Text Analytics for Customer Engagement in Social Media

Richard Gruss

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Business Information Technology

Alan S. Abrahams, Chair

Weiguo (Patrick) Fan

Christopher W. Zobel

G. Alan Wang

Onur Seref

March 21, 2018

Blacksburg, VA

Keywords: text analytics, customer engagement, social media, natural language processing, machine learning, pharmacovigilance

Text Analytics for Customer Engagement in Social Media

Richard Gruss

ABSTRACT (academic)

Businesses have recognized that customers provide value to the firm beyond transactions, and leveraging this value through relationships in social media is a new area of interest for both academics and practitioners. Recent research has investigated how businesses can best manage their online presence on platforms not fully under their control, such as Facebook, YouTube, Instagram, TripAdvisor, and Yelp, among others. This dissertation extends the literature of customer engagement in social media through four contributions. First, we propose a framework that foregrounds the textual artifacts involved in online communication. Second, we develop a novel method for discovering the elements of successful Business to Customer (B2C) messages in online communities. Third, we propose a method, validated through experimentation, for finding critical product feedback in Customer to Customer (C2C) communications. Finally, we demonstrate that a set of novel numerical features can enhance the discovery of product defect mentions in C2C communications. We conclude by proposing a research agenda suggested by the framework that will further enhance our understanding of the complex customer interactions that characterize business in the era of social media.

Text Analytics for Customer Engagement in Social Media

Richard Gruss

ABSTRACT (general audience)

Businesses have recognized that customers provide value to the firm beyond transactions, and leveraging this value through relationships in social media is a new area of interest for both academics and practitioners. Recent research has investigated how businesses can best manage their online presence on platforms not fully under their control, such as Facebook, YouTube, Instagram, TripAdvisor, and Yelp, among others. This dissertation extends the literature of customer engagement in social media through four contributions. First, we propose a framework that foregrounds the textual artifacts involved in online communication. Second, we develop a novel method for discovering the elements of successful Business to Customer (B2C) messages in online communities. Third, we propose a method, validated through experimentation, for finding critical product feedback in Customer to Customer (C2C) communications. Finally, we demonstrate that a set of novel numerical features can enhance the discovery of product defect mentions in C2C communications. We conclude by proposing a research agenda suggested by the framework that will further enhance our understanding of the complex customer interactions that characterize business in the era of social media.

ACKNOWLEDGEMENTS

It would be impossible to overstate my gratitude to Alan Abrahams for his guidance, mentorship, and friendship over all these years. His energy and patience have been a wellspring of inspiration without which this would never have been accomplished. When I consider the thousands of emails we exchanged over the years, I realize that his future biographer has his work cut out for him in assembling the collected correspondence.

I would also like to thank my committee members Patrick Fan, Alan Wang, Chris Zobel, and Onur Seref for the time they dedicated to helping me refine and clarify my work. I feel lucky to have been guided by such a team of eminent researchers.

Thanks also to Cliff Ragsdale, who did so much to see me through the program. His customary quiet wisdom turned into fierce advocacy at just the right moments.

Many thanks also to my Math Emporium friends, Terri Bourdon, whose reassurance helped so much in the dark times, and Peter Haskell, whose enlightened vision of what a university is made it possible for a humble coder to pursue his curiosities.

I also owe a debt of gratitude to my Radford family, who never let me go a day without an update: Dale Henderson, Tal Zarankin, Iain Clelland, Danylle Kunkel, Wil Stanton, Angela Stanton, Maneesh Thakkar, Gary Schirr, Hooshang Beheshti, Vernard Harrington, Jae Jeong, Jerry Kopf, Steve Childers, Tom Lachowicz, Brooklyn Cole, Shu Wang, and Jonathan Preedom.

Most of all, I would like to thank my amazing wife Laura, whose love and support was unfailing year after year, and Rose, Lily, and Danny, who had to forgo so much while waiting for Dad to look up from his computer.

NOTE

The PamTag collaborative tagging system developed by the author to support major aspects of the dissertation work was a core component of the Text Analytics Suite for Consumer Product Safety Surveillance, which won the international INFORMS Information Systems Society Design Science Award, in Seoul, Korea, December 2017.

The author also contributed similar text analytic work to the following publications:

Nasri, L., Baghersad, M., **Gruss**, R., Marucchi, N. S. W., Abrahams, A. S., & Ehsani, J. P. (2018). An investigation into online videos as a source of safety hazard reports. Journal of Safety Research.

Mummalaneni, V, R **Gruss**, D Goldberg, Ehsani J, A Abrahams. "Social media analytics for quality surveillance and safety hazard detection in baby cribs." *Safety Science*, *104*, 260–268.

Adams, D. Z., **Gruss**, R., & Abrahams, A. S. (2017). Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. *International Journal of Medical Informatics, 100*, 108-120.

Law, D., **Gruss**, R., & Abrahams, A. S. (2017). Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications*, 67, 84-94.

Winkler, M., Abrahams, A. S., **Gruss**, R., & Ehsani, J. P. (2016). Toy safety surveillance from online reviews. *Decision Support Systems, 90*, 23-32.

TABLE OF CONTENTS

Introduction vii

Text Analytics viii
Customer Engagement xi
Social Media xiv
A Framework for Text Analytics for Customer Engagement in Social Media xvi
Description of the Studies xvii

Chapter 1. Text Analytics for B2C Interactions: Engaging Customers in Online Communities 1

1.1 Introduction 1
1.2 Related Work 3
1.3 Theoretical Foundations and Hypothesis Development 8
1.4 Case Study 1: User Engagement with Academic Libraries on Facebook 11
1.5 Case Study 2: User Engagement with Restaurants on Facebook 23
1.6 Results 28
1.7 Discussion 38
1.8 Limitations and Future Work 39
1.9 Conclusions and Implications for Research and Practice 41

Chapter 2. Text Analytics for C2C Interactions: Application in Pharmacovigilance 43

2.1 Introduction 44
2.2 Related Work 48
2.3 Data and Methods 52
2.4 Results 59
2.5 Discussion 69
2.6 Conclusions 73
Appendix 2.A. Detailed list of Amazon.com categories and subcategories 76
Appendix 2.B. Examples of online reviews indicative of safety concerns 77
Appendix 2.C. Most commonly selected unigrams, bigrams, and trigrams 78
Acknowledgements 80

Chapter 3. Text Analytics for C2C Interactions: Numeric Information Extraction

3.1 Introduction 81
3.2 Theoretical Background 85
3.3 Related Work 92
3.4 Methodology 95
3.5 Results 112
3.6 Numerical Attributes for Information Retrieval Systems 121
3.7 Limitations and Future Work 123
3.8 Conclusions 124
Appendix 3.A. Tagging Protocol 125
Appendix 3.B. Number classification tables 127
Appendix 3.C. Sample word contexts from training set 132

Conclusion 140

References 144

INTRODUCTION

The emergence of the term "customer engagement" (CE) in both practitioner and academic literature reflects the inadequacy of the older term "customer relationship management" (CRM) to fully capture the complex nature of interacting with customers through social media in addition to traditional channels (Hennig-Thurau et al., 2010). Social media provide business an unprecedented opportunity to pursue extra-transactional relationships with customers that add value for both parties (V Kumar & Pansari, 2016). Whereas customers were once considered as "exogenous entities" whose value derived solely from expected lifetime transactions, there has been a paradigmatic shift to a new attitude in which customer value manifests in a variety of ways as they take on roles such as brand advocate, product champion, troubleshooter, and co-innovator (Lusch, Liu, & Chen, 2010). In this view, customers are stakeholders with whom businesses need to maintain productive conversations to stay competitive, through both retaining existing customers and acquiring new ones (Bijmolt et al., 2010; Vivek, Beatty, & Morgan, 2012).

Several benefits follow from the intimate-yet-public relationship with customers in social media, such as increased loyalty, new idea generation, enhanced feedback, and information exchange. But there are serious risks as well (Alvarez & Fournier, 2016). Dissatisfied customers spread negative word of mouth (WOM) quickly, which has been shown to have significant negative effects on stock price and cash flows (Luo, 2009). The extremity of the risks and benefits has created problems for businesses as they struggle to learn exactly how to negotiate these relationships (Vivek et al., 2012). Some businesses have found that commitment from customers in the relationship is not as strong as they would like (Alvarez & Fournier, 2016), and it has even been suggested that businesses are unwelcome intruders in social media (Fournier & Avery, 2011). In an age where dissatisfied customers can publicly rant or even create viral brand parodies (Vanden Bergh, Lee, Quilliam, & Hove, 2011), businesses are in grave danger of losing control of the

conversation, compromising a carefully-crafted brand image (Libai et al., 2010; Mangold & Faulds, 2009).

Because of the potential benefits and risks, CE has been identified by the Marketing Science Institute as a research priority (Hollebeek, Conduit, & Brodie, 2016). The Conceptualization of CE has been a complicated process, as discussed below under "Definitions", but there is general agreement that CE refers to a variety of business-customer interactions that create value for both parties. This dissertation examines CE interactions in social media. Because conversations are conducted in language, we concentrate on the text artifacts of business-customer interactions, applying the methods and insights of text analytics.

In the sections that follow, we provide working definitions for our chief constructs, present our overall research framework, and summarize the three studies that make up this dissertation.

Definitions

To ensure precision and clarity about our terminology, we provide our operational definitions for each of our three key constructs in our title: "text analytics", "customer engagement", and "social media".

"TEXT ANALYTICS"

Although sometimes used interchangeably with "text mining" (TM) and "natural language processing" (NLP) (Aggarwal & Zhai, 2012), the term "text analytics" (TA) has become the de facto standard term in industry for the derivation of structured data from unstructured text (Ittoo, Nguyen, & van den Bosch, 2016). The major business-oriented text processing software vendors--SAS, IBM, Microsoft, Clarabridge--use the term throughout their documentation. NLP and "computational linguistics" (CL) are conceptually related to TA, but these terms are most often used within computer science, where the primary research

goal is to discover new algorithms for specific tasks (Cambria & White, 2014; Nirenburg & McShane, 2016). Tasks that are actively being researched in NLP include text categorization (G. Chen, Ye, Xing, Chen, & Cambria, 2017; Zheng, Tang, Zhang, & Tang, 2017) , text clustering (Abualigah et al., 2017; Xu et al., 2017), concept/entity extraction (Z. Jiang, Zhang, & Li, 2017), sentiment analysis (Rosenthal, Farra, & Nakov, 2017; Shelke, Deshpande, & Thakare, 2017), document summarization(Cao, Li, Li, & Wei, 2017; Yinfei Yang, Bao, & Nenkova, 2017), knowledge based population (KBP) (Chaganty, Paranjape, Liang, & Manning, 2017; Min, Freedman, & Meltzer, 2017), and question answering (QA) (Yuyu Zhang, Dai, Kozareva, Smola, & Song, 2017). An overview of the central tasks for NLP research can be found in the historical tracks for the Text Analysis Conferences hosted by the National Institutes of Standards and Technology (NIST) (https://tac.nist.gov/tracks/index.html).

"Text analytics" and "text mining", by contrast, are concerned with applying these algorithms to the creation of analyzable data from text to answer particular operational questions (Khan, Khan, Vorley, & Vorley, 2017). Research in this area attempts to establish which approaches work well on what kinds of problems, and to determine the benefits and tradeoffs involved in implementing solutions (Aggarwal & Zhai, 2012). In this dissertation, we adopt the following definition of TA that is consistent with recent TA research: "Text analytics: the application of NLP and CL techniques for the purpose of answering strategic business questions."

Businesses have been investing heavily in text analytics. Across industries, 30% of projects aimed at generating insights are text-based, and in some industries that figure is as high as 50% (Müller, Debortoli, Junglas, & vom Brocke, 2016). Text analytics for business was initially motivated by large volumes of internal documents. It has been estimated that 80% of an organization's internal data is "unstructured" (Castellanos, Castillo, Lukyanenko, & Tremblay, 2017), mostly in the form of text, and many recent studies have been directed at transforming text into actionable insights. For example, insurance companies have

analyzed claims to predict fraud, healthcare providers have analyzed electronic health records to aid in diagnosis, and pharmaceutical companies have analyzed scientific literature to enhance their product development processes (Müller et al., 2016).

The availability of social media and its potential strategic benefits have motivated research into text analytics on external text documents from the Web (Forman, Ghose, & Wiesenfeld, 2008; Ittoo et al., 2016). Recent applications of text analytics in social media have included discovering the sources of customer dissatisfaction (Gu & Ye, 2014) and satisfaction (Xiang, Schwartz, Gerdes, & Uysal, 2015), analyzing competitor strategy (He, Zha, & Li, 2013), assessing the impact of customer WOM (Goh, Heng, & Lin, 2013), gauging brand sentiments (Mostafa, 2013), and detecting product defects (Abrahams, Fan, Wang, Zhang, & Jiao, 2015; Abrahams, Jiao, Wang, & Fan, 2012; Adams, Gruss, & Abrahams, 2017; Goldberg & Abrahams, 2017; Winkler, Abrahams, Gruss, & Ehsani, 2016) , to name only a few.

Table 1 below highlights some recent work in business-oriented text analytics.

Table 1 . Recent business applications of text analytics.

