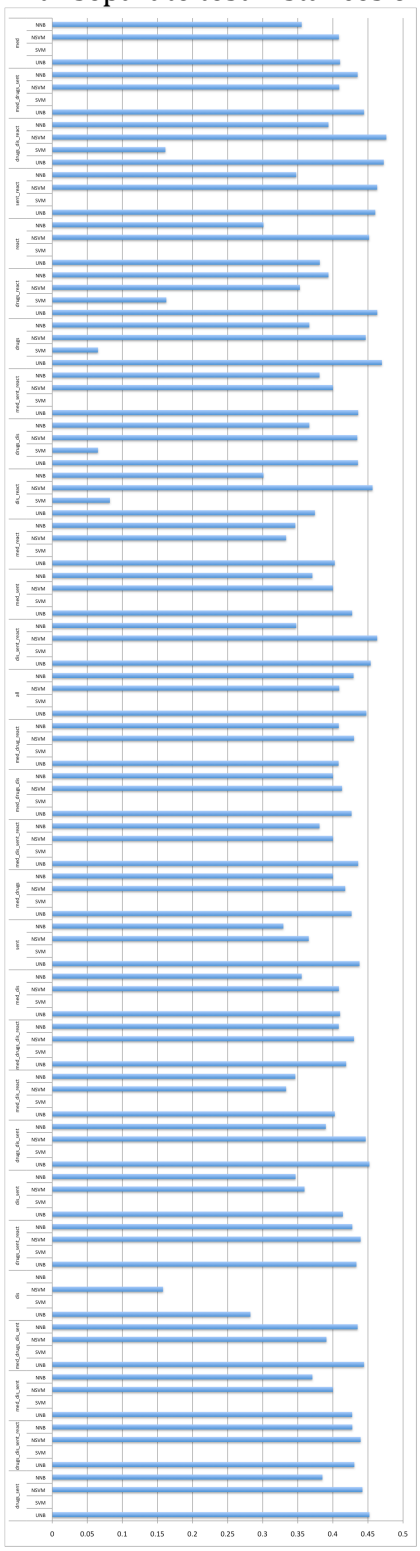


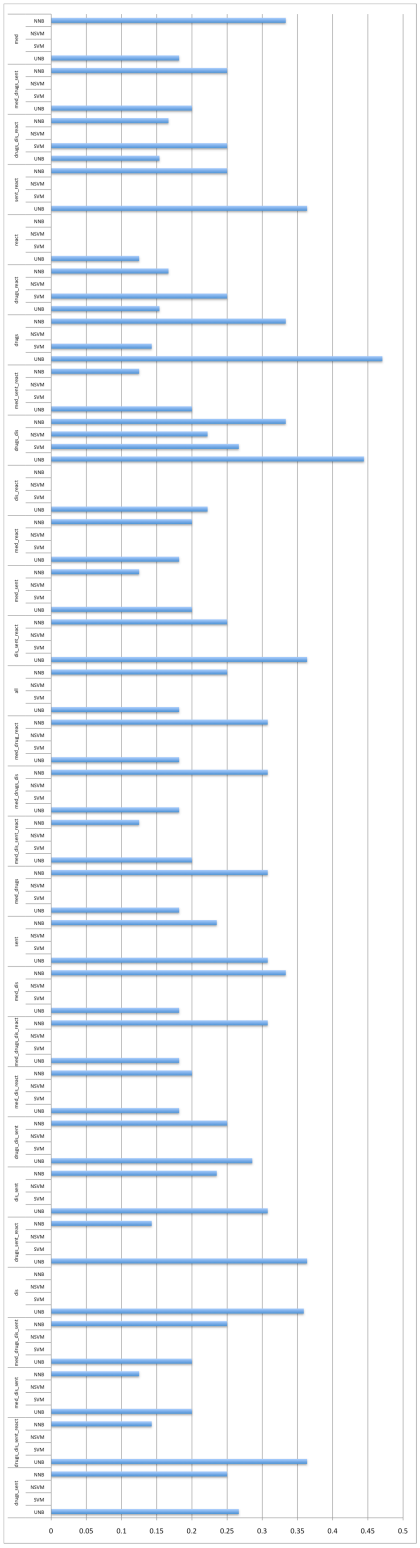
The graph below depicts the accuracy experiments for cost weighted classification algorithms.



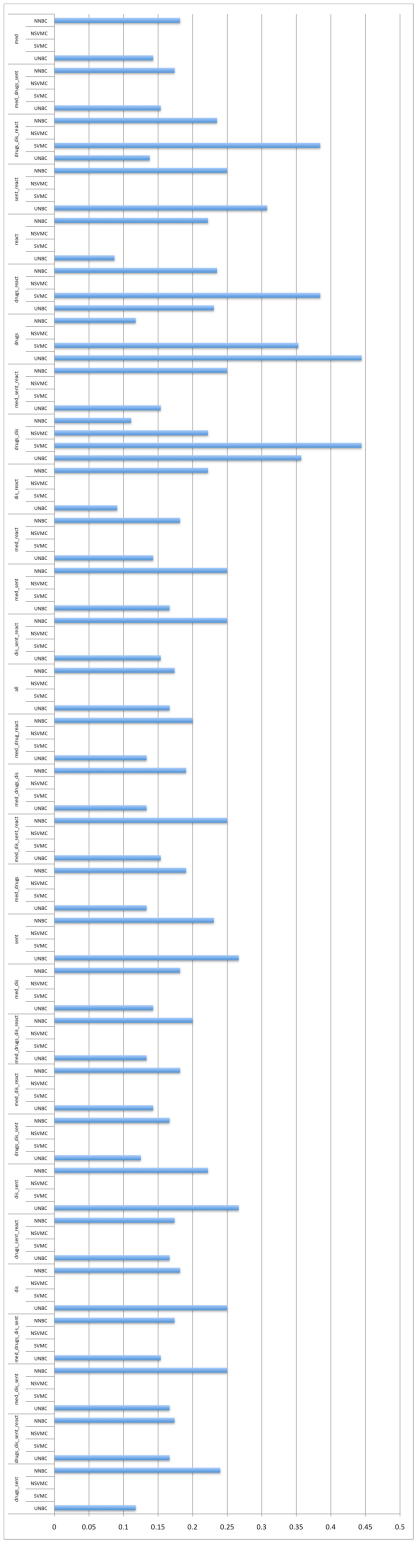
Appendix E: F1 Graphs

The graph on the left depicts the cross validation experiments and classification with separate test instances on the right.



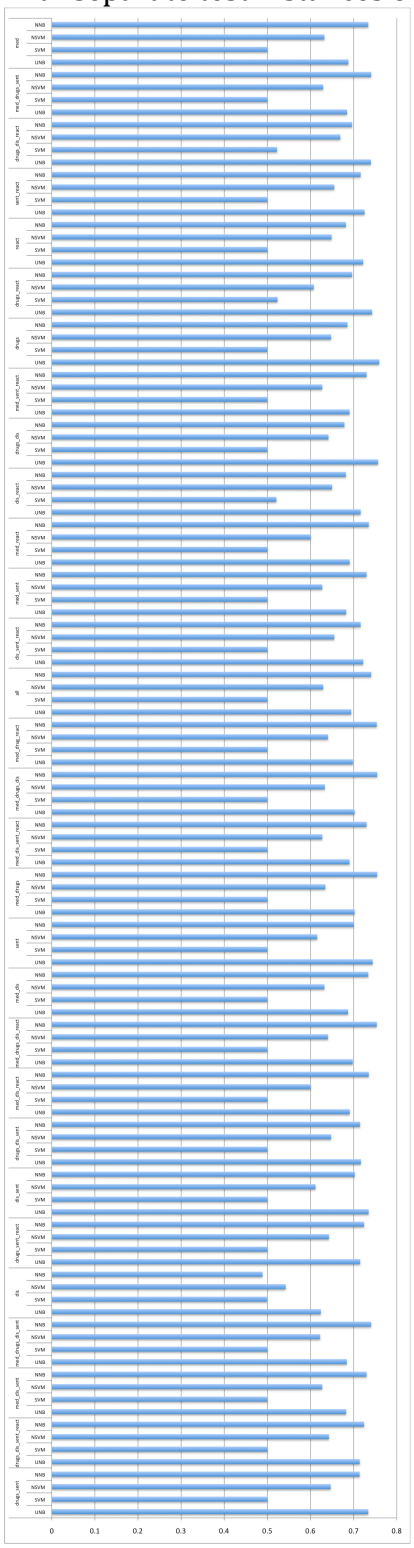


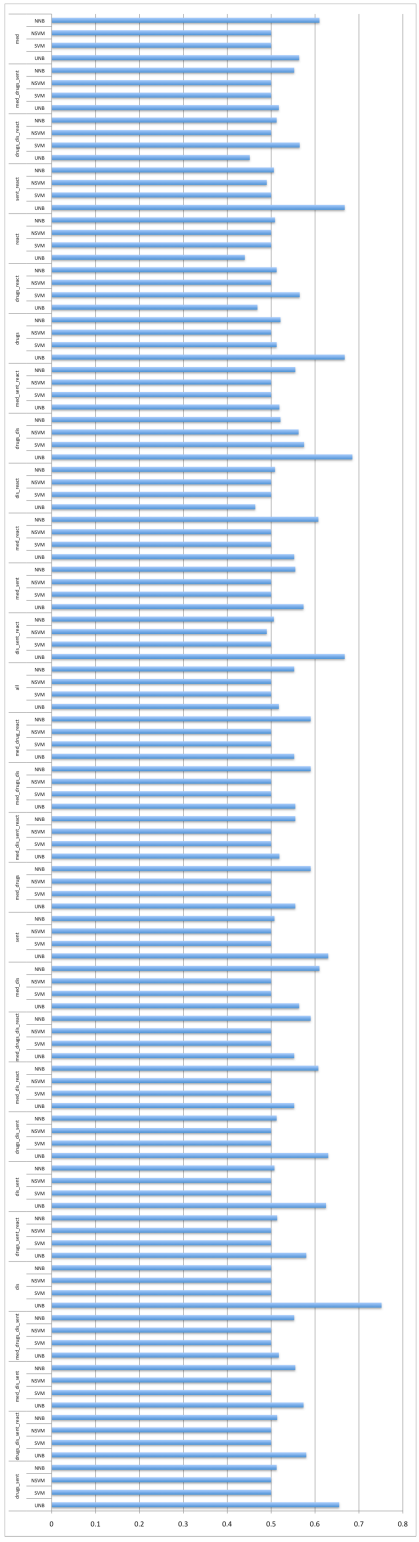
The graph below depicts the F1 experiments for cost weighted classification algorithms.



