ABSTRACT

ELMESSIRY, ADEL MAGDI. Natural Language Techniques for Decision Support Based on Patient Complaints. (Under the direction of Dr. Munindar P. Singh).

Complaining is a fundamental human characteristic that has prevailed throughout the ages. We normally complain about something that went wrong. Patient complaints are no exception; they focus on problems that occurred during the episode of care. The Institute of Medicine estimated that each year thousands of patients die due to medical errors. The number of patient deaths associated with medical errors was $98,000$ in 1984. It would be reasonable to assume that this figure should have been considerably reduced given the tremendous scientific progress achieved over the past three decades. The facts are quite contrary; an epidemic of patient harm in hospitals has developed. In 2013, a staggering $210,000$ deaths per year were associated with preventable harm in hospitals, while the number of premature deaths associated with preventable harm was estimated at more than $400,000$ per year [James, 2013].

Healthcare in the United States is not monolithic. Rather, it is a broad and ever-increasing range of options. This rapid growth can quickly overwhelm care providers and results in significant medical errors. Although human life is the ultimate casualty, the monetary impact is another aspect of the underlying problem. The cost stemming from countless wasted hours is staggering. Healthcare expenditure constitutes a considerable percentage of the global gross domestic product. The United States comes in at the top of the range with 17.5%, which has more than tripled from 5% in 1960. Healthcare malpractice payouts are a \$3.6 billion component of \$55.8 billion healthcare risk cost and the overall $\$2,799$ billion healthcare expenditure.

An early precursor of malpractice is a patient complaint that can result in an adverse action report, which is an action taken against a practitioner's clinical privileges or medical staff membership in a healthcare entity. Adverse actions have been consistently on the rise over the past ten years. The current mitigation strategy depends on human coders who analyze patient feedback to classify the complaints to help a more advanced team build an intervention plan. Early intervention is critical to prevention and the systematic improvement of the healthcare system. Without the ability to scale the intervention, the tragic loss of human life and the astronomical financial cost will only continue to rise. Nevertheless, due to the growth in the generated patient feedback, scaling the current approach requires automating the initial triage process. However, due to the complexity and diversity of the linguistic representations, building a computational tool for this task is challenging.

This dissertation focuses on the problem of developing an understanding of patient complaints to enable a robust approach requiring minimal human supervision to build a patient complaint analysis tool that can be used across healthcare providers. To systematically study this problem, we have identified three important tasks that are critical for building such a tool:

(1) Establishing the urgency concept by (a) defining patient complaint urgency and (b) building a framework to predict urgency. (2) Automating the current systems by (a) automatically mapping patient complaints to a sentiment-based model and (b) accurately inferring the class for each complaint. (3) Exploring grammatical analysis by (a) defining a method to build a domain-specific dependency based rules for feature extraction and reduction and (b) applying those rules on the ground truth dataset to predict patient complaint classification.

Accordingly, for the first task, we implement a set of classifiers to model urgency and predict subsequent complaints. In the second task, we propose a novel mapping method that maps complaint terms to the linguistic inquiry and word count domains (LIWC) and implement a classification tool that employs supervised learning. In the final task, we build on the grammatical dependency models and develop a set of domain-based rules enabling the extraction of more meaningful features. For all of these tasks, we evaluate the effectiveness of our approaches on a real-life dataset collected by the Vanderbilt Center for Patient and Professional Advocacy and validated against trained human coders.

Results show that our approach is quite useful in identifying patient complaint's urgency. Furthermore, our LIWC-based classifiers yield greater overall accuracy than traditional methods in classifying a limited set of categories. The results we obtained using domain-specific grammatically extracted features are promising. Compared to basic unigram features, our method produced superior results across all categories, showing a weighted F-Measure gain from 0.41 to 0.82.

Overall, this dissertation indicates the potential benefits of applying natural language techniques in analyzing patient complaints to support decision making in healthcare.

Natural Language Techniques for Decision Support Based on Patient Complaints

by
Adel Magdi ElMessiry

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2016

APPROVED BY:

Dr. David L. Roberts	Dr. Christopher G. Healey
Dr. Lynsey K. Romo	Dr. Munindar P. Singh Chair of Advisory Committee

DEDICATION

To my parents, children, and wife.
Without whom, I have no Life.

BIOGRAPHY

Adel ElMessiry grew up in Alexandria, Egypt, where two of the old world seven wonders were (the library and the lighthouse). Following in the family's tradition, he earned his Bachelor's Degree in Electric and Electronics Engineering from Alexandria Faculty of Engineering, where his father is a tenured professor. Driven by the quest for knowledge, he continued his graduate study at North Carolina State University, where he earned his Master's Degree in Computer Engineering in 1998. Since then, he has been striving to improve the state of the health care industry through advancing massive online health care education systems. He combines his interests in semantic web and trust networks with his healthcare information technology career. He recently stepped down as the CTO of InVivoLink, a healthcare technology innovation leader to focus on his research. He joined the research team led by Professor Dr. Munindar P. Singh and plans to continue academic research after graduation.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisor Professor Munindar P. Singh, without whom, my journey would be lost. I had the pleasure to take two courses with him in my Master's, at which point I realized the depth and wealth of knowledge I have access to in him. He guided my long journey to conduct research carefully and methodically with enthusiasm. Having him on my side through my Ph.D. studies was both an encouragement and an inspiration. I would like to express my sincere appreciation to my committee members: Prof. Christopher Healey, for guiding me through this research and all of the insightful comments; Prof. David L. Roberts, for his through reviews and ever uplifting comments; Prof. Lynsey K. Romo for her enriching comments and advice.

I am also sincerely grateful to Camille Cox, Kathy Luca, Margery Page, Prof. Douglas Reeves, Prof. George Rouskas, Andrew Sleeth, Prof. David Thuente, and Prof. Mladen A. Vouk for their advice and support throughout my studies. I am especially indebted to my father Prof. Magdi ElMessiry at Alexandria University, who initiated my first steps in research when I was in high school and who guided me to my first publication. He has remained a constant source of guidance and encouragement throughout my research.

I had the privilege to collaborate with an outstanding group of students at the Multiagent Systems and Service-Oriented Computing Laboratory especially Dr. Zhe Zhang, Dr. Xibin Gao, Dr. Anup Kalia, and Dr. Pradeep Murukannaiah. A special thanks to Dr. Zhe Zhang who provided me a great deal of advice and knowledge in my research.

I can not understate the important impact which the Center for Patient and Professional Advocacy (CPPA) had on my work and all the help provided by Dr. William O. Cooper, Dr. James W. Pichert, Dr. Thomas F. Catron, Dr. Jan Karrass, Anna G. Eldridge and Nik Zakrzewski.

I have to mention both Jessica Harthcock and Prof. Larry Bridgesmith for their support.

I like to thank my family, Azza, Malak, Kenzy and Adam for loving and supporting me at all times (go clean your rooms!). Last, but the most important, none of this would have been possible at all were it not for my parents. My mom has made countless sacrifices over the years to provide me with the best of everything. She is always my guarding angel. My dad has always encouraged me to pursue my goals and inspired me to follow in his footsteps.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The Hippocratic Oath is an oath historically taken by physicians, part of the Hippocratic oath shown in Figure 1.1. It is one of the most widely known of Greek medical texts and is considered to be a rite of passage for practitioners of medicine in many countries in a modern version [Markel, 2004]. It is interesting that the main part of it could be translated as "As to diseases, make a habit of two things to help, or at least, to do no harm." The recent statistics indicate that an epidemic of harm is growing in healthcare.

Figure 1.1. The 3rd-century papyrus Oxyrhynchus 2547 showing a fragment of the Hippocratic oath.

1.1 Medical Error

To error is human. However, the size of medical error reveals a growing epidemic of preventable harm to patients [Stokowski, 2016]. A "medical error" may or may not cause harm to the patient and be defined an error such as:

- An unintended act (either of commission or omission)
- An act that does not achieve its intended outcome
- The failure of a planned action to be completed (an error of execution)
- The use of a wrong plan to achieve an aim (an error of planning)
- Deviation from the process of care

Makary and Daniel [2016] focused only on preventable lethal events and found that medical error was the third cause of deaths in the United States, as shown in Table 1.1. The rising

Table 1.1. Top Five Causes of Death, United States 2013

Rank	Cause	Cases in Thousands
1	Heart Disease	611K
2	Cancer	585K
3	**Medical Error**	**251K**
4	COPD	149K
5	Suicide	41K

numbers signify the human impact of the problem we are facing.

1.2 Healthcare Expenditure

The second aspect of healthcare is the dramatic increase in healthcare expenditure. Since 1970 till 2010, healthcare expenditure grew by a staggering $34,795\%$ as shown in Table 1.2. This is evident in the many new treatments and technologies available to patients. A genuine measure of medical efficacy is patient improvement and satisfaction with healthcare service. However, this increase in expenditure is not reflected in the rising number of patient complaints nor is it showing in the surprisingly high number of patient deaths due to medical error. It is interesting to note that healthcare investment as a percentage of healthcare expenditure decreased consistently in the same period. A New, more efficient approach to mitigating healthcare risk is required.

The direct source we have of patient perception of the healthcare system is hidden within their complaint.

Table 1.2. National Healthcare Expenditures in Billions

Item	1970	1980	1990	2000	2010
National Health Expenditures	$74.6	$255.3	$721.4	$1369.7	$2595.7
Investment	$7.5	$19.9	$47.3	$83.3	$142.7
Investment % of Expenditure	10%	8%	7%	6%	5%

1.3 Complaints

Merriam-Webster defines a complaint as:

- Expression of grief, pain, or dissatisfaction
- Something that is the cause or subject of protest or outcry
  - A bodily ailment or disease
  - A formal allegation against a party

Complaints are an inherent part of human activity. Almost in all periods of recorded history, complaints have appeared in the literature. The complaints of Khun-Anup, an ancient Egyptian farmer who stumbled upon the property of the noble Rensi, son of Meru, were so eloquent that they were recorded on an extended set of papyrus to be preserved for the ages, as shown in Figure 1.2. The ancient Babylonian Nanni has inscribed his complaint on a clay tablet against Ea-nasir, for sending a shipment of copper ore of an inferior grade. The 1750 B.C. complaint is on display at the British Museum.

In current times we still complain, but we increasingly record our complaints in electronic textual format. In consumer research, customer feedback is used to find ways of improving service quality. Healthcare is catching up with the move to the Internet. According to [Fox and Duggan, 2013], one in four internet users has read or watched someone else's experience about health or medical issues in the previous 12 months. A subset of those users leaves feedback regarding their healthcare experience. Patient complaints represent a rich source of feedback regarding their healthcare experience. British Parliamentary Under-Secretary of State for Health[1] Lord Howe said: "*Every complaint holds valuable information on how patients feel about their care.*

[1] In office 6 May 2010 , May 2015.

Figure 1.2. Papyrus capturing the complaints of Khun-Anup.

Complaints can be the earliest symptom of a problem within an organization, and the NHS[2] *should use them to learn from and improve their service [Gold, 2013].*" This feedback is used to detect patterns associated with the service provider. An intervention plan can be generated to address those patterns and mitigate the risk of escalation. The intervention is based on the ability to classify patient complaints based on the object of the complaint. The total number of complaints received by the NHS was 162,019 in 2012-2013, which is about 430 a day. For the period from January to March 2016, the NHS started with 15,075 complaints brought forward from the previous period and 30,782 new written complaints in the fourth quarter[3]. At the end of quarter four, 19,722 complaints remained unresolved. These are carried forward to the next quarter. The statistics show the level of despair the system is facing, where 64% of the received complaints are still unsolved. This task is currently performed manually due to the nature of the dataset, which presents a challenge to automating the process.

1.4 Challenges

The current manually intensive solution is both limited and very costly. Scaling patient complaint modeling requires in-depth Natural Language Processing (NLP) of the complaint text. This

[2]The National Health Service (NHS) is the publicly funded healthcare system for England. It is the largest and the oldest single-payer healthcare system in the world.

[3]The fourth quarter ends in March.

problem is challenging due to the following reasons:

- The data come from multiple people from various affiliated institutions. The authors (patient advocates in different institutions) have different writing styles and input formats for reporting the patient complaints as well as the actions taken (if any) in response to a complaint.
- Different users may contribute to the same complaint. These can include the patient, a family member of the patient, a physician, or a nurse. Each user reflects his or her point of view and these points of view can be mutually contradictory.
- Variation in patients' ages can lead to discrepancies in the expressions used and the length of the complaint text.
- Large size of the data. A typical hospital handles thousands of patients, all of whom can potentially contribute unsolicited feedback. Unlike surveys where the hospital would request the information and control the format, unsolicited feedback is initiated by the patient, the patient's family, and in some cases, by the care provider.
- Time sensitivity, due to the medical nature of the domain. The system needs to triage the data per client and decide quickly if there is an issue. Time sensitivity is a factor that increases the difficulty of classification due to the downstream effect. If an issue is not correctly classified, the correct intervention plan will not be generated and the underlying issue may escalate and result in a malpractice suit driven by other patients.

1.5 Contributions

This dissertation focuses on the cross-disciplinary problem of developing an understanding of patient complaints and the robust approaches requiring minimal human supervision. The objective is to propose and implement a patient complaint analysis tool to be used for a wide range of healthcare providers on different scales that can address the challenges discussed above. We start with the basic understanding of patient complaints attempting a binary classification, followed by more advanced sentiment-based modeling of the patient complaint text. Finally, we explore domain-specific grammatical dependency feature extraction to model patient complaints. The major contributions of this dissertation are as follows:

- Study urgent patient complaints requiring physician action and how the patient text differs in urgent verses not urgent complaints. Investigate the benefits of using NLP features for modeling urgent patient complaints. Construct an urgency model, then implement an array of machine-learning models for urgent patient complaints classification. Compare the performance of the machine-learning models and select the best-performing one.

- Patient safety is a priority in healthcare organizations. Patient complaints classified as *Safety of Environment* issue have a very time sensitive impact in the healthcare domain. To that extent, we expand the research to investigate the benefits of using sentiment features for patient complaints modeling focusing on *Care Related*, *Safety of Environment*, and *No Complaint* classifications.

- Propose a novel approach of mapping complaints into sentiment vectors utilizing domain-specific enhanced Linguistic Inquiry and Word Count (LIWC) dimensions. Demonstrate and implement a machine-learning model for patient complaints classification based on the proposed approach and compare the results to current approaches.

- To accommodate the disparity in the used language and style, we explore using domain-specific grammatical dependency for feature extraction. We propose a method to extract domain-specific terms which we use to construct a set of grammatical dependency rules.

- Demonstrate machine-learning models for patient complaint classification using our rules to achieve high results as compared with basic features only.

- Implement all the developed methods and techniques on a real live medical dataset establishing both the validity of the approach and the efficacy of the methods of improving the current state of the art systems.

1.6 Organization

The rest of this dissertation is organized as follows: we continue our background work in Chapter 2 to lay the foundation for the following chapters and reduce the common sections. We elaborate through the concept of *Urgency* in Chapter 3 and demonstrate the robustness of the concept and our ability to predict it based on the patient complaint text. Chapter 4 expands on how we have implemented a sentiment-based model to extract the complaint sentiments and then use them to classify the complaints with high accuracy. Chapter 5 presents the more advanced grammatical dependency model and how we have used domain-specific terms along with dependency rules to reduce the extracted feature set into a smaller, more representative subset allowing higher prediction. We conclude with Chapter 6 in which we present an overview of our approaches and our future research directions.

Chapter 2

Related Work

In this chapter, we review the basic components of our work. Mainly, we consider healthcare risk mitigation, patient complaints, sentiment classification, and machine learning.

2.1 Healthcare Risk Management

Like any complicated business, healthcare involves various levels of risk to both the healthcare providers and patients. The risk in healthcare acquires additional emphasis due to the critical nature of the healthcare domain not only on the direct patient (the consumer) but also on the patient's friends and family extending to the local community. Consider a patient who is mistakenly treated and released with a contagious condition. The patient will impact everyone in contact with the patient until the case is discovered and treated. The financial impact on the community, let alone the healthcare organization, can be huge. For this and other reasons, government regulations have been put in place and enforced with organizations such as the Joint Commission on Accreditation of Healthcare Organizations (JCAHO), to assure the adequate level of risk management is set in place. Healthcare risk management program scope must cover all sources of risk and liabilities including: patient care, medical staff, employee, property, financial and other [Carroll et al., 1997]. Our research falls under the first two risk categories:

Patient care, which covers many related issues such as:

- HIPAA and other privacy regulations require maintaining confidentiality and appropriate release of patient medical information.
- Protection against neglect and abuse whether by staff, other patients, or visitors.
- Assuring that the patient is informed about and consents to the medical treatment.
- Upholding nondiscriminatory treatment of patients.
- Providing access to care concerns.

- Protecting patient valuables from loss or damage.

Medical staff, which the relevant risks would include:

- Peer review and performance improvement.
- Implementing medical staff disciplinary proceedings.
- Identifying and treating impaired physicians who pose a threat to the patient safety or other employee safety.

2.2 Patient Complaints

Most consumer business organizations pay considerable attention to consumer feedback. Consumer satisfaction has been linked to repeated business and loyalty [Bendall-Lyon and Powers, 2001; Luca and Atuahene-Gima, 2007]. The healthcare industry has been relying primarily on peer-to-peer reviews and steering committees. However, an increasing emphasis on evidence and outcome-based medicine is changing the industry. While evidence-based medicine is the process of systematically finding, appraising, and using relevant research findings as the basis for clinical decisions, outcome-based medicine focuses on closing the loop between the implemented treatment and the results obtained not only clinically but also as an impact on the patient [Rosenberg and Donald, 1995]. This renewed focus on the patient in healthcare industry emphasized the view of a patient as a healthcare consumer. Unlike business consumers, the patients have different types of feedback. Four forms of "patient-reported information" have been identified [Schlesinger et al., 2015], namely:

- patient-reported outcomes measuring self-assessed physical and mental well-being
- patient experience surveys, similar to business consumer surveys
- narrative accounts describing encounters with clinicians in patients' own words
- complaints/grievances signaling patients' distress when treatment or outcomes fall short of expectations

As a consumer, the patient generates two types of feedback: solicited and unsolicited. The solicited feedback is normally conducted as a survey. The survey could be medical in nature or general. In both cases, the care provider requests feedback from the patient. The unsolicited feedback, on the other hand, is a spontaneous feedback, which takes the form of a complaint in most cases.

A complaint is a negative feedback by definition, which has been shown to have a richer textual description. Patients and their friends and families are in a unique position to observe the interactions in the medical setting. Furthermore, reported observations by patients and

families to healthcare organizations in the form of spontaneous complaints of unprofessional behavior exhibited by a surgeon has been shown to be an indication of surgical team performance degradation [Catron et al., 2015]. For forthcoming reasons, patient complaints have been mined for clues to help identify developing risks and thus create intervention programs to mitigate them. The question now is how are those patient complaints processed and used?