Reference	Application	Text Data Set	Techniques
(Kronrod & Danziger, 2013)	how customers respond to figurative language in reviews	online reviews of hotels, content manipulated	content analysis, experiments
(Dong, Liao, Xu, & Feng, 2016)	predicting which firms will commit fraud	Twitter	combining social media data (word topic, and sentiment features) with financial ratios
(Sarker et al., 2016)	discovering prescription drug abuse	Twitter	n-gram features, supervised classification
(He, Tian, Chen, & Chong, 2016)	compare use to discover customer experience of drug stores online	Facebook sites of Walgreens, CVS, and Rite Aid	sentiment analysis
(Packard & Berger, 2017)	the impact of specific endorsement language in reviews	book reviews, hotel reviews	manipulated language in randomized controlled experiment
(Poon, Lam, & Moon, 2017)	competitive analysis of the fashion	Facebook, Twitter, Google, Baidu,	branded keywords, named entity

	industry	Weibo, Uwants, B2C websites	recognition (NER), sentiment analysis
(Xiang, Du, Ma, & Fan, 2017)	discovering different representations of the hotel industry across platforms	online reviews from TripAdvisor, Expedia, and Yelp	linguistic characteristics, semantic features, sentiment
(Xiang, Schwartz, & Uysal, 2017)	classifying hotels based on how users comment	online hotel reviews	classification
(R. Chen & Xu, 2017)	discovering customer priorities in products	online reviews of cameras	aspect-oriented sentiment analysis
(Shi, Guan, Zurada, & Manikas, 2017)	aviation safety monitoring	text in incident reports	latent semantic analysis, data stream learning
(Huizinga, Ayanso, Smoor, & Wronski, 2017)	deriving new insurance products	Twitter	association rule mining
(Y. Wang & Xu, 2018)	auto insurance fraud detection	text descriptions of accidents	LDA and deep learning
(Abrahams et al., 2015; Abrahams, Jiao, Fan, Wang, & Zhang, 2013; Abrahams et al., 2012; Adams et al., 2017; Goldberg & Abrahams, 2017; Law, Gruss, & Abrahams, 2017; Mummalaneni, Gruss, Goldberg, & Abrahams, 2018; Winkler et al., 2016)	product defect discovery	discussion forum postings	term prevalence metrics, document scoring

"CUSTOMER ENGAGEMENT"

The presence of businesses on social media opened the door to new interactions with customers. Customers can interact with the business (B2C) or with each other (C2C), and in both cases, there are opportunities to create value.

The term "customer engagement" in business originated in practitioner literature (Hollebeek, 2011), but by 2011 it had attracted sufficient academic interest to warrant a special edition of the Journal of Service Research (Volume 14, Issue 3). Several articles in this issue took the view that since "customer engagement" was new to academic research, it

was necessary to conceptualize the term precisely and to propose a research agenda. Brodie, Hollebeek, Jurić, & Ilić (2011) sought to clarify CE's relationship with several overlapping concepts (relationship marketing, customer relationship management, brand loyalty) and proposed that CE refers to "a customer's emotional and cognitive investment in interacting with a business" (Brodie et al., 2011).

Sashi (2012) sought to develop a theoretical framework of customer engagement founded on real-world business practice. Businesses are motivated to develop long-term relationships with customers, and as the relationship matures, it goes through several stages: connection, interaction, satisfaction, retention, loyalty, advocacy, and finally, engagement. It is no coincidence that interest in customer engagement coincided with the advent of social media: "Its rise in the consciousness of managers has paralleled the emergence of new technologies and tools that enable greater interactivity among individuals and organizations" (Sashi, 2012).

Vohra & Bhardwaj(2016) argued that although "customer engagement" is a multidimensional construct, customer interactions form the backbone, whether B2C or C2C. Customers add value in several ways by interacting with firms on social media (B2C). They can serve as "co-innovators" by providing feedback or providing recommendations for product improvements (Aral, Dellarocas, & Godes, 2013), a process sometimes called "collaborative product development" (Mangold & Faulds, 2009).

Customer interactions with each other (C2C) are also an important aspect of engagement (Bijmolt et al., 2010; Libai et al., 2010). The critical motivating fact upon which several authors agree is that customers add value to the firm through customer-to-customer behaviors (Brodie et al., 2011; Van Doorn, 2011), as was previously proposed in (Verhoef, Reinartz, & Krafft, 2010) and subsequently expanded upon in (Jaakkola & Alexander, 2014). Customers turn to social media to express their brand enthusiasm or offer their expertise in product technical details (Aral et al., 2013). The importance of C2C communication has been greatly magnified by social media (Mangold & Faulds, 2009). (Van Doorn et al., 2010)

identifies several C2C engagement behaviors, including word of mouth activity (WOM), recommendations, helping other customers, blogging, writing reviews, and even engaging in legal action.

In 2016, a special edition of the Journal of Marketing Management (Volume 32, Issue 5-6) extended the research into CE with several pieces that produced clarified definitions and validated scales. (Hollebeek et al., 2016) argued that there is now consensus that CE has cognitive, emotional, and behavioral manifestations, and that both business to customer (B2C) and customer-to-customer (C2C) interactions are critical elements to be researched.

While some sources refer to CE as a psychological state, we take a process-oriented view, and adopt the following as our working definition of "customer engagement", which comes from (Brodie et al., 2011) and (Hollebeek et al., 2016): "Customer engagement is the voluntary non-transactional, value-adding interaction of a customer with either the business or other customers of the business." Social media is an essential enabler of these interactions.

Does proficient online customer engagement lead to improved business outcomes? Some empirical results have suggested so. Customer engagement has been shown to have a direct and positive effect on word of mouth (M. Zhang, Hu, Guo, & Liu, 2017) and on customer stickiness (M. Zhang, Guo, Hu, & Liu, 2017). (Braojos, Benitez, & Llorens-Montes, 2017) found that organizations that have strong social media capability and e-business capability see improved organizational performance. Companies that do little advertising see shareholder valuations rise when they initiate customer engagement activities (Beckers, van Doorn, & Verhoef, 2017), although firms with an already-stable reputation see little benefit. While most studies have confirmed the co-creative value of rich customer interactions, one interesting counter-intuitive study found that designs that incorporated community input showed less variety and lower quality (Hildebrand, Häubl, Herrmann, & Landwehr, 2013). So the nature and value of the customer input must also be kept in mind.

"SOCIAL MEDIA"

There is little question that social media have changed the way businesses interact with stakeholders: "Social media have fundamentally transformed the relationship among firms, employees, and consumers. They have changed the norms of behavior at various levels and have introduced a bewildering range of new opportunities and challenges. All stakeholders must, therefore, learn how to optimally use this new set of tools to meet their respective objectives" (Aral et al., 2013).

Social media studies are still fairly new to academia (Ngai, Tao, & Moon, 2015), and researchers are still struggling to deal with several data quality challenges associated with working with social media data. Care must be taken when drawing conclusions about real-world phenomena from social media data, due to several factors, including selection bias (Hu, Pavlou, & Zhang, 2006), system-gaming (Malbon, 2013), endogenous variables (Aral et al., 2013), and unobserved correlates (Hartmann et al., 2008). (Aral et al., 2013) summarized some of the practices that should be employed when predicting real outcomes from social media (or any non-experimental) data: using panel data to control for unobserved heterogeneity, difference-in-difference methods, matched sample estimation, and instrumental variables. Differentiating helpful information from junk has been the objective of several social media studies (Connors, Mudambi, & Schuff, 2011; Hong, Xu, Wang, & Fan, 2017; Y. Liu, C. Jiang, Y. Ding, et al., 2017). More fundamentally, user-generated content is rife with non-standard grammar, spelling, punctuation, and abbreviations (Salloum, Al-Emran, Monem, & Shaalan, 2017), posing challenges for standard text mining techniques such as sentiment or topic analysis.

Given the noisy nature of social media data, then, throughout the three studies that constitute this dissertation, we seek to minimize validity and reliability threats through a variety of research best practices: completely randomized samples, strict standards for

inter-rater reliability, holdout test data sets, verification of findings on supplementary case studies, high confidence levels, and various methods of stable predictive model induction.

We adopt a specialized form of the broad definition of "social media" from the Oxford English Dictionary (https://en.oxforddictionaries.com/definition/social_media): "Social media: Websites and applications that enable users to create and share content or to participate in social networking." Although the term "social media" encompasses a vast range of online interaction tools, we confine our focal social media to three venues in which B2C and C2C interactions take place: brand communities, online reviews, and discussion forums.

Brand Communities: Brand communities, whether managed by the company or spontaneously created by fans, serve two strategic goals: to provide a consistent and differentiated brand image, and to promote the brand by stimulating WOM from fans (Gummerus, Liljander, Weman, & Pihlström, 2012). Brand communities can take on a variety of formats (M. Zhang, M. Hu, et al., 2017) but for our purposes, by "brand community" we mean a Facebook fan page owned and managed by the business.

Online Reviews: Online reviews have attracted a great deal of attention in recent research, (for a thorough review see (Zhao, Stylianou, & Zheng, 2017)). We consider online reviews "social media" because they are specifically for the sharing of content among users, even though the social graphs connecting users on online review sites are often minimal.

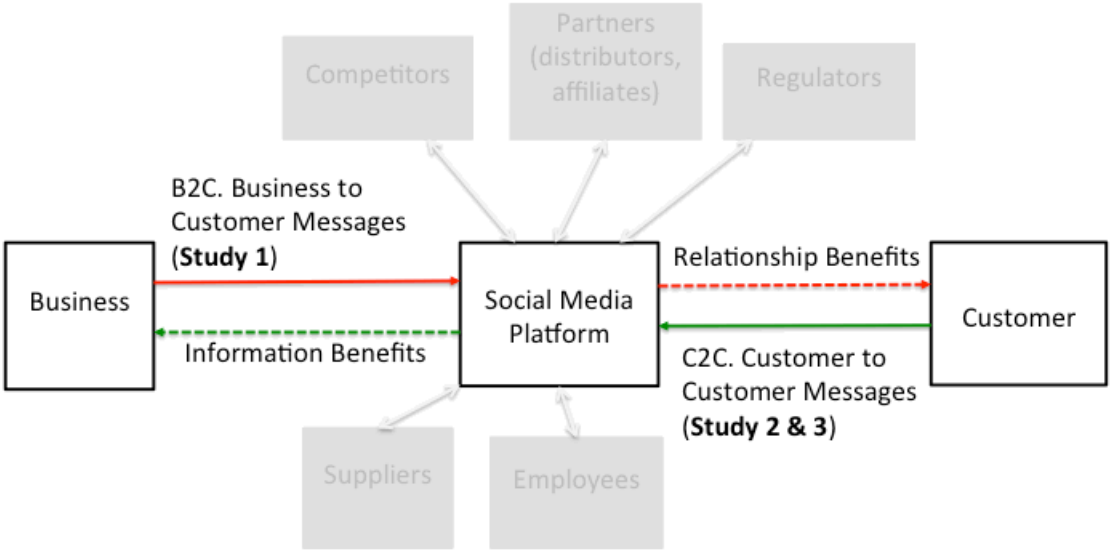
Discussion forums. Organized around specific interests, discussion forums allow users to interact and share expertise. Discussion forums have drawn interest because of their use in education (Loncar, Barrett, & Liu, 2014; Onah, Sinclair, & Boyatt, 2014) and for what they sometimes reveal about human behavior (Cheng, Danescu-Niculescu-Mizil, & Leskovec, 2015). In this dissertation, we examine messages from three auto enthusiast discussion forums: Honda-Tech.com, ToyotaNation.com, and ChevroletForum.com.

A Framework for Text Analytics for Customer Engagement in Social Media

Managing stakeholder relationships on social media platforms poses a challenge for businesses (Kaplan & Haenlein, 2010). Maintaining relationships with customers in social media is difficult for businesses because of the multiplicity of channels through which they need to operate that are beyond their control (Hennig-Thurau et al., 2010). Figure 1 presents our framework for modeling these channels. We concentrate here on engagement with customers, and leave for later work the analysis of engagement with other stakeholders, such as competitors (He et al., 2013), regulators (Adams et al., 2017), suppliers (Swani, Milne, & P. Brown, 2013), and employees (Leonardi, Huysman, & Steinfield, 2013).

Each chapter of this dissertation addresses a specific question involving a text artifact within this framework, concentrating along the customer engagement channels. Our framework is a modified form of the CE framework in (Brodie et al., 2011) that foregrounds the text artifacts of communication. In the studies described below, we examine the effective elements of B2C messages (Study 1) and the information value of C2C message (Study 2 and Study 3).

Figure 1. A Framework for Text Analytics for Customer Engagement in Social Media.



Descriptions of the Studies

In this dissertation, we present three studies that take a text analytics approach to engaging with customer on social media. We study four different industries (higher education, hospitality, pharmaceuticals, and automotive), three online venue genres (Facebook for outbound organizational messaging, Amazon.com for online reviews, and two automotive enthusiast discussion forums), and two message categories (Business-to-Customer and Customer-to-Customer).

In Study 1, we examine the drivers of customer interest in organizational postings (B2C) on Facebook by testing whether "community building" messages prompt a greater response than other kinds of messages. Miller and Tucker (2013) examined the complexities of how firms interact with various stakeholders on social media, and argued that further work should study what types of communications evoke what type of response. In this study, we take up this task and investigate the drivers of customer interest in interacting with organizations on Facebook. Guided by social belongingness theories, we test whether appeals to belonging to a community evoke a high level of response from

customers, while controlling for several other factors. We find that messages with belongingness appeals are consistently associated with higher likes, but not necessarily with comments or shares. We discuss what this implies for practice, and suggest future studies for how to trigger more comments and shares.