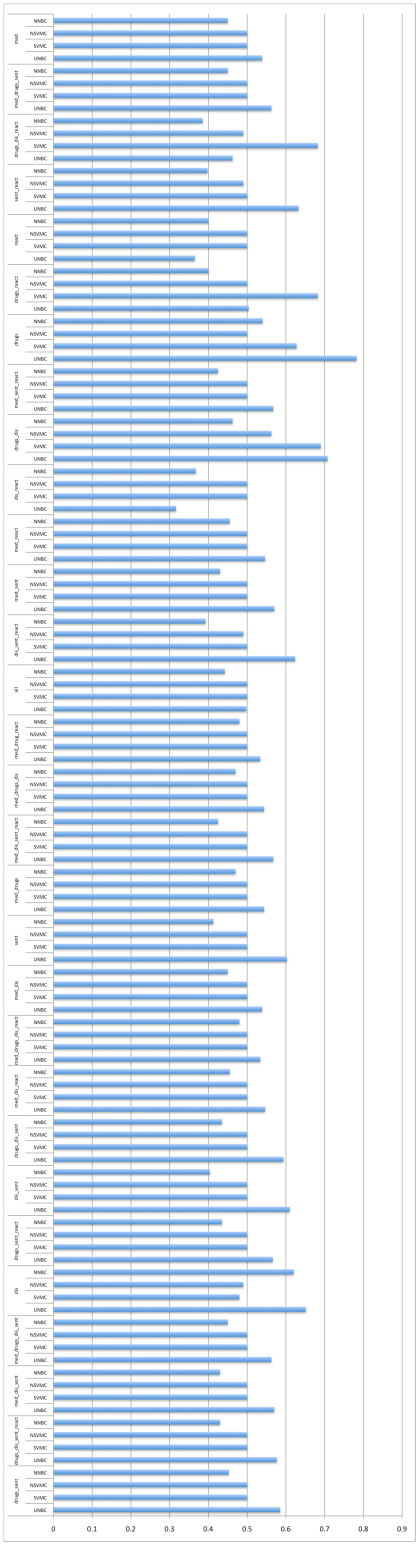
Appendix F: AUC Graphs

The graph on the left depicts the cross validation experiments and classification with separate test instances on the right.



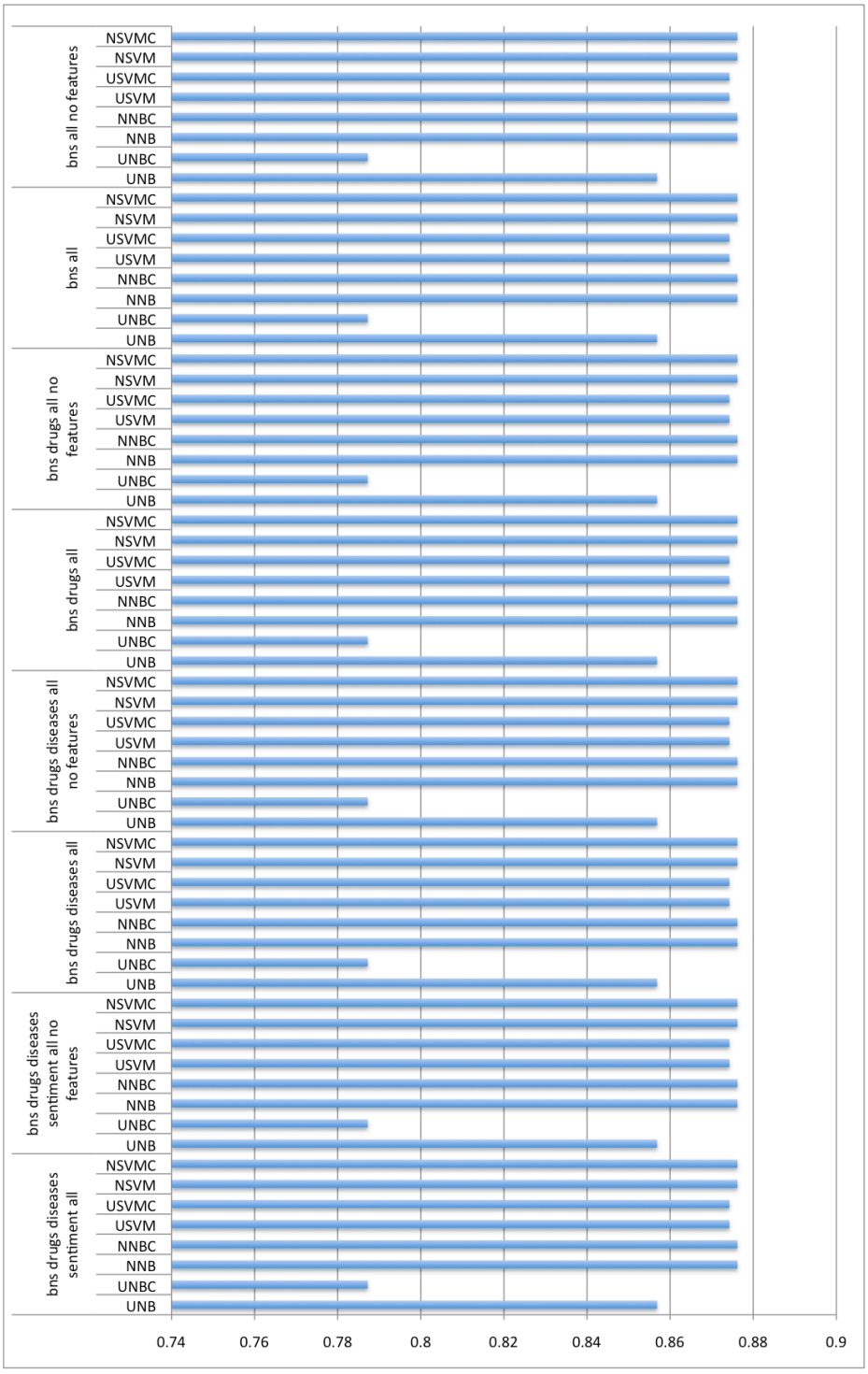


The graph below depicts the AUC experiments for cost weighted classification algorithms.