2.3 Patient Complaint Classification

The fundamental step in creating an intervention program is patient complaint classification. Patient complaint classification is not an easy task; extensive training and preparation are required before even a human coder can start. A well-known program conducted at Vanderbilt University Medical Center, PARS, is a process that uses patient comments and complaints about their healthcare experiences to promote a kinder, safer, more reliable healthcare environment while addressing malpractice risk [Hickson et al., 2010]. PARS uses aggregate patient complaints data to identify providers associated with higher risk than the Risk Scores national benchmark [Stimson et al., 2010]. Preparing human coders for classifying patient reports in the PARS process takes the following phases:

Initial Training: The novice coders undergo 2 to 3 days of training to go over the coding manual and protocols. From there, they work in Training PARS for the next two weeks (on average).

Training Dataset Classification: Novice coders start by classifying a training version of the patient complaints. Experienced coders review their reports and provide extensive feedback.

Supervised Classification: Once novice coders are ready, they move to coding in a real dataset. They code in small patches of 5 to 10 reports and stop coding until experienced coders can review and give feedback.

Independent Classification: The learning curve is much steeper when a coder does not have knowledge of hospital structure and systems. Reading comprehension level comes next after previous medical experience. On average, a coder would require two months before independently classifying patient complaints. Even then they are constantly asking questions and requesting experienced coders to review their work.

The preceding shows how elaborate and demanding manual coding is.

The entire team of coders conducts reliability evaluations approximately every six months. For the date range of patient complaint reports included in the present study, the results of each inter-coder agreement reliability test are presented in Table 2.1. The inter-coder agreement is high due to the extensive training they undergo on the PARS classification.

Table 2.1. Reliability of coders.

Test Date	Number of Coders	Median alpha	Minimum alpha	Maximum alpha
July 2011	11	0.65	0.48	0.78
January 2012	10	0.75	0.65	0.83
January 2013	12	0.79	0.63	0.95
July 2013	13	0.85	0.64	1.00
December 2013	11	0.90	0.58	0.98

An experienced coder takes on average 12 minutes to classify an average patient complaint. Complaints can take anywhere from five to thirty minutes to classify depending on how long and complicated the patient complaint is. To put this in perspective, the smallest dataset we used contains 1500 complaints. It would require an experienced coder on average 18,000 minutes or 300 hours, or more than thirty-seven business days to classify the used dataset after they have been fully trained.

2.4 Natural Language Processing in Healthcare

Researchers increasingly have been looking at the use of NLP in healthcare, [Friedman and Hripcsak, 1999] predicted that if accurate clinical information were available electronically, automated applications could be developed using this information to improve patient care and lower healthcare costs. They also recognize the difficulties that prevented early adoption due to the complex nature of the medical data itself and the lack of a clear conceptual understanding. Another early usage of NLP was introduced in [Popowich, 2005] to automate healthcare claims processing which accesses both structured and unstructured information associated with medical insurance claims. They used call center logs to extract concepts with NLP techniques that act as indicators of recoverable claims. The goal is to determine whether the claims should be paid or flagged for potential fraud or abuse.

Computerized clinical decision support (CDS) aims to aid decision-making of healthcare providers and the public through easily accessible health-related information when it is needed. CDS recently renewed interest in the development of advanced NLP methods. NLP can facilitate adopting free-text information to drive CDS, representing clinical knowledge and CDS interventions in standardized formats and leveraging clinical narrative [Demner-Fushman et al., 2009]. They also noted that the past 45 years of CDS research had not been translated into widespread use and daily practice. Discharge summaries are a standard practice in the healthcare industry. Discharge summaries normally contain the following information:

Diagnosis: The diagnosis will immediately guide the practitioner's care plan.

Past Medical History: (Secondary Diagnoses) Known issues contributing to provide comprehensive care for this patient.

Medications (and allergies) This list is used as a guide for the admission orders at the nursing home.

Procedures and Significant findings: Contains the test results which were important in patient outcomes and medical management decisions.

Reason for admission and hospital treatment course: This section is free text and captures the story associated with the patient's hospitalization. It requires conveying the relevant details in a succinct, cohesive manner. It consumes the most time to develop and generally includes:

- The present patient condition.
- Key milestones that led to the current condition.
- Factors or events that affected management during the course of hospitalization.

Outstanding Issues: The requirements needed to provide continued care.

Follow-up appointments: Scheduling the next contact point with the patient and acquiring patient contacts.

Melton and Hripcsak [2005] showed that using NLP is an effective technique for detecting a broad range of adverse events in discharge summary and outperformed traditional and previous automated adverse event detection methods. More recently, [Wagholikar et al., 2012] used NLP to develop a computerized clinical decision support system (CDSS) for cervical cancer screening. They demonstrated that NLP could be effectively utilized to process free text in the EMR and provide key support recommendations. It is clear that NLP is playing an ever-increasing role in the healthcare industry.

More recently, research has been adding more emphases on social media text. As a community, we have evolved our interactions to adopt online platforms such as Twitter. Twitter is an online social networking service that enables users to send and read short, 140-character messages called "tweets". Only registered users can read and post tweets; however, everyone can read them. Twitter is especially interesting due to the open API (Application Program Interface) they provide for connecting to the services. This API allows researchers to collect tweets easily and thus process them. Twitter became more of an open social way for society to express itself instantaneously on a global scale. Naturally, tweets found their way into healthcare as patients express their health observations regarding themselves and their communities.

The work in [Collier et al., 2010] showed that text mining for global health surveillance is an emerging technology which is gaining increased attention from public, private and government organizations. Such implementation can be found in [Signorini et al., 2011], where they used a set of pre-specified search terms[1] in addition to keywords representing public concern regarding disease transmission[2] and finally, disease countermeasures[3] in combination with consumer concerns about pork consumption[4] to track levels of disease activity and public concern in the U.S. during the Influenza A H1N1 Pandemic. They demonstrate that Twitter traffic can be used not only descriptively, i.e., to track users interest and concerns related to H1N1 influenza, but also to estimate disease activity in real time, i.e., one to two weeks faster than current practice allows.

These successes motivated more research such as the use of a simple semantic-filtering, such as negation, hashtags, emoticons, humor, and geography, to enhance Twitter data analysis [Doan et al., 2012]. An example is presented for tracking influenza-like-illnesses (ILI). Their results indicate that simple NLP-based enhancements to existing approaches to mine Twitter data can achieve an improvement of 3.98% over the previous state-of-the-art method.

kalyanam et al. [2015] used Twitter to discern between facts and fabrications during the Ebola epidemic. The main contribution of the study is a proposed method to differentiate between credible and speculative tweets using hashtags as category indications. Tweets, blogs, and other social artifacts are general in nature. In contrast, the one direct unsolicited direct feedback is patient complaints. We need to understand how patient complaints are currently used.

2.5 Sentiment Analysis

One important application of natural language processing is sentiment analysis, or opinion mining, in which, the main focus is on how to identify and extract the attitude of the author with respect to some topic or the overall contextual polarity of a document. Early work on sentiment analysis considered the document level, attempting to determine whether the document overall is positive or negative [Pang et al., 2002]. However, if the document contains multi-perspectives, like addressing a questionnaire or different aspects of an experience, the document may contain more than one sentiment. Wilson et al. [2005] tackled this issue by focusing on phrase level contextual polarity. They considered the phrase in which a word appears polarity versus the word's prior polarity. Due to the popularity of sentiment analysis, lexical resources were explicitly created for supporting sentiment classification and opinion mining applications such as SENTIWORDNET [Baccianella et al., 2010].

[1] Flu, swine, influenza, vaccine, tamiflu, oseltamivir, zanamivir, relenza, amantadine, rimantadine, pneumonia, h1n1, symptom, syndrome, and illness.

[2] Travel, trip, flight, fly, cruise and ship.

[3] Wash, hand, hygiene and mask.

[4] Pork and bacon.

Our approach is based on the observation that realizing what compels the patient to complain is the key to modeling the patient. A sentiment can be defined as an attitude, thought, or judgment prompted by a feeling [Munezero et al., 2014]. Attempting to model patient complaints in a non-sentiment approach ignores the fundamental drive behind the complaint inception, which explains why when faced with a challenging dataset, we can obtain better classification accuracy incorporating sentiments.

The sentiment analysis of text [Liu and Zhang, 2012] is challenging because humans express their sentiments in subtle, domain-specific ways. We adopt Pennebaker and colleagues' Linguist Inquiry and Word Count (LIWC) approach [Pennebaker and King, 1999a] as a way to determine the sentiment-relevant words in a complaint. LIWC provides subjective dictionaries that have been validated through previous studies. LIWC has been shown to tap into the psychological meaning of words [Tausczik and Pennebaker, 2010]. Furthermore, the studies conducted by [Schultheiss, 2013] and [Kahn et al., 2007] demonstrated the causal validity of LIWC-based emotive scores by documenting their sensitivity to motive arousal. Measuring consumer satisfaction is shown in [Ren and Quan, 2012] to be possible through linguistic-based emotion analysis and recognition which extend the LIWC approach. Extensive validation has been conducted on the use of LIWC techniques in diverse domains, e.g., [Donohue et al., 2014; Smith, 2010].

2.6 Machine Learning

Machine learning can be defined as the algorithms that give computers the ability to learn without being explicitly programmed [Samuel, 1959]. Classification is a subset of machine learning in which the focus is on learning how to classify a dataset. A wide variety of classifiers exists that range between general classifiers to domain-specific ones. We adopt the following classifiers in our research:

Bagging Was suggested by [Breiman, 1996] as a method for generating multiple versions of a predictor and using these to get an aggregated predictor. If number of versions is more than one, bootstrap replicates of the learning set are generated and then used as new learning sets.

Boosting Aggregates a set of weak learners (classifiers that perform slightly better than random) to create a strong learner by weighting them appropriately [Schapire, 1990].

GLMNET An implementation of the Lasso and elastic-net regularized generalized linear models. GLMNet is popular for domains with large databases [Friedman et al., 2010].

Max Entropy Is a probabilistic classifier that selects the model with maximum entropy from among a set of models and uses it to classify the data [Osborne, 2002].

Neural Network A family of models inspired by the central nervous systems of animals, in particular, the brain [Venables and Ripley, 2002]. They are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Weights are used to tune the relation between the nodes based on experience, rendering the network capable of learning [Burke et al., 1994]. Neural networks gained popularity in text classifications across different languages due to it's flexibility and adaptability [Li and Park, 2007] [Harrag and El-Qawasmah, 2009].

Random Forests Selects the best performing from among multiple learning algorithms so as to improve predictions. Significant improvements in classification accuracy can be gained from growing an ensemble of trees and letting them vote for the most popular class. Care must be taken with setting the size of the forest so that a time bound for the solution can be achieved [Breiman, 2001].

SLDA Scaled Linear Discriminant Analysis expresses one dependent variable as a linear combination of other variables similar to ANOVA, but with the difference that SLDA assumes continuous independent variables and categorical dependent labels. SLDA is widely used in image and pattern recognition [Martínez and Kak, 2001].

SVM Support Vector Machines divides the dataset via a set of hyperplanes during the learning phase and maps new data to fall behind one of the hyperplanes. SVM has been used for text classification [Peng et al., 2008].

Chapter 3

Triaging Patient Complaints Requiring Physician Action by Modeling Patient Complaint Text

3.1 Introduction

In this chapter, we start our research by examining urgent patient complaints that require physician action. Patient complaints are an important source of information for health care organizations regarding the quality of care [Hayden et al., 2010]. Patients are uniquely positioned to make observations about the care they receive, particularly when health care professionals or organizations fail to meet or exceed their expectations. When patients and family members share their observations, service recovery can take place. Put simply, service recovery is the process by which organizations attempt to "make right" what went wrong for patients and families. Beside regulation by the Centers for Medicare & Medicaid Services (CMS) requirements to address patient complaints, service recovery is an important practice enabling the organization of addressing:

Moral Motivation demonstrate the commitment to deliver safe, compassionate, quality care.

Marketing Motivation helps rebuild patient confidence and improves patient retention. Distinguishing the organization and builds a recurring loyal customer base.

Financial Motivation may reduce revenue loss and risk associated with dissatisfied patients and families to improve the organization's bottom line.

Some patient complaints contain information that may necessitate urgent action on the part of the health care organization and/or the health care professional. For example, patients' descriptions of behaviors that are mandated by law, regulation, or policy to be investigated or

require institutional action must be identified and responded to in a timely manner [Pichert et al., 2008]. Such mandated events could include assertions of possible sexual boundary violations, discrimination, touching or grabbing a patient, descriptions of possible impairment by drugs or alcohol and other serious events. Similarly, assertions of other significant safety risks (e.g. asked to sign a consent on a wrong side) or threats to share the complaints with lawyers and/or the media may necessitate engagement by various organizational leaders to address and mitigate potential risk.

However, many healthcare organizations receive thousands of complaints a year [Catron et al., 2015; Bendall-Lyon and Powers, 2001; Luca and Atuahene-Gima, 2007]. Manual review of these complaints by trained coders has been shown to be reliable and valid [Pichert et al., 2013], but is time-consuming and may occur some weeks or months after the complaint is received. In addition, scalability of human coding presents logistical and time challenges. Thus, there is a need to triage patient complaints to identify complaints that necessitate urgent action on the part of the health care organization or the professional that is timely, reliable, and scalable. We describe a study in which we implemented several well-known machine learning classifiers to optimally detect urgent patient complaints using data from the Patient Advocacy Reporting System (PARS add the TM), a national program which draws data from multiple hospitals patient complaint reporting systems to identify professionalism concerns and malpractice risk among health care professionals [Pichert et al., 2013].

3.1.1 Problem, Challenges, and Approach in Brief

We posit that complaint text can be used to discern the urgency of a complaint and thus correctly and efficiently triage a new complaint. The problem we address is how to detect whether a given patient complaint is urgent. We seek to achieve accuracy that is on par with existing manual approaches.

Our problem is challenging due to the following factors:

- Physician urgency is not easy to characterize. What may sound urgent in one case, may not hold true in the other. The vocabulary used to describe urgent complaints overlaps with that used for non-urgent ones, partly because of the common effect of the medical setting.

- Time sensitivity in classification vitally important. Ignoring an urgent complaint could escalate the underlying concern dramatically and result in increased expenses. However, a false positive would result in wasted time and effort.

- Text for a single complaint may gather multiple perspectives, including the patient, the patient's family, friends, and care providers. These parties have different and possibly conflicting objectives.

Our approach pairs extracted physician-related actions from resolved patient complaints along with features extracted from those complaints to classify complaints as urgent or non-urgent.

For our comparisons, we (1) implement a framework that employs six well-known classifiers and (2) experiment with two methods of feature extraction from complaint text.

3.2 Related Work

The bulk of the textual artifacts in healthcare can be found in two main sources: clinical and non-clinical. *Clinical* textual artifacts are largely entries in the medical chart, comments on the case, or physician notes. They tend to be consciously made, well-structured, and focus on treatment (including diagnoses) of the patient. *Nonclinical* textual artifacts include unsolicited patient feedback and often revolve around complaints. The text is variable, may contain abbreviations, and may extend beyond the actual treatment or diagnosis.

Previous research has focused on clinical textual artifacts [Baud et al., 1992]. Recent research demonstrates the possibility to apply Natural Language Processing (NLP) of electronic medical records to identify postoperative complications [Murff et al., 2011]. Bejan and Denny [2014] showed how to identify treatment relationships in clinical text using a supervised learning system that is able to predict whether or not a treatment relation exists between any two medical concepts mentioned in clinical notes.

Cui et al. [2014] explored a large number of consumer health questions. For each question, they selected a smaller set of the most relevant concepts adopting the idea of the Term Frequency-Inverse Document Frequency (TF-IDF) metric. Instead of computing the TF-IDF based on the terms, they used Concept Unique Identifiers (CUIs). Their results indicate that we can infer more information from patient comments than commonly thought. However, questions are short and limited, whereas patient complaints are rich and elaborate.

Sakai et al. [2016] concluded that how risk assessment and classification is configured is often a decisive intervention in the reorganization of the work process in emergency services. They demonstrate the textual analysis of feedback provided by nurses can expose the sentiment and feelings of the emergency workers and help improve the outcomes.

Temporal information in discharge summaries has been successfully used [Zhou et al., 2006] to classify encounters, enabling the placement of data within the structure to provide a foundational representation upon which further reasoning, including the addition of domain knowledge, can be accomplished.

Additional research [Garla et al., 2011] extended the clinical Text Analysis and Knowledge Extraction System (cTAKES) with a simplified feature extraction, and the development of both rule and machine-learning based document classifiers. The resulting system, the Yale cTAKES Extensions (YTEX) can help classify radiology reports containing findings suggestive of hepatic

decompensation. A recent systematic literature review of 85 articles focusing on the secondary use of structured patient records showed that EHR data structuring methods are often described ambiguously and may lack clear definition as such [Vuokko et al., 2015].

3.3 Nature of Urgent Complaints

We now briefly describe PARS since we use it both as a source of motivation and the ground truth with which to compare our approach. The goal of tools such as PARS, is to promote a kinder, safer, more reliable healthcare environment while addressing malpractice risk [Hickson et al., 2010]. PARS uses aggregate patient complaints to identify professionals associated with higher risk of malpractice than the national benchmark [Stimson et al., 2010].

When a patient at a hospital has an issue that they cannot resolve on their own, also spontaneous, they contact the Office of Patient Relations or the Office of Patient Affairs at the hospital. A patient advocate helps understand the concern. Once the concern is resolved, the patient advocate enters notes about that issue into a computer system. Once a month, all of the hospitals and medical centers send all those narratives of patient complaints to PARS. Professional coders classify patient complaints.

For the objective of this research, we group complaints into two main categories, as below.

Urgent: complaints that require further medical escalation, most likely will involve a physician interaction:

- *Patient states he came in for a test and his IV infiltrated. He states he is extreme pain, was dizzy on the way home. He states his arm is swollen, black and blue and he is angry. Patient kept repeating he should call his lawyer. Patient reported blood all over his shirt. He states he wants to speak to someone who can give him some answers quick.*
- *Patient states during office visit to... Dr. XX that he abruptly changed her medications and sexually harassed her.*
- *Patient reported that Dr. XX snickered and laughed when sexual orientation* was discussed. *Patient Felt he was mistreated because of his sexual orientation.*

Not Urgent: complaints that concern billing or requesting information (or are not a complaint at all). They normally do not require medical escalation, and can be handled by the staff:

- Pt indicates need for medical records from past treatment at ZZ. Action Taken: SL to send pt copy of Release of Information form to fill out and return to ZZ Medical Records department.
- *I had to wait 2 hours to be seen. My time is valuable, too.*

- *Mr. XX called wanting to speak to someone about the pool hours at Sports Medical Center.*

An intervention plan is created to address the risks through local peers and leadership. The intervention ranges between awareness, authority and administrative disciplinary action [Su et al., 2010].

3.3.1 Dataset

We implement and present a comparison of several well-known machine learning classifiers to detect patient urgency. We compare these classifiers using a real-life dataset containing 14,335 patient complaints associated with 768 physicians that were collected by the Patient Advocacy Reporting System (PARS) developed at Vanderbilt and associated institutions.

3.4 Approach

No single term or attribute signifies complaint urgency. Therefore, we approach the problem as one of clustering text into one of two clusters. Documents are commonly represented as a sparse vector over the entire feature set consisting of all distinct terms over all documents. Two major drawbacks are (1) high dimensionality, i.e., a large number of features; and (2) feature sparsity, i.e., features appearing in only a few documents [Aggarwal and Yu, 2000].