In Study 2, we examine how businesses can use online reviews for product quality management. The experiences that customers write about in online reviews, when leveraged correctly, provide the "co-innovation" that is a fundamental benefit of engagement (Aral et al., 2013). The high volume of online reviews makes reading them a formidable undertaking, so reliable automated methods of filtering out noise are in demand. In this study, we apply a text analytic methodology that has elsewhere proven effective in finding safety defects in consumer products to the task of automatically locating adverse drug reactions among thousands of online reviews for over-the-counter medications. We find that the method is effective at differentiating reviews that contain adverse drug reactions from those that do not. We also find that the machine learning process of lexicon generation that we apply is outperformed by lexicons created by groups of participants who brainstorm lists from scratch. This has important implications for practice, which we discuss in depth in Chapter 2.

In Study 3 we continue our inquiry into automated methods of using C2C messages for product quality management. In this study, we propose and evaluate a procedure for extracting, identifying, and binning numerical attributes. We then demonstrate a practical application of these numerical attributes by showing that they enhance the reliability of automated defect discovery. We further propose a decision support system based around using numerical attributes for post-market quality control.

The three studies that constitute this dissertation are standalone research papers, but they all investigate the potential of text analytic methods for managing customer interactions in social media. As the importance of social media continues to grow,

understanding how to derive maximum benefit from the previously non-existent opportunities to engage with customers will be a critical element of business strategy.

CHAPTER 1

Text Analytics for B2C Interactions: Engaging Customers in Online Communities

Abstract

Many organizations have recognized that engaging with customers on social media is now a critical element of their public relations strategy, but many struggle to find the optimal way to use these new tools. In this paper, we examine the success factors of online posts from organizations for evoking a high level of response from group members. We contribute to the literature a unique set of features that have a significant impact on customer engagement and support our findings with a large data set from two different industries. In accordance with theories involving the need for belongingness, we find that appeals to a sense of community belonging have a significant impact on user engagement. Specifically, we discover that communities have idiomatic vocabularies consisting of "activation words" that are especially effective at engaging group members on social media. This has both theoretical implications, as it constitutes a large-scale, real-world confirmation of belongingness hypotheses, and managerial implications, as it suggests best practices for maintaining an online presence.

1.1. Introduction

Organizations are increasingly looking for ways to develop long-term, extra-transactional relationships with stakeholders on social media (Hoffman & Fodor, 2010). The value of a single Facebook "like" to a business is rising (Sashi, 2012), and interest has been growing in fully understanding what prompts likes, comments, and shares (Johnson, Safadi, & Faraj, 2015). Changes in Facebook's newsfeed selection algorithm now mean that organizational content that has not received many likes, comments, or shares could drop out

of a user's view altogether, making it more important to have content that prompts a response from users.

Previous work investigates Facebook "liking" behavior by focusing on the users, exploring online behaviors through the theoretical lens of psychology and sociology. Other studies examine liking behavior by focusing on the content of the posts themselves. In this study, we examine a large body of Facebook posts from organizations in an attempt to discover what linguistic, visual, sentimental, and timing elements contribute to a high level of customer engagement, manifested as an uncommonly large number of likes, comments, and shares. In particular, we are interested in whether appeals to a sense of community—either to an abstract concept of commonality or to a specific community with its own specialized vocabulary—help to trigger the viral audience response that has become the holy grail of marketing (Royo-Vela & Casamassima, 2011). In this paper, we identify various classes of community-related activation words and assess their impact on user engagement. Since our primary interest is on community, we conduct two case studies on organizations that have identifiable communities to whom they cater: academic libraries, (whose community is the university) and local restaurants in Boulder, Colorado. These two case studies allow us to investigate whether our hypotheses hold across distinct industries, academic non-profit, and hospitality for-profit.

A key construct in our study is "community building", which requires some definition and clarification. We take as our starting point a content analysis of library tweets (Stvilia & Gibradze, 2014) in which community building was found to be one of nine categories of posts, and was defined as follows:

> "The category community building included general tweets promoting the library as a place to receive research support, study, or hang out and have fun. This category also included tweets providing emotional support and congratulating students on various achievements (e.g., completing exams), as well as congratulating the library or a specific department for being recognized or achieving a high ranking in a national poll. In general, these types of tweets used a higher rate of affective terms."

We use this as our starting point, and make an extension inspired from (Phillips, 2011), who examined ways that libraries build rapport with users. Effective libraries tap into the shared norms, values, concerns, and symbols of the university: "The university setting not only creates a context for messages, but also offers a mutual set of experiences and values shared by libraries and students." Our operational definition of a "community building" post is one that has no particular relevance to the organization's main function, and serves only to create a common personal connection among audience members through shared symbols or concerns. Strictly functional posts from libraries would include those that announce collections, events, services, or hours. Following (Stvilia & Gibradze, 2014), community building posts are not strictly utilitarian, but rather promote community or provide emotional support, such as cheering on a team, wishing good luck on exams, or exhortations to stay safe in inclement weather, for example. One key endeavor of this study is to use the methods of text mining to build a set of features that quantify the extent of "community-buildingness" of a post, and then to evaluate whether these posts in fact motivate greater levels of user engagement.

The rest of this paper is organized as follows. In section 1.2 we review related work with social media engagement in general, and engagement on Facebook in particular. In section 1.3 we discuss the theories that lead to our hypotheses. We then present the background and methodology for our two cases studies in section 1.4 and 1.5, with results and discussion in section 1.6. Finally, in section 1.7, we discuss limitations and future work.

1.2. Related Work

The literature on Facebook engagement can be broadly divided into two categories: studies that concentrate on user characteristics and those that concentrate on post content. In

section 1.2.1, we examine previous research into user characteristics, and in section 1.2.2 we discuss the research focusing on post textual content.

1.2.1 The Influence of User Characteristics on Engagement

In examining user characteristics, Wallace et al. (2014) proposed a taxonomy of Facebook liking personalities which vary in the extent to which they are motivated by genuine interest, image creation, incentive, social tie strength, homophily, and other characteristics.

Within an online community, perceived similarity and trust of other members are found to increase user willingness to comment or share (Kozinets, 2002). Some studies investigated user motivation based on the specific benefits they derive from the online relationship. Royo-Vela & Casamassima (2011) determined that belonging to a virtual brand community increases satisfaction and affective commitment. Coulter et al. (2012) found that organizations are successful insofar as they offer the right kinds of relationship benefits, specifically providing the customer with recognition, friendship, and a feeling of usefulness.

Some studies of Facebook behavior find that an important motivation behind liking is the creation of personal identity: by liking content we are affirming and presenting our values and beliefs (Brandtzaeg & Haugstveit, 2014). Based solely on likes, computer algorithms are able to accurately predict personal attributes such as substance use, health, gender, race and political views (Youyou, Kosinski, & Stillwell, 2015). Users are aware of their self-image as they use Facebook, and respond to brands whose image is congruent with their own (Wallace, Buil, & de Chernatony, 2014).

1.2.2 The Influence of Post Content on Engagement

Other studies have focused on the post content. Swani et al. (2013) examined likes in the context of B2B transactions, and found that posts are more effective if they include corporate names and emotional terms and avoid hard-sell tactics. In a content analysis of 100,000 unique messages across 800 companies, (Lee, Hosanagar, & Nair, 2014) determined that emotional and philanthropic content increases engagement, but informative content such as prices and availability reduces it. In the non-profit arena, (Gaby & Caren, 2012) found that highly emotional messages that emphasize confrontation and solidarity are most effective at attracting followers to a political movement. Information-sharing behavior on social media has also been shown to be effected by emotional content (Stieglitz & Dang-Xuan, 2013).

Previous success with text analytics suggests that this is a mature and promising approach to analyzing online text. Several studies contribute to a taxonomy of text features that are successful in analyzing a variety of content for many tasks.

Table 1.2.1 summarizes these features and their usage in prior work.

Table 1.2.1: Summary of Text Features Frequently Used for Text Analytic Tasks.

Category	**Features**	**Prior Work**
Lexical	words, phrases, noun phrases, named entities	(Abbasi & Chen, 2008; Yulei Zhang, Dang, Chen, Thurmond, & Larson, 2009)
Stylistic	total words, characters per word, number of unique words, words per sentence, sentences per paragraph, readability	(Abbasi & Chen, 2008)
Social	number of posts, credibility, expertise, influence	(G. A. Wang, Jiao, Abrahams, Fan, & Zhang, 2013; Yulei Zhang et al., 2009)
Sentiment	subjectivity, positivity, negativity	(Stone, Dunphy, Smith, & Ogilvie, 1968)
Distinctive terms	industry-specific dictionaries	(Abrahams et al., 2015)
Product features	industry-specific categories	(Abrahams et al., 2015)
Semantic cues	concept classes	(Stone et al., 1968)

1.2.3 Problem Statement

Despite the value of customer engagement on social media and the potential of textual analysis to arrive at important insights, few studies have conducted textual analyses of the large volume of post content available to the public to explore the drivers of user engagement (see Table 1.2.2 for a research summary). We add to the literature by addressing this shortage using two large data sets from distinct industries and collecting a variety of textual attributes to build our model. We also extend the literature by taking a multi-level view, investigating both organizational features and content together. We propose that organizations seek to create an online community, with its own vocabulary and customs, and that rhetorical moves that seek to reinforce this community lead to increased user engagement. We therefore investigate both the attributes of the organizations and the content of their online messages, and how these tie to distinctive activation words from the user's community.

Table 1.2.2. Summary of social media engagement research.

Work	Focus	Methods	Findings
(S. Zhang, Jiang, & Carroll, 2010)	User	50 scenarios in interviews with 10 participants	Users dynamically construct identities through social media behavior.
(Gerolimos, 2011)	Content	Content analysis of 3513 posts from organizations	Engagement is not widespread, and activity is low.
(Phillips, 2011)	Content	Content analysis of 439 posts from 17 organizations	Low engagement, despite methods of community engagement.
(Coulter et al. 2012)	User	276 surveys to individuals	Social and entertainment benefits mediate Community Engagement Behaviors and Transaction Engagement Behaviors.
(Gaby & Caren, 2012)	Content	Content analysis of 100 posts	People respond to the messages that are confrontational.
(Swani et al., 2013)	Content	Content analysis of 1143 posts from 193 organizations	Brand name mentions increase engagement, while hard-sell tactics decrease engagement
(De Vries & Carlson, 2014)	User	Survey, 404 brand pages	Users like on the basis of 'hedonic' gratifications.
(Wallace, Buil, de Chernatony, et al., 2014)	User	Surveys of 438 users	There is a typology of Facebook personalities.
(Brandtzaeg & Haugstveit, 2014)	User	Content analysis of 405 surveys	Users have six levels of commitment to their liking.
(Lee et al., 2014)	Content	Textual analysis of 100,000 posts from 800 organizations, tagging with Mechanical Turk	Emotional and philanthropic content increases engagement, while information decreases engagement.
(Johnson et al., 2015)	Content	Textual analysis of 1 year of posts and surveys	Leaders' language is more readable and positive.
(Eranti & Lonkila, 2015)	User	Survey of 26 students	There are varied motives for liking beyond surface support.
This study	Multi-level: Organization, Content	Textual analysis of 51,760 posts from 100 organizations	Organizations and users are co-creators of an online community, and acts of community building by a posting organization leads to increased engagement.

1.3. Theoretical Foundations and Hypothesis Development

The need to belong is a fundamental and universal human motivator that explains much of interpersonal behavior (Baumeister & Leary, 1995). Other drivers, such as the need for power or the need for achievement, which have appeared frequently in motivational literature, may in fact be expressions of the more basic need for validation and recognition from others. The absence of a feeling of belonging has been associated with unhappiness and depression (Myers, 1992), and one recommended practice of clinical psychology is to show patients how to form social connections (Brehm, 1987).

While the belongingness hypothesis has theoretical roots in evolutionary biology and modern psychoanalysis, belongingness as a basic need has also been demonstrated empirically. Several studies have shown how readily people form strong attachments with a group. Relations & Sherif (1961) observed that feelings of inter-group antagonism dissolved when two groups were merged into one and individuals took on an identification with the new group. Intergroup discrimination, it has been suggested, is more about securing a "positive ingroup identity" than about competition (Turner, 1975). Many of the most joyous human rituals are those that involve commitments to reinforced group bonds (Baumeister & Leary, 1995). Individuals continuously monitor their level of social belonging, to the extent that if effects their information processing (Gardner, Pickett, & Brewer, 2000).

Social psychology provides a rich source of theory for understanding behavior in social media, and the belongingness literature is especially apt. People join virtual communities for social reasons such as friendship and support (Ridings & Gefen, 2004). One study investigating the reasons why people join online communities (Gangadharbatla, 2008) included the "need to belong," along with self-efficacy, need for cognition, and

collective self-esteem. Belongingness survey items were derived from (T. Leary, 1958) and (M. R. Leary, Kelly, Cottrell, & Schreindorfer, 2007). These included items such as "I want other people to accept me," "I do not like being alone," and "I try hard to stay in touch with my friends." The need to belong was the second-strongest significant factor determining willingness to join, after collective self-esteem.

There are, however, reasons why social media may not be satisfying the belongingness need. Baumeister & Leary (1995) differentiate between "belonging" and "mere affiliation," and contends that two criteria are necessary to satisfy the belongingness need: an ongoing relationship of mutual concern, and frequent contact. Social media specialize in providing frequent contact, but whether online interactions represent stable relationships of mutual concern is questionable. Still, people approach relationships with acquaintances as potential incipient bonds (Baumeister & Leary, 1995), so may behave online as though these relationships have the potential to develop. It has been observed that even when explicit support behaviors are not provided, belonging to a group reduces stress (S. Cohen & Wills, 1985).