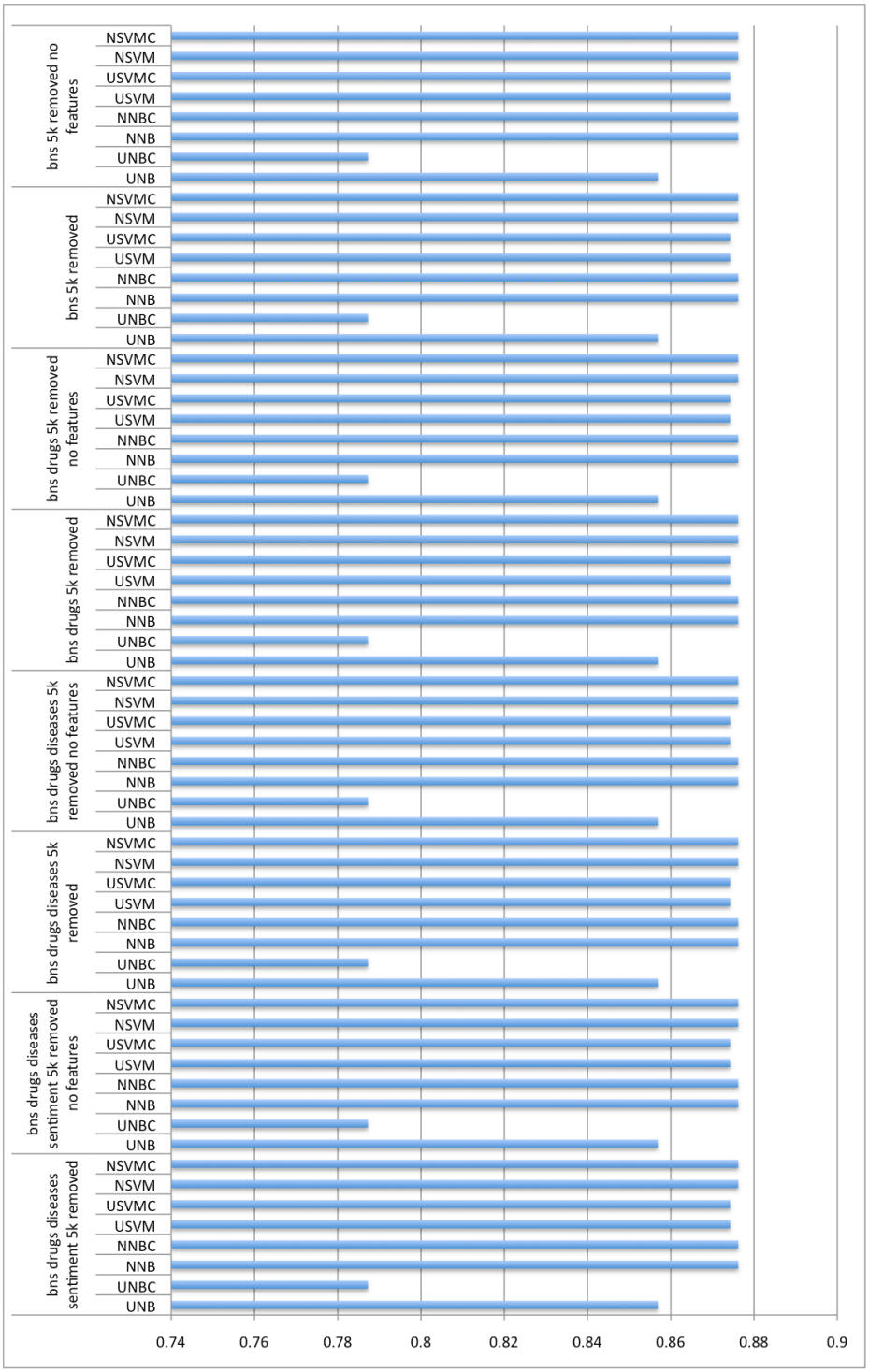


Appendix G: Accuracy Graphs

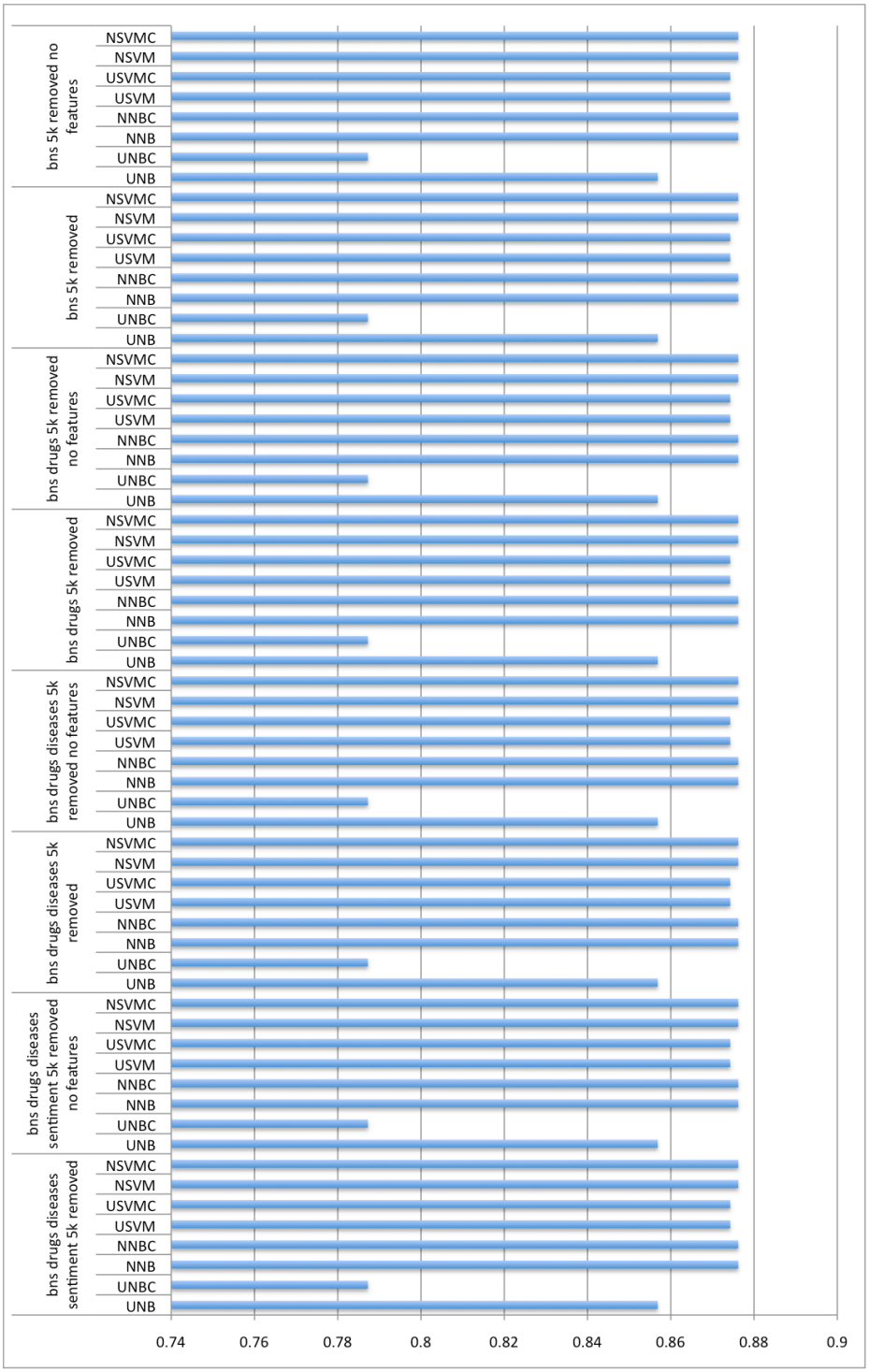
The graph below depicts the accuracy over all 15,000 word features selected from the sampled lexicon.



The graph below depicts the accuracy over all 10,000 word features selected from the sampled lexicon.

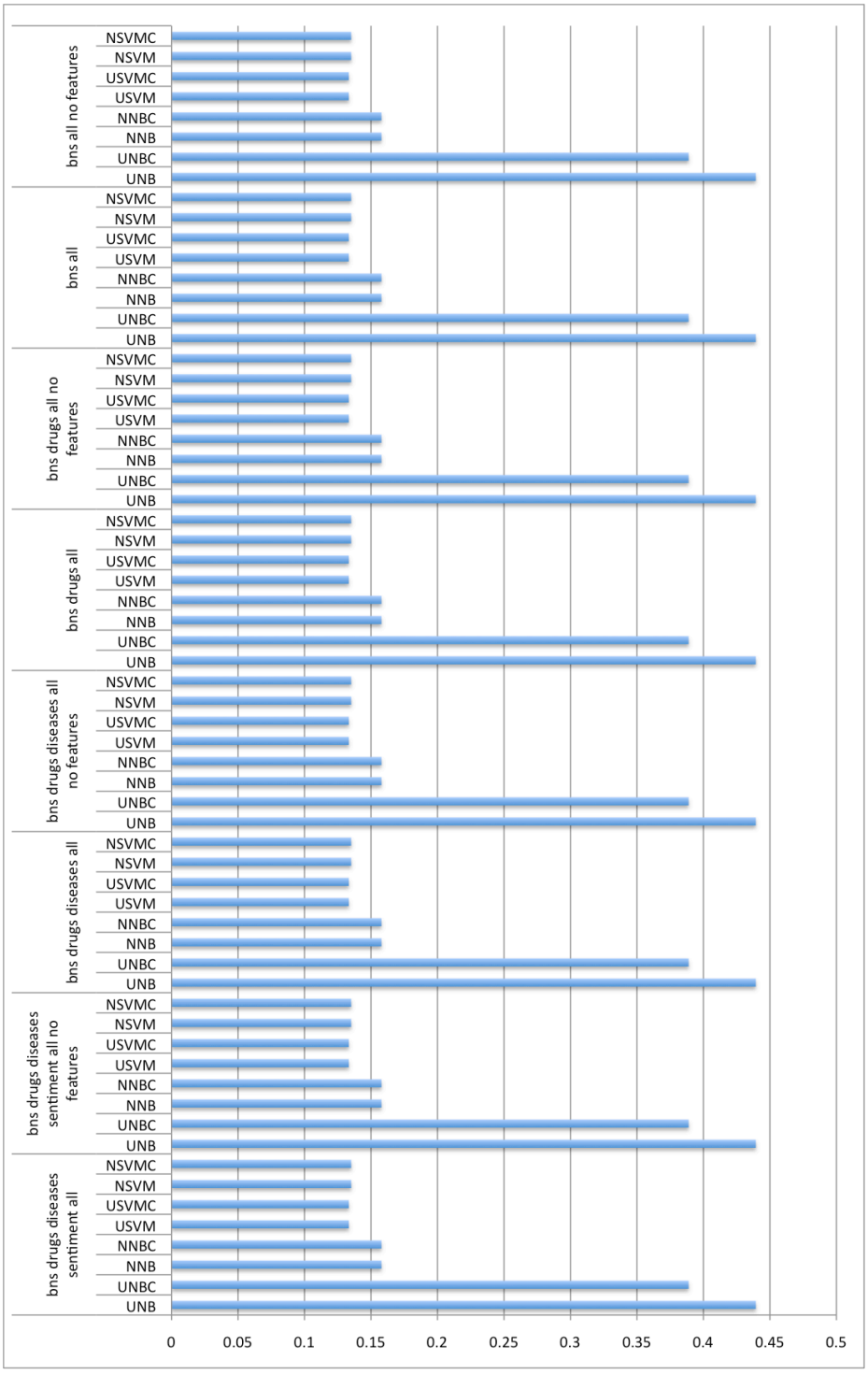


The graph below depicts the accuracy over all 5,000 word features selected from the sampled lexicon.