Accordingly, we implement a framework that consists of the following steps:

- Preprocess the documents to remove stop words and numbers and to perform stemming.
- Run Monte Carlo cross-validation [Xu and Liang, 2001] using ten splits, for each we:
  - Randomly sample the training and testing dataset from our corpus.
  - Extract features through generating sparse representation of the documents based on TF or TF-IDF.
  - Reduce features by removing sparse terms.
  - Learn a model to predict the labels.
- Compute the average accuracy, sensitivity, and specificity for each classifier.
- Select the best performing classifier.

Figure 3.1 illustrates the former steps.

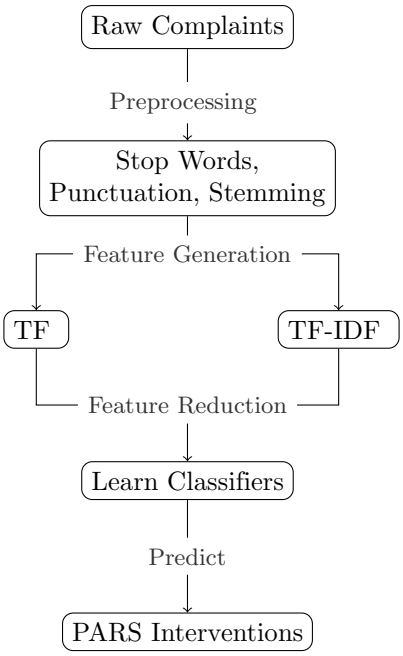

Figure 3.1. Physician urgent complaint classification steps.

3.4.1 Feature Extraction

The first step is to convert patient complaints to a set of representative features. Wilcox and Hripcsak [Wilcox and Hripcsak, 2003] show that domain knowledge representation can vary between task-specific and representation-specific knowledge. Medical knowledge is specific to the conditions being identified and is essential to classifying clinical reports. As in our case, Wilcox and Hripcsak emphasize attribute or feature extraction. Generating medically relevant features requires an understanding of the medical report or the underlying meaning of the text. Our approach associates medical relevance with feature relevance to the document.

We compare two methods for feature extraction, namely: TF (Term Frequency) and TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF seeks to emphasize the importance of a word to a document in a collection or corpus [Rajaraman et al., 2012].

TF-IDF is widely used in information classification and retrieval [Dumais et al., 1998]. The idea is simply to multiply the TF (Term Frequency) with IDF (Inverse Document Frequency) computed with respect to the entire corpus as shown in Equation 3.1.

$$\mathrm{TFIDF}\,(t) = \mathrm{tf}\,(t,d) * \log \frac{N}{n_t} \tag{3.1}$$

where $\mathrm{tf}\,(t,d)$ counts the frequency by which term t appears in document d, N is the total number of documents in the corpus, and n_t is the number of documents in which the term t appears.

The idea of incorporating IDF is to reduce the weight on words that occur frequently in each document but are not sufficiently selective. For example, words "doctor" and "nurse" would occur too commonly in patient complaints to be useful for retrieval or selection.

We adopt TF-IDF for feature extraction as follows:

- Generate a vocabulary of unique terms.
- Generate term frequency per document.
- Generate inverse document weight per term.
- Replace the frequency with the TF-IDF weights using Equation 3.1.

The result is a sparse vector representation of the document.

3.4.2 Feature Reduction

Feature reduction aims at reducing the number of features while maintaining the underlying meaning of the document. A smaller number of representative features can maintain a comparable level of prediction performance while reducing noise and unnecessary processing. Both TF and TF-IDF generate a large number of features, the majority of which are not relevant in predicting the urgency. To reduce the number of features, we remove sparse features. The sparsity of a term is defined as the percentage of documents that this term occurs to the entire corpus, as shown in Equation 3.2.

$$\text{Sparsity} = \frac{n_t}{N} \tag{3.2}$$

where n_t is the number of documents in which the term t appears, and N is the total number of documents in the corpus. A term with 0.9 sparsity means the term appears in at least 90% of the documents, while a term with 0.99 sparsity appears in at least 99% of the documents.

We repeated the Monte Carlo cross-validation training and prediction while varying the sparsity from 0.9 to 0.99 to assess the minimum number of features to select and still maintain the desired prediction performance. The following example shows some selected word stem features organized into four groups for illustration purposes:

Financial acct, charg, close, bill and call.

Medical cardiac, cardiology, complications, injuri and coronari.

Facility center, clinic, access, action and assist.

Care complaint, concern, attach, avail and care.

3.4.3 Classifier Selection

The final step is to assess the best classifier to employ for our problem at hand. Due to the special nature of the problem, selecting a classifier prospectively is difficult. We implement a supervised learning framework to capture the relation between patient text and the resultant physician action. The models then can detect whether the complaint is urgent or not. Our framework supports six well-known classifiers. We used RTextTools [Jurka et al., 2014] as the library to implement the following classifiers: Boosting, GLMNET, Max Entropy, Random Forests, SLDA, and SVM. Section 2.6 provides a detailed description of each classifier.

After experimenting with the above classifiers on the same dataset, we select the best overall performing classifier.

3.5 Evaluation

We divided the dataset into a training and a testing dataset. We used one of the six classifiers to learn a model over the mapped dataset. We then used the testing dataset to validate the accuracy of our classifiers. Accuracy is defined by Equation 3.3, where TP, FP, TN, and FN refer to true and false positive and negative.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \tag{3.3}$$

Sensitivity captures how many patients with a condition are detected—i.e., the avoidance of false negatives as in Equation 3.4.

$$\text{Sensitivity} = (\text{TP})/(\text{TP} + \text{FN}) \tag{3.4}$$

Specificity captures how many patients without a condition are not detected—i.e., the avoidance of false positives shown in Equation 3.5.

$$\text{Specificity} = (\text{TN})/(\text{TN} + \text{FP}) \tag{3.5}$$

3.5.1 Results

We first report our full ten-splits results for each classifier predictions. Figure 3.2 shows the results obtained using TF extracted features. We experimented with changing the sparsity from 0.9 to 0.99 to reduce the number of selected features. The prediction accuracy either slightly improved or remained steady with the reduced number of features except in the random forests case where the accuracy peaked and dropped slightly at the end of the range.

The case is a bit different with results obtained using TF-IDF extracted features, as shown in Figure 3.3. The prediction of all classifiers improved notably as we reduced the number of

selected features. The gap between the best performing classifier using TF-IDF and the rest of the classifiers is more pronounced as well. Since results are generally better at higher sparsity, we report the detailed results at sparsity of 0.99 with accuracy, sensitivity, and specificity for both TF and TF-IDF in Table 3.1 over each of the six classifiers we implemented.

Table 3.1. Classifiers TF versus TF-IDF Accuracy, Sensitivity, and Specificity using ten-splits Monte Carlo cross-validation at 0.99 Sparsity.

Classifier	TF			TF-IDF		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SLDA	0.76	0.72	0.80	0.74	0.66	0.83
SVM	0.79	0.71	0.86	0.75	0.67	0.82
GLMNET	0.76	0.71	0.81	0.75	0.64	0.86
MAX ENTROPY	0.77	0.71	0.83	0.77	0.69	0.84
BOOSTING	0.70	0.85	0.55	0.73	0.82	0.64
RANDOM FORESTS	0.80	0.74	0.87	0.82	0.76	0.87

The best performing classifier was Random Forests with an 82% accuracy using TF-IDF for feature generation, followed by SVM classifier which achieved 79% accuracy using the simpler TF for feature generation.

3.5.2 Limitations

Modeling patient complaint urgency is a challenging problem. We limited feature extraction to TF and TF-IDF which, although generating robust results still poses the question of exploring deeper analysis. We can see the advantage of exploring more advanced NLP methods to dive into the underlying language structure and reduce the noise. Another limitation of our work is our focus on the binary classification we have used. Urgent complaints requiring physician actions are not all the same, rather, some may be treatment options, environmental issues, physician behavioral issues or competency questions. It would be interesting to expand our scope to address those issues. Finally, our dataset size is adequate for an exploratory study. However, a wider more inclusive dataset would cement the results.

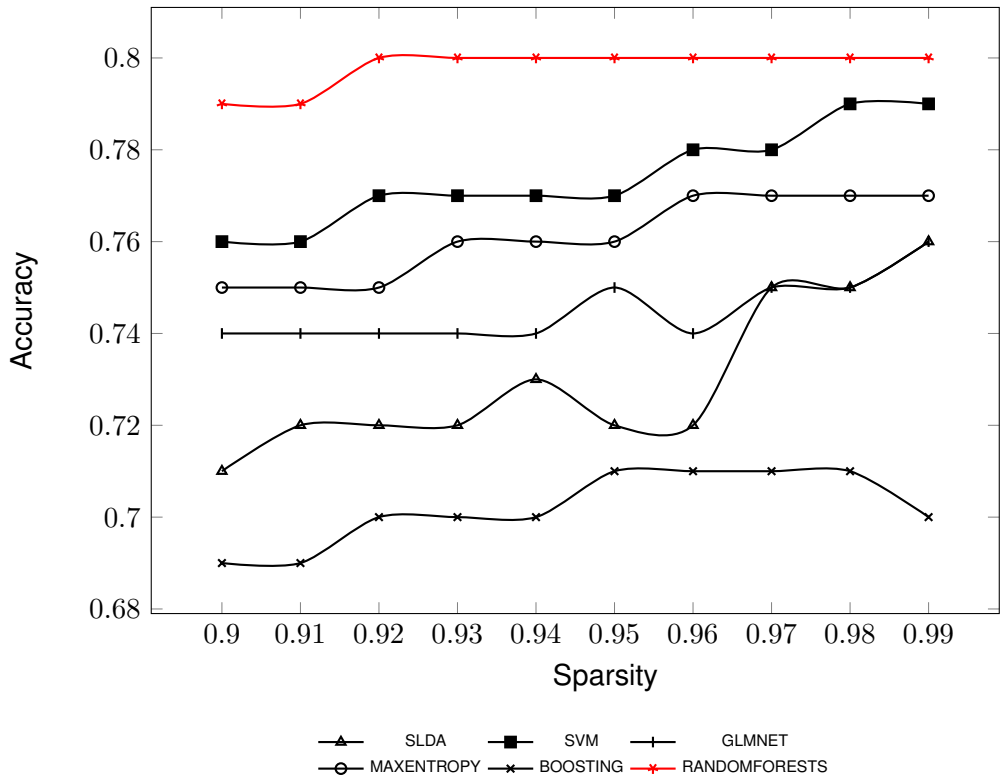

Figure 3.2. Urgency detection TF ten-splits Monte Carlo cross-validation accuracy.

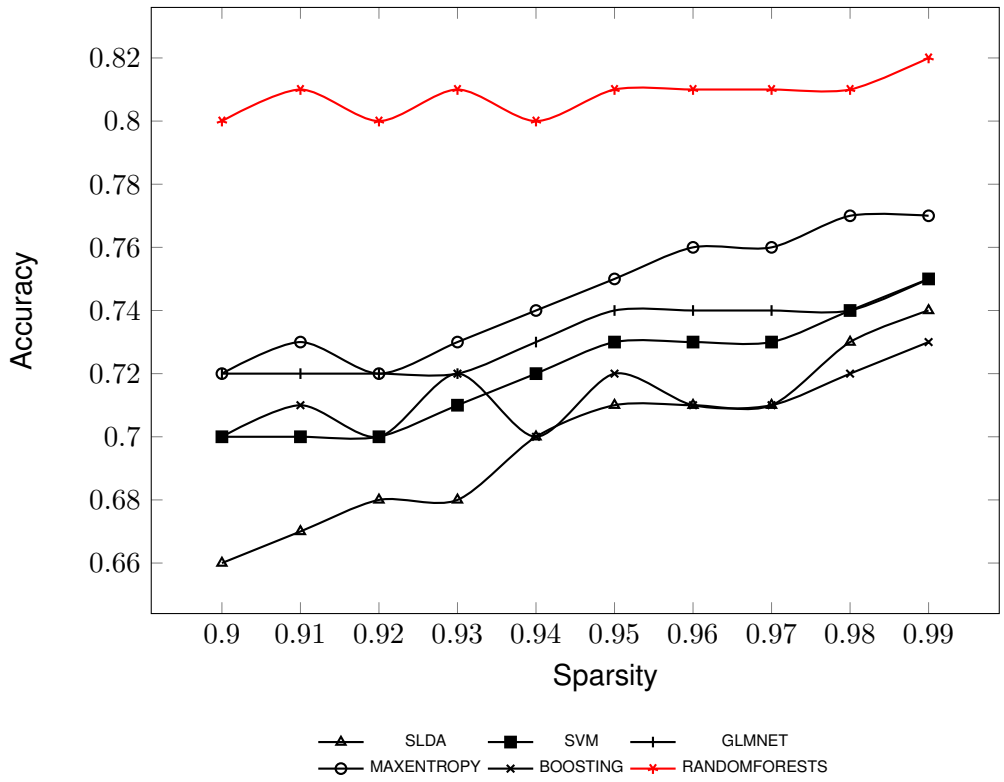

Figure 3.3. Urgency detection TF-IDF ten-splits Monte Carlo cross-validation accuracy.

Chapter 4

Using Sentiment Analysis for Classifying Patient Complaints

4.1 Introduction

We continue our research by addressing the deeper classification of patient complaints. Classifying patient complaints is made difficult by the complexity of linguistic representation. Current practice relies upon manual classification, which limits scalability. An automatic approach can potentially improve response time and scale, thereby enhancing opportunities to promote physician accountability for safe and respectful care. When patients and family members share their observations, in the form of a patient complaint, service recovery can take place. Put simply; service recovery is the process by which organizations attempt to "make right" what went wrong for patients and families [Hayden et al., 2010]. In this regard, how can we measure, track, and improve healthcare quality of service?

Following NPDB Research Statistics [Bank, 2014], we define *adverse action* as "(1) An action taken against a practitioner's clinical privileges or medical staff membership in a healthcare entity, or (2) A licensure disciplinary action." As Fig. 4.1 shows, the number of adverse actions has gone up, which generates avoidable cost and drag to process, triage, document, and respond to them. Although the total dollar amounts paid per year on malpractice for all of United States went down from 4.8 billion dollars in 2004 to 3.6 billion in 2014, the decline has mostly leveled off in the last five years. This recent plateau indicates that current risk mitigation and intervention techniques have reached their capacity and will not scale to meet the increase in adverse action reports. New approaches are needed to enable further reduction.

In commerce, customer feedback is a valuable resource for validating the quality of service and improve customer satisfaction [Bendall-Lyon and Powers, 2001; Luca and Atuahene-Gima, 2007]. Healthcare is changing through the wealth of user-generated textual artifacts. According

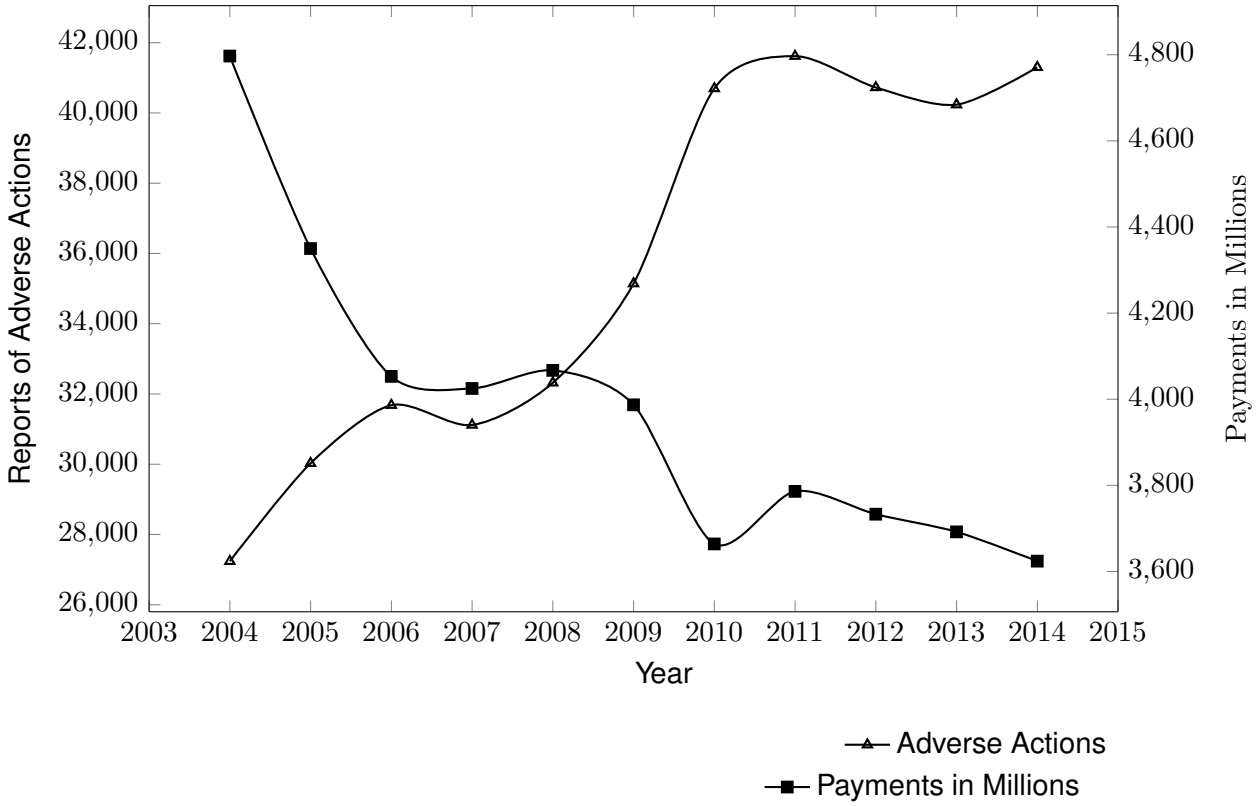

Figure 4.1. Numbers of adverse actions and malpractice payments in the US over the last decade or so [Commission, 2008].

to [Fox and Duggan, 2013], one in four Internet users has read or watched someone else's experience in health matters in the previous year. A subset of those users leaves feedback on their healthcare experience. Zhao et al. [2014] show how the sentiment included in healthcare posts can reflect the user emotions at the time of posting. Using Natural Language Processing (NLP) has been shown to be an effective technique for detecting a broad range of adverse events in discharge summary and outperformed traditional and previous automated adverse event detection methods [Melton and Hripcsak, 2005].

Patient feedback regarding the quality of service has significant ramifications on how patient outcomes are assessed and thus on healthcare provider financial viability. The most valuable and direct feedback are patient complaints. Patient complaints can help healthcare organizations to identify unsafe and dissatisfying behaviors as well as avoidable variability in performance. Reported observations by patients and families to health care organizations in the form of spontaneous complaints of unprofessional behavior exhibited by the surgeon has been shown to be an indication of surgical team performance degradation [Catron et al., 2015]. The Patient Advocacy Reporting System (PARS) developed at Vanderbilt and other medical centers is a system

used to capture, classify, and address such complaints. Effective and efficient classification is crucial to subsequent responses. Current manual techniques are expensive and slow. We develop an approach that automates the classification and produces human-competitive results.