Belongingness needs can be stimulated, linguistically, in both abstract and concrete ways. In the abstract, language that refers more generally to togetherness, community, a common collective, etc. would be gratifying a need to belong. These kinds of expressions are reminders of the social aspects of interacting online, and its potential for needs-fulfillment. We therefore propose hypothesis 1:

*H1: Language that **emphasizes a feeling of community** is related to higher user engagement.*

Belongingness language can also be more concrete, with references to specific

community symbols, rituals, shared secrets, in-jokes, etc., more directly invoke the unique elements of the community. To derive this hypothesis, we draw on the literature of brand community marketing.

A definition of "brand community" was proposed in (Muniz & O'Guinn, 2001) as: "a specialized, non-geographically bound community, based on a structured set of social relations among admirers of a brand… These brand communities exhibit three traditional markers of community: shared consciousness, *rituals and traditions*, and a sense of moral responsibility. "

Brand communities have a heavy social component surrounding the imagery and rituals of a brand, and people join because they consider themselves "part of the family" (McAlexander, Schouten, & Koenig, 2002). These communities involve a rational admiration of a brand combined with the emotion of shared experiences. Advertising aimed at brand community formation uses myth, shared images and encodings, such as Apple's cult of "independent thinking" and Harley Davidson's ethos of individualism and brotherhood (Kilambi, Laroche, & Richard, 2013). Bagozzi & Dholakia (2006) observed that brand communities have a "well-developed social identity" and that brand admiration and socializing intermingle. In this study, we elaborate a method of discovering the shared language of a particular community, and hypothesize that this language provokes user engagement. Therefore:

H 2: Words that are **unique to a particular brand community** are related to higher user engagement.

1.4. Case Study 1: User Engagement with Academic Libraries on Facebook

We test our hypotheses first using a case study in which we examine the liking, commenting, and sharing behaviors toward the posts of 100 academic libraries. In the sections that follow, we will discuss our case study background and motivation (1.4.1) and methodology (1.4.2). Results of both case studies are reported and discussed in section 1.6.

1.4.1 Background and Motivation

Academic libraries are an especially appropriate subject for our case study because our interest is primarily in community factors, and as integral components of a university, they come with a ready-made community of users. Academic libraries have a foundational role in university community as educators (Kuh & Gonyea, 2003), curators of authoritative knowledge(Campbell, 2006), and guides for the information age (Budd, 1998).

Like most organizations, academic libraries began their Internet presence with web sites, and Web 1.0 proved to be inadequate for academic library purposes. Usability became an issue early on (Heinrichs, Lim, Lim, & Spangenberg, 2007) and Web portals for libraries were found to be underutilized by students (Y. H. Chen, 2011). An early investigation of Facebook use by academic libraries (Gerolimos, 2011) found that public engagement with post content is low, and argued that digital millennials are not interested in using this mode of interaction with libraries. Of 3513 posts from 20 libraries, 2228 had no feedback at all and 3191 had no comments. The 477 comments in total were mostly from staff. As of 2012, academic libraries were starting to adopt Facebook and Twitter, but were hindered by lack of technical expertise and perceived student interest (Chu & Du, 2013). Nevertheless, over 90% had Facebook pages by 2013 were actively using at least one form of social media to engage with users (Peacemaker, Robinson, & Hurst, 2016). A recent study found that text mining was

an effective approach to understanding how academic libraries can use Twitter to improve services(Al-Daihani & Abrahams, 2016).

Attempts by universities to interact with students on social media have yielded lukewarm responses in the past (Gerolimos, 2011), but it is unclear whether the problem is in the nature of the student-university relationship, the structure of social media technology, or the approaches that have been tried by university entities. Social media users do not avoid commercial-brand Facebook pages, but they tend to interact most strongly with brands that provide an opportunity for self-expression (Wallace, Buil, de Chernatony, et al., 2014).

Several analyses of postings from academic libraries have observed that many have adopted community-building as a strategy. One study conducted a content analysis of 17 Facebook pages to examine the ways they build rapport with students (Phillips, 2011), and found that an emphasis on community, defined as both the university and the surrounding area, was a standard rhetorical practice. A recent analysis of social media postings confirmed that libraries are engaging with the university community with the goal of establishing a connection (Harrison, Burress, Velasquez, & Schreiner, 2017).

It is our goal in this study to evaluate the response from users to these kinds of postings.

1.4.2 Methodology

To test our hypotheses in our first case study, we downloaded a large set of Facebook posts from academic libraries, extracted their basic numerical features, supplemented their features with some derived language attributes, and built a hierarchical generalized linear model to assess the relative effects of all of the features together. The details are described in the sections below.

1.4.2.1 Data Set

Using the Facebook API, 51,760 posts from highest-ranking 100 academic libraries from English-speaking countries as identified by the ShanghaiRanking Consultancy ("Academic Ranking of World Universities,") were downloaded on March 25, 2015. Among the posts, 48,225 contained text in either the message, name, description or caption fields. The placement of these text fields is shown in Figure 1.4.1 below. The earliest post was dated August 9, 2007, and the latest March 25, 2015. In all, we had 150,456 likes and 12,272 comments.

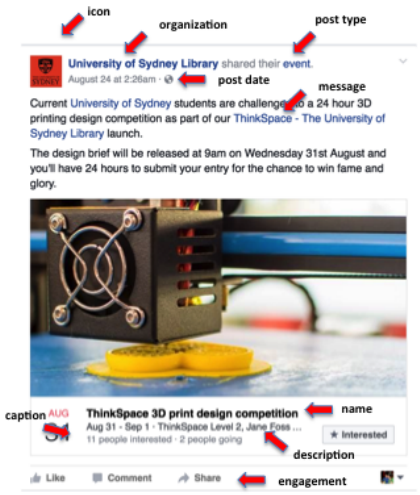
Since our objective was to isolate the effects of community-oriented language, we controlled for several post attributes that would likely also contribute to higher engagement. In all, we extracted 7 features of the posts, and 6 features of the posting organization (the library). Definitions and descriptive statistics of the *post* features follow in section 1.4.2.2, and definitions and descriptive statistics for the *organization* features are discussed in 1.4.2.3. Features addressing *community* (the intersection of post and organization) are introduced and described in 1.4.2.4, and finally 1.4.2.5 summarizes all post, organization, and community features incorporated in our regression models.

1.4.2.2 Post Features

Previous studies have suggested features that are related to higher engagement, including word count/readability (Johnson et al., 2015), media type(Cvijikj & Michahelles, 2013), and posting time of day (Cvijikj & Michahelles, 2013). We therefore collected the following features of each post as controls and computed descriptive statistics for each. A summary of all variables appears below in Table 1.6.1.

P1. Word count. The total number of words in the concatenated message, name, description and caption fields. See Figure 1.4.1 for the placement of each of these fields in a typical post.

Figure 1.4.1. Visible post features



P2. Picture. Whether a picture was included in the post.

P3. Type. Provided by the Facebook API, one of link (45%), photo(24%), status (23%), event(5%), video (2%), music (<1%), or note (<1%). Due to low numbers, the music and note categories were merged into an "other" category.

P4. Hours old. Hours between the posting date and the download date of March 25, 2015.

P5. Readability. This metric uses the Automated Readability Index (Senter & Smith, 1967), which estimates the maximum grade level in the American school system that could

comfortably read the text, and is defined as: 4.71 * (characters/words) + 0.5 * (words/sentences) – 21.43. The mean and median across all posts indicate a college-level reading audience (15-16), but a small number of outliers appear due to the difficulties inherent in automatically identifying sentence boundaries (Negi, Rauthan, & Dhami, 2010). Research indicates that readability has an impact on message appeal (Johnson et al., 2015).

P6. Time of day. Post timestamps were downloaded from Facebook as UTC time (0 offset from GMT) and converted to local time to retrieve accurate time-of-day readings. Our selected libraries operate in 8 different time zones. Cutoff times were as follows: morning (14%)=5 am- noon, afternoon(38%)=noon-6pm, evening(29%)= 6pm-10pm, night(19%)=10pm-5am.

P7. Contains URL. This flag indicates whether the post contains a link to external content. Facebook displays linked content in way that entices users to click, so we control for this variable.

1.4.2.3 Organization features

Organization-level features determine the potential audience of each post. When a user likes a page, the posts from that page appear in the user's News Feed. If a user no longer engages with posts from that organization, the posts stop appearing in the New Feed, so monotonous content across posts has an effect on the exposure of an individual post. If an organization posts too frequently, a user is likely to hide the organization. Organizations that have only recently started posting are likely to have fewer fans and less experience formulating effective messages. We therefore collect the organization-level aggregate features summarized in Table 1.4.3 and described below.

01. *Page likes*. The number of users who liked the library's page, and thus receive the library's posts in their news feed.

02. *Total posts*. Count of all posts of any type.

03. *Days posting*. Number of days between the oldest post and the most recent post.

04. *Number of active users*. The number of distinct users who have engaged with a post. This is an estimate of how evenly-spread activity is among users who may have seen a library's posting activity. A small number indicates "groupy-ism", where most activity is generated by a small number of dedicated fans.

05. *Average monthly posts*. The average number of posts created per month.

06. *Lexical diversity*. Lexical Diversity, which measures the number of different words used across a body of text, serves as a rough estimate of the heterogeneity of a library's postings. Defining each library's corpus as the entirety of its posts concatenated together, we used the Measure of Textual Lexical Diversity (MTLD) to represent the diversity in the language. The MTLD is less sensitive to text length than the standard Type-to-Token Ratio (TTR), which is a legitimate concern given the large range of corpus sizes (see Table 1.4.3). The MTLD reports the average length of a sequence of tokens that maintains a TTR above a certain threshold, the default being 0.72 (McCarthy & Jarvis, 2010). Although the MTLD is still somewhat sensitive to text length, it nevertheless provides a reliable measure of language variety. If a library tends to post similarly-worded updates on a narrow range of topics, this number will be low.

1.4.2.4 Community Features

So that we could test our hypotheses regarding community language, we derived three community fields (C1, C2, and C3) using methods from Natural Language Processing (NLP). C1 and C2 address Hypothesis H1, whereas C3 addresses Hypothesis H2.

C1. *Category: Community Building*

Since our interest is in community, it is helpful to identify posts that were motived by Community Building. A prior content analysis, by Stvilia and Gibradze (Stvilia & Gibradze, 2014), proposes a taxonomy for how academic libraries use Twitter consisting of 7 categories of tweets: Event, Resource, Community Building, Operations Update, Study Support, Q&A, Survey, Staff, and Club. Although this classification scheme was derived from an analysis of Tweets, and has not been methodologically verified on Facebook, we reason that an organization's public relations intent remains constant across all channels, even if the delivery might be tailored to the specific medium. We adapted Stvilia and Gibradze's taxonomy, and trained and tested our own classifier based on this taxonomy, because we were examining a different text medium.

To create a training set, we had 13 undergraduate business students each manually code a random set of 200 posts from a random set of 3,000 from the original set. A random 10% of each coder's posts were checked against an authority tagger for accuracy, and tags from coders who fell below 60% agreement threshold were discarded. All conflicts were resolved by a authority-then-majority-rule scheme. A preliminary round of coding revealed that four categories--Event, Resource, Community Building, and Operations Update--constituted over 90% of all posts, so the remaining categories were dropped in the main coding round. A total

of 5,182 distinct judgments were submitted on the 3,000 posts, with an overall agreement rate of 63% (Cohen's κ=.53, indicating "moderate" agreement according to Landis and Koch (Landis & Koch, 1977). A derived binary field that indicated whether a post was community building or not had an inter-rater agreement of 78% (Cohen's κ=.73, indicating "substantial" agreement (Landis & Koch, 1977)). The derived binary field was set to 1 if the post was Community Building, or 0 if the post was any other category. Of the 2,983 usable labeled posts, 698 (23%) of them were Community Building. We combined this 698 with a random 698 non-Community Building to create a balanced training data set. We used this shuffled balanced training data to create a Stochastic Gradient Descent classifier to classify the posts as community building or other. On an 80/20 train/test split, the classifier achieved an accuracy of 67% (precision=.67, recall=.72, f1=.69) on the target "Yes" class. The most informative Porter-Stemmed (Jones, 1997) terms for the community building posts appear in Listing 1.4.1. Of the 42,022 posts that contained text, 882 were classified as Community Building.

Listing 1.4.1. Most informative 50 terms that differentiate library community building posts.

happi, welcome, congratul, thank, luck, parti, good, fun, winner, everyon, great, go, hit, day, fan, did, final, what, your, give, where, wish, exam, around, season, year, newest, spring, nation, pick, bag, instagram, say, visit, halloween, here, color, friday, hello, love, thing, comment, like, break, valentine, award, freshmen ,support, impact, fall

C2. *Community Orientedness*

As a measure of community-orientedness, we also scored the posts using a relevant subset of the word categories from the Harvard General Inquirer (Stone et al., 1968). The Harvard General Inquirer is a dictionary-based text analysis tool maps words and phrases to 182 semantic and sentiment categories derived from social science theories. Some categories are general, such as *Postiv* (terms with an optimistic tone), but some are quite specific, such as *Aquatic* (terms describing locations in and around water). Although there is some overlap

between some of the broader and the more particular categories, the Harvard GI categories are designed to be reasonably orthogonal. The words are divided into 26 categories and 182 subcategories, and we selected those most relevant to community orientation, as described in Table 1.4.1.

Table 1.4.1. Breakdown of variable C2: selected word categories from the Harvard General Inquirer.