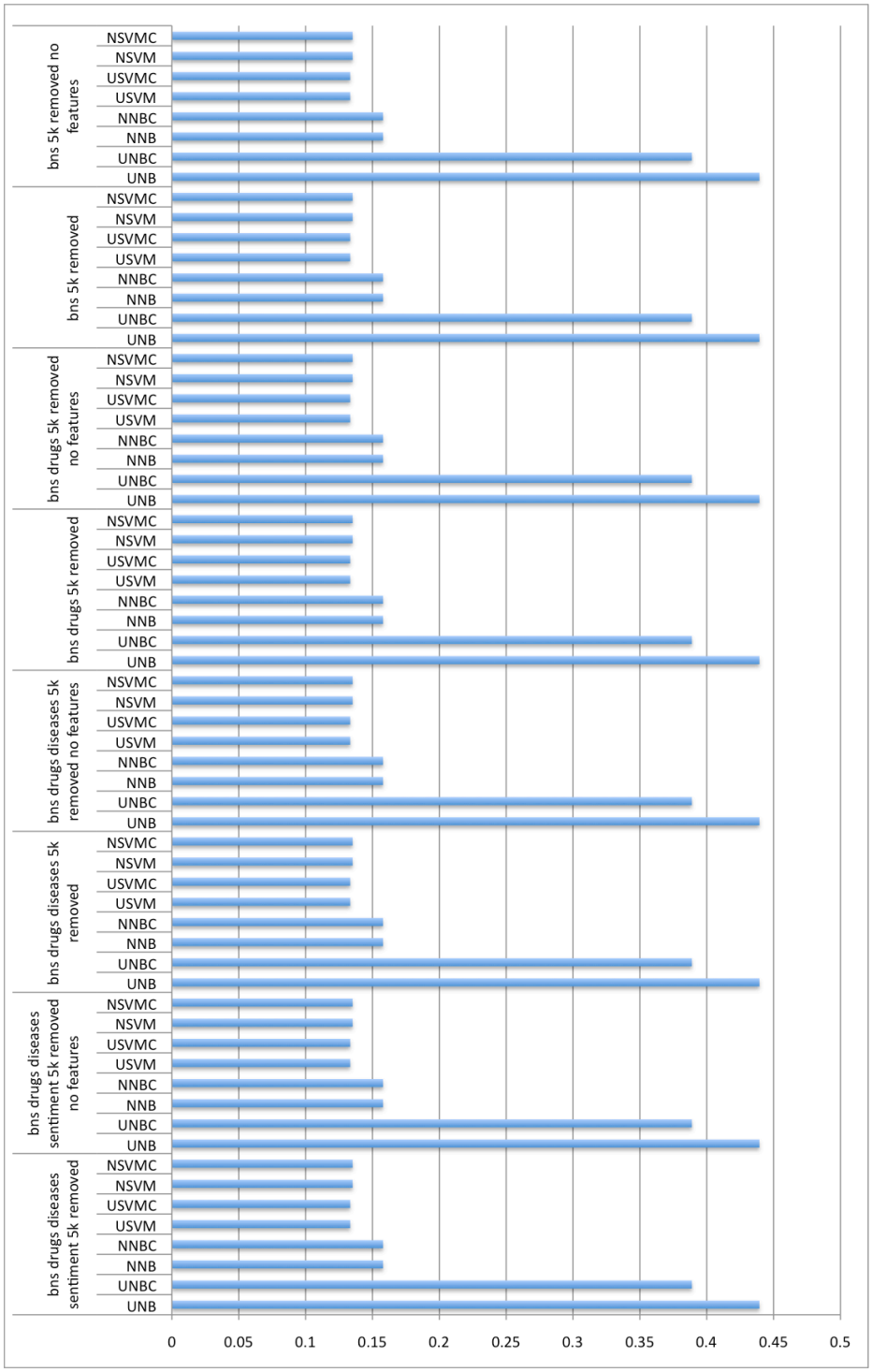


Appendix H: F1 Graphs

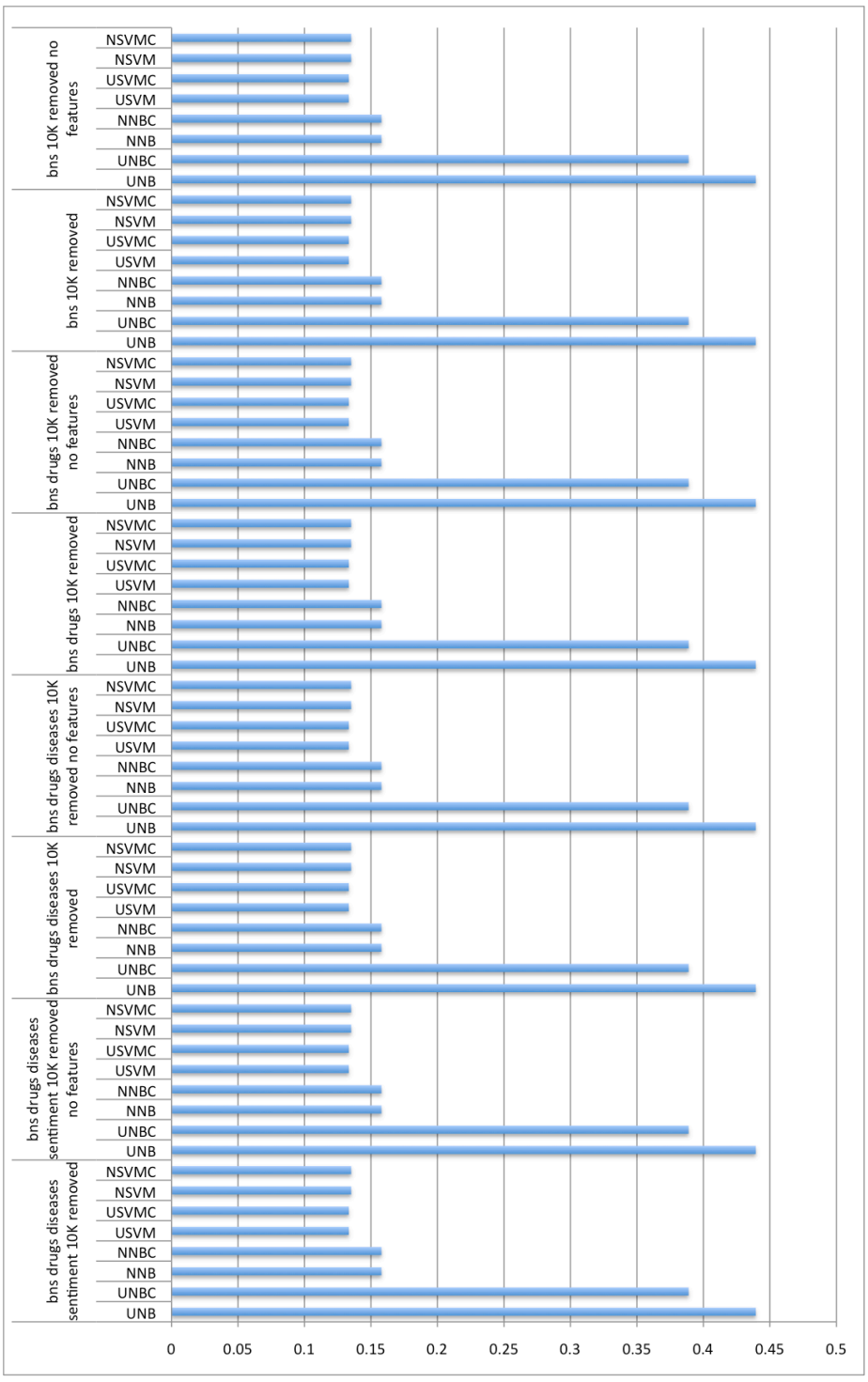
The graph below depicts the F1 scores over all 15,000 word features selected from the sampled lexicon.



The graph below depicts the F1 scores over all 10,000 word features selected from the sampled lexicon.

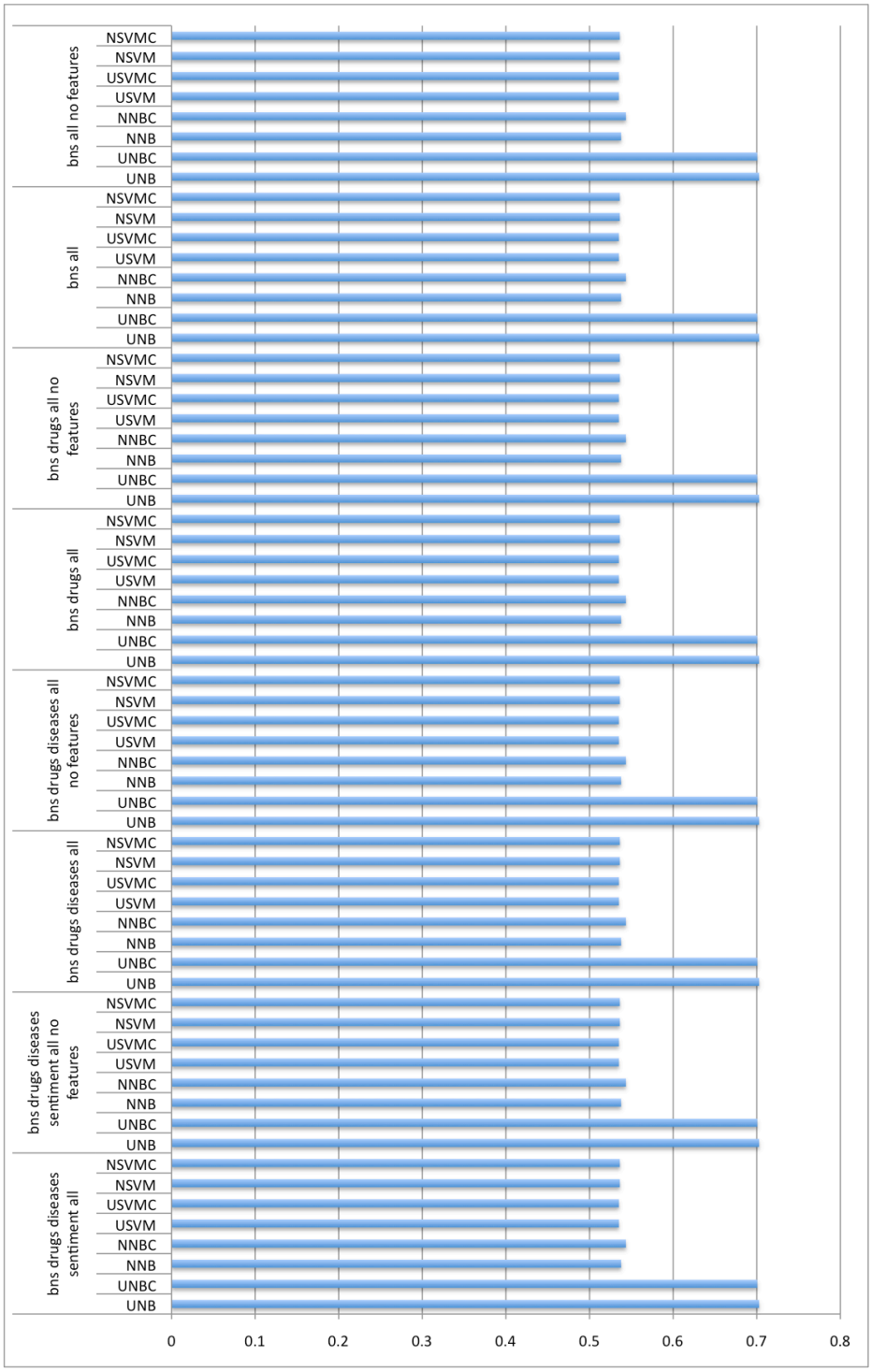


The graph below depicts the F1 scores over all 5,000 word features selected from the sampled lexicon.