Self-expression by patients and their families has been shown to reveal their health experience [Liehr et al., 2002]. It is natural to explore how the emotions detected from such text can be used to reveal the main topics expressed in the patient complaints. An interesting question arises: do those comments reflect deeper meaning? Can we infer the underlying emotions? An established approach is the Linguistic Inquiry and Word Count (LIWC) [Pennebaker et al., 2001] which provides a dictionary of words representing the emotional content of the text. We posit that the complaint types have distinct distributions over the underlying dimensions, which we can use to classify them.

4.1.1 Problem and Motivation

The technical problem we address is how to take real-life patient complaint data and classify it into one of a small number of categories. The determination of the categories is an essential step in responding to patient complaints. In the current process, this step is carried out manually through extensively trained personnel. The problem we address is to perform the classification automatically with accuracy that is competitive with manual approaches.

This problem is challenging for the following reasons:

- Variable data sources: The data come from multiple people from multiple affiliated institutions. The authors (patient advocates in different institutions) have different writing styles for reporting the patient complaints as well as the actions taken (if any) in response to a complaint.

- Multiple data sources: Different users may contribute to the same complaint. These can include the patient, a family member of the patient, a physician, or the nurse. Each user reflects his or her point of view and these points of view can be mutually contradictory.

- Generational differences: Variation in patients' ages can lead to discrepancies in the expressions used and the length of the complaint text.

- Size of the data: A typical hospital handles thousands of patients, all of whom can potentially contribute unsolicited feedback. Unlike surveys where the hospital would request the information and control the format, unsolicited feedback is initiated by the patient, the patient's family, and sometimes the provider.

- Time sensitivity in classification: It is important to classify a feedback so as to produce a correct intervention plan and respond before the problem escalates leading to a potential malpractice lawsuit.

4.1.2 Analyzing Sentiment in Text

Our approach is based on the observation that realizing what compels the patient to complain is the key to modeling the patient's concerns. A sentiment can be defined as an attitude, thought, or judgment promoted by a feeling [Munezero et al., 2014]. Attempting to model patient complaints in a nonsentiment approach ignores the fundamental drive behind the complaint inception, which explains why when faced with a challenging dataset, we can obtain better classification accuracy incorporating sentiments.

The sentiment analysis of text [Liu and Zhang, 2012] is challenging because humans express their sentiments in subtle domain-specific ways. We adopt Linguist Inquiry Word Count (LIWC) [Pennebaker and King, 1999b] as a way to determine the sentiment-relevant words in a complaint. LIWC provides dictionaries that have been validated through previous studies. LIWC has been shown to tap into the psychological meaning of words [Tausczik and Pennebaker, 2010]. Additionally, previous studies [Jurka et al., 2014; Schultheiss, 2013] demonstrate the causal validity of LIWC-based emotive scores by documenting their sensitivity to motive arousal. Measuring consumer satisfaction is shown to be possible through linguistic-based emotion analysis and recognition based on LIWC [Ren and Quan, 2012].

4.2 Materials and Methods

PARS is a tool and a process that uses patient comments and complaints about their healthcare experiences to promote a kinder, safer, more reliable healthcare environment while addressing malpractice risk [Hickson et al., 2010; Stimson et al., 2010].

Patients with unresolved issues contact the Office of Patient Relations or the Office of Patient Affairs at the hospital. A patient advocate helps resolve the issue and documents the complaint in the system. Narratives of patient complaints are sent to Vanderbilt CPPA. Professional coders mark the relevant parts of the complaint narratives so that they can be shared with the care provider. Examples of the categorizations are as below. (Care and Treatment are the most common complaints.)

Care and Treatment *I needed to be examined . . . Dr. XX never touched me except to shake hands., I feel Dr. XX has put off my surgery too long with no good reason.*

Safety of Environment *Wife states when nurse started IV that the patient bled and there was a large amount of blood on the floor and he had blood on his gown.*

No Complaint *Mr. XX called wanting to speak to someone about the pool hours at Sports Medical Center.*

An intervention plan is created to address the risks through local peers and leadership. In most cases, the plan has three levels [Pichert et al., 2008]:

Level I: Awareness A confidential interaction with the care provider to make sure the provider is aware of the risk.

Level II: Authority An authority intervention is provided via an authority figure such as the Department Chair.

Level III: Administrative An administrative disciplinary process is initiated to assure organization risk mitigation.

By implementing the PARS recommendation, the organization can detect and address risk before it escalates [Su et al., 2010].

4.2.1 Data Composition

We examined a longitudinal dataset containing 7,400 complaints associated with 457 physicians collected from June 2002 to September 2014. We observed that many complaints have a standard beginning expressing the department, date, and the last person updating the record. Such boilerplate consumes a lot of characters. For this reason, to ensure that we had substantial information content, we restricted our dataset to comments of 150 characters or more. This dataset contained 3954 complaints, 16 labeled SAFETY, 3504 labeled CARE, and 434 labeled NO COMPLAINT.

4.2.2 Our Approach

In order to analyze the sentiment content of patient complaints we first enriched the LIWC dimensions with three simple healthcare specific dimensions, then we developed a mapping function using the LIWC lexicon. The simple healthcare specific dimensions help account for the specific domain that we are addressing. The mapping generates a sparse representation of the patient complaint in the form of a vector containing the hits for each of enriched LIWC's 67 dimensions. Attempting to learn a model using all of the 67 dimensions is exhaustive. Our problem can then be viewed as a feature extraction followed by feature reduction, where the extracted features are the LIWC dimensions that we seek to reduce for the purpose of learning a model that can predict the patient complaint classification. The desired solution will need to determine the minimal number of features to obtain the best prediction performance.

. Our approach consists of four steps:

- Enhance with domain-specific dimensions.
- Extract features through mapping to LIWC dimensions.
- Reduce features to select the relevant LIWC dimensions.
- Learning a model to predict the labels.

Next, we describe each step in detail.

4.2.3 Healthcare Specific Domains

LIWC is based on ordinary language whereas patient complaints tend to use specialized words. We select terms based on their TF-IDF weights [Jing et al., 2002], additionally selecting medical terms. Our steps are:

- Generate a TF-IDF vector of the entire corpus.
- Select the top 250 TF-IDF weighted terms.
- Remove any term that appeared in the LIWC terms.
- From the remaining, select words that have medical meaning.

As in LIWC, we retain misspelled words with high TF-IDF weight such as surgeri. The resulting list partitions into three dimensions:

Patient contains words such as patient, patients, pt

Medical Personnel contains words such as physician, nurse

Treatment contains words such as surgery, procedure, diagnosis

We include these dimensions in the lexicon we use below.

4.2.4 Feature Extraction: Mapping to LIWC Dimensions

Feature extraction focuses on computing a representation of the input data that facilitates learning a model of it. Using LIWC dimensions for feature extraction has been well established in work done by [Osherenko and André, 2007]. We consider the 64 dimensions defined in the LIWC framework and the three domain-specific dimensions we generated above (total of 67 dimensions) as the source of our features. The dimensions, in essence, are lists of words (including their variants and derivatives) organized according to their psycholinguistic significance. For example, the *Negemo (Negative Emotions)* dimension includes words connoting negative emotion, such as "alone," "anger," "stress," "sufferer," "suffering,".... Likewise the *Family* dimension includes family-related words, such as "bro," "brother," "daughter," "ex," "hubby," "husband," ....

We map each patient complaint to the LIWC dimensions. Specifically, our mapping function measures the frequency with which a certain dimension is repeated in a complaint using Equation 4.1.

$$f_d(C_n) = \sum w_{d,n} \tag{4.1}$$

where $w_{d,n}$ counts a word that appears both in dimension d and complaint C_n.

Using the above, we generate a compact representation of each complaint as a vector: each element in this vector equals the mapping to a particular LIWC dimension as Equation 4.1 explains. In general, we would consider not all, but only some selected dimensions.

To understand the mapping, consider the following complaint with respect to six LIWC dimensions, {Posemo (Positive Emotions), Negemo, Family, Money, Relig, Home}:

> My father and hubby almost slipped on the wet floor, even my bro who is well built, had a hard time to control himself. So stupid, especially that we have a huge bill to pay.

The underlined words occur in the selected LIWC dimensions: "father," "hubby," "bro" in *Family*; "stupid" in *Negemo*; and "bill," "pay" in *Money*. Therefore, this complaint maps to the six-element vector (over the selected dimensions):

$$\langle 0, 1, 3, 2, 0, 0 \rangle$$

If we were to append *Relativ* and *Space* to our list of dimensions, "room" in *Relativ* and "floor" in *Space* would be relevant, producing an eight-element vector

$$\langle 0, 1, 3, 2, 0, 0, 1, 1 \rangle$$

Notice that some words may appear in two or more LIWC dimensions. Our mapping counts such a word in each dimension with the intuition that such a word indicates multiple dimensions are relevant. However, as we explain below, we select the dimensions themselves so as to reduce the features thus countering any systematic redundancy across the dimensions.

4.2.5 Feature Reduction: Selecting the Relevant LIWC Dimensions

In general, having too many features can lead to an over-fitted model. This problem is exacerbated by features that are mutually correlated. For this reason, we adopt the well-known idea of feature reduction [Nunes et al., 2004; Kohavi and John, 1997].

We adopt a variation of steepest-ascent hill climbing [Russell and Norvig, 2009] to select a small set of dimensions as our features. Specifically, we identify three LIWC dimensions that yield good prediction as follows. The idea is to compute the accuracy resulting from several sets of three dimensions on the training data and then use the best of those sets on the test data. That is, the objective function for our hill climbing is prediction accuracy using a model based on the selected features.

We maintain five best sets of dimensions. We begin from the five singletons of dimensions that give the best prediction accuracy. For each of these five singletons, we add to it each of the

remaining 66 dimensions in turn and evaluate the accuracy of the resulting pairs. From these 5×66 pairs, we select the best five pairs. Next, for each of these five singletons, we add to it each of the remaining 65 dimensions in turn and evaluate the accuracy of the resulting triples. From these 5×65 triples, we select the best five triples.

Although our algorithm does not ensure finding the optimum set of dimensions, it presents a good tradeoff by producing a good enough set of dimensions without dealing with the combinatorial explosion of considering all possibilities. In general, we could proceed in this manner to find the best five quadruples, and so on. However, we find that having six or more dimensions is not helpful.

We have implemented our method using R [R Core Team, 2014] and RTextTools [Jurka et al., 2014].

4.2.6 Predicting the Labels

We divided the dataset into a training and a testing dataset. We used a Naïve Bayes classifier to learn a model over the mapped LIWC dimensions. We then used the testing dataset to validate the accuracy of our classifier. Accuracy is defined by Equation 3.3, where TP, FP, TN, and FN refer to true and false positive and negative

Fig. 4.2 provides an overview of our process.

4.2.7 Evaluation Methods

We evaluate our work in two parts. We examine the validity of our first hypothesis, namely, that the complaint text carries a sentiment-related meaning. To that end, we investigate the occurrence of LIWC dimensions in the dataset. We map the complaints to the LIWC dimensions as in Equation 4.1. Then, for every label, we compute the ratio of the occurrences of the Money, Family, Posemo, Negemo, Anx and Anger LIWC dimension to the total number of LIWC dimension occurrences, as described in Equation 4.2.

$$p_d = \frac{\sum_{n=1}^{n=k} f_d(C_n)}{\sum_{n=1}^{n=k} \sum_{i=1}^{i=m} f_i(C_n)} \tag{4.2}$$

Where k is the total number of complaints and m is the number of LIWC dimensions. We expect to see a pattern that can be attributed to the underlying meaning.

Next, we compare our approach to an array of classifiers, implemented using RTextTools, including: Boosting, GLMNET, Max Entropy, Neural Network, Random Forests, SLDA, and SVM. Section 2.6 provides a detailed description of each classifier.

We manually select the best performing classifier per label, to obtain the best performance as a baseline and compare it with our LIWC-based model.

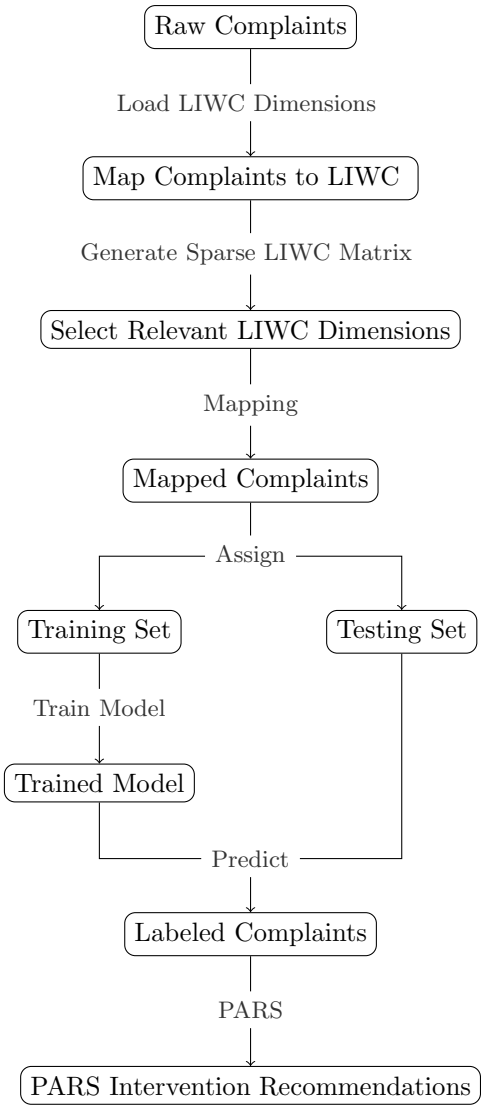

Figure 4.2. Classification and evaluation process overview.

Finally, we compare our results to using only TF-IDF (Term Frequency–Inverse Document Frequency) for feature extraction. TF-IDF is designed to emphasize the importance of a word to a document in a collection or corpus [Rajaraman et al., 2012], for feature extractions.

TF-IDF has been widely used in information classification and retrieval [Dumais et al., 1998]. The idea is simply to combine the TF (Term Frequency) with IDF (Inverse Document Frequency) by multiplying the frequency of a term in a document by the inverse document frequency of this term in the documents of the corpus. A commonly used example is the word "the". If we only use TF, the word "the" will score highly while it does not convey any significant meaning for predicting the label. Now, contrast that with counting the number of documents in the corpus which the word "the" appears in, which is equally high and almost in all documents.

IDF for a term takes the logarithm of the total number of documents divided by the number of documents that this word appears in, as shown in Equation 3.1.

In the case of the word "the" and assuming it will appear in all documents, that ratio is 1 and thus will generate an IDF value of 0. The IDF will then cancel the weight of the word "the". Our approach consists of four steps:

- Preprocess the documents to remove stop words, numbers and perform stemming.
- Feature extraction through generating TF-IDF sparse representation of the documents.
- Feature reduction by removing sparse terms.
- Learning a model to predict the labels.

We randomly select the training and testing dataset. The feature selection is performed on the training dataset independently of the testing dataset. We repeat the entire process 10 folds, reselecting the features each time. We take the average prediction of the 10 folds.

4.3 Results

Our first hypothesis that complaint text classes emphasize the LIWC dimensions differently was supported. As Fig. 4.3 shows, *CARE RELATED* is more evenly distributed whereas *SAFETY OF ENVIRONMENT* emphasizes *FAMILY*. Interestingly, *SAFETY OF ENVIRONMENT* emphasizes *Family* more than *NO COMPLAINT does*. We attribute this to the fact that the environment impacts, not the patient, but visitors, who tend to be the patient's family.

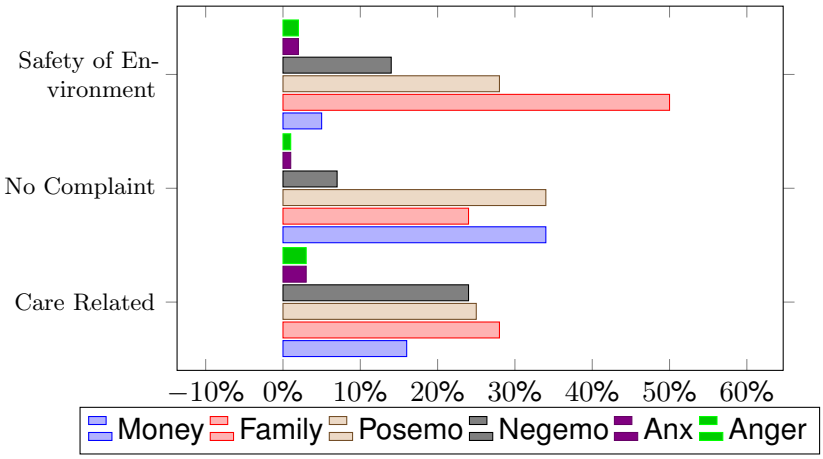

Figure 4.3. LIWC dimension analysis.

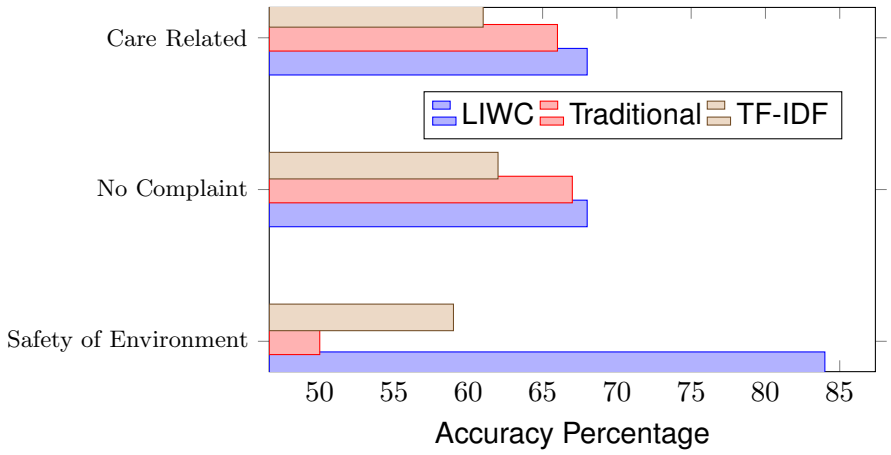

Figure 4.4. Comparison of LIWC-based, traditional and TF-IDF features.

Using TF-IDF alone failed to yield significant improvements in most of the labels, except in the case of Safety. An explanation for this result is that the bag of words uses term frequency as the base representation of the document, which is ineffective when the terms with most impact are localized to only a subset of documents. The IDF component helps by emphasizing terms that are more informative. We see that mapping the complaints to LIWC dimensions helps predict complaint labels better than with traditional features, which ignore sentiment and consider only unigrams. Our approach mitigates the shortcomings of using only TF-IDF and further enhances the feature extraction by using the dimension, which unlike unigram terms, combines adjacent words in one dimension that conveys a sentiment. The *SAFETY OF ENVIRONMENT* label shows the most gain in accuracy. Further, the high accuracy associated with *NO COMPLAINT* prediction has a huge impact on scaling PARS as it removes unnecessary processing. These results coupled with the ease of applying our method demonstrate the potential benefits of our work, as Fig. 4.4 shows.

We hypothesize that our approach yields a higher mean value for the above-mentioned quality metrics (i.e., accuracy, sensitivity, and specificity) than classification approaches that disregard sentiment in patient complaint. We produce three pairs of hypotheses, one for each metric.