Word Category	Subcategory	Frequent examples from our collection
C2.1: ***basic language universals***	*Affil*: words indicating affiliation or supportiveness	fellowship, partnership, advocacy, community, support, mother, friendly, bride, dictate, colleague, subscribe
C2.2: ***words reflecting the language of a particular institution.***	*Academ:* 153 words relating to academic, intellectual or educational matters, including the names of major fields of study	proctor, teacher, theory, graduate, theoretical, historian, instructor, statistics, college, physicist, typography, professor, education, academic, taught, mathematics
C2.3: ***words referring to roles, collectivities, rituals, and forms of interpersonal relations, often within one of these institutional contexts.***	*Ritual*: 134 words for non-work social rituals	ceremony, funeral, ritual, commendation, inauguration, entertainment, drama, holiday,experience, banquet, marriage, festival, perform, reception, wedding, birthday, debut, formal, commemorate, celebrate, nominate
	SocRel: 577 words for socially-defined interpersonal processes	participate, defend, accompany, attract, monitor, preside, attach, acknowledge, enroll, woo, receive, intervene, regulate
C2.4: ***references to places, locations and routes between them.***	*Social*: 111 words for created locations that typically provide for social interaction and occupy limited space	university , library, own, hospital, cabin, dairy, settlement, community, dwell, garage, palace, college

		suite, site, lawn, airport, headquarters, , seminary, treasury, estate, urban, birthplace, ranch, barn, institute, home, gallery, village, church, hotel, museum, chapel, stadium, heaven, bureau, restaurant, tent, cafe, campus, castle, theatre, capitol, fort, residence, lodge, theater, arena,
C2.5: ***Pronouns reflecting an "I" vs. "we" vs. "you" orientation, as well as names***	*Our:* 6 pronouns referring to the inclusive self ("we", etc.)	let's, us, our, ourselves, us, we

Harvard GI scores consist of the raw count of terms in the document from each category normalized for document length. In general, scores for the Harvard GI categories were exponentially distributed, and therefore the values used in our analysis were logarithms with Laplace smoothing (Manning & Raghavan).

C3*. Brand community centric score.* To quantify the extent to which a post appeals to the sense of shared brand community, we define the community as the library's university and scored each post on its density of university-specific terminology. We obtained a list of these "activation words" by crawling the Wikipedia pages[1] of all of the universities and selecting the

[1] The Wikipedia pages were selected for this analysis because a pilot study indicated that the "About Us" sections of the university webpages were far less differentiated. The sum of the TF*IDF scores for the top 100 terms on the "About us" pages averaged only .511(M=.508, Min=.244, Max=.882, St. Dev=.130), whereas the sum of the top 50 terms on the Wikipedia pages averaged 3.90 (M=3.90, Min=2.65, Max=4.84, St. Dev=.430). Promotional webpages clustered around a set of standard topics

50 terms with the highest TF*IDF. So that words with identical roots but different cases or moods (e.g., "announcement" vs. "announcing") would be considered the same, all terms were stemmed using the Porter Stemmer (Jones, 1997).

The resulting "activation word" list for the University of Virginia, for example, included the terms in Listing 1.4.2:

Listing 1.4.2: Activation words for the University of Virginia.

*uva***,** *virginia***,** *jefferson, charlottesville, college, state, school, rotunda, cavalier, lawn darden, nation, acc, access, alderman, public, new, president, men, one, sullivan, washington, jeffersonian, society, commonwealth, honor, year, thomas, center, poe, ncaa, american, randolph us, monroe, mcintire, unit, hereford, monticello, research, fraternity, law, rector, first, rank*

While most of terms in Listing 1.4.2 are distinctive and highly recognizable to a University of Virginia student, some terms, such as *college*, *state*, and *school* are common anywhere. Some investigation reveals that there are reasons why the UVA pages used these terms statistically more frequently than the others, however. Most of the university's students are enrolled in "The College" as opposed to the other 11 schools at the university, so the terms "college" and "school" carry more semantic suggestion in this context. Also, the university takes pride in being consistently ranked among the best state schools in the country, which explains why the term *state* is so prominent (108 occurrences in the Wikipedia entry).

Each post is assigned a "brand community centric score" calculated as the sum of the TF*IDF weights for all the activation words appearing in the post. Descriptive statistics for this metric appear below. The TF*IDF is used as a weight because larger values imply a higher degree of distinctiveness to the term.

(university(4405), research(2173), students(2939), campus(2049), new(1028), faculty(921), world(849), community(758), academic(607)), whereas Wikipedia presented a fuller picture of the university community, including traditions, symbols, and mascots.

A summary of descriptive statistics and correlations for post and organization features appears below in Tables 1.4.2 and 1.4.3.

Table 1.4.2. Descriptive statistics and correlations for academic library post attributes.

Variable	Mean	Median	StDev	log likes	log comments	log shares	P1	P2	P5	P6	P7	C2.1	C2.2	C2.3	C2.4	C2.5	C2.6	C3
log_likes	0.74	0.69	0.91	1.00														
log_comments	0.13	0	0.37	0.33	1.00													
log_shares	0.23	0	0.53	0.55	0.25	1.00												
P1. word count	51	43	61	-0.04	-0.04	0.02	1.00											
P2. contains picture	0.64	1	0.48	0.25	-0.03	0.20	0.09	1.00										
P5. readability	16.6	15.8	6.9	-0.07	-0.12	-0.01	0.21	0.13	1.00									
P6. hours old	18597	17610	13567	-0.22	0.02	-0.17	-0.07	-0.33	-0.04	1.00								
P7. contains url	0.37	0	0.48	-0.08	-0.10	0.00	0.06	0.08	0.08	-0.08	1.00							
C2.1 log_Affil	0.39	0	0.45	0.04	0.01	0.02	0.07	0.03	-0.03	-0.04	-0.02	1.00						
C2.2 log_Academ	0.59	0.73	0.47	0.05	0.00	0.03	0.03	0.00	0.10	0.00	0.01	0.03	1.00					
C2.3 log_Ritual	0.1	0	0.26	0.01	-0.01	0.02	0.03	0.02	0.00	-0.01	0.01	0.06	0.00	1.00				
C2.4 log_SocRel	0.19	0	0.33	-0.02	-0.01	0.00	0.11	0.00	0.01	-0.04	0.00	0.45	0.05	0.02	1.00			
C2.5 log_Social	0.45	0.5	0.45	0.08	0.03	0.04	0.01	0.00	0.05	0.03	-0.03	0.01	0.67	0.00	0.02	1.00		
C2.6 log_Our	0.2	0	0.35	0.04	0.02	0.01	0.02	0.00	-0.12	-0.05	-0.02	0.47	-0.05	0.01	0.16	-0.06	1.00	
C3. community centric score	0.66	0.21	0.93	0.04	-0.05	0.06	0.34	0.14	0.27	-0.07	0.09	0.06	0.18	0.03	0.04	0.15	-0.02	1.00

Table 1.4.3. Descriptive statistics and correlations for academic library organization attributes.

	Mean	Median	St.Dev	total_likes	total_shares	total comments	O1	O2	O3	O4	O5	O6
total_ likes	1504	950	2303	1.00								
total_shares	339	205	679	0.93	1.00							
total_comments	580	469	389	0.56	0.53	1.00						
O1. page_likes	1769	1198	1860	0.59	0.43	0.47	1.00					
O2. total_posts	517	415	361	0.48	0.48	0.98	0.37	1.00				
O3.days_posting	1904	2013	525	0.17	0.18	0.24	0.24	0.24	1.00			
O4. active_users	674	415	946	0.79	0.67	0.33	0.51	0.21	0.16	1.00		
O5. monthly_posts	8	6	5	0.36	0.36	0.87	0.24	0.90	-0.16	0.12	1.00	

O6. lexical_diversity	90	91	13	-0.04	0.01	-0.08	-0.24	-0.10	0.04	-0.04	-0.12	1.00

1.5. Case Study 2: User Engagement with Restaurants on Facebook

To further test whether community-oriented language enhances engagement, we collected data for a second case study and performed similar feature extraction. We downloaded posts from the 1,625 restaurants in Boulder, Colorado with Facebook pages. Boulder presents an apt case because it is large enough (pop. 97,000) to have several restaurants active on Facebook, but still small enough to have a distinct community identity. Of the 1,625 restaurants, 428 posted at least once, and all available API fields of 174,706 posts were downloaded on May 25, 2016. Posting dates ranged from October 1, 2006 to May 15, 2016.

A previous analysis of 982 Facebook messages from restaurants (Kwok & Yu, 2013) found that the words that provoke the most likes are those that mention menu items, special occasions, or commitment to the community. The words that gained the least attention described marketing campaigns or promotions (such as *winner* or *ticket).* Additionally, this study found that "conversational" messages—general well-wishing posts somewhat related to our "community building" posts—received more attention than those that explicitly referenced products or promotions. A survey of restaurant Facebook fans (Kang, Tang, & Fiore, 2014) revealed that users participate for social-psychological and hedonic benefits, which in turn increase brand trust and commitment, rather than for strictly monetary benefits. We extend this nascent stream by exploring the use of community language among restaurants on Facebook.

Descriptive statistics for restaurants are summarized in Table 1.5.2 and Table 1.5.3. Type (P3) breakdown is link (13%), photo(46%), status (35%), event(3%), video (2%), music (<1%), or note (<1%). Time of day breakdown is morning (46%)=5 am- noon, afternoon(35%)=noon-6pm, evening(13%)= 6pm-10pm, night(6%)=10pm-5am.

We followed the same procedures described in 4.2.4 to train a new classifier for the restaurant posts. We had 13 undergraduate business students manually code a random set of 200 posts from a random set of 3,000 from the original set. The categories for the restaurant posts were the same as for the libraries, with some additions: *Event, Menu, Community Building, Operations Update, Promotion, and Accolades/Press*. A random 10% of each coder's posts were checked against an authority tagger for accuracy, and tags from coders who fell below a 60% agreement threshold were discarded. A total of 5,035 distinct judgments were submitted on the 3,000 posts, with an overall agreement rate of 58% (Cohen's κ=.51, indicating "moderate agreement"). All conflicts were resolved by a authority-then-majority-rule scheme A derived binary field indicating whether a post was community building had an inter-rater agreement of 77% (Cohen's κ=.73, indicating "substantial" agreement(Landis & Koch, 1977)). Of the 3000 labeled posts, 918 of them were Community Building. We combined this 918 with random 918 non-Community Building and used this training data to create a Stochastic Gradient Descent classifier to classify the posts as Community Building or other. On an 80/20 train/test split, the classifier achieved an accuracy of 72% (precision=.7, recall=.9, f1=.8) on the target "Yes" class. The most informative Porter-Stemmed (Jones, 1997) terms are listed in Listing 1.5.1. Of the 160,581 posts that contained a message, 11,475 were classified as Community Building.

Listing 1.5.1. Most informative 50 terms that differentiate community building posts.

happi, thank, love, everyon, family, success, weekend, celebr, friend, big, photo, congratul, wonder, halloween, friday, birthday, excit, time, bronco, air, warm, spring, help, cute, nation, much, nice, we, inspire, share, great, heart, memori, pic, better, thing, easter, sunday, timeline, click, graduat, weather, championship, fun, father, snow, beauty

Table 1.5.1 shows the Harvard GI categories and the frequent terms from the restaurants data set.

Table 1.5.1. Harvard GI terms prevalent in restaurant case study.

Word Category	Subcategory	Frequent examples from our collection
C2.1: ***basic language universals***	*Affil*: words indicating affiliation or supportiveness	love, thank, support, help, like, join, give, offer, share, community, visit, welcome, friends, companion, festival, participate, gift, host, care, hand, congratulations, appreciate, collaborate, cheer, heart, benefit, receive, partner, trust, comfort, inspire, peace, brother, common, associate
C2.2: ***words reflecting the language of a particular institution.***	*Academ:* 153 words relating to academic, intellectual or educational matters, including the names of major fields of study	school, learn, graduate, educate, art, course, study, teach, classics, read, research, senior, wise, student, history, major, college, literature, philosophy, degree, teacher, coast, knowledge, museum, wisdom, professor, lecture, scholar, academic, grade
C2.3: ***words referring to roles, collectivities, rituals, and forms of interpersonal relations, often within one of these institutional contexts.***	*Ritual*: 134 words for non-work social rituals	play, festivities, celebrate, bar, holiday, market, game, visit, dance, experience, meet, birthday, contest, racce, custom, football, concert, magic, entertain, feast, derby, parade, vacation, ceremony, affair, ritual, retreast, jubilee, honeymoon
	SocRel: 577 words for socially-defined interpersonal processes	love, join, help, call, support, care, accept, participate, guide, meet, inspire, accept, challenge, invite, believe, lead, volunteer, entertain, involve, influence, encourage, satisfy, inovle, assist, charm, resist, treasure, attach

C2.4: ***references to places, locations and routes between them.***	*Social*: 111 words for created locations that typically provide for social interaction and occupy limited space	house, restaurant, bar, community, cafe, market, home, shop, school, center, town, hotel, room, stage, base, ranch, theater, office, site, station, palace, gallery, workshop, heaven, church, campus
C2.5: ***Pronouns reflecting an "I" vs. "we" vs. "you" orientation, as well as names***	*Our:* 6 pronouns referring to the inclusive self ("we", etc.)	we, us, our, let's, ourselves

To gather brand community terms (C3), we used a similar method as for the library case, this time contrasting Boulder's Wikipedia page with those of 20 other comparable cities and towns in Colorado. Boulder brand terms appear in Listing 1.5.2.