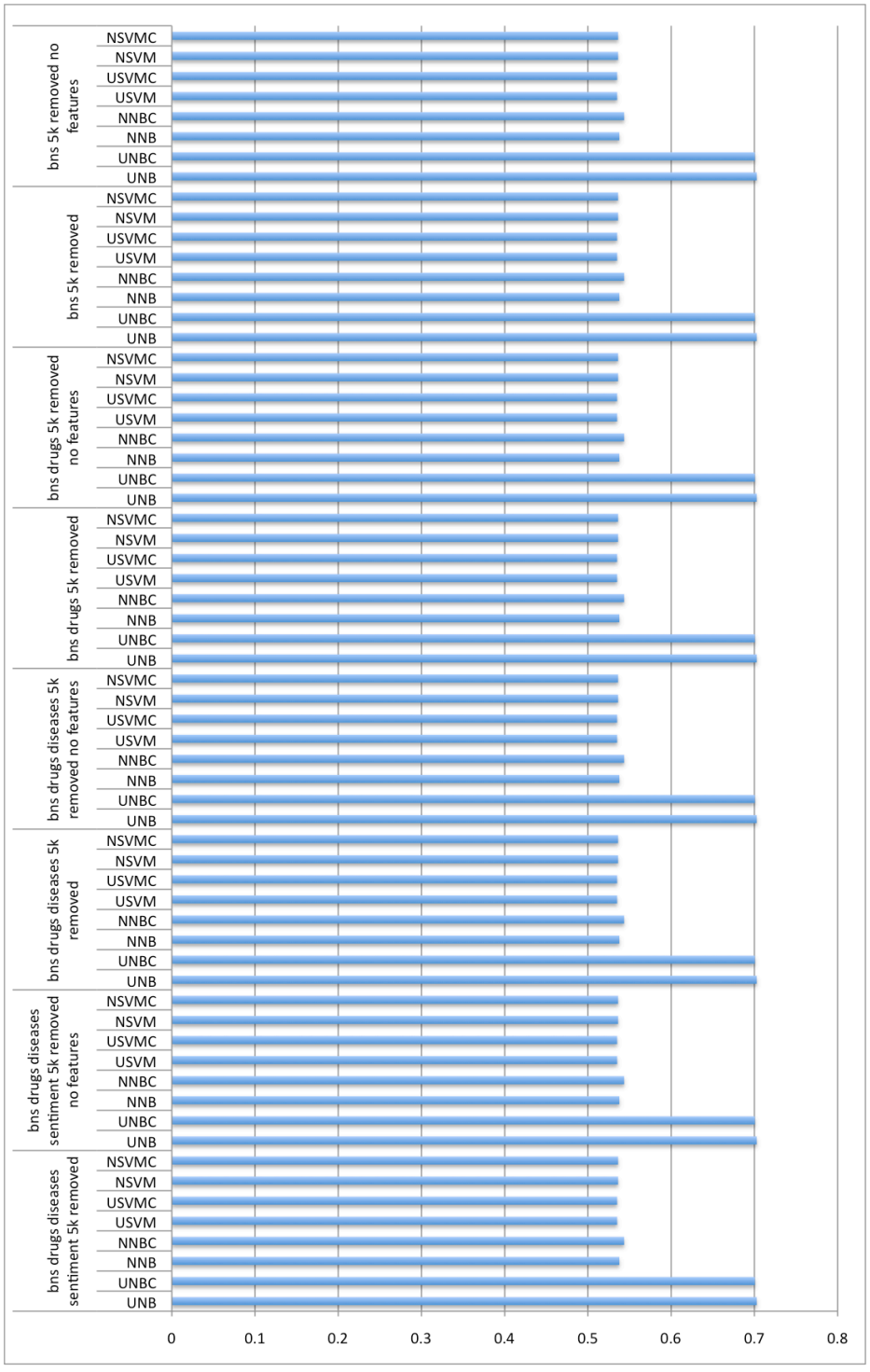


Appendix I: AUC Graphs

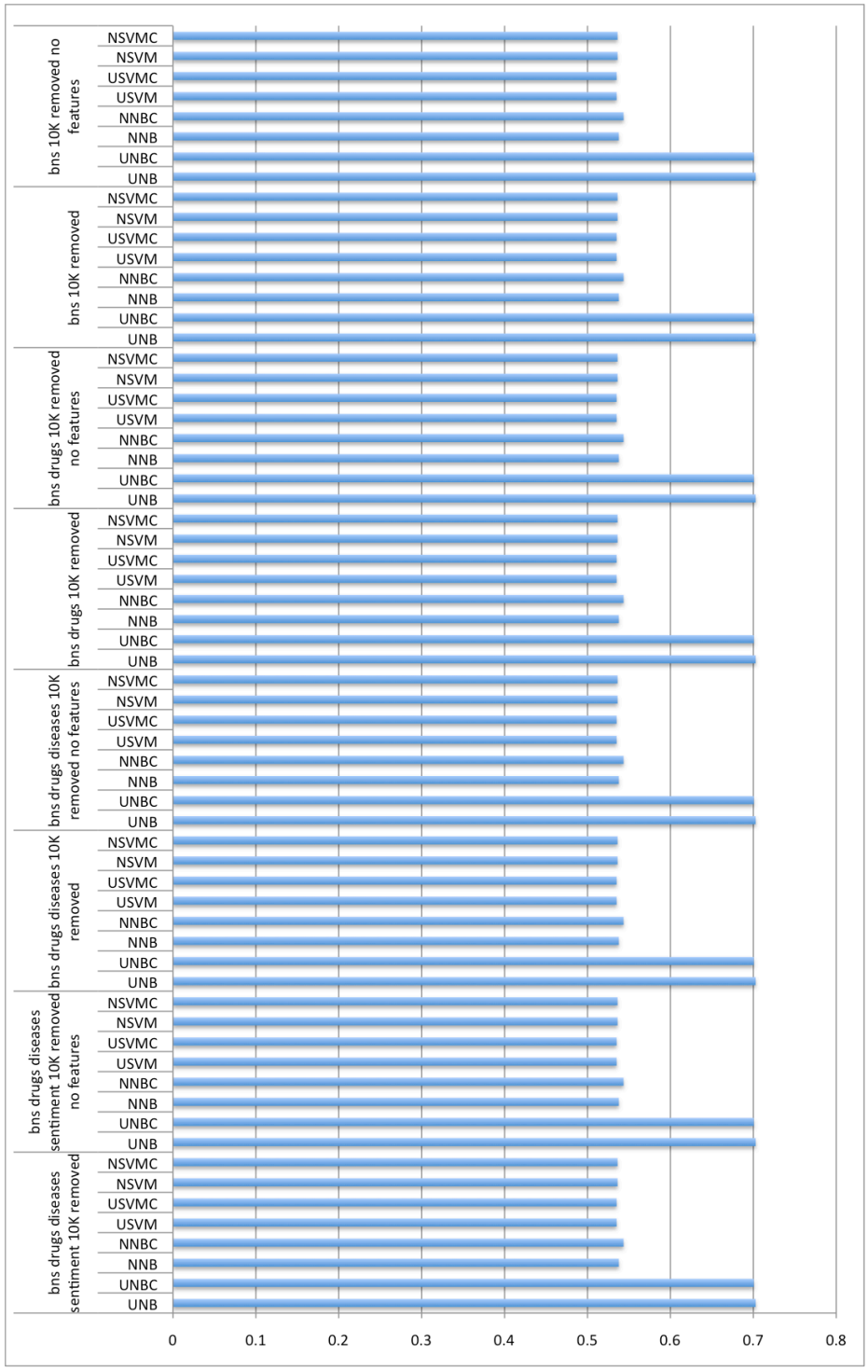
The graph below depicts the AUC scores over all 15,000 word features selected from the sampled lexicon.



The graph below depicts the AUC scores over all 10,000 word features selected from the sampled lexicon.



The graph below depicts the AUC scores over all 5,000 word features selected from the sampled lexicon.



Appendix J: Potential Problematic Drug List from False Positives

Below are the drugs and the associated scores of drugs identified as false positives from the three top-performing classifiers. Next to the drug name, the Pos column denotes the number of times a classifier marked the drug as a false positive. Occ indicates the number of occurrences, or number of times the drug was classified. Class indicates the number of different types of classifiers (1-3) that predicted a false positive for a drug. Score indicates the linear combination of Pos, Occ, and Class resulting in a score of the confidence in prediction. Drugs withdrawn from market are highlighted in yellow. Interesting results are highlighted in green.

Drug	Pos	Occ	Class	Score
clozapine OR Clozaril OR FazaClo	31	64	3	45.047
fludarabine OR Fludara OR Oforta	29	61	3	41.361
methylphenidate OR Concerta OR Daytrana OR Metadate CD OR Metadate ER OR Methylin OR Methylin ER OR Ritalin OR Ritalin LA OR Ritalin-SR	25	50	3	37.500
morphine OR Astramorph PF OR Avinza OR Duramorph OR Infumorph OR Kadian OR MS Contin OR MSIR OR Morphine IR OR Oramorph SR OR RMS OR Roxanol	14	38	3	15.474
meloxicam OR Mobic	15	50	3	13.500
Extraneal	10	36	3	8.333
aripiprazole OR Abilify OR Abilify Discmelt	9	30	3	8.100
evening primrose OR Evening Primrose Oil OR Primrose Oil	17	56	1	5.161
quetiapine OR Seroquel OR Seroquel XR	15	52	1	4.327
trazodone OR Desyrel OR Desyrel Dividose OR Oleptro	14	46	1	4.261
(acetaminophen AND diphenhydramine) OR Anacin P.M. Aspirin Free OR Coricidin Night Time Cold Relief OR Excedrin PM OR Headache Relief PM OR Legatrin PM OR Mapap PM OR Midol PM OR Percogesic Extra Strength OR Sominex Pain Relief Formula OR Tylenol PM OR Tylenol Severe Allergy OR Tylenol Sore Throat Nighttime OR Unisom with Pain Relief	12	34	1	4.235
thalidomide OR Thalomid	13	44	1	3.841
clomipramine OR Anafranil	13	44	1	3.841
vigabatrin OR Sabril	13	50	1	3.380
risperidone OR Risperdal OR Risperdal Consta OR Risperdal M-Tab	11	36	1	3.361
eszopiclone OR Lunesta	11	36	1	3.361

omega-3 polyunsaturated fatty acids OR Animi-3 OR EPA Fish Oil OR Fish Oil OR Icar Prenatal Essential Omega-3 OR Lovaza OR Marine Lipid Concentrate OR MaxEPA OR MaxiTears Dry Eye Formula OR MaxiVision Omega-3 Formula OR Mi-Omega OR Mi-Omega NF OR Omacor OR Omega-500 OR Proepa OR Sea-Omega OR Sea-Omega 30 OR Sea-Omega 70 OR TheraTears Nutrition	13	52	1	3.250
influenza virus vaccine, live, trivalent OR FluMist	10	31	1	3.226
verapamil OR Calan OR Calan SR OR Covera-HS OR Isoptin OR Isoptin SR OR Verelan OR Verelan PM	12	46	1	3.130
acetazolamide OR Diamox OR Diamox Sequels	12	47	1	3.064
(conjugated estrogens AND medroxyprogesterone) OR Premphase OR Prempro	11	46	1	2.630
methylprednisolone OR A-methapred OR Depo-Medrol OR Medrol OR Medrol Dosepak OR MethylPREDNISolone Dose Pack OR Solu-Medrol	11	47	1	2.574
nifedipine OR Adalat OR Adalat CC OR Afeditab CR OR Nifediac CC OR Nifedical XL OR Procardia OR Procardia XL	10	41	1	2.439
temazepam OR Restoril	11	50	1	2.420
sulindac OR Clinoril	10	42	1	2.381
calcipotriene OR Dovonex	6	32	2	2.250
vancomycin OR Lyphocin OR Vancocin OR Vancocin HCl OR Vancocin HCl Pulvules	10	45	1	2.222
insulin lispro OR Humalog OR Humalog KwikPen OR Humalog Pen	9	38	1	2.132
albuterol OR AccuNeb OR Airet OR ProAir HFA OR Proventil OR Proventil HFA OR Ventolin OR Ventolin HFA OR Volmax OR Vospire ER	9	39	1	2.077
amlodipine OR Norvasc	9	40	1	2.025
dronabinol OR Marinol	5	38	3	1.974
nicardipine OR Cardene OR Cardene IV OR Cardene SR	8	33	1	1.939
hydromorphone OR Dilaudid OR Dilaudid-HP OR Exalgo OR Palladone	9	42	1	1.929
losartan OR Cozaar	7	26	1	1.885
irinotecan OR Camptosar	9	43	1	1.884
(dyphylline AND guaifenesin) OR COPD OR Difil G OR Dilex-G OR Dy-G OR Dyfilin GG OR Dyflex-G OR Dyphylline GG OR Jay-Phyl OR Lufyllin-GG OR Panfil G	8	34	1	1.882
doxepin OR Adapin OR Prudoxin OR Silenor OR Sinequan OR Zonalon	9	44	1	1.841
filgrastim OR Neupogen	8	36	1	1.778