Null Hypothesis 1 *Our approach using LIWC dimensions yields the same mean metric (accuracy, sensitivity, specificity) as approaches that disregard sentiment.*

Alternative Hypothesis 1 *Our approach yields greater mean metric (accuracy, sensitivity, specificity) than approaches that disregard sentiment.*

To evaluate the null hypothesis, we generate Table 4.1; wherein we use the approaches ignoring sentiments prediction as the expected values for each label and the LIWC predictions as the observed values. Then we conduct the one-tailed Chi-squared test [Pearson, 1900]. We

set our significance level α to be 0.005. We obtain a p-value of 0.03^{-10}. We thus reject the null hypothesis.

Table 4.1. Chi-squared test per-label data.

Label	Traditional Expected	LIWC Observed
Safety of Environment	9	13
No Complaint	247	308
Care Related	2028	1809

As Table 4.2 shows, we exceed accuracy achieved by the best traditional approach (SLDA) for each category label. Our approach yields slightly greater accuracy than the best traditional approaches for NO COMPLAINT and CARE RELATED. For SAFETY OF ENVIRONMENT, our approach yields a 68% improvement, from 50% to 84%. For SAFETY OF ENVIRONMENT, a false negative is much riskier than a false positive. For this reason, it helps to compute the classifier sensitivity. Our approach predicts SAFETY OF ENVIRONMENT with a sensitivity of 0.96. That is, the probability of a false negative, (1 Sensitivity), is 0.04. For the other two labels, the results are mixed though on balance accuracy favors our approach, as described above. Table 4.3 shows our Sensitivity and specificity computed over ten folds.

Table 4.2. Prediction accuracy per label. Accuracy is computed over ten folds.

Label	Accuracy (Ours)	LIWC Dimensions	Accuracy (SLDA)	Gain
Safety	0.84	Treatment, MedPerson, Patient, Ipron, Conj, Space	0.5	**68%**
Care	0.68	Treatment, MedPerson, Patient, Funct, SheHe, Negemo	0.66	**3%**
No Complaint	0.68	Treatment, MedPerson, Patient, Negemo, Inhib, Body	0.67	**1.5%**

Table 4.3. Prediction sensitivity and specificity per label, computed over ten folds.

Label	Our Approach		SLDA Approach	
	Sensitivity	Specificity	Sensitivity	Specificity
Safety	0.96	0.68	0	1
Care	0.60	0.75	0.97	0.29
No Complaint	0.75	0.66	0.30	0.97

Chapter 5

Domain-Specific Dependency (DSD) Feature Extraction for Patient Complaint Classification

The problem we address in this chapter is how to create a deeper understanding of patient complaints so that we can automate the classification process. Our method is to extract a domain-specific grammar dependency set of rules. We use those rules to extract a more representative set of features, which we then use to learn a set of classifiers. Finally, we employ the learned classifiers to predict the testing dataset labels. Patient complaints are critical and direct patient textual feedback. In order to formulate a correct intervention plan, patient complaints must be correctly classified. State of the art approaches rely on human coders but such approaches are costly in time and effort. Using basic features fails to produce F-Measure above 41% for any label due to the dataset challenges. We introduce a novel approach for patient complaint classification. Our approach incorporates domain-specific dependency (DSD) for grammatical analysis and feature extraction. We build a framework to extract part of speech tags as well as dependencies, then used them for feature extraction. We show how using the domain-specific, grammatically extracted features can produce superior results. Finally, we learn, test, and compare eight well-known classifiers to obtain Weighted F-measure above 80%.

5.1 Introduction

Patient opinions, feedback, and complaints are gaining increased importance as an instrument for effective healthcare quality assessment and service recovery [Hayden et al., 2010]. This is true of single payer systems, as in the United Kingdom, where the National Health Service (NHS)

provides services such as *Online Patient Opinion*[1] feedback service with the ultimate goal to improve the NHS. In the USA, The Joint Commission encourages patient and family reporting of concerns about their experiences as one way to promote quality and safety. Patient Advocacy provides a similar role of capturing patient feedback. Patient advocacy aims to improve facility healthcare quality [Baldwin, 2003] by analyzing patient complaints.

In both cases, the distillation of knowledge from this unstructured information proves to remain a difficult and complex task. Text complexity, domain linguistic structure, and other factors prevent the simple analysis in healthcare domain. In this study, we explore an approach to patient complaints analysis via machine learning on using part of speech and dependency-based bigrams. We demonstrate the rules to extract and reduce the feature set as well as compare four widely used classification models.

The Patient Advocacy Reporting System (PARS) is a representative existing approach by which healthcare professionals can capture, classify, and address patient complaints. PARS uses aggregate patient complaints to identify professionals associated with higher risk than the national benchmark [Stimson et al., 2010]. Effective and efficient classification is crucial to subsequent responses. However, current techniques are manual, which presents a challenge in terms of scalability [Hickson et al., 2010].

Importantly, classification is nontrivial. A well-established system, employs inputs from an array of specialists in fields including social behavior, educational psychology, and bio-statistics to build the classification process [Pichert et al., 2013]. Human coders are required to undergo training for three or more months to become proficient in classifying complaints. Subsequently, they are evaluated to guard against any performance change.

Using basic feature extraction techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) [Rajaraman et al., 2012] fail to produce results. For the forthcoming reasons, a more advanced approach is required to provide faster, automated classification of patient complaints with comparable accuracy to human coders.

Contributions Our contribution is two-fold. First, we propose that using domain-specific dependency for feature extraction would produce superior quality features and reduce noise, which enables our selected classifiers to correctly learn and predict the labels. Second, we implement our proposed framework that achieves promising results with real life data.

5.2 Related Work

Healthcare initiatives, such as meaningful use and outcome-based medicine, are inducing healthcare professionals to address patient concerns proactively in order to improve the quality of

[1]http://www.nhs.uk/aboutNHSChoices/aboutnhschoices/partners/patient-opinion/Pages/patient-opinion.aspx.

service and reduce readmission [Hsiao et al., 2011]. Consumer feedback is a significant source of information that provides service quality improvement path [Bendall-Lyon and Powers, 2001; Luca and Atuahene-Gima, 2007]. Patient feedback regarding the quality of service has significant ramifications on how patient outcomes are assessed and thus on healthcare professional financial viability. Healthcare organizations can gain insight from patient complaints about unsafe and dissatisfying behaviors and variability in performance, e.g., of surgical teams [Catron et al., 2015; Baldwin, 2003].

An opinion mining and sentiment analysis paradigm to automatically analyze patient opinions about the UK's NHS is presented in [Howard and Cambria, 2013]. They use an emotion categorization model, an affective semantic network and a language visualization and analysis system. The goal is to turn patient on-line contributions into useful information about the perceived quality of many UK hospitals.

Much of the previous use of Natural Language Processing (NLP) in healthcare has been applied to clinical textual artifacts, e.g., [Baud et al., 1992; Murff et al., 2011; Bejan and Denny, 2014].

Nonclinical patient complaint text presents different challenges, such as variation in style as well as the background and training of the author. Some patient complaints combine reports from different sources, including the patient and the patient's family and friends. Despite this, patient originated text has been shown to reveal important observations regarding their health experience [Liehr et al., 2002]. Traditional techniques relying solely on basic features such as those unigrams generated by TF (Term Frequency) or TF-IDF alone are confused by the dataset challenges resulting in poor performance. We develop an approach that automates detecting required physician action based on patient complaint.

Our approach enhances grammatical dependency with domain-specific terms to enhance feature extraction. The features are used to train an ensemble of well-known classifiers to predict subsequent patient complaints. We use the combined F-measure to automatically select the best performing classifier and thus, to triage new patient complaints.

5.3 Task Description

In this study, our task is to classify patient complaints into one of seven classes in order for the intervention plan to be constructed.

5.3.1 Feature Description

The features we are interested in are the PARS complaint classification.

When a patient at a healthcare center or hospital has an issue that they cannot resolve on their own, they contact the Office of Patient Relations or the Office of Patient Affairs at the

hospital. A patient advocate helps to understand and resolve the issue and enters notes into the facility database, which is subsequently sent to PARS. Trained coders mark relevant parts of complaint narratives to share with the care professional. Examples of the categorizations are as below. (Care and Treatment are the most common complaints.)

Communication

I tried asking questions . . . Dr. XX doesn't explain well . . . gives short answers.

Care and Treatment

I needed to be examined . . . Dr. XX never touched me except to shake hands., *I feel Dr. XX has put off my surgery too long with no good reason.*

Concern for Patient or Family

Dr. XX was rude. I was 7 minutes late and apologized. Dr. XX looked at the clock and said, That's 7 minutes I won't be spending with you.

Accessibility and Availability

I had to wait 2 hours to be seen. My time is valuable, too.

Safety of Environment

Wife states when nurse started IV that the patient bled and there was a large amount of blood on the floor and he had blood on his gown.

Money or Payment Issues

Dr. XX made no diagnosis so I went to a good doctor in [another town] who did exploratory surgery and found my trouble . . . I should not be responsible for the bill for my visit to Dr. XX.

No Complaint

Mr. XX called wanting to speak to someone about the pool hours at Sports Medical Center.

An intervention plan is created to address the risks through local peer, managers, and leadership.

5.3.2 Challenges Description

This problem is challenging for the following reasons:

- The data come from multiple people who are affiliated with different institutions. The authors (patient advocates in different institutions) have different writing styles for reporting the patient complaints as well as the actions taken (if any) in response to a complaint.

- Different users may contribute to the same complaint. These can include the patient, a family member of the patient, a physician, or the nurse. Each user reflects his or her point of view and these points of view can be mutually contradictory.
- Variation in patients' age can lead to discrepancies in the expressions used and the length of the complaint text.
- Time sensitivity in classification: It is important to classify a feedback so as to produce a correct intervention plan and respond before the problem escalates leading to possible medical errors and potential malpractice lawsuit.

5.4 Dataset Description

We consider 1,500 patient complaints randomly selected from a larger corpus spanning multiple years. We use a manual classification as our gold standard. The PARS describes seven categories of complaints, which we adopt.

We divide the complaints into a training dataset containing roughly 80% of all the selected records and the remaining 20% dataset is used for testing. The training data and test data are randomly selected ten times, and the results of their average performances in the experiments are reported.

As shown in Table 5.1, most of the complaints in this dataset are classified as *Care and Treatment*, *Communication*, and *Accessibility and Availability*.

The dataset is divided into training and testing dataset by randomly selecting each data point to eliminate overfitting or lucky selection. The entire process is repeated ten fold to further guarantee the result's robustness.

Table 5.1. Dataset Composition

Classification	N
Accessibility and Availability	242
Care and Treatment	570
Communication	399
Concern for Patient/Family	99
Money or Payment Issues	63
No Complaint	94
Safety of Environment	33

5.5 Approach

In this chapter, we present a novel approach to incorporate domain-specific dependency-based features in learning the models as opposed to basic unigram features extracted via TF-IDF. We performed the same pre-processing (removing punctuation, numbers stop words, and stemming) in both cases. We learned the exact same eight classifiers using the same data structure. The basic approach consists of two main steps:

- Feature Extraction, using bigram and TF-IDF weights
- Classifier Selection

Our approach is comprised of three steps, namely:

- Feature Extraction
- Domain-Specific Feature Reduction
- Classifier Selection

Our goal is to explore grammatically based learning and how can we use it to classify patient complaints, the overall process is outlined in Figure 5.1.

In the following section, we will expand on the details of our approach.

5.5.1 Feature Extraction

Conventional feature extraction addresses the text as tokens. Most widely used are unigrams and bigrams. To illustrate, consider the following real-life example:
"*Pt states that she fell and was brought into ER.*"

Unigrams

Unigram features are simply a bag of words extracted by separating text by spaces and noise characters. In our example, the words "Pt, states, that, she, fell, and, was, brought, into, ER" are all distinct unigram features.

Bigrams

Bigram features consist of two consecutive words in the text. In our example, "Pt states", "states that", "that she", "she fell", "fell and", "and was", "was brought", "brought into", "into ER" are distinct bigram features. These features are capable of incorporating some contextual information.

We are more interested in grammatically based features. In our example, the part of speech (POS) tags are shown in Table 5.2 extracted using *coreNLP* which is an R wrapper for Stanford CoreNLP [Manning et al., 2014].

Table 5.2. Part of Speech Tags.

Token	Word	POS Tag
1	Pt	JJ
2	states	NNS
3	that	IN
4	she	PRP
5	fell	VBD
6	and	CC
7	was	VBD
8	brought	VBN
9	into	IN
10	ER	NN
11	.	DOT

Bi-tagged

Bi-tagged features differ from bigrams in that they are extracted based on part of speech (POS) rules. This selection generates features that contain mostly adjectives and adverbs which are considered more sentiment bearing.

Dependency features

Dependency features describe syntactic relations between words in a sentence, provide linguistic analysis and can be useful for a sentiment analysis model. It has been shown in literature that syntactic patterns are effective for subjective detection [Zhang et al., 2009; Missen et al., 2013]. We used dependency parse tree, implemented using the Stanford Parser [Manning et al., 2014], to extract dependency relations from texts. In our example, the dependency features are shown in Table 5.4.

5.5.2 Domain-Specific Feature Reduction

Not all features contribute equally to prediction. A large number of features can add noise and reduce performance. We need to reduce the number of features and select a representative feature subset. To accomplish this goal, we extract rules that consider domain impact. In the following subsections, we will outline how we use TF-IDF (Term Frequency–Inverse Document Frequency) to extract domain-specific terms that we can infuse our rules with.

Table 5.3. Rules to Extract POS Features.

Rule	First Term	Second Term	Source
1	JJ	NN/NS	Tureny
2	RB/RBR/RBS	JJ	Tureny
3	JJ	JJ	Tureny
4	NN/NNS	JJ	Tureny
5	RB/RBR/RBS	VB/VBD/VBG	Tureny
6	VBN	NN/NNS	Agarwal
7	VB/VBG/VBP	JJ/JJR/JJS	Agarwal
8	JJ	VBN/VBG	Agarwal
9	RB/RBR/RBS	RB/RBR/RBS	Agarwal

Table 5.4. Extracted Dependency.

Dependency	governor	dependent	type
1	ROOT	brought	root
2	states	Pt	amod
3	brought	states	nsubj
4	fell	that	mark
5	fell	she	nsubj
6	states	fell	dep
7	fell	and	cc
8	fell	was	conj:and
10	brought	ER	nmod:into
9	ER	into	case

Domain-Specific

Each domain has specific characteristics that define the language used in the domain. TF-IDF has been widely used in information classification and retrieval [Dumais et al., 1998]. The idea is simply to combine the TF (Term Frequency) with IDF (Inverse Document Frequency) by multiplying the frequency of a term in a document by the inverse document frequency of this term in the documents of the corpus. A commonly used example is the word "the". If we only use TF, the word "the" will score highly while it does not convey any significant meaning for predicting the label. Now, contrast that with counting the number of documents in the corpus which the word "the" appears in, which is equally high and almost in all documents. IDF for a term takes the logarithm of the total number of documents divided by the number of documents that this word appears in, as shown in Equation 3.1.

In case of the word "the", and assuming it will appear in all documents, that ratio is 1 and thus will generate an IDF value of 0. The IDF will then cancel the weight of the word "the". The research presented in [Jing et al., 2002] demonstrated a feature selection approach based on ranking the terms using their TF-IDF weights and then selecting top n features. The justification is that features with higher TF-IDF weights would carry more information content and thus, would be more suitable for predicting the label. We use TF-IDF to generate a the most impact-full terms and then we select those that have more relevant association with the heathcare domain. This list is then used in constructing the domain-specific terms that would contribute more to our dataset. Our analysis indicates that the terms "pt", "patient", "dr", "doctor" have emphasized importance in our dataset.

Reduction Rules

Sentiment analysis has been widely used in NLP to help understand the underlying meaning of the text rather than the basic unigram/bigram approach. This is extremely important in our case as the dataset we face is challenging and requires deeper analysis. We follow the work presented in [Turney, 2002; Hatzivassiloglou and McKeown, 1997; Agarwal et al., 2015] where it has been shown that the use of POS-based information to extract sentiment rich features generates good predictive results. Turney [2002] presented five rules to extract POS-based features consisting of an adjective or an adverb. Agarwal [2015] observed that verbs can also contain sentiment information that is useful for sentiment analysis. They extended the rule set to extract more sentiment bearing features by adding four more rules. These rules are shown in Table 5.3. We adopted a dependency based approach and generated three rules that are enriched with domain-specific terms, as shown in Table 5.5. We focus the rule on three main aspects of the dependency relation:

nsubj

It is the nominal subject which is the syntactic subject of a clause. The governor of this

Table 5.5. Rules to Extract DSD Features.

Rule	Governor	Dependent	Type
1	–	–	nsubj
2	–	–	dobj
3	patient/pt/dr/doctor	–	–
4	–	patient/pt/dr/doctor	–

relation might not always be a verb: when the verb is a copular verb, the root of the clause is the complement of the copular verb, which can be an adjective or noun. We selected this dependency type as it expresses the doer of an action. In a complain situation, it is natural to focus on the doer of the action.

dobj

It is the direct object of a verb which is the noun phrase which is the (accusative) object of the verb. In complaints, the direct object is the one affected by the complaint, it is natural to select this relation type.

Domain Terms

Those are the domain-specific terms we obtained from our *TF-IDF* analysis. We select any pair in which the governor or dependent is one of those domain selected terms. The reasoning is that those are the main actors in the healthcare domain and thus, it reasonable to select them as features.

Applying those rules to the example above, we extract three features, namely:

states, Pt using rule number 3.

brought, states using the first rule, the dependency type is nsubj.

fell, she the same rule applies here, the dependency type is nsubj.

5.5.3 Classifier Selection

Next, we learn an array of eight well-known classifiers using RTextTools [Jurka et al., 2014], namely: Bagging, Boosting, GLMNET, Max Entropy, Neural Network, Random Forests, SLDA, and SVM. Section 2.6 provides a detailed description of each classifier.