Listing 1.5.2. Top scoring Boulder terms.

boulder, cruiser, pearl, camera, mindi, flatiron, jump, mork, goat, cubould, pumpkin, paladin, weed, gather, ride, wellb, runner, polar, etown, smokeout, elk, yamagata, brittani, bower, danish, dushanb, eldorado, plung, ear, climb

Descriptive statistics for restaurants appear in Table 1.5.2 and 1.5.3

Table 1.5.2. Descriptive statistics for restaurant post attributes.

Variable	Mean	M	StDev	log likes	log comts	log shares	P1	P2	P5	P6	P7	C2.1	C2.2	C2.3	C2.4	C2.5	C2.6	C3
log_likes	1.12	1.09	1.06	1.00														
log_comments	0.31	0	0.55	0.38	1.00													
log_shares	0.21	0	0.48	0.49	0.26	1.00												
P1. word count	36	25	50	0.04	-0.02	0.08	1.00											
P2. contains picture	0.63	1	0.48	0.36	-0.03	0.21	0.05	1.00										
P5. readability	12.92	11.72	11.4	0.03	-0.09	0.03	0.51	0.05	1.00									
P6. hours old	25632	24330	15148	-0.26	0.15	-0.14	-0.12	-0.48	-0.13	1.00								
P7. contains url	0.09	0	0.28	-0.04	-0.07	0.04	0.11	0.16	0.06	-0.05	1.00							
C2.1 log_Affil	0.52	0.63	0.48	-0.08	-0.05	-0.06	0.40	-0.07	-0.06	0.15	0.16	1.00						
C2.2 log_Academ	0.07	0	0.19	-0.14	-0.03	-0.05	0.36	0.05	0.12	-0.02	0.18	0.24	1.00					
C2.3 log_Ritual	0.15	0	0.28	-0.05	-0.07	0.10	0.24	0.07	0.07	0.05	0.33	0.38	0.38	1.00				
C2.4 log_SocRel	0.22	0	0.33	-0.10	-0.06	-0.03	0.33	-0.14	-0.03	0.05	0.05	0.50	0.18	0.15	1.00			
C2.5 log_Social	0.21	0	0.36	-0.05	-0.02	-0.02	0.24	0.03	0.07	0.14	0.17	0.13	0.17	0.05	-0.02	1.00		
C2.6 log_Our	0.33	0	0.39	-0.03	-0.01	-0.09	0.32	-0.12	-0.13	0.13	-0.01	0.62	0.09	0.12	0.34	0.10	1.00	
C3. community centric score	0.0035	0	0.0069	0.02	-0.05	0.06	0.29	0.10	0.19	-0.10	0.16	0.17	0.13	0.05	0.10	0.03	0.15	1.00

Table 1.5.3. Descriptive statistics for restaurant attributes.

	Mean	Median	St.Dev	total likes	total shares	total comts	O1	O2	O3	O4	O5	O6
total_ likes	2421	450	5210	1.00								
total_shares	216	27	578	0.90	1.00							
total_comments	335	65	823	0.85	0.74	1.00						
O1. page_likes	1365	358	5925	0.55	0.68	0.33	1.00					
O2. total_posts	405	118	742	0.78	0.72	0.86	0.27	1.00				
O3.days_posting	1439	1209	1819	0.16	0.12	0.14	0.04	0.23	1.00			
O4. active_users	758	163	1854	0.89	0.88	0.71	0.80	0.63	0.12	1.00		
O5. monthly_posts	8	5	11	0.72	0.64	0.71	0.27	0.84	0.00	0.58	1.00	
O6. lexical_diversity	91	96	31	0.23	0.20	0.10	0.11	0.18	0.22	0.18	0.14	1.00

1.6. Results

All features described in the previous sections are summarized in Table 1.6.1. To enable comparison of coefficient magnitudes, variables were centered on their mean and divided by their standard deviation. Regressions were conducted using the glmer package in R version 3.1.0.

Table 1.6.1: Summary of Variables

Source	Input Variable	Type
Post	P1. word count	numeric, scaled
	P2. Picture	Indicator
	P3. Type	factor w/7 levels
	P4. hours old	numeric, scaled
	P5. Readability	numeric, scaled
	P6. time of day	factor w/4 levels
	P7. contains url	Indicator
	C1. Category=Community Building	Factor w/2 levels
	C2.1. Affil	numeric, scaled
	C2.2 Academ	numeric, scaled
	C2.3. Ritual	numeric, scaled
	C2.4. SocRel	numeric, scaled
	C2.5. Social	numeric, scaled
	C2.6. Our	numeric, scaled
	C3. brand community centrism	numeric, scaled
Organization	O1. page likes	numeric, scaled
	O2. total posts	numeric, scaled
	O3. days posting	numeric, scaled
	O4. active users	numeric, scaled
	O5. average monthly posts	numeric, scaled
	O6. lexical diversity	numeric, scaled

Since our variables exist at two different levels—that is, post attributes are "nested" within organization attributes—simple regression is insufficient to describe the relationship between the covariates and the independent variables. Models that attempt to merge the levels by aggregating lower level attributes or disaggregating higher level attributes either

ignore particular differences at the lower level or fail to account for shared variance from the higher level (Woltman, Feldstain, MacKay, & Rocchi, 2012). We therefore employ multi-level modeling.

Multi-level models (also known as hierarchical linear models, linear mixed effects models, or nested models) are an extension to generalized linear models to account for data with a hierarchical grouping structure. By hierarchical, we mean that observations are organized within Level 1 groups, Level 1 groups are organized within Level 2 groups, and so on. We need to model this hierarchical structure in order to accurately assess effects that are present at each hierarchical level and are common to all observations in that level's groups. Mixed effects models contain both random and fixed effects. Fixed effects are the traditional regression coefficients, while random effects allow for idiosyncratic variation caused by differences in organizations. This allows us to determine the size of the inter-organization variation. Through these random and fixed effects, mixed effects models can better estimate the covariance structure for grouped data, which results in a model that better meets the regression assumptions.

The principle assumption under this model is that that the odds of a post being engaged with can be expressed as a linear function of the covariates. The interpretation for the *j*th coefficient β_j is that if all else is held constant, a unit increase in the *j*th variable increases the log odds additively by β_j . Note that this is the same as increasing the odds multiplicatively by e^{β_j}. From there it can be seen that if $\beta_j > 0$ then increases in the *j*th predictor variable correspond to higher odds of engagement and if $\beta_j < 0$ then increases in the *j*th predictor variable correspond to lower odds of engagement. Of the original 42,020 academic library posts with text, 42,012 were processed successfully by the Harvard General Inquirer, so only these are used in the analysis. Of the original 174,706 restaurant posts, 173,510 were successfully processed and used.

We created three random intercept logistic regressions for each case study: one each for predicting high likes, high comments, and high shares, where "high" implies the 90th percentile. The link function for these models is the logit function, so that the output is a decision whether the post belongs to the high category given a probability threshold. The model for high likes is detailed in Listing 1.6.1. Similar models were created for high comments and high shares.

Listing 1.6.1. Random intercept model for high likes.

$$\log\left(\frac{prob\_high\_likes_i}{1-prob\_high\_likes_i}\right) = \beta_0 + \beta_1 word\_count_i + \beta_2 type_i + \beta_3 hours\_old + \beta_4 readability_i + \beta_5 time\_of\_day + \beta_6 contains\_url_i + \beta_7 community\_building_i + \beta_8 \log\_Affil_i + \beta_9 \log\_Academ_i + \beta_{10} \log\_Ritual_i + \beta_{11} \log\_Soc\,Rel_i + \beta_{12} \log\_Social_i + \beta_{13} Our_i + \beta_{14} community\_orientation_i + \beta_{15} fans_j + \beta_{16} total\_posts_j + \beta_{17} days\_posting_j + \beta_{18} active\_users_j + \beta_{19} monthly\_av_j + \beta_{20} lexical\_diversity_j + b_{0j} + \varepsilon_i,$$
$$\varepsilon \sim N(0,\sigma^2), b_0 \sim N(0,\sigma^2)$$

For each case study, we shuffled the data set and created an 80/20 train-test split (33,610 training instances for libraries, _ for restaurants). With the training data, we first constructed a hierarchical generalized linear model consisting only of non-experimental post and organizational attributes, with the organization random effect. We then ran a second GLM consisting of only our experimental community features. Finally, we ran a third full model. Diagnostics for each model are reported below. All model diagnostics—precision, recall, F1, and AUC—were calculated on the holdout test split, and were computed at a .5 probability threshold..

Results are reported in Tables 1.6.2 (likes), 1.6.3 (comments), and 1.6.4 (shares). Because regression using large sample sizes has the potential to detect effect sizes that are statistically significant but pragmatically unimportant, we follow (Lin & Lu, 2011) in

providing a practical interpretation of the coefficients in our models. For example: the .609 coefficient on the Community Building variable (C1) for libraries in Table 1.6.2 for the full model indicates that posts that are specifically aimed at building community among its readers have a .609 increase in log-odds over other posts in achieving high likes, or an increase in odds of a factor $e^{.609}$=1.84, meaning 84% increase in odds. Likewise, the .15 coefficient on the scaled C2.5 variable indicates that than increase of 1 standard deviation over the mean in the use of *Social* vocabularies increases the odds of high likes by $e^{.15}$ = 1.16 , or 16%. The interpretation of the coefficient for C3 (.148 for libraries) is heavily determined by the distinctiveness of the community-specific vocabulary. A single highly distinctive term with a TF*IDF of .9 increases the likelihood of high likes by $e^{(.9)(.118)}$ or 11%. Four community-centric terms of average TF*IDF (.07) increase the likelihood of high likes by $e^{(.07)(4)(.118)}$ or about 3.4%. To check for multicollinearity, we report Variance Inflation Factors (VIFs) for all variables in each full model. High VIFs (>10) appear in the library case for likes and shares for the time of day categorical variable (P3). This is likely due to the fact that these are dummy variables (Wissmann & Toutenburg, 2011), where the reference category (night) only constitutes 19% of cases. Since P3 is a control variable, this is not a problem. The high VIF does not cause a bias in the coefficients of our hypothesized variables. In fact, it does not cause a bias in the coefficients of P3, but rather an increase in standard errors, making significance harder to achieve (Allison, 2012). VIFs near the 10 threshold appear for comments in the library case for O2 and O4, but we ignore it for the same reasons.

For likes, the results show that C1 is significant ($p < .0001$) for predicting high likes for both academic libraries and restaurants, although the effect is larger for libraries. Likewise, C3 is a significant predictor of likes for both industries, with larger effect among libraries. C2.5 and C2.6 are the only variable in C2 that are significant and positive across both case studies. When we add the community variables to the model, AUC increases from

.851 to .866, and a Chi-square test indicates that the full model is significantly better, so we conclude that community oriented language has an impact on liking behavior.

For comments, C1 is significant for both case studies, but C3 is only significant for the libraries case. None of the C2 variables are consistently positive and significant across both case studies. This suggests that commenting behavior is not especially effected by community-oriented language. Our models predict high comments with a reasonable degree of accuracy (AUC .728 and .727), but most of the predictive power derives from the P3, P6, O1 and O4.

For shares, C3 is significant for both industries but C1 is only significant for restaurants. Interestingly, several of the General Inquirer word categories are significant for restaurants (C2.3, C2.4, and C2.5) but not for libraries. Only C2.5 and C2.6 are positive and significant across both case studies.

Table 1.6.2. Regression with dependent variable *High Likes*.

Dependent Variable: High Likes								
	Academic Libraries				Restaurants			
Variable	Model 1 Controls	Model 2 Community Vars	Model 3 Full Model	VIF	Model 1 Controls	Model 2 Community Vars	Model 3 Full Model	VIF
(Intercept)	-3.817	-2.40	-3.781		-4.416	-3.335	-4.294	
P1. word_count	-0.182***		-0.223***	1.39	0.168***		0.150***	1.44
P3. type=event	-1.074**		-1.117**	1.05	-0.722***		-0.659***	1.10
P3. type=link	0.459***		0.392***	3.16	1.192***		1.143***	1.46
P3. type=other	-11.200		-11.279	1.00	2.005***		1.972***	1.00
P3. type=photo	1.862***		1.827***	2.91	1.941***		1.904***	1.50
P3. type=video	1.084***		0.991***	1.33	1.216***		1.191***	1.08
P4. hours old	-0.540***		-0.541***	1.07	-0.259***		-0.220***	1.28
P5. readability	-0.059*		-0.075**	1.30	-0.116***		-0.128***	1.38
P6. time of day=afternoon	0.607***		0.578***	15.90	0.323***		0.152**	5.28
P6. time of day=evening	0.589**		0.577**	4.82	0.458***		0.302***	2.72
P6. time of day=morning	0.599***		0.559***	16.61	0.327***		0.146**	5.31
P7. contains url=yes	-0.558***		-0.544***	1.02	-0.446***		-0.457***	1.15
O1. page likes	0.340**		0.347**	1.65	0.405***		0.416***	1.12
O2. total posts	0.647		0.637	9.01	-0.149		-0.106	4.01
O3. days posting	-0.071		-0.075	2.23	0.133		0.128	1.29
O4. active users	0.360***		0.349***	1.36	0.745***		0.740***	1.32
O5. monthly avg posts	-0.890*		-0.869*	8.10	0.051		0.056	4.09
O6. lexical diversity	0.285***		0.291***	1.10	0.169		0.161	1.07
C1. category=Community Building		0.885***	0.609***	1.03		0.582***	0.398***	1.03
C2. 1 Affil		0.042	0.027	1.56		-0.019	-0.005	1.60
C2.2 Academ		-0.037	-0.016	1.95		-0.054***	-0.044***	1.05
C2.3 Ritual		-0.037*	-0.062**	1.01		-0.336***	-0.262***	1.14
C2.4 SocRel		-0.133***	-0.128***	1.24		-0.027*	-0.005	1.22
C2.5 Social		0.221***	0.150***	1.95		0.050***	0.040***	1.07
C2.6 Our		0.108***	0.077***	1.334		0.174***	0.133***	1.42
C3. community centric score		0.03988	0.148	1.394		0.075***	0.095***	1.16
AIC	18534	18335	18364		66521	66521	60924	
BIC	18702	18570	18600		66618	66618	61197	
logLik	-9247	-9139	-9154		-33250	-33250	-30434	
deviance	18494	18279	18308		66501	66501	60868	

Precision	0.611	0.553	0.628		0.546	0.546	0.534	
Recall	0.251	0.107	0.264		0.024	0.024	0.087	
F1	0.346	0.179	0.372		0.046	0.046	0.149	
AUC	**0.851**	**0.771**	**0.866**		**0.792**	**0.792**	**0.840**	
classification metrics are at .5 likelihood threshold								

Full model is better fit, Pr (>ChiSqr) < .0001

***** p < .001**
**** p < .001**
*** p < .05**

** **P2. Picture** removed due to high VIF.
** **P3. Type**: Base level="status". video and note levels were merged into "other."
** **P6. Time of day:** Base level="night."