(acetaminophen AND butalbital AND caffeine) OR Alagesic OR Alagesic LQ OR Anolor 300 OR Dolgic LQ OR Dolgic Plus OR Esgic OR Esgic-Plus OR Fioricet OR Geone OR Margesic OR Medigesic OR Repan OR Zebutal	9	46	1	1.761
trovafloxacin OR Trovan	9	46	1	1.761
ziprasidone OR Geodon	8	38	1	1.684
oxycodone OR ETH-Oxydose OR OxyContin OR OxyIR OR Oxyfast OR Percolone OR Roxicodone OR Roxicodone Intensol	8	38	1	1.684
clonazepam OR Klonopin OR Klonopin Wafer	7	30	1	1.633
rofecoxib OR Vioxx	9	50	1	1.620
flecainide OR Tambocor	8	40	1	1.600
methotrexate OR Rheumatrex Dose Pack OR Trexall	7	31	1	1.581
hydroxyzine OR Atarax OR Hyzine OR Vistaril	7	31	1	1.581
lorazepam OR Ativan	6	24	1	1.500
guaifenesin OR Altarussin OR Amibid LA OR Drituss G OR Duratuss G OR GG 200 NR OR Ganidin NR OR Guaifenesin LA OR Guaifenex G OR Guaifenex LA OR Hytuss OR Liquibid OR Mucinex OR Mucinex for Kids OR Muco-Fen 1200 OR Organidin NR OR Q-Bid LA OR Robitussin Chest Congestion OR Scot-Tussin Expectorant OR Tussin	8	43	1	1.488
pneumococcal 7-valent vaccine OR Prevnar	8	43	1	1.488
pramipexole OR Mirapex	8	44	1	1.455
selegiline OR Eldepryl OR Emsam OR Zelapar	6	26	1	1.385
glyburide OR DiaBeta OR Glycron OR Glynase OR Glynase PresTab OR Micronase	6	27	1	1.333
nortriptyline OR Aventyl Hydrochloride OR Pamelor	7	38	1	1.289
chlordiazepoxide OR Librium	7	38	1	1.289
(metformin AND rosiglitazone) OR Avandamet	5	21	1	1.190
acetic acid OR Acetasol OR Acid Jelly OR Fem pH OR Klout OR Relagard	7	42	1	1.167
metoprolol OR Lopressor OR Metoprolol Succinate ER OR Toprol-XL	5	22	1	1.136
indomethacin OR Indocin OR Indocin IV OR Indocin SR	6	33	1	1.091
benztropine OR Cogentin	5	23	1	1.087
vitamin e OR Alpha E OR Amino-Opti-E OR Aquasol E OR Aquavite-E OR Centrum Singles-Vitamin E OR E Pherol OR E-400 Clear OR Nutr-E-Sol	5	24	1	1.042
azithromycin OR Azasite OR Azithromycin Dose Pack OR Zithromax OR Zmax	5	25	1	1.000
metformin OR Fortamet OR Glucophage OR Glucophage XR OR Glumetza OR Riomet	6	37	1	0.973
pilocarpine OR Isopto Carpine OR Pilocar OR Pilopine HS OR Pilostat OR Salagen	6	37	1	0.973

lisdexamfetamine OR Vyvanse	5	26	1	0.962
propranolol OR Inderal OR Inderal LA OR InnoPran XL	6	38	1	0.947
sibutramine OR Meridia	5	27	1	0.926
zolpidem OR Ambien OR Ambien CR OR Edluar OR Zolpimist	5	28	1	0.893
enoxaparin OR Lovenox	5	30	1	0.833
phenytoin OR Dilantin OR Phenytek OR Phenytoin Sodium, Prompt	4	20	1	0.800
ubiquinone OR CoQ10 OR Coenzyme Q10 OR LiQsorb OR Liquid Co-Q10 OR NutraDrops OR QuinZyme	5	36	1	0.694
dexmethylphenidate OR Focalin OR Focalin XR	5	40	1	0.625
hyoscyamine OR A-Spaz OR Anaspaz OR Cystospaz OR Hyospaz OR Hyosyne OR IB-Stat OR Levbid OR Levsin OR Levsin/SL OR Levsinex OR Levsinex SR OR NuLev OR Spasdel OR Symax Duotab OR Symax FasTab OR Symax SL OR Symax SR	5	42	1	0.595
ropinirole OR Requip OR Requip Starter Kit OR Requip XL	4	27	1	0.593
tizanidine OR Zanaflex	4	29	1	0.552
raloxifene OR Evista	4	32	1	0.500
acetaminophen OR Acephen OR Actamin OR Adprin B OR Anacin Aspirin Free OR Apra OR Atasol OR Bromo Seltzer OR Childrens ElixSure OR Childrens Silapap OR Childrens Tylenol OR Dolono OR Ed-APAP OR Elixsure Fever/Pain OR Febrol Solution OR Feverall OR Genapap OR Genebs OR Infants Tylenol OR Jr. Tylenol OR Mapap OR Mapap Arthritis Pain OR Mapap Childrens OR Mapap Infant Drops OR Mapap Meltaway OR Mapap Rapid Release Gelcaps OR Mapap Rapid Tabs OR Pain-Eze OR Q-Pap OR Q-Pap Extra Strength OR Silapap Childrens OR Silapap Infants OR St. Joseph Aspirin-Free OR Tactinal OR Tempra OR Tempra Quicklets OR Tycolene OR Tylenol OR Tylenol 8 Hour OR Tylenol Arthritis Pain OR Tylenol Extra Strength OR Tylenol GoTabs OR Tylenol Sore Throat Daytime OR Tylophen OR Uniserts OR Vitapap	5	50	1	0.500
erythromycin OR A/T/S OR Akne-Mycin OR E-Mycin OR E.E.S. Granules OR E.E.S.-200 OR E.E.S.-400 OR E.E.S.-400 Filmtab OR Emcin Clear OR Emgel OR Ery Pads OR Ery-Tab OR EryPed OR Eryc OR Erycette OR Eryderm OR Erygel OR Erymax OR Erythra-Derm OR Erythrocin OR Erythrocin Lactobionate OR Erythrocin Stearate Filmtab OR Ilosone OR Ilotycin OR PCE Dispertab OR Roymicin OR Staticin OR T-Stat OR Theramycin Z	4	35	1	0.457
anakinra OR Kineret	3	22	1	0.409

black cohosh OR Menopause Support	3	26	1	0.346
atorvastatin OR Lipitor	3	30	1	0.300
(sulfamethoxazole AND trimethoprim) OR Bactrim OR Bactrim DS OR Cotrim OR SMZ-TMP DS OR Septra OR Septra DS OR Sulfatrim OR Sulfatrim Pediatric	3	32	1	0.281
hydrocortisone OR A-Hydrocort OR Ala-Cort OR Ala-Scalp HP OR Anucort-HC OR Anumed-HC OR Anusol-HC OR Aquanil HC OR Beta HC OR Caldecort OR Cetacort OR Colocort OR Cortaid OR Cortaid Intensive Therapy OR Cortaid Maximum Strength OR Cortaid with Aloe OR Cortalo with Aloe OR Cortef OR Cortenema OR Corticaine OR Cortifoam OR Cortizone for Kids OR Cortizone-10 OR Cortizone-10 Anal Itch Cream OR Cortizone-10 Intensive Healing Formula OR Cortizone-10 Plus OR Cortizone-5 OR Dermarest Dricort OR Dermarest Eczema Medicated OR Dermarest Plus Anti-Itch OR Dermtex HC OR Genasone/Aloe OR Gly-Cort OR Gynecort Maximum Strength OR Hemorrhoidal HC OR Hemril-30 OR Hemril-HC Uniserts OR Hycort OR Hydrocortisone 1% In Absorbase OR Hydrocortisone with Aloe OR Hydrocortone OR Hytone OR Instacort OR Itch-X Lotion OR Locoid OR Locoid Lipocream OR MD Hydrocortisone OR Massengill Medicated Soft Cloth OR Neutrogena T-Scalp OR NuCort with Aloe OR NuZon OR Nutracort OR Pandel OR Preparation H Anti-Itch Cream Hydrocortisone 1% OR Procto-Kit 1% OR Procto-Kit 2.5% OR Procto-Pak 1% OR ProctoCare-HC OR Proctocort OR Proctocream-HC OR Proctosert HC OR Proctosol-HC OR Proctozone HC OR Proctozone-H OR Recort Plus OR Rectasol-HC OR Rederm OR Sarnol-HC OR Scalacort OR Scalpicin OR Solu-Cortef OR Texacort OR Tucks Hydrocortisone Anti-Itch Ointment OR U-Cort OR Westcort	3	34	1	0.265
ampicillin OR Principen	3	36	1	0.250
meperidine OR Demerol OR Meperitab	2	33	2	0.242

psyllium OR Alramucil OR Cilium OR Fiber Eze OR Fiberall OR Genfiber OR Hydrocil OR Konsyl OR Konsyl Orange Sugar-free OR Konsyl for Kids OR Konsyl-D OR Konsyl-Orange OR Laxmar OR Laxmar Orange OR Laxmar Sugar Free OR Maalox Daily Fiber Therapy OR Metamucil OR Metamucil Berry Burst Smooth Texture Sugar Free OR Metamucil Orange Coarse Milled Original Texture OR Metamucil Orange Smooth Texture OR Metamucil Orange Smooth Texture Sugar Free OR Metamucil Smooth Texture OR Metamucil Unflavored Coarse Milled Original Texture OR Metamucil Unflavored Smooth Texture Sugar Free OR Modane Bulk OR Natural Fiber Therapy OR Perdiem Fiber Powder OR Reguloid OR Serutan OR Syllact OR V-Lax	3	39	1	0.231
aprepitant OR Emend OR Emend 3-Day	3	39	1	0.231
erlotinib OR Tarceva	2	23	1	0.174
cisapride OR Propulsid	2	24	1	0.167
modafinil OR Provigil	2	27	1	0.148
tranylcypromine OR Parnate	2	30	1	0.133
(hydrocortisone AND pramoxine) OR Analpram E OR Analpram-HC OR Enzone OR Epifoam OR HC Pramoxine OR Hydropram OR Novacort OR Pramosone OR Proctofoam HC OR Rectocort HC OR Zone-A OR Zone-A Forte OR Zypram	2	31	1	0.129
povidone iodine OR Betadine OR Betadine Aerosol Spray OR Minidyne OR Pharmadine OR Polydine	2	37	1	0.108
methionine OR Me-500	2	47	1	0.085
atovaquone OR Mepron	1	20	1	0.050
(atropine AND diphenoxylate) OR Lomocot OR Lomotil OR Lonox OR Vi-Atro	1	21	1	0.048
tacrolimus OR Prograf OR Protopic	1	22	1	0.045
ginkgo OR Ginkgo Biloba	1	24	1	0.042
polyethylene glycol 3350 OR ClearLax OR GaviLAX OR GlycoLax OR MiraLax	1	24	1	0.042
digoxin OR Digitek OR Lanoxicaps OR Lanoxin	1	26	1	0.038
omeprazole OR Prilosec OR Prilosec OTC	1	27	1	0.037
ritonavir OR Norvir	1	27	1	0.037
cefuroxime OR Ceftin OR Zinacef	1	28	1	0.036
fat emulsion OR Intralipid OR Liposyn II	1	28	1	0.036
diltiazem OR Cardizem OR Cardizem CD OR Cardizem LA OR Cardizem SR OR Cartia XT OR Dilacor XR OR Diltia XT OR Diltzac OR Taztia XT OR Tiazac	1	29	1	0.034
pimecrolimus OR Elidel	1	30	1	0.033
leflunomide OR Arava	1	30	1	0.033
metaxalone OR Skelaxin	1	30	1	0.033