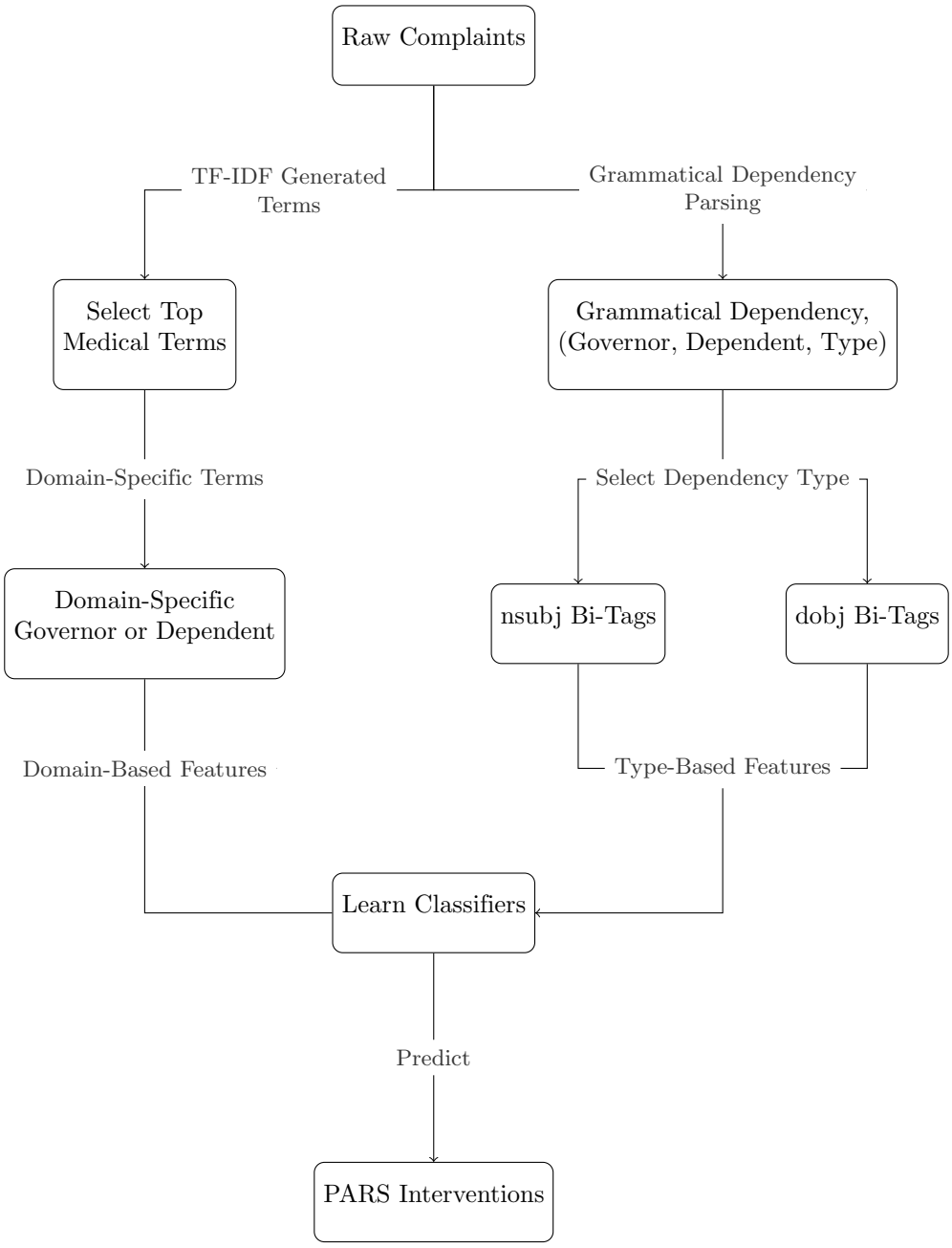

Figure 5.1. Domain-specific dependency overall steps.

5.6 Evaluation Metrics

To evaluate our approach, we used the testing dataset to compare both the accuracy and F-measure of the learned classifier using our feature extraction rules to those using basic feature extraction methods. Accuracy is defined by Equation 3.3.

Sensitivity captures how many patients with a condition are detected—i.e., the avoidance of false negatives. We compute sensitivity using Equation 3.4.

Specificity captures how many patients without a condition are not detected—i.e., the avoidance of false positives. We compute specificity using Equation 3.5.

F-measure has been gaining more acceptance as a single measure of the three types of errors in information retrieval, combining substitutions, deletions, and insertions errors in one easy to use value [Makhoul et al., 1999]. The definition of F-measure combines both precision and recall as defined in Equation 5.1.

$$\text{F-measure} = (2 * \text{P} * \text{R})/(\text{P} + \text{R}) \tag{5.1}$$

In the healthcare domain, precision is known as *positive predictive value*, while recall is known as *sensitivity* [Sasaki et al., 2007]. We will use the F-measure as it perfectly matches the healthcare and NLP intersection. In addition to the per-label *F-measure*, we compute the *Weighted F-measure* which takes into consideration the label distribution in the dataset as shown in Equation 5.2.

$$\text{Weighted F-measure} = (\sum_{l=1}^{l=L} \text{F-measure}_l * n_l)/(\sum_{l=1}^{l=L} n_l) \tag{5.2}$$

where $F - measure_l$ is the F-measure for label l, while L is the total number of labels and n_l is the number of complaints classified as label l.

5.7 Results and Discussions

A considerable amount of research illustrated the grammatical evolution which is not just unique to a certain language, but rather across a multitude of languages [Bybee et al., 1994]. The grammatical behavior in a text reflects a unique pattern that carries the intent behind the communication. Our results in this paper support this broad meaning and take advantage of it in order to improve patient complaint classification. We conducted our experiment twice, once using basic bigram with TF-IDF weights for feature extraction. The second, using grammatical dependency for feature extraction followed by domain-specific rules for feature reduction. As shown in Table 5.6, the basic approach is confused with the noise in the dataset. It performed very poorly and failed to produce any results in the *Safety of Environment* label. In contrast with our DSD approach, which produced almost perfect results in the same label. The only label

where the basic approach outperformed our DSD was in the *No Complaint*. The insight here is that the grammatical composition of a text is affected by the conveyed message. It would make sense that each complaint label would be different enough to enable our DSD learned classifiers of correctly modeling the label. The exception to this would be *No Complaint* label, as it would encompass varying messages and would not follow a consistent grammatical composition. For example, it could be parsing the establishment or commending the service which does not relate to the patient or the doctor and thus would not carry enough features to allow the classifier of learning them. By the same token, the basic approach was confused in all the labels. However, the basic approach results for the *No Complaint* label, although low, are better than the DSD approach. We can attribute that to the ability of TF-IDF to emphasize the unigram tokens and the lack of a distinct grammatical pattern for the DSD to learn. This issue is mitigated in the weighted F-Measure is computed by taking into consideration the label frequency, which helps balance the overall performance of the approach and classifier.

Our results confirm those insights, the overall F-measure performance of the classifiers is shown in Figure 5.11, while the details of each classifier results are shown in Figures:5.2, 5.3, 5.4, 5.5, 5.6, 5.7, 5.8 and 5.9. We note that all classifiers achieved extremely high scores for the *Safety of Environment* label which is a measure of the feature quality produced by our rules. We obtain good results across all the other labels with the *No Complaint* exception.

An interesting observation is that in the case of *BAGGING* and *SLDA* classifiers, the performance was relatively low for the *Concern for Patient or Family* label. One explanation is that those classifiers may be suffering from variable permutation importance as discussed in [Strobl et al., 2009]. This is especially interesting due to the low frequency of the label even though, other labels such as *Safety of Environment* and *Money or Payment Issues*. This problem is overcome in *Random Forests* where the random selection of splitting variables allows predictor variables, that were otherwise outplayed by a stronger competitor, to enter the ensemble: If the stronger competitor cannot be selected, a new variable has a chance to be included in the model. This new variable may reveal interaction effects with other variables that otherwise would have been missed.

Individual label performance can be misleading as it ignores the relative frequency of the label in the dataset. When we applied the weighted F-measure, computed by weighting the result by the relative frequency of the label, our DSD approach clearly outperforms the basic approach in each classifier and we can see superior results from the *GLMNET*, *MAXENT*, *RF* and *SVM* classifiers as shown in Figure 5.10.

Table 5.6. Domain-Specific Dependency Features and Basic Features Classifier Results

Classification	BAGGING		BOOSTING		GLMNET		MAXENT		NNET		RF		SLDA		SVM	
	DSD	Basic	DSD	Basic	DSD	Basic	DSD	Basic	DSD	Basic	DSD	Basic	DSD	Basic	DSD	Basic
Accessibility and Availability	0.76	0.09	0.76	0.21	**0.89**	0.1	0.88	0.25	0.84	0.11	0.88	0	0.63	0.21	0.85	0.21
Care and Treatment	0.68	0.57	0.76	0.52	0.82	0.56	**0.85**	0.5	0.84	0.38	0.81	0.59	0.84	0.52	0.85	0.57
Communication	0.55	0.39	0.6	0.33	0.83	0.35	0.81	0.38	0.82	0.26	**0.83**	0.41	0.6	0.34	0.78	0.39
Concern for Patient or Family	0.23	0.19	0.73	0.15	0.81	0.21	0.81	0.28	0.74	0.17	**0.85**	0.06	0.36	0.16	0.77	0.25
Money or Payment Issues	0.55	0	0.6	0.15	**0.96**	0.22	0.94	0.2	0.89	0.1	0.97	0	0.59	0.33	0.9	0.24
No Complaint	0.16	0.45	0.22	0.45	0.18	0.41	0.33	0.49	0.24	0.28	0.17	0.5	0.25	0.41	0.45	**0.51**
Safety of Environment	1	0	1	0	1	0	**1**	0	0.97	0	1	0	0.91	0	1	0
Over All	**0.6**	0.38	**0.68**	0.36	**0.8**	0.37	**0.82**	0.39	**0.8**	0.26	**0.8**	0.37	**0.66**	0.37	**0.81**	0.41

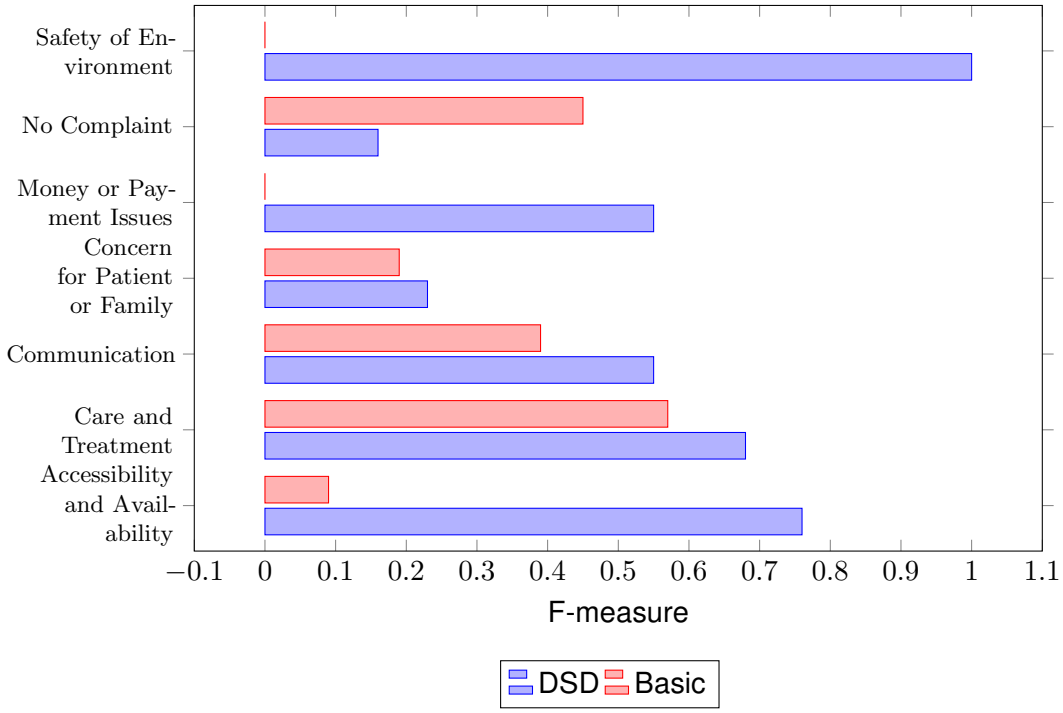

Figure 5.2. Results of the BAGGING classifier.

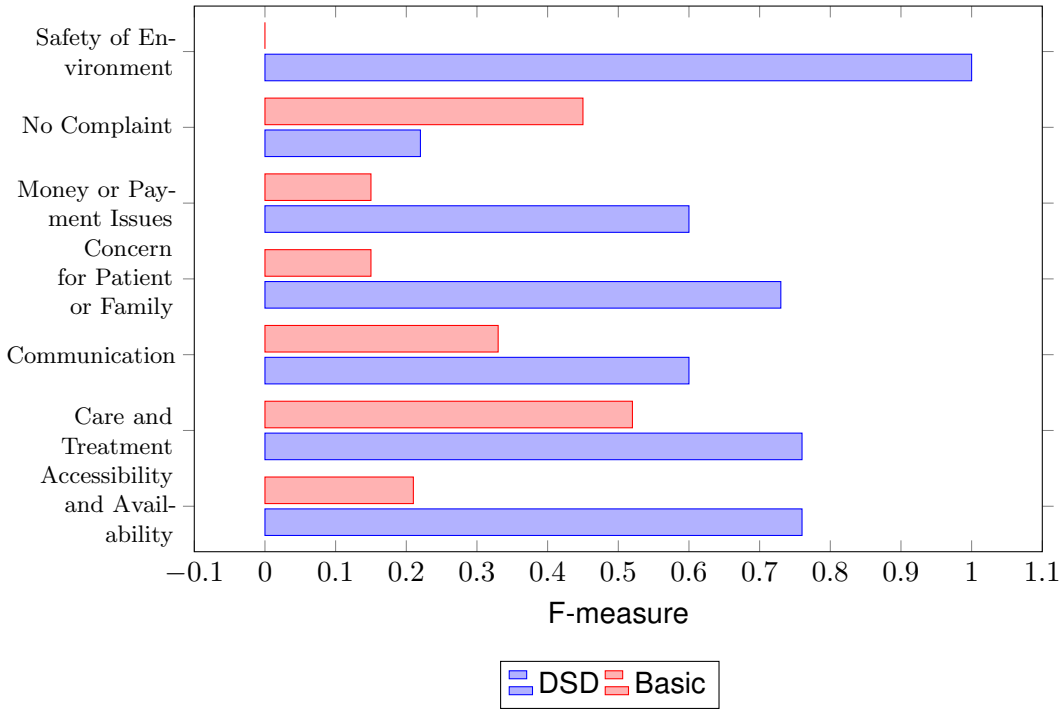

Figure 5.3. Results of the BOOSTING classifier.

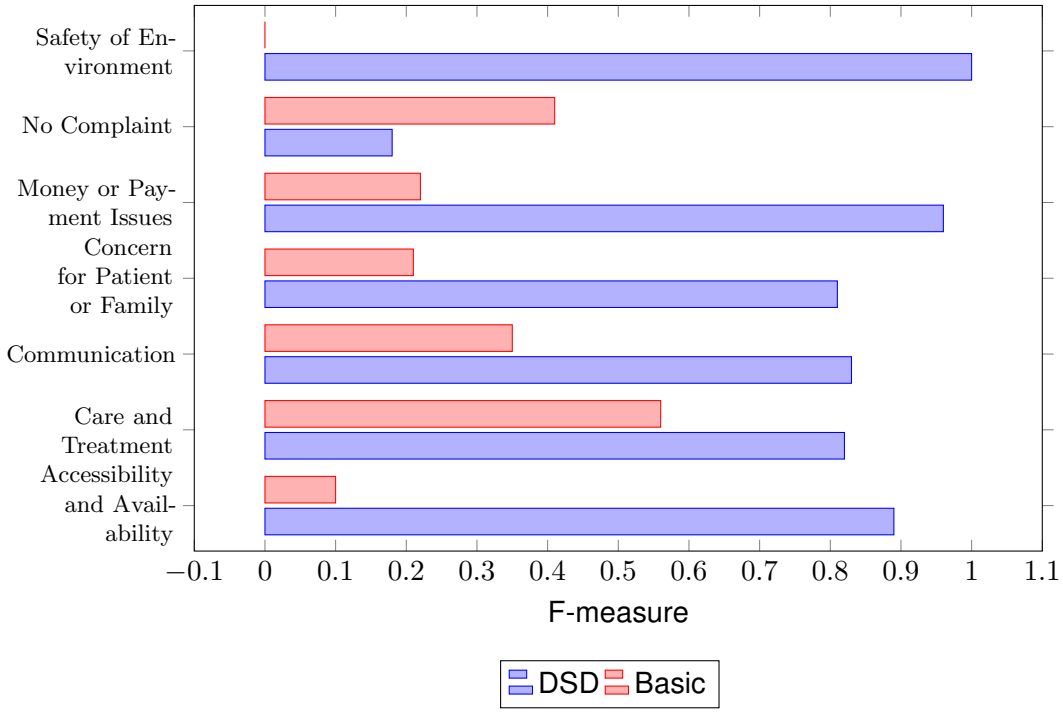

Figure 5.4. Results of the GLMNET classifier.

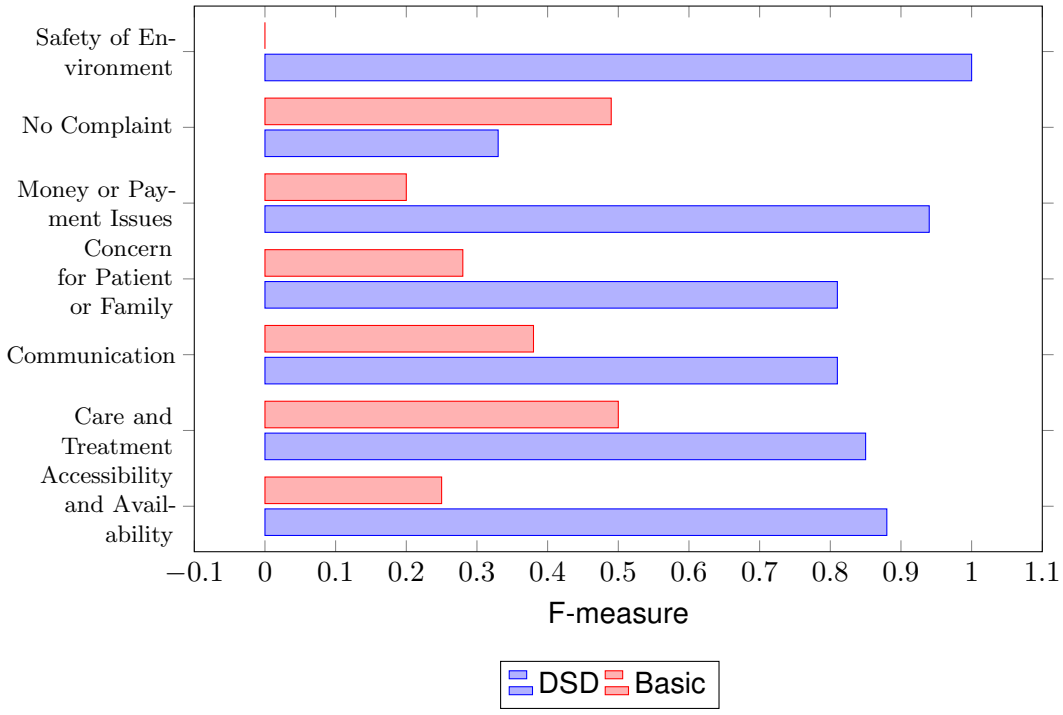

Figure 5.5. Results of the MAX Entropy classifier.