Table 1.6.3. Regression with dependent variable *High Comments*.

Dependent Variable: High Comments								
	Academic Libraries				Restaurants			
Variable	Model 1 Controls	Model 2 Community Vars	Model 3 Full Model	VIF	Model 1 Controls	Model 2 Community Vars	Model 3 Full Model	VIF
(Intercept)	-1.258	-2.052	-1.243		-2.523	-2.443734	-2.532	
P1. word_count	-0.007***		-0.011	1.28	0.199***		0.197***	1.44
P3. type=event	-1.442***		-1.440***	1.11	-1.655***		-1.609***	1.10
P3. type=link	-0.614***		-0.622***	1.92	-0.030		-0.015	1.46
P3. type=other	0.088		-0.090	1.00	0.627		0.691	1.00
P3. type=photo	0.201***		0.187***	1.72	0.589***		0.604***	1.50
P3. type=video	-0.058		-0.077	1.12	0.418***		0.446***	1.08
P4. hours old	0.083***		0.080***	1.16	0.399***		0.399***	1.28
P5. readability	-0.142***		-0.152	1.33	-0.249***		-0.195***	1.38
P6. time of day=afternoon	0.411***		-0.425***	8.18	0.217***		0.204***	4.28
P6. time of day=evening	-0.292**		-0.300	2.82	0.233***		0.227***	2.72
P6. time of day=morning	-0.475***		-0.492***	8.54	0.167***		0.153***	4.31
P7. contains url=yes	-0.495***		-0.488***	1.02	-0.419***		-0.412***	1.15
01. page likes	0.331***		0.338***	1.12	0.247***		0.252***	1.12
O2. total posts	0.353		0.336	10.01	-0.173		-0.182***	4.01
O3. days posting	-0.112		-0.111	1.29	-0.002		-0.002***	1.29
O4. active users	0.209***		0.204***	1.32	0.544***		0.544***	1.32
O5. monthly avg posts	-0.605**		-0.582**	10.70	0.035		0.024***	4.09
O6. lexical diversity	0.244***		0.250***	1.07	0.046		0.050***	1.07
C1. category=Community Building		0.368***	0.065	1.03		0.179***	0.080*	1.03
C2. 1 Affil		0.049*	0.041	1.60		-0.094***	-0.083***	1.60
C2.2 Academ		-0.052*	-0.038	1.05		-0.081***	-0.064***	1.05
C2.3 Ritual		-0.039*	-0.036*	1.14		0.130***	0.085***	1.14
C2.4 SocRel		-0.051**	-0.044*	1.22		-0.023*	-0.011	1.22
C2.5 Social		0.178***	0.120***	1.07		0.005	0.014	1.0

								7
C2.6 Our		0.031	-0.006	1.42		0.123***	0.115***	1.42
C3. community centric score		0.182***	0.030	1.16		-0.036***	0.020	1.16
AIC	23966	24774	23934		88546	90837	88316	
BIC	24134	24858	24170		88741	90934	88589	
logLik	-11963	-1237	-11939		-44253	-45408	-44130	
deviance	23925	24753	23878		88506	90817	88260	
Precision	0.522	0.514	0.529		0.476	0.578	0.481	
Recall	0.031	0.016	0.032		0.014	0.006	0.014	
F1	0.058	0.031	0.061		0.027	0.012	0.028	
AUC	**0.725**	**0.684**	**0.728**		**0.726**	**0.694**	**0.727**	
classification metrics are at .5 likelihood threshold								

Full model is better fit, Pr (>ChiSqr) < .0001

***** p < .001**
**** p < .001**
*** p < .05**

***** Indicates significance at the 99.9% level
**** P2. Picture** removed due to high VIF.
**** P3. Type**: Base level="status". video and note levels were merged into "other."
**** P6. Time of day:** Base level="night."

Table 1.6.4. Regression with dependent variable *High Shares*.

Dependent Variable: High Shares								
	Academic Libraries				Restaurants			
Variable	Model 1 Controls	Model 2 Community Vars	Model 3 Full Model	VIF	Model 1 Controls	Model 2 Community Vars	Model 3 Full Model	VIF
(Intercept)	-3.365	-1.510	-3.322		-2.990	-1.961	-2.918***	
P1. word_count	0.036*		0.024	1.40	0.241***		0.201***	1.44
P3. type=event	0.125		0.085	1.05	-1.412***		-1.351***	1.14
P3. type=link	0.958***		0.929***	3.19	1.053***		1.011***	1.97
P3. type=other	-10.969		-11.020	1.00	2.019***		2.002***	1.00
P3. type=photo	1.504***		1.487***	2.94	1.372***		1.364***	1.88
P3. type=video	1.314***		1.292***	1.35	1.344***		1.329***	1.16
P4. hours old	-0.540***		-0.537***	1.13	-0.190***		-0.159***	1.13
P5. readability	-0.024		-0.033	1.25	-0.016		-0.018	1.25
P6. time of day=afternoon	0.863***		0.835***	19.09	0.491***		0.396***	6.09
P6. time of day=evening	0.679***		0.667***	3.41	0.405***		0.321***	3.41
P6. time of day=morning	0.989***		0.959***	18.14	0.506***		0.406***	6.14
P7. contains url=yes	-0.202***		-0.200***	1.20	-0.007		-0.021	1.20
01. page likes	0.008		0.009	1.12	0.248***		0.257**	1.12
O2. total posts	1.127***		1.112***	3.92	0.090		0.125	3.92
O3. days posting	-0.118		-0.116	1.27	0.015		0.012	1.27
O4. active users	0.247**		0.244**	1.34	0.394***		0.390***	1.34
O5. monthly avg posts	-1.121***		-1.096***	3.98	-0.105		-0.102	3.98
O6. lexical diversity	0.246***		0.248***	1.08	0.130*		0.111	1.08
C1. category=Community Building		0.129	0.043	1.03		0.192***	0.084***	1.03
C2. 1 Affil		0.014	0.000	1.68		-0.001	0.004***	1.68
C2.2 Academ		0.033	0.040	1.04		0.016*	0.008***	1.04
C2.3 Ritual		0.018	0.000	1.04		-0.267***	-0.196***	1.04
C2.4 SocRel		-0.004	-0.012	1.22		0.036***	0.047***	1.22
C2.5 Social		0.051**	0.042*	1.07		0.051***	0.030***	1.07
C2.6 Our		0.079***	0.057***	1.47		0.130***	0.104***	1.47
C3. community centric score		0.087***	0.056**	1.15		0.131***	0.081***	1.15
AIC	29056	31707	29019		103051	109238	102421	
BIC	29225	31791	29255		103246	109335	102693	
logLik	-14508	-15843	-14481		-51505	-54609	-51182	

deviance	29016	31687	28963		103011	109218	102365	
Precision	0.577	0.603	0.564		0.591	0.583	0.577	
Recall	0.197	0.140	0.200		0.123	0.071	0.130	
F1	0.294	0.228	0.295		0.203	0.127	0.211	
AUC	**0.790**	**0.697**	**0.792**		**0.777**	**0.722**	**0.781**	
classification metrics are at .5 likelihood threshold								

Full model is better fit, Pr (>ChiSqr) < .0001
***** p < .001**
**** p < .001**
*** p < .05**

***** Indicates significance at the 99.9% level.
**** P2. Picture** removed due to high VIF.
**** P3. Type**: Base level="status". video and note levels were merged into "other."
**** P6. Time of day:** Base level="night."

1.7. Discussion

We observe first that the addition of the community features improves the predictive model in the case of likes (AUC rises from .851 to .866 for libraries, .792 to .840 for restaurants) but not substantially for comments or shares. Although Chi-square difference tests indicate in all cases that the community features improve prediction, the practical magnitude of the difference is negligible except in the case of likes. All three of our operationalizations of community-oriented language (C1, C2, and C3) have a significant positive impact on likes across both industries, but only C3 drives shares across both industries, and only C1 affects comments consistently across both industries. While our community features have a significant effect on likes, our selected set of control variables dominate the predictability of high likes. The largest significant positive predictors of likes across both industries are O2 (page likes/fans), P3 (photos or videos), and P6 (time of day—afternoon is better). The largest significant positive predictors of comments across both industries are O2 (page likes/fans) and P3 (photos). Community building (C1) drives comments among restaurants, but not libraries. Sharing is driven by P3 (photo or video) and time of day (P6) for both

industries, consistent with (Kwok & Yu, 2013). Potential audience (O1) heavily affects sharing among restaurants, but not libraries.

The difference between liking vs. comments and shares may be a product of how unanticipated a message is. A community building message is by its nature familiar, and although it might be satisfying, it will not necessary be shared. Messages of commonality are not controversial, and therefore are not likely to generate comments.

Because C1 and C2 have a significant positive relationship with likes across both case studies, we conclude that H1 is supported and generalizable across industries, but only consistently for likes.

C3 is positive and significant across both case studies for likes, so we conclude that H2 is also supported, but only consistently for likes. The vocabulary, images, and symbols that define the brand either of the university or of the town trigger a liking response from readers, consistent with (Angela Hausman, Kabadayi, & Price, 2014). However, brand-centric language does not provoke comments or shares. Fans appear to gain their belongingness satisfaction through a minimum of effort.

1.8. Limitations and Future Work

This study has some limitations that are an inescapable part of working with publicly-available data, such as that available through the Facebook API. It is difficult to assess exactly how many people have seen a given post. Even the potential audience (in page likes) and the "reach" metric only give an approximation of the number of views: we also need information about individual users' habits, such as log-in frequency and number of friends, since it is easy for a post to be buried in a long news feed.

Another difficulty associated with this kind of analysis is the sheer volume of data. In dealing with word category, for example, we confined ourselves to a small number of

categories in the interest of having regressions that converged in a reasonable amount of time, but we would like to consider a greater sample of the 182 available word categories. We would also like to consider the effect of some other variables, such as grammatical structures or image content.

In our regressions, we control for time by introducing the "hours old" covariate (P4). Although most likes occur on the first day or two after a post appears, it is possible that a post may accumulate likes over time as a wider audience explores an organization's older content. One possible enhancement to the current study would be to model Facebook reactions as time series and observe the rate at which different content accumulates likes, comments, and shares. Previous studies have seen strong predictive performance by modeling social media data as time series (Asur & Huberman, 2010; L. T. Nguyen, Wu, Chan, Peng, & Zhang, 2012).

One potentially fruitful follow-up to this study would be an experiment in which organizations identify their activation terms and test how their usage effects the number of likes their posts receive. Our case study finds that organizations are already making an effort build communities around a brand (Kang et al., 2014; Stvilia & Gibradze, 2014), and so they must perceive the rewards, but a study in which messages can be modified with specific manipulations would help quantify this effect and further validate the patterns we have observed.

Ultimately, organizations want to engage with their clientele to promote their products or services, and it is difficult to judge the value of online engagement with respect to that larger goal. Some previous research questions whether Facebook engagement truly captures brand loyalty, since even enthusiasts tend not to participate in Facebook brand communities (Lin & Lu, 2011). Testing whether increases in engagement are in fact helping libraries meet their larger organizational goals requires field research. Survey-based studies that go beyond text analytics and investigate operational improvements could help validate

the effectiveness of the library's online presence. Target metrics could include growth in event attendance and resource utilization, or reduction in the number of routine questions handled by staff members. To our knowledge, the effectiveness of Facebook messages in driving restaurant business has not been investigated, although much attention has been given the effects of online reviews on restaurants (Z. Zhang, Ye, Law, & Li, 2010).

1.9. Conclusions and Implications for Research and Practice

Predicting the response that a Facebook post will have is difficult because of the variety of factors involved. This paper presents an approach to isolating some of those factors and assessing their relative contributions, focusing especially on those pertaining to *communities* in general, and a *brand community* in particular. We find that language that appeals to the social belongingness need has a small but significant impact on engagement. We also find that each community has a set of *activation words* unique to its brand, and that these terms likewise contribute significantly to engagement. Although this community-oriented language plays a role, we find that it is joined by a variety of other factors that play a comparable or larger role, such as multimedia, the time of day, the variety of material discussed, the posting experience of the organization, and the percentage of the fan base that regularly engages.