(acetaminophen AND aspirin AND caffeine) OR Anacin Advanced Headache Formula OR Excedrin OR Excedrin Extra Strength OR Excedrin Menstrual Complete OR Excedrin Migraine OR Genace OR Goodys Extra-Strength Headache Powders OR Goodys Extra Strength OR Goodys Headache Powders OR Supac OR Vanquish	1	31	1	0.032
rotavirus vaccine OR RotaTeq OR Rotarix	1	31	1	0.032
ondansetron OR Zofran OR Zofran ODT	1	32	1	0.031
zaleplon OR Sonata	1	32	1	0.031
dobutamine OR Dobutrex	1	32	1	0.031
nitroglycerin OR Minitran OR Nitrek OR Nitro TD Patch-A OR Nitro-Bid OR Nitro-Dur OR Nitro-Time OR NitroMist OR NitroQuick OR Nitrocot OR Nitrogard OR Nitrol Appli-Kit OR Nitrolingual Pumpspray OR Nitrostat OR Nitrostat Tablets OR Transderm-Nitro	1	32	1	0.031
atenolol OR Tenormin	1	33	1	0.030
isoniazid OR Nydrazid	1	33	1	0.030
cerivastatin OR Baycol	1	33	1	0.030
famotidine OR Heartburn Relief OR Leader Acid Reducer OR Mylanta AR OR Pepcid OR Pepcid AC OR Pepcid AC Chewable Tablets OR Pepcid AC Maximum Strength OR Pepcid AC Maximum Strength Tablets OR Pepcid Oral Suspension OR Pepcid RPD	1	33	1	0.030
pegfilgrastim OR Neulasta	1	35	1	0.029
oxybutynin OR Ditropan OR Ditropan XL OR Gelnique OR Oxytrol OR Urotrol	1	38	1	0.026
(acetaminophen AND dichloralphenazone AND isometheptene mucate) OR Amidrine OR Diacetazone OR Duradrin OR Epidrin OR Iso-Acetazone OR Isocom OR Midrin OR Migquin OR Migrapap OR Migratine OR Migrazone OR Migrin-A	1	38	1	0.026
azathioprine OR Azasan OR Imuran	1	41	1	0.024
ramelteon OR Rozerem	1	42	1	0.024
allopurinol OR Aloprim OR Zyloprim	1	45	1	0.022
penicillamine OR Cuprimine OR Depen OR Depen Titratabs	1	46	1	0.022
naloxone OR Narcan	1	46	1	0.022

Appendix K: Potential Problematic Drug List from False Positives

Below are the drugs and the associated scores of drugs identified as false positives from the three top-performing classifiers. Next to the drug name, the Pos column denotes the number of times a classifier marked the drug as a false positive. Occ indicates the number of occurrences, or number of times the drug was classified. Class indicates the number of different types of classifiers (1-3) that predicted a false positive for a drug. Score indicates the linear combination of Pos, Occ, and Class resulting in a score of the confidence in prediction. Drugs withdrawn from market are highlighted in yellow. Interesting results are highlighted in green. Withdrawn drugs were tested separately and not counted as negative examples; their classification score is shown for comparison.

Drug	Pos	Occ	Class	Score
methylphenidate OR Concerta OR Daytrana OR Metadate CD OR Metadate ER OR Methylin OR Methylin ER OR Ritalin OR Ritalin LA OR Ritalin-SR	30	34	3	79.412
morphine OR Astramorph PF OR Avinza OR Duramorph OR Infumorph OR Kadian OR MS Contin OR MSIR OR Morphine IR OR Oramorph SR OR RMS OR Roxanol	13	38	3	13.342
quetiapine OR Seroquel OR Seroquel XR	14	31	2	12.645
indomethacin OR Indocin OR Indocin IV OR Indocin SR	19	37	1	9.757
sibutramine OR Meridia	17	34	1	8.500
meloxicam OR Mobic	17	35	1	8.257
vigabatrin OR Sabril	14	31	1	6.323
losartan OR Cozaar	13	28	1	6.036
oxycodone OR ETH-Oxydose OR OxyContin OR OxyIR OR Oxyfast OR Percolone OR Roxicodone OR Roxicodone Intensol	13	30	1	5.633
doxepin OR Adapin OR Prudoxin OR Silenor OR Sinequan OR Zonalon	14	37	1	5.297
aripiprazole OR Abilify OR Abilify Discmelt	13	32	1	5.281
guaifenesin OR Altarussin OR Amibid LA OR Drituss G OR Duratuss G OR GG 200 NR OR Ganidin NR OR Guaifenesin LA OR Guaifenex G OR Guaifenex LA OR Hytuss OR Liquibid OR Mucinex OR Mucinex for Kids OR Muco-Fen 1200 OR Organidin NR OR Q-Bid LA OR Robitussin Chest Congestion OR Scot-Tussin Expectorant OR Tussin	13	34	1	4.971
enoxaparin OR Lovenox	13	35	1	4.829
risperidone OR Risperdal OR Risperdal Consta OR Risperdal M-Tab	12	30	1	4.800
tizanidine OR Zanaflex	12	30	1	4.800
Extraneal	13	36	1	4.694
nortriptyline OR Aventyl Hydrochloride OR Pamelor	12	31	1	4.645

zolpidem OR Ambien OR Ambien CR OR Edluar OR Zolpimist	12	34	1	4.235
clozapine OR Clozaril OR FazaClo	10	24	1	4.167
selegiline OR Eldepryl OR Emsam OR Zelapar	10	25	1	4.000
nicardipine OR Cardene OR Cardene IV OR Cardene SR	11	31	1	3.903
(acetaminophen AND diphenhydramine) OR Anacin P.M. Aspirin Free OR Coricidin Night Time Cold Relief OR Excedrin PM OR Headache Relief PM OR Legatrin PM OR Mapap PM OR Midol PM OR Percogesic Extra Strength OR Sominex Pain Relief Formula OR Tylenol PM OR Tylenol Severe Allergy OR Tylenol Sore Throat Nighttime OR Unisom with Pain Relief	11	33	1	3.667
eszopiclone OR Lunesta	10	28	1	3.571
(dyphylline AND guaifenesin) OR COPD OR Difil G OR Dilex-G OR Dy-G OR Dyfilin GG OR Dyflex-G OR Dyphylline GG OR Jay-Phyl OR Lufyllin-GG OR Panfil G	11	36	1	3.361
nifedipine OR Adalat OR Adalat CC OR Afeditab CR OR Nifediac CC OR Nifedical XL OR Procardia OR Procardia XL	10	31	1	3.226
clomipramine OR Anafranil	10	31	1	3.226
lisdexamfetamine OR Vyvanse	10	32	1	3.125
dexmethylphenidate OR Focalin OR Focalin XR	9	26	1	3.115
filgrastim OR Neupogen	11	39	1	3.103
pneumococcal 7-valent vaccine OR Prevnar	9	28	1	2.893
pramipexole OR Mirapex	10	35	1	2.857
sulindac OR Clinoril	9	29	1	2.793
albuterol OR AccuNeb OR Airet OR ProAir HFA OR Proventil OR Proventil HFA OR Ventolin OR Ventolin HFA OR Volmax OR Vospire ER	10	36	1	2.778
thalidomide OR Thalomid	10	36	1	2.778
(sulfamethoxazole AND trimethoprim) OR Bactrim OR Bactrim DS OR Cotrim OR SMZ-TMP DS OR Septra OR Septra DS OR Sulfatrim OR Sulfatrim Pediatric	9	30	1	2.700
ampicillin OR Principen	9	30	1	2.700
acetazolamide OR Diamox OR Diamox Sequels	9	31	1	2.613
phenytoin OR Dilantin OR Phenytek OR Phenytoin Sodium, Prompt	9	31	1	2.613
pilocarpine OR Isopto Carpine OR Pilocar OR Pilopine HS OR Pilostat OR Salagen	9	32	1	2.531
verapamil OR Calan OR Calan SR OR Covera-HS OR Isoptin OR Isoptin SR OR Verelan OR Verelan PM	10	40	1	2.500
(acetaminophen AND butalbital AND caffeine) OR Alagesic OR Alagesic LQ OR Anolor 300 OR Dolgic LQ OR Dolgic Plus OR Esgic OR Esgic-Plus OR Fioricet OR Geone OR Margesic OR Medigesic OR Repan OR Zebutal	9	34	1	2.382

influenza virus vaccine, live, trivalent OR FluMist	8	27	1	2.370
ubiquinone OR CoQ10 OR Coenzyme Q10 OR LiQsorb OR Liquid Co-Q10 OR NutraDrops OR QuinZyme	9	35	1	2.314
(conjugated estrogens AND medroxyprogesterone) OR Premphase OR Prempro	8	28	1	2.286
temazepam OR Restoril	8	28	1	2.286
clonazepam OR Klonopin OR Klonopin Wafer	8	30	1	2.133
evening primrose OR Evening Primrose Oil OR Primrose Oil	8	30	1	2.133
methylprednisolone OR A-methapred OR Depo-Medrol OR Medrol OR Medrol Dosepak OR MethylPREDNISolone Dose Pack OR Solu-Medrol	8	30	1	2.133
trazodone OR Desyrel OR Desyrel Dividose OR Oleptro	8	30	1	2.133
fludarabine OR Fludara OR Oforta	6	19	1	1.895
ziprasidone OR Geodon	7	26	1	1.885
omega-3 polyunsaturated fatty acids OR Animi-3 OR EPA Fish Oil OR Fish Oil OR Icar Prenatal Essential Omega-3 OR Lovaza OR Marine Lipid Concentrate OR MaxEPA OR MaxiTears Dry Eye Formula OR MaxiVision Omega-3 Formula OR Mi-Omega OR Mi-Omega NF OR Omacor OR Omega-500 OR Proepa OR Sea-Omega OR Sea-Omega 30 OR Sea-Omega 70 OR TheraTears Nutrition	7	28	1	1.750
benztropine OR Cogentin	7	28	1	1.750
insulin lispro OR Humalog OR Humalog KwikPen OR Humalog Pen	8	39	1	1.641
lorazepam OR Ativan	8	40	1	1.600
amlodipine OR Norvasc	7	31	1	1.581
vancomycin OR Lyphocin OR Vancocin OR Vancocin HCl OR Vancocin HCl Pulvules	7	31	1	1.581
flecainide OR Tambocor	7	31	1	1.581
vitamin e OR Alpha E OR Amino-Opti-E OR Aquasol E OR Aquavite-E OR Centrum Singles-Vitamin E OR E Pherol OR E-400 Clear OR Nutr-E-Sol	7	31	1	1.581
metoprolol OR Lopressor OR Metoprolol Succinate ER OR Toprol-XL	7	34	1	1.441
irinotecan OR Camptosar	6	27	1	1.333
(hydrocortisone AND pramoxine) OR Analpram E OR Analpram-HC OR Enzone OR Epifoam OR HC Pramoxine OR Hydropram OR Novacort OR Pramosone OR Proctofoam HC OR Rectocort HC OR Zone-A OR Zone-A Forte OR Zypram	5	20	1	1.250
propranolol OR Inderal OR Inderal LA OR InnoPran XL	6	29	1	1.241
acetic acid OR Acetasol OR Acid Jelly OR Fem pH OR Klout OR Relagard	6	30	1	1.200
chlordiazepoxide OR Librium	6	38	1	0.947