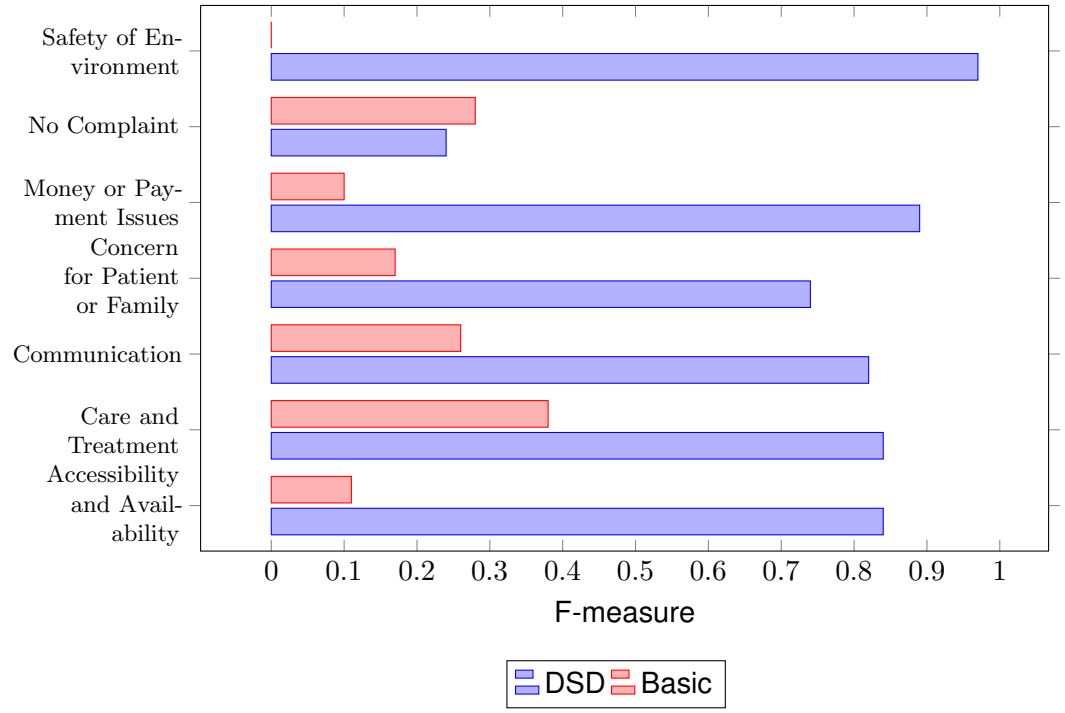

Figure 5.6. Results of the Nural NET classifier.

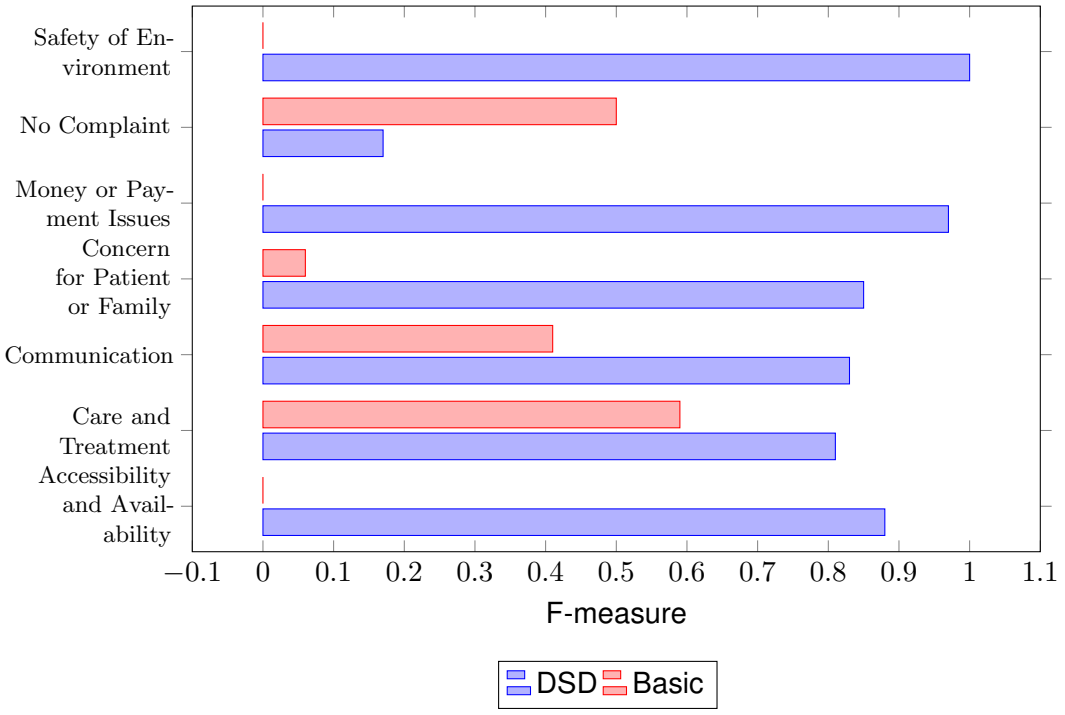

Figure 5.7. Results of the RF classifier.

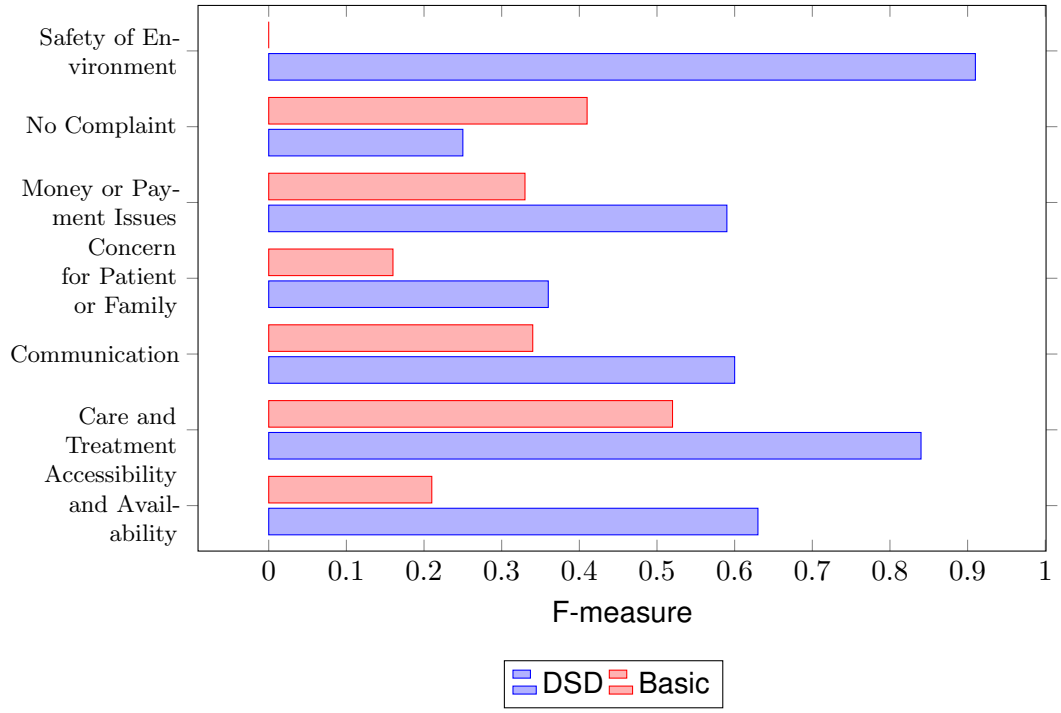

Figure 5.8. Results of the SLDA classifier.

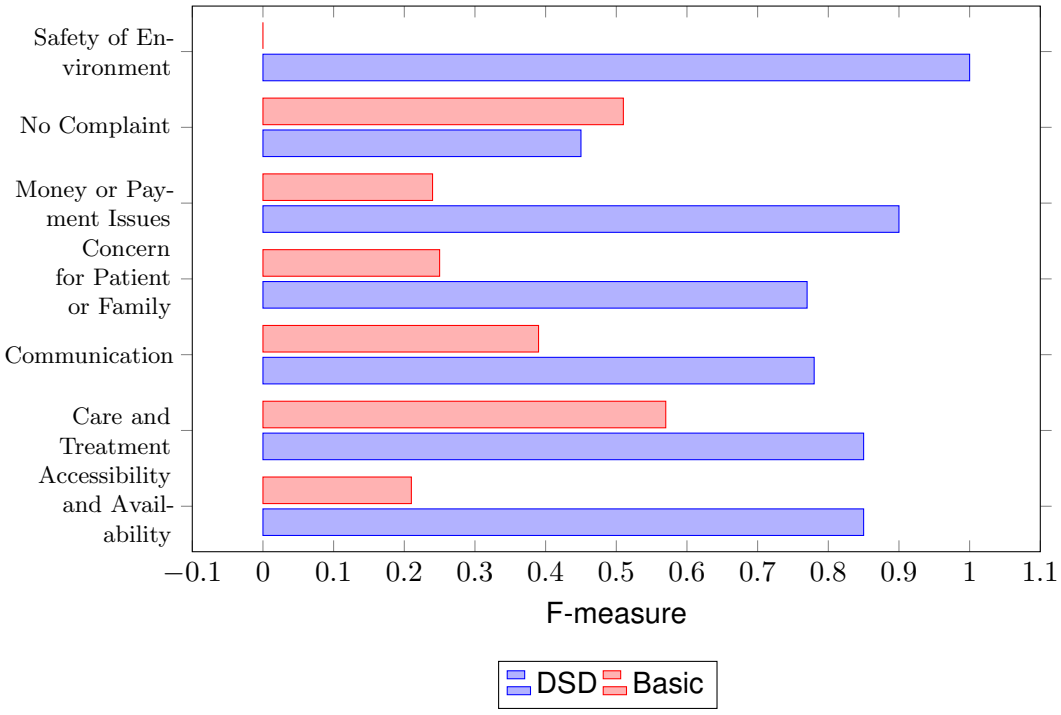

Figure 5.9. Results of the SVM classifier.

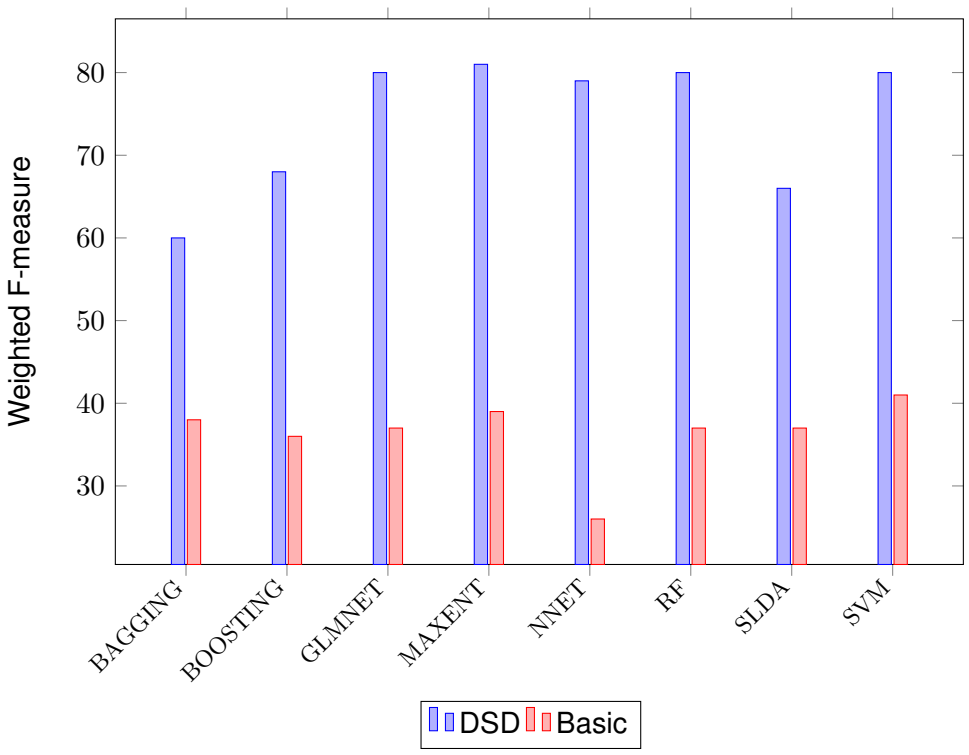

Figure 5.10. Weighted F-measure for domain-specific dependency and basic features.

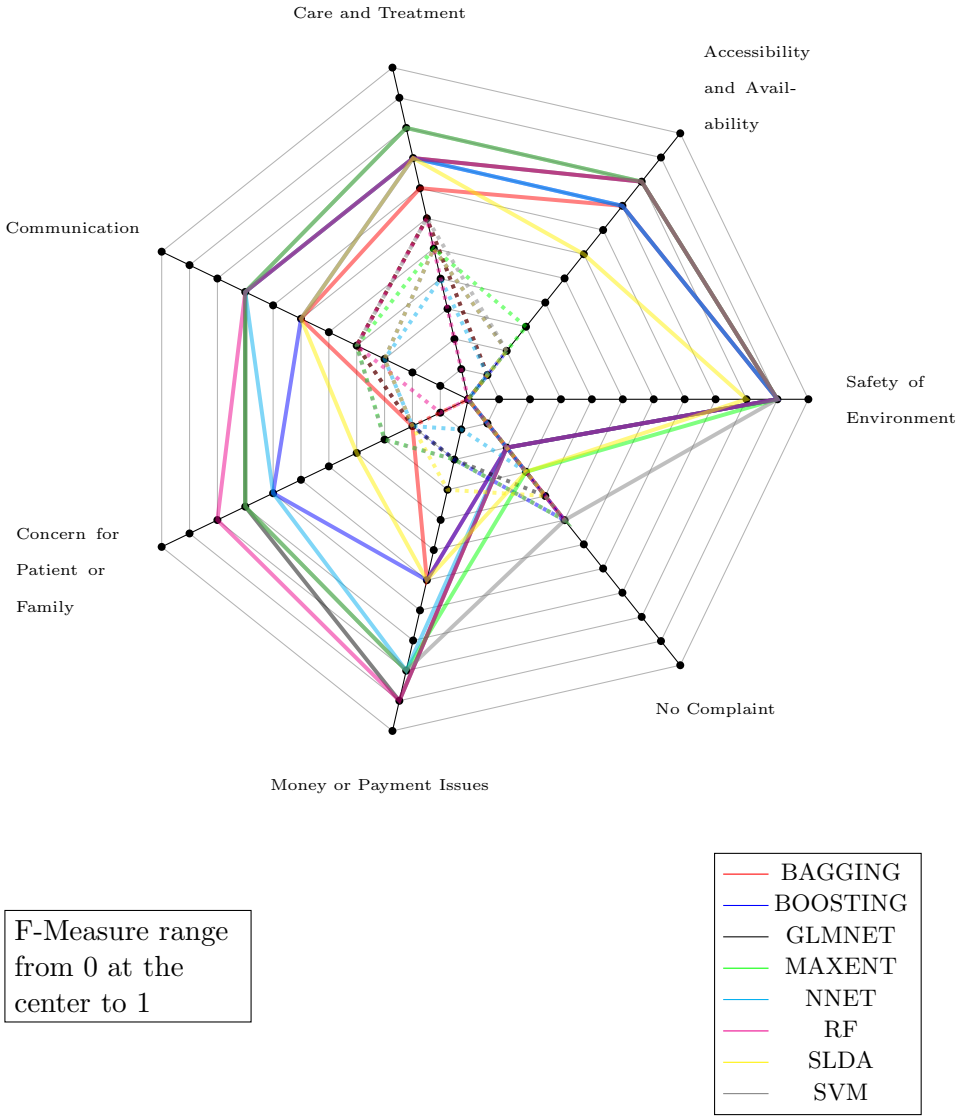

Figure 5.11. Overall F-measure for domain-specific dependency and basic features across labels.

Chapter 6

Conclusions and Future Work

Our time is marked by the expansion of human online activities. Online text is becoming the preferred medium for daily human activities. This is true of mainstream domains such as healthcare. Healthcare represents a substantial size of spending across the globe and especially in the United States[1]. In conjunction with the scientific research to improve healthcare treatment, considerable focus remains on improving healthcare patient experience. Research shows there is a link between patient outcomes and patient experience. Besides having solicited evaluation forms, the only valuable source of insight into the problems within healthcare organizations are unsolicited patient complaints. Patient complaints are critical and direct patient textual feedback has the potential to affect behavioral changes throughout the healthcare system and lead to better healthcare outcomes for the patients. In order to formulate a correct intervention plan, patient complaints must be correctly classified. Current state-of-the-art approaches rely on human coders but such approaches are costly in time and effort. Healthcare domains stand to benefit considerably from analyzing patient complaints to predict required intervention. In this dissertation, we set out to bridge the gap between healthcare and computer science domains. Our goal is to understand, model, and predict the classification of patient complaints relying on an array of NLP, AI concepts, and techniques that we have extended to adopt the healthcare domain.

We start with the basic task of triaging a patient's complaints into an urgent case requiring physician action or not. The concept of urgency is closely coupled with physician-related actions, whether it is a treatment urgency or physician behavior urgency. Our experiments show that a machine-learning approach can be quite effective in identifying patient complaints that indicate urgency. It is interesting that using term sparsity to reduce the feature set size provides robust improvement until we arrive at a point where the terms are too few to provide any meaningful discrimination between the labels and thus, the prediction accuracy falls. Adding IDF in combination with TF helps remove features that do not contribute significant information as

[1]Projected to be 17.5% of GDP for 2016.

compared to TF alone. Our results agree with prior research, e.g., Liu et al. [2003] and Cho and Lee [2016], showing improved results with a reduced (and hence a more representative) set of features.

Our specific findings are that the best-performing classifier was Random Forests with an 82% accuracy using TF-IDF for feature generation, followed by SVM classifier which achieved 79% accuracy using the simpler TF for feature generation. Adopting our automated approach would lead to the identification of urgent patient complaints much faster than any manual approach and thereby, enable early mitigation before the issue escalates and consumes more resources.

Intervention plans are very detailed and complicated, which requires a deeper level of classification. We built on our success and demonstrate that inferring the complaint sentiment leads to improved classification accuracy. We map each complaint to a vector based on an enhanced Linguistic Inquiry and Word Count (LIWC) lexicons and to train a Naïve Bayes classifier over those vectors. We compare it to both, Term Frequency–Inverse Document Frequency (TF-IDF) and the best case results of any classifier over bag of words features. Our approach is less computationally demanding, but produces better results, than traditional approaches, which disregard sentiment. Our classifier yields 3% greater accuracy overall than traditional approaches. For the *Safety of Environment* label, our classifier an accuracy of 84% (compared to 50% for traditional) and a sensitivity of 0.96 (compared to 0.00 for traditional).

We attribute the success of our approach to its incorporating the sentiment implicitly conveyed in the complaints. As medicine is adopting evidence-based practices, there are expanding efforts, e.g., pSCANNER [Ohno-Machado et al., 2014], on building and using patient-centric clinical databases to support medical research. However, to understand and avoid adverse medical actions, we need not only clinical information but also information about patients subjective reactions to a treatment and associated interactions with people.

Finally, we dive deeper into the actual language structure. We introduce a novel approach for patient complaint classification. Our approach incorporates domain-specific dependency (DSD) for grammatical analysis and feature extraction. We build a framework to extract part of speech tags as well as dependencies, then used them for feature extraction. We demonstrate how using the domain-specific, grammatically extracted features can produce superior results as compared to basic unigram features. Finally, we learn, test, and compare eight well-known classifiers to obtain Weighted F-measure above 80% as compared to using basic features, which fails to produce F-Measure above 41% for any label due to the dataset challenges. We are able to achieve a best case weighted F-measure of 0.82[2] as compared to only 0.41[3] using basic features.

We conclude that using domain-specific dependency for feature extraction produces superior quality features and reduces noise, which enables our selected classifiers to correctly learn and predict the labels. We attribute this results to a better modeling of the actual language structure

[2]Classified with MAXENT.
[3]Classified with SVM.

that the patient is using rather than simple terms.

Cross-disciplinary research is fundamentally difficult as it attempts to blend the well-established line between the domains. However, the result is often remarkable. In our case, contributing to the patient care improvement and enabling new applications of NLP in healthcare is well worth it.

6.1 Contributions

This dissertation focuses on the problem of developing an understanding of patient complaints and the robust approaches requiring minimal human supervision so as to build a patient complaint analysis tool that can be used for a wide range of healthcare providers on different scales that can address the challenges discussed above. We start with the basic understanding of patient complaints attempting a binary classification, followed by more advanced sentiment-based modeling of the patient complaint text. Finally, we explore domain-specific dependency feature extraction to model patient complaints. The major contributions of this dissertation are as follows.

- Study urgent patient complaints, requiring physician action and how the patient text differed in both. Investigate the benefits of using NLP features for modeling urgent patient complaints. Construct an urgency model then implements an array of machine learning models for urgent patient complaints classification. Compare the performance of the machine learning models and select the best-performing one.

- Alleviate the *Safety of Environment* detection problem in patient complaints. The *Safety of Environment* has a very time sensitive impact in the healthcare domain. To that extent, we expand the research to Investigate the benefits of using sentiment features for a patient complaints modeling focusing on *Care Related*, *Safety of Environment* and *No Complaint* classification.

- Propose a novel approach of mapping complaints into sentiment vectors utilizing domain-specific enhanced linguistic Inquiry and Word Count (LIWC) dimensions. Demonstrate and implement a machine learning model for patient complaints classification based on the proposed approach and compare the results to current approaches.

- To accommodate the disparity in the used language and style, we explore using domain-specific grammatical dependency for feature extraction. We propose a method to extract domain-specific terms which we use to construct a set of grammatical dependency rules.

- Demonstrate eight machine learning models for patient complaint classification using our rules to achieve significantly higher results as compared with basic unigram features using the same models.

6.2 Future Work

The results of domain-specific LIWC and domain-specific grammatical dependency open up many interesting directions for future work. We would like to continue our work and build new healthcare-specific rules exploring healthcare specific linguistic structures that can help further improve the performance and accuracy of our model across larger dataset. In this section, we briefly introduce three directions.

6.2.1 Temporal Analysis

Time plays a role in how humans communicate. Language evolves to meet the expression needs. New patterns, abbreviations, and terminology are added continuously. Medieval English is very different than ours. For example: "*Students must suffer to earn their degree.*" would translate to "*Students wilt suff'r to earneth their grise.*" [4]. This is still true over a shorter time span. Not all the expressions used hold the same meaning, nor are they used anymore. It would be interesting to investigate the difference in complaint language over a period of ten years and how that affects classifier learning. Our aim would be to account for language evolution and maintain classification accuracy.

6.2.2 Geolocation Analysis

Dialects vary by location. According to The Nationwide Speech Project [Clopper and Pisoni, 2006], six different regional varieties of American English have been considered: New England, Mid-Atlantic, North, Midland, South, and West. The speech samples obtained from each talker include isolated words, sentences, passages, and samples of interview speech. Understanding the differences would help create a more general model that accounts for them.

6.2.3 Diagnosis and Demographics

Patient diagnosis plays an important role in selecting the subset of terms used to describe and communicate condition, treatment plans and procedures. Patient demographics contribute to how they express themselves. The socioeconomic background shapes the style of patient writing. We propose to segregate patient complaints by diagnoses and age to compare the complaint language used.

6.2.4 Deep Domain-Specific Grammatical Dependency

Our domain-specific grammatical dependency approach showed the potential achieved when considering the language structure as well as the domain important terms. We strongly think

[4]Translated using: http://lingojam.com/EnglishtoShakespearean.

that exploring more rules to select the features as well as other techniques to extract those domain important terms will allow us to achieve higher results, especially in the *No Complaint* label.

6.2.5 Detecting Physician Behavioral Anomalies

Not all complaints carry the same weight, an indecent that happens once is not as significant as a persistent pattern. However, a physician that is commonly abrupt may not be taken as seriously as a more elaborate physician. We would like to establish a physician-specific profile based on each physician complaints, with the aim to compare each new complaint against the physician profile. The system would be able to detect physician behavioral anomalies based on the new complaints and escalate the complaints.

IRB Approval

This research was reviewed and approved by the Vanderbilt Medical Center Institutional Review Board and the North Carolina State University Institutional Review Board.

REFERENCES

Basant Agarwal, Soujanya Poria, Namita Mittal, Alexander Gelbukh, and Amir Hussain. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. *Cognitive Computation*, 7(4):487–499, 2015.

Charu C Aggarwal and Philip S Yu. *Finding generalized projected clusters in high dimensional spaces*, volume 29. ACM, 2000.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

Moyra A Baldwin. Patient advocacy: a concept analysis. *Nursing Standard*, 17(21):33–39, 2003.

National Practitioner Data Bank. Medical malpractice payment reports (mmpr) and adverse action reports (aar) reported on all practitioners by location for the years 2004-2014. `http://www.npdb.hrsa.gov/resources/npdbstats/npdbStatistics.jsp#ContentTop`, 2014. Accessed: 12/1/2015.

RH Baud, Anne-Marie Rassinoux, and Jean R Scherrer. Natural language processing and semantical representation of medical texts. *Methods of information in medicine*, 31(2):117–125, 1992.

Cosmin A Bejan and Joshua C Denny. Learning to identify treatment relations in clinical text. In *AMIA Annual Symposium Proceedings*, volume 2014, page 282. American Medical Informatics Association, 2014.

Dawn Bendall-Lyon and Thomas L Powers. The role of complaint management in the service recovery process. *Joint Commission Journal on Quality and Patient Safety*, 27(5):278–286, May 2001.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

H.B. Burke, D.B. Rosen, and P.H. Goodman. Comparing artificial neural networks to other statistical methods for medical outcome prediction. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 4, pages 2213–2216, Orlando, Jun 1994.

Joan Bybee, Revere Perkins, and William Pagliuca. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. University of Chicago Press, 1994.

Robert Carroll, American Society for Healthcare Risk Management, et al. Risk management handbook. *Burlington: The University of Vermont*, 1997.

Thomas F. Catron, Oscar D. Guillamondegui, Jan Karrass, William O. Cooper, Barbara J. Martin, Roger R. Dmochowski, James W. Pichert, and Gerald B. Hickson. Patient complaints and adverse surgical outcomes. *American Journal of Medical Quality*, page 1062860615584158, 2015.

Heeryon Cho and Jong-Seok Lee. Data-driven feature word selection for clustering online news comments. In *2016 International Conference on Big Data and Smart Computing (BigComp)*, pages 494–497. IEEE, 2016.

Cynthia G Clopper and David B Pisoni. The nationwide speech project: A new corpus of american english dialects. *Speech communication*, 48(6):633–644, 2006.

Nigel Collier, Reiko Matsuda Goodwin, John McCrae, Son Doan, Ai Kawazoe, Mike Conway, Asanee Kawtrakul, Koichi Takeuchi, and Dinh Dien. An ontology-driven system for detecting global health events. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 215–222. Association for Computational Linguistics, 2010.

Joint Commission. Joint commission requirements. `http://www.jointcommission.org/standards_information/tjc_requirements.aspx`, 2008. Accessed: 07/01/2015.

Licong Cui, Shiqiang Tao, and Guo-Qiang Zhang. A semantic-based approach for exploring consumer health questions using umls. In *AMIA Annual Symposium Proceedings*, volume 2014, page 432. American Medical Informatics Association, 2014.

Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5): 760–772, 2009.

Son Doan, Lucila Ohno-Machado, and Nigel Collier. Enhancing twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses. In *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on*, pages 62–71. IEEE, 2012.

William A. Donohue, Yuhua Liang, and Daniel Druckman. Validating LIWC dictionaries: The Oslo I Accords. *Journal of Language and Social Psychology*, 33(3):282–301, June 2014.

Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155. ACM, 1998.

Susannah Fox and Maeve A. Duggan. Peer-to-peer health care, 2013. http://www.pewinternet.org/2013/01/15/peer-to-peer-health-care/.

Carol Friedman and George Hripcsak. Natural language processing and its future in medicine. *Academic Medicine*, 74(8):890–5, 1999.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Vijay Garla, Vincent Lo Re, Zachariah Dorey-Stein, Farah Kidwai, Matthew Scotch, Julie Womack, Amy Justice, and Cynthia Brandt. The yale ctakes extensions for document classification: architecture and application. *Journal of the American Medical Informatics Association*, 18 (5):614–620, 2011.

Anthony Gold. Nhs hospital complaints. `http://www.lexology.com/library/detail.aspx?g=9570c37b-e324-4507-bbdf-b5faee8be068`, 2013. Accessed: 11/1/2016.

F. Harrag and E. El-Qawasmah. Neural network for arabic text classification. In *Proceedings of the Second International Conference on the Applications of Digital Information and Web Technologies.*, pages 778–783, Aug 2009.

Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

Anna C Hayden, James W Pichert, Jodi Fawcett, Ilene N Moore, and Gerald B Hickson. Best practices for basic and advanced skills in health care service recovery: A case study of a re-admitted patient. *The Joint Commission Journal on Quality and Patient Safety*, 36(7): 310–318, 2010.

Gerald B Hickson, Anna C Caruso-Hayden, and James W Pichert. The PARS® program: How unsolicited patient comments can be used to promote a safer healthcare environment, address unprofessional conduct and reduce unnecessary malpractice risk. In *Annual Meeting of the American Health Lawyers Association*, Seattle, 2010.

Newton Howard and Erik Cambria. Intention awareness: improving upon situation awareness in human-centric environments. *Human-centric Computing and Information Sciences*, 3(1): 1–17, 2013.

Chun-Ju Hsiao, Esther Hing, Thomas C Socey, and Bill Cai. Electronic health record systems and intent to apply for meaningful use incentives among office-based physician practices: United states, 2001–2011. *system*, 18(17.3):17–3, 2011.

John T James. A new, evidence-based estimate of patient harms associated with hospital care. *Journal of patient safety*, 9(3):122–128, 2013.

Li-Ping Jing, Hou-Kuan Huang, and Hong-Bo Shi. Improved feature selection approach tfidf in text mining. In *Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on*, volume 2, pages 944–946. IEEE, 2002.

Timothy P. Jurka, Loren Collingwood, Amber E. Boydstun, Emiliano Grossman, and Wouter van Atteveldt. *RTextTools: Automatic Text Classification via Supervised Learning*, 2014. URL `http://CRAN.R-project.org/package=RTextTools`. R package version 1.4.2.

Jeffrey H. Kahn, Renee M. Tobin, Audra E. Massey, and Jennifer A. Anderson. Measuring emotional expression with the linguistic inquiry and word count. *The American Journal of Psychology*, 120(2):263–286, Summer 2007.

Janani Kalyanam, Sumithra Velupillai, Son Doan, Mike Conway, and Gert Lanckriet. Facts and fabrications about ebola: A twitter based study. *arXiv preprint arXiv:1508.02079*, 2015.

Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

Cheng Hua Li and Soon Cheol Park. Neural network for text classification based on singular value decomposition. In *Proceedings of the 7th IEEE International Conference on Computer and Information Technology*, pages 47–52, Oct 2007.

Patricia Liehr, Ryutaro Takahashi, Chie Nishimura, Lorraine Frazier, Iwao Kuwajima, and James W. Pennebaker. Expressing health experience through embodied language. *Journal of Nursing Scholarship*, 34(1):27–32, March 2002.

Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, chapter 13, pages 415–463. Springer US, 2012.

Tao Liu, Shengping Liu, Zheng Chen, and Wei-Ying Ma. An evaluation on feature selection for text clustering. In *Icml*, volume 3, pages 488–495, 2003.

Luigi M De Luca and Kwaku Atuahene-Gima. Market knowledge dimensions and cross-functional collaboration: Examining the different routes to product innovation performance. *Journal of Marketing*, 71(1):95–112, January 2007.

Martin A Makary and Michael Daniel. Medical errorthe third leading cause of death in the us. *BMJ*, 353:i2139, 2016.

John Makhoul, Francis Kubala, Richard Schwartz, Ralph Weischedel, et al. Performance measures for information extraction. In *Proceedings of DARPA broadcast news workshop*, pages 249–252, 1999.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

Howard Markel. i swear by apolloon taking the hippocratic oath. *N Engl J Med*, 350(20): 2026–2029, 2004.

Aleix M. Martínez and Avinash C. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, Feb 2001.

Genevieve B. Melton and George Hripcsak. Automated detection of adverse events using natural language processing of discharge summaries. *Journal of the American Medical Informatics Association*, 12(4):448–457, 2005.

Malik Muhammad Saad Missen, Mohand Boughanem, and Guillaume Cabanac. Opinion mining: reviewed from word to document level. *Social Network Analysis and Mining*, 3(1):107–125, 2013.

Myriam D. Munezero, C. Suero Montero, Erkki Sutinen, and John Pajunen. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transaction on Affective Computing*, 5(2):101–111, April–June 2014.

Harvey J Murff, Fern FitzHenry, Michael E Matheny, Nancy Gentry, Kristen L Kotter, Kimberly Crimin, Robert S Dittus, Amy K Rosen, Peter L Elkin, Steven H Brown, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *Jama*, 306(8):848–855, 2011.

C.M. Nunes, Jr. Britto, Ad.S., C.A.A. Kaestner, and R. Sabourin. An optimized hill climbing algorithm for feature subset selection: evaluation on handwritten character recognition. In *Frontiers in Handwriting Recognition, 2004. IWFHR-9 2004. Ninth International Workshop on*, pages 365–370, Oct 2004. doi: 10.1109/IWFHR.2004.18.

Lucila Ohno-Machado, Zia Agha, Douglas S Bell, Lisa Dahm, Michele E Day, Jason N Doctor, Davera Gabriel, Maninder K Kahlon, Katherine K Kim, Michael Hogarth, et al. pscanner: patient-centered scalable national network for effectiveness research. *Journal of the American Medical Informatics Association*, 21(4):621–626, 2014.

Miles Osborne. Using maximum entropy for sentence extraction. In *Proceedings of the the Association for Computational Linguistics (ACL) Workshop on Automatic Summarization*, pages 1–8, Philadelphia, 2002.

Alexander Osherenko and Elisabeth André. Lexical affect sensing: Are affect dictionaries necessary to analyze affect? In *Affective Computing and Intelligent Interaction*, pages 230–241. Springer, 2007.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *Language*, 10(July):79–86, 2002. URL `http://arxiv.org/abs/cs/0205070`.

Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302):157–175, 1900.

Tao Peng, Wanli Zuo, and Fengling He. SVM based adaptive learning method for text classification from positive and unlabeled documents. *Knowledge and Information Systems*, 16(3): 281–301, September 2008.

James W. Pennebaker and Laura A. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312, December 1999a.

James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999b.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

James W Pichert, Gerald Hickson, and Ilene Moore. Using patient complaints to promote patient safety. 2008.

James W Pichert, Ilene N Moore, Jan Karrass, Jeffrey S Jay, Margaret W Westlake, Thomas F Catron, and Gerald B Hickson. An intervention model that promotes accountability: Peer messengers and patient/family complaints. *Joint Commission Journal on Quality and Patient Safety*, 39(10):435–446, 2013.

Fred Popowich. Using text mining and natural language processing for health care claims processing. *ACM SIGKDD Explorations Newsletter*, 7(1):59–66, 2005.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, 2014. URL `http://www.R-project.org`.

Anand Rajaraman, Jeffrey D. Ullman, Jeffrey David Ullman, and Jeffrey David Ullman. *Mining of massive datasets*, volume 77. Cambridge University Press Cambridge, 2012.

Fuji Ren and Changqin Quan. Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing. *Information Technology and Management*, 13(4):321–332, December 2012.

William Rosenberg and Anna Donald. Evidence based medicine: an approach to clinical problem-solving. *BMJ: British Medical Journal*, 310(6987):1122, 1995.

Stuart Jonathan Russell and Peter Norvig. *Artificial intelligence: a modern approach (3rd edition)*. Prentice Hall, 2009.

Andressa Midori Sakai, Mariana Angela Rossaneis, Maria do Carmo Fernandez Lourenço Haddad, and Denise da Silva Scaneiro Sardinha. Feelings of nurses in the reception and risk classification evaluation in the emergency room. 2016.

Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

Yutaka Sasaki et al. The truth of the f-measure. *Teach Tutor mater*, pages 1–5, 2007.

Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, July 1990.

Mark Schlesinger, Rachel Grob, and Dale Shaller. Using patient-reported information to improve clinical practice. *Health services research*, 50(S2):2116–2154, 2015.

Oliver C. Schultheiss. Are implicit motives revealed in mere words? Testing the marker-word hypothesis with computer-based text analysis. *Frontiers in Psychology*, 4, October 2013.

Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467, 2011.

Charles D. Smith. *Palestine and the Arab-Israeli conflict: A history with documents*. Bedford/St. Martin's, Boston, 2010.

C.J. Stimson, James W. Pichert, Ilene N. Moore, Roger R. Dmochowski, M. Bernadette Cornett, Angel Q. An, and Gerald B. Hickson. Medical malpractice claims risk in urology: An empirical analysis of patient complaint data. *The Journal of Urology*, 183(5):1971–1976, May 2010.

Laura A Stokowski. Who believes that medical error is the third leading cause of hospital deaths? *Medscape, May*, 26, 2016.

Carolin Strobl, James Malley, and Gerhard Tutz. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 14(4):323, 2009.

Chunke Su, Meikuan Huang, and Noshir Contractor. Understanding the structures, antecedents and outcomes of organisational learning and knowledge transfer: A multi-theoretical and multilevel network analysis. *European Journal of International Management*, 4(6):576–601, 2010.

Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, March 2010.

Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0.

Riikka Vuokko, Päivi Mäkelä-Bengs, Hannele Hyppönen, and Persephone Doupi. Secondary use of structured patient data: Interim results of a systematic review. *Digital Healthcare Empowering Europeans: Proceedings of MIE2015*, 210:291, 2015.

Kavishwar B. Wagholikar, Kathy L. MacLaughlin, Michael R. Henry, Robert A. Greenes, Ronald A. Hankey, Hongfang Liu, and Rajeev Chaudhry. Clinical decision support with automated text processing for cervical cancer screening. *Journal of the American Medical Informatics Association*, 19(5):833–839, 2012.

Adam B Wilcox and George Hripcsak. The role of domain knowledge in automating medical text report classification. *Journal of the American Medical Informatics Association*, 10(4): 330–338, 2003.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

Qing-Song Xu and Yi-Zeng Liang. Monte carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11, 2001.

Ziqiong Zhang, Qiang Ye, Rob Law, and Yijun Li. Automatic detection of subjective sentences based on chinese subjective patterns. In *Cutting-Edge Research Topics on Multiple Criteria Decision Making*, pages 29–36. Springer, 2009.

Kang Zhao, John Yen, Greta Greer, Baojun Qiu, Prasenjit Mitra, and Kenneth Portier. Finding influential users of online health communities: a new metric based on sentiment influence. *Journal of the American Medical Informatics Association*, 21(e2):e212–e218, 2014. ISSN 1067-5027. doi: 10.1136/amiajnl-2013-002282.

Li Zhou, Genevieve B Melton, Simon Parsons, and George Hripcsak. A temporal constraint structure for extracting temporal information from clinical narrative. *Journal of biomedical informatics*, 39(4):424–439, 2006.