hydroxyzine OR Atarax OR Hyzine OR Vistaril	5	29	1	0.862
modafinil OR Provigil	4	23	1	0.696
acyclovir OR Zovirax OR Zovirax Cream OR Zovirax Ointment	5	36	1	0.694
methotrexate OR Rheumatrex Dose Pack OR Trexall	4	24	1	0.667
metformin OR Fortamet OR Glucophage OR Glucophage XR OR Glumetza OR Riomet	4	27	1	0.593
levothyroxine OR Levothroid OR Levoxyl OR Synthroid OR Tirosint OR Unithroid	4	30	1	0.533
meperidine OR Demerol OR Meperitab	4	32	1	0.500
azithromycin OR Azasite OR Azithromycin Dose Pack OR Zithromax OR Zmax	4	33	1	0.485
tacrolimus OR Prograf OR Protopic	4	35	1	0.457
(chondroitin AND glucosamine) OR Cosamin DS OR Osteo Bi-Flex OR Osteo Bi-Flex Double Strength OR Osteo Bi-Flex Triple Strength OR Pryflex OR Relamine OR Schiff Move Free OR Schiff Move Free Caplets	4	36	1	0.444
hydrocortisone OR A-Hydrocort OR Ala-Cort OR Ala-Scalp HP OR Anucort-HC OR Anumed-HC OR Anusol-HC OR Aquanil HC OR Beta HC OR Caldecort OR Cetacort OR Colocort OR Cortaid OR Cortaid Intensive Therapy OR Cortaid Maximum Strength OR Cortaid with Aloe OR Cortalo with Aloe OR Cortef OR Cortenema OR Corticaine OR Cortifoam OR Cortizone for Kids OR Cortizone-10 OR Cortizone-10 Anal Itch Cream OR Cortizone-10 Intensive Healing Formula OR Cortizone-10 Plus OR Cortizone-5 OR Dermarest Dricort OR Dermarest Eczema Medicated OR Dermarest Plus Anti-Itch OR Dermtex HC OR Genasone/Aloe OR Gly-Cort OR Gynecort Maximum Strength OR Hemorrhoidal HC OR Hemril-30 OR Hemril-HC Uniserts OR Hycort OR Hydrocortisone 1% In Absorbase OR Hydrocortisone with Aloe OR Hydrocortone OR Hytone OR Instacort OR Itch-X Lotion OR Locoid OR Locoid Lipocream OR MD Hydrocortisone OR Massengill Medicated Soft Cloth OR Neutrogena T-Scalp OR NuCort with Aloe OR NuZon OR Nutracort OR Pandel OR Preparation H Anti-Itch Cream Hydrocortisone 1% OR Procto-Kit 1% OR Procto-Kit 2.5% OR Procto-Pak 1% OR ProctoCare-HC OR Proctocort OR Proctocream-HC OR Proctosert HC OR Proctosol-HC OR Proctozone HC OR Proctozone-H OR Recort Plus OR Rectasol-HC OR Rederm OR Sarnol-HC OR Scalacort OR Scalpicin OR Solu-Cortef OR Texacort OR Tucks Hydrocortisone Anti-Itch Ointment OR U-Cort OR Westcort	3	34	1	0.265
erlotinib OR Tarceva	3	38	1	0.237
(metformin AND rosiglitazone) OR Avandamet	3	42	1	0.214

(acetaminophen AND tramadol) OR Ultracet	2	20	1	0.200
tranylcypromine OR Parnate	2	23	1	0.174
oxybutynin OR Ditropan OR Ditropan XL OR Gelnique OR Oxytrol OR Urotrol	2	25	1	0.160
(acetaminophen AND dextromethorphan AND doxylamine AND pseudoephedrine) OR All-Nite Multi-Symptom Cold/Flu Relief OR NyQuil OR NyQuil Multi-Symptom OR Nyquil Cold Medicine OR Tylenol Severe Cold & Flu Night Time	2	26	1	0.154
black cohosh OR Menopause Support	2	28	1	0.143
tetracycline OR Ala-Tet OR Sumycin OR Topicycline	2	28	1	0.143
(acetaminophen AND dichloralphenazone AND isometheptene mucate) OR Amidrine OR Diacetazone OR Duradrin OR Epidrin OR Iso-Acetazone OR Isocom OR Midrin OR Migquin OR Migrapap OR Migratine OR Migrazone OR Migrin-A	2	28	1	0.143
glyburide OR DiaBeta OR Glycron OR Glynase OR Glynase PresTab OR Micronase	2	28	1	0.143
lactobacillus acidophilus OR Acidophilus OR Acidophilus Extra Strength OR Bacid OR Flora-Q OR Flora-Q 2 OR Novaflor OR RisaQuad OR Superdophilus	2	30	1	0.133
sulfasalazine OR Azulfidine OR Azulfidine EN-tabs OR Sulfazine	2	30	1	0.133
ropinirole OR Requip OR Requip Starter Kit OR Requip XL	2	30	1	0.133
dronabinol OR Marinol	2	30	1	0.133
digoxin OR Digitek OR Lanoxicaps OR Lanoxin	2	31	1	0.129
doxycycline OR Adoxa OR Adoxa CK OR Adoxa TT OR Alodox OR Avidoxy OR Doryx OR Doxy 100 OR Doxy 200 OR Monodox OR Oracea OR Oraxyl OR Periostat OR Uracil OR Vibra-Tabs OR Vibramycin	2	32	1	0.125
raloxifene OR Evista	2	32	1	0.125
dobutamine OR Dobutrex	2	32	1	0.125
erythromycin OR A/T/S OR Akne-Mycin OR E-Mycin OR E.E.S. Granules OR E.E.S.-200 OR E.E.S.-400 OR E.E.S.-400 Filmtab OR Emcin Clear OR Emgel OR Ery Pads OR Ery-Tab OR EryPed OR Eryc OR Erycette OR Eryderm OR Erygel OR Erymax OR Erythra-Derm OR Erythrocin OR Erythrocin Lactobionate OR Erythrocin Stearate Filmtab OR Ilosone OR Ilotycin OR PCE Dispertab OR Roymicin OR Staticin OR T-Stat OR Theramycin Z	2	34	1	0.118
hyoscyamine OR A-Spaz OR Anaspaz OR Cystospaz OR Hyospaz OR Hyosyne OR IB-Stat OR Levbid OR Levsin OR Levsin/SL OR Levsinex OR Levsinex SR OR NuLev OR Spasdel OR Symax Duotab OR Symax FasTab OR Symax SL OR Symax SR	2	36	1	0.111
fat emulsion OR Intralipid OR Liposyn II	2	36	1	0.111

anakinra OR Kineret	2	37	1	0.108
acetaminophen OR Acephen OR Actamin OR Adprin B OR Anacin Aspirin Free OR Apra OR Atasol OR Bromo Seltzer OR Childrens ElixSure OR Childrens Silapap OR Childrens Tylenol OR Dolono OR Ed-APAP OR Elixsure Fever/Pain OR Febrol Solution OR Feverall OR Genapap OR Genebs OR Infants Tylenol OR Jr. Tylenol OR Mapap OR Mapap Arthritis Pain OR Mapap Childrens OR Mapap Infant Drops OR Mapap Meltaway OR Mapap Rapid Release Gelcaps OR Mapap Rapid Tabs OR Pain-Eze OR Q-Pap OR Q-Pap Extra Strength OR Silapap Childrens OR Silapap Infants OR St. Joseph Aspirin-Free OR Tactinal OR Tempra OR Tempra Quicklets OR Tycolene OR Tylenol OR Tylenol 8 Hour OR Tylenol Arthritis Pain OR Tylenol Extra Strength OR Tylenol GoTabs OR Tylenol Sore Throat Daytime OR Tylophen OR Uniserts OR Vitapap	2	41	1	0.098
cisapride OR Propulsid	2	41	1	0.098
omeprazole OR Prilosec OR Prilosec OTC	1	17	1	0.059
rotavirus vaccine OR RotaTeq OR Rotarix	1	21	1	0.048
atorvastatin OR Lipitor	1	23	1	0.043
fibrinogen OR RiaSTAP	1	25	1	0.040
methionine OR Me-500	1	26	1	0.038
atenolol OR Tenormin	1	27	1	0.037
telithromycin OR Ketek OR Ketek Pak	1	27	1	0.037
aprepitant OR Emend OR Emend 3-Day	1	27	1	0.037
valsartan OR Diovan	1	27	1	0.037
azathioprine OR Azasan OR Imuran	1	27	1	0.037
leflunomide OR Arava	1	28	1	0.036
spironolactone OR Aldactone	1	30	1	0.033
(acetaminophen AND aspirin AND caffeine) OR Anacin Advanced Headache Formula OR Excedrin OR Excedrin Extra Strength OR Excedrin Menstrual Complete OR Excedrin Migraine OR Genace OR Goodys Extra-Strength Headache Powders OR Goodys Extra Strength OR Goodys Headache Powders OR Supac OR Vanquish	1	30	1	0.033
pimecrolimus OR Elidel	1	30	1	0.033
fexofenadine OR Allegra OR Allegra ODT	1	30	1	0.033

calcium carbonate OR Alka-Mints OR Ami-Lac OR Amitone OR Cal-Gest OR Calcarb OR Calci Mix OR Calci-Chew OR Calcium Concentrate OR Calcium Liquid Softgel OR Calcium Oyster Shell OR Caltrate OR Chooz OR Extra Strength Mylanta Calci Tabs OR Icar Prenatal Chewable Calcium OR Maalox Childrens Relief OR Maalox Regular Strength OR Mylanta Child OR Nephro Calci OR Os-Cal 500 OR Oysco 500 OR Oyst Cal 500 OR Oyster Cal OR Oyster Calcium OR Oyster Shell OR Pepto Childrens OR Rolaids Extra Strength Softchews OR Titralac OR Tums E-X 750 OR Tums Kids OR Tums Regular Strength OR Tums Ultra 1000	1	32	1	0.031
polyethylene glycol 3350 OR ClearLax OR GaviLAX OR GlycoLax OR MiraLax	1	32	1	0.031
darbepoetin alfa OR Aranesp	1	33	1	0.030
bevacizumab OR Avastin	1	40	1	0.025
(acetaminophen AND oxycodone) OR Endocet OR Magnacet OR Narvox OR Percocet OR Percocet 10/325 OR Percocet 10/650 OR Percocet 2.5/325 OR Percocet 5/325 OR Percocet 7.5/325 OR Percocet 7.5/500 OR Perloxx OR Primalev OR Roxicet OR Tylox OR Xolox	1	42	1	0.024
belladonna OR Belladonna Tincture	1	43	1	0.023
dimenhydrinate OR Dramamine OR Driminate OR Travel-Eze	1	44	1	0.023

Drug	**Pos**	**Occ**	**Class**	**Score**
trovafloxacin OR Trovan	33	100	1	10.89
hydromorphone OR Dilaudid OR Dilaudid-HP OR Exalgo OR Palladone	33	100	1	10.89
rofecoxib OR Vioxx	32	100	1	10.24
cerivastatin OR Baycol	2	100	1	0.04