Our findings have a number of implications for research and practice. Firstly, researchers should be cognizant of the important impact of implicit, derived features – specifically, community-specific activation words – on response variables. Careful consideration of supplementary features, such as these, which are not explicit in the source data, is essential for improving model performance, particularly in predicting engagement, as illustrated here. Further attention needs to be paid to constructing and evaluating novel derived features which may be helpful to model performance. Derived features are innovative mergers of post data with additional data sources, such as Wikipedia entries and

corporate websites (semi-structured text), or word category dictionaries and custom taxonomies (structured data). Established social science and marketing science theories can be a source of valuable derived features. In our study, *community-specific words* derived from Wikipedia entries, and generic *community-oriented terms* from the Harvard General Inquirer dictionary, provided a fertile source of activation words. The need for belongingness and the suggestions of Brand Community Brand, were the social science and marketing science theories that inspired these derived features in our study. Practitioners, likewise, should be aware of the importance of such derived features. Specifically, practitioners may find that incorporation of community-specific activation words may activate and engage their user base, supplementing the stimulating effects of multimedia, appropriate post timing, and concise and varied post content, on user engagement.

CHAPTER 2

Text Analytics for C2C Interactions: Application in Pharmacovigilance

Abstract

Millions of patients are hospitalized each year because of Adverse Drug Reactions (ADR), and researchers are seeking ways to promptly discover effects that had remained hidden before the drug was approved and marketed. Electronic health records (EHRs), published biomedical research, and clinical reports have been recognized as rich text-based sources of early warning signals, and recent research has started investigating the potential of social media as well. Recent studies have validated the efficacy of text mining approaches to pharmacovigilance in social media. In this study, we test whether a text mining methodology that has proven successful in identifying hazards in consumer products in online reviews can be applied to the discovery of ADRs. Since lexicon generation is a key step to this methodology as well as several others in pharmacovigilance, we also test two methods of lexicon creation: one driven by statistical term prevalence, the other by manual curation by individuals and groups. We find that our methodology is effective at differentiating online reviews with and without ADRs. We also find that the top quantile of manually-curated lists outperform statistical term prevalence (supervised machine learning) on a variety of ADR classification metrics. Additionally, we find that groups outperform individuals, and that there is no correlation between list size and classification performance.

Keywords: Pharmacovigilance, Text Mining, Natural Language Processing, Information Retrieval, Machine Learning

2.1 Introduction

Adverse Drug Reactions (ADRs), "noxious" and "unintended" physical responses to a normal dose of a medication (WHO, 1972), pose a serious public health threat. A meta-analysis in 1998 determined that in a single year, over 2 million patients are hospitalized because of ADRs, resulting in approximately106,000 fatalities (Lazarou, Pomeranz, & Corey, 1998). Timely and accurate post-marketing discovery of ADRs has become a public health priority (Harpaz et al., 2012; Sarker et al., 2015), especially in an environment of increasing globalization and free trade (Vijay Kumar, 2013).

Pharmacovigilance (PV) is the "detection, assessment, understanding, and prevention" of adverse reactions to medication (WHO). In the traditional approach to PV, Phase I-III clinical trials are conducted before regulator approval. After approval, regulatory agencies additionally require Phase IV clinical trials, but these are limited in scope, and may not be adequate to fully establish the safety of the drug (Harpaz et al., 2012).

While drug companies are required by law to report ADRs observed during clinical trials, post-marketing surveillance occurs in the US largely through voluntary reporting by healthcare professionals in Spontaneous Reporting Systems (SRSs). Widely-used SRSs include the US Food and Drug Administration's Adverse Event Reporting System (FAERS) and VigiBase, maintained by the WHO (Harpaz et al., 2012). SRSs have several shortcomings, however. Because reporting is voluntary, reports may be censored or omitted (Edwards & Lindquist, 2011). The prevalence of ADRs in SRSs has been shown to be underestimated (X. Liu & Chen, 2013). The under-reporting rate is considerable, especially for recently-marketed drugs (Alvarez-Requejo et al., 1998).

Because SRSs likely contain incomplete information, health researchers have been exploring alternative text-based data sources for ADRs. Sources for text mining approaches

to ADR surveillance have included electronic health records (EHRs) (Aramaki et al., 2010; Friedman, 2009; X. Wang, Hripcsak, Markatou, & Friedman, 2009), biomedical literature (Tafti et al., 2017), and clinical reports (Gurulingappa, Fluck, Hofmann-Apitius, & Toldo, 2011; Gurulingappa, Mateen‐Rajpu, & Toldo, 2012). EHRs and clinical reports present challenges because they require substantial preprocessing and contain symptoms and diagnoses, and not necessarily ADRs (Harpaz et al., 2012). Also, there are access issues across healthcare providers due to privacy concerns.

Recently, social media have proven a rich resource for PV (Abbasi et al., 2014; Harpaz et al., 2012; Leaman et al., 2010; X. Liu & Chen, 2013; Nikfarjam & Gonzalez, 2011; Sarker & Gonzalez, 2015). Social media are especially promising because represent uncensored personal accounts, and thus depict a more authentic "patient perspective" (Bhattacharya et al., 2017). Timeliness is another advantage, to the extent that some have proposed an "event-driven" epidemiology with social media (Hartley, 2014). However, research into ADR detection in social media is still in its "infancy" (Sarker & Gonzalez, 2015).

That social media may provide an accurate view of ADRs is somewhat evidenced by the fact that the prevalence of ADR mentions in social media is approximately the same as that reported by official sources (T. Nguyen et al., 2017). Social media may also be a quality source of information because there is a large volume of easily-accessible patient-reported experiences available on sites such as Ask a Patient, DailyStrength, Yahoo Health and Wellness, and PatientsLikeMe (Chou, Hunt, Beckjord, Moser, & Hesse, 2009). Social media have been shown to beneficially augment other data sources for discovering ADRs (Sarker & Gonzalez, 2015) and drug-drug-interaction (DDI) (Vilar, Friedman, & Hripcsak, 2017). The utility of social media for PV is further established by the fact that regulatory groups are publishing guidelines for managing ADR complaints in social media (Sarker et al., 2015).

Several methodologies for applying text mining methods to PV have been tested, including lexicon-based approaches (Leaman et al., 2010), controlled vocabularies (Benton et al., 2011), association rules (Nikfarjam & Gonzalez, 2011), sentiment analysis with topic extraction on heterogeneous corpora (Sarker & Gonzalez, 2015), and feature selection with ensembles of classifiers (J. Liu, Zhao, & Zhang, 2016).

This study extends this stream by testing the application of the smoke-term based approaches of (Abrahams et al., 2015; Abrahams et al., 2012; Adams et al., 2017; Goldberg & Abrahams, 2017; Law et al., 2017; Winkler et al., 2016). In this method, 1,2,and 3-word phrases that are significantly more common in the text of interest are found using supervised machine learning on hand-tagged data. Each term is assigned a weight based on a prevalence metric (Fan, Gordon, & Pathak, 2005), and new documents are scored based on the weighted sum of the "smoke terms" present. This method has proven effective in finding safety hazards in social media postings about a variety consumer products.

In addition to testing the effectiveness of the smoke term approach in finding ADRs, we also tested how well this method works against manually-curated term lists and lists created by group brainstorming. In our experiment, we had 72 non-experts study a corpus of online reviews from three categories of over the counter (OTC) medicines on Amazon.com. Then, using whatever resources available—online health databases, dictionaries of side effects, etc.—participants generated a list of 1,2,and 3-word terms that might function as effective search terms to find reviews with ADRs. Participants were also asked to assign real-valued weights from 0 to 1 to each term based on how important they judged that term to be. We then had these individuals work in teams of 2,3, and 4 to create a group list. Finding an effective set of search terms in OTC medicine reviews is challenging for a layperson because are no obvious keywords that are guaranteed to produce results.

There are innumerable possible drug reactions and side effects, expressed in language ranging from clinically precise to emotional and colloquial.

We tested the ADR classification performance of the smoke term list (the "computer list"), the individual lists, and the group lists on a holdout set of 6000 reviews, only 212 of which contained ADRs (3.5%). We found that the top quartile of human-curated lists significantly outperformed the computer-generated list on a variety of classification metrics, achieving an AUC of over 70% ((Hanley & McNeil, 1982) see section 2 for a definition). Additionally, we found that groups outperformed individuals. Thus our human lists generally outperformed those of the computer, even achieving performance comparable to that of lists assembled by experts (Leaman et al., 2010) and of systems with more elaborate algorithms (Nikfarjam & Gonzalez, 2011).

We focused our investigation around three research questions: 1. Can ADR's be discovered using a methodology that designed to find product safety defects? 2. Do manually-curated lexicons provide better search terms than those generated using smoke-term prevalence scores? 3. Is medical expertise a necessary prerequisite for quality lexicon creation?

To our knowledge, no research into ADR discovery in social media has used online reviews. We contend that reviews are a good source of information because of the intentionality behind a review. In health forums, it is not always clear whether a patient is actually taking a drug or is simply giving advice (Leaman et al., 2010). In a review, the understanding is that the author has used a product and is providing a frank evaluation to another person.

This chapter continues our examination of how C2C messages provide businesses with informational value for co-creation (Brodie et al., 2011; Libai et al., 2010; Sawhney, Verona, & Prandelli, 2005). Finding reliable and useful information in C2C is difficult due to

the volume of text generated every day, and automated methods are challenged by the peculiarities of expression and haphazard spelling (Pimpalkhute, Patki, Nikfarjam, & Gonzalez, 2014). PV presents a case study of how businesses or regulators can apply text mining methods to finding useful and important information in C2C online communication.

2.2 Related Work

A comprehensive review of recent innovations in ADR detection in social media can be found in (Sarker et al., 2015). In this work, Sarker reviews 22 studies and lays out a systematic pathway to ADR monitoring using social media, making the critical observation that lexicon-based methods remain the most popular. A more recent summary can be found in (J. Liu, Zhao, & Wang, 2017). Rather than reiterating the content of these two reviews, in this section, we highlight some of the text mining approaches that are relevant to our study because they entail lexicon construction.

Classifiers, programs that determine whether a text document is a member of a certain target class, are evaluated on a standard set of metrics, which we define here. *Precision* is the proportion of true positives to the sum of true positives and false positives (precision = TP/(TP+FP)). *Recall* is the proportion of true positives to the sum of true positives and false negatives (recall = TP/(TP+FN)). Since precision decreases as recall increase, a helpful metric that combines recall and precision is *F1*, which is calculated as 2*precision*recall/(precision+recall). When documents are assigned a probability of belonging to the target class, as they are in our case, precision and recall will vary according to the cutoff threshold for classifying in the target class. For example, a cutoff of .1 (10% probability) might yield a low precision but a very high recall, because there would be so many false positives, whereas a cutoff of .9 might show a high precision but a low recall, due to false negatives. A metric that describes a more comprehensive assessment at all cutoff

thresholds is AUC (Bradley, 1997). AUC is the area under the ROC curve, a plot of true positives (y axis) against false positives (x axis). Random guessing would yield a ROC curve that is a straight diagonal line and an AUC of .5. Any AUC above .5 is an improvement over a baseline of guessing.

The typical task for PV studies is to locate drug-reaction pairs within unstructured user-generated content (UGC). This consists of finding drug mentions, which is made especially difficult by misspellings (Pimpalkhute et al., 2014; Sarker et al., 2015), and then pairing the mention with a reaction, often expressed in idioms nowhere near standard medical terminology (Leaman et al., 2010; T. Miller, Leroy, Chatterjee, Fan, & Thoms, 2007). We note that the first part of this task is rendered unnecessary in our study by the fact that Amazon.com reviews are already grouped under product names; therefore, our sole challenge is to identify symptoms that are cause for concern.

Recent text analytic methods for finding ADRs in social media include association rule mining (Harpaz et al., 2012; Nikfarjam & Gonzalez, 2011), lexicon-based concept extraction (O'Connor et al., 2014), and synset expansions combined with "change phrases" (Patki et al., 2014). Error analyses across several experiments indicate that recurring reasons for false negatives include misspellings and inconsistent descriptions of symptoms (J. Liu et al., 2016). Most methods used supervised classification (Bian, Topaloglu, & Yu, 2012; Ginn et al., 2014; K. Jiang & Zheng, 2013; Sarker & Gonzalez, 2015; M. Yang, Wang, & Kiang, 2013).

One of the first attempts to apply text mining methods to finding ADRs in social media was (Leaman et al., 2010). Examining text from DailyStrength, a platform for patient support groups, this study confirmed that UGC contained extractable and useful information. This method made use of four sources of term features: 1) the UMLS (United Medical Language System) Metathesaurus COSTART (Coding Symbols for a Thesaurus of Adverse Reaction

Terms) developed by the FDA, which contains 3787 concepts, 2) the SIDER side effect resource, consisting of 888 drugs linked with 1450 adverse reaction terms, 3) the Canada Drug Adverse Reaction Database, with 10,192 drugs with 3279 adverse reactions, and 4) a manual set of colloquial terms from DailyStrength. (We note that participants in our study reported using 1-3 as well). Using these terms, their classifier was able to achieve 78% precision, 70% recall, 74% F1. [2]

Benton et al. (2011) used a similar lexicon-based approach to finding ADRs on a set of breast cancer message boards. In addition to painstaking anonymization, this work introduced a novel step of mapping common expressions onto a controlled vocabulary. While they were unable to achieve high recall, they found frequent instances of the most commonly-occurring ADRs.

Another study tested association rule mining (Hipp, Güntzer, & Nakhaeizadeh, 2000) to a data set from DailyStrength (Nikfarjam & Gonzalez, 2011). Like (Leaman et al., 2010), this work affirmed that despite the vagaries of informal communication of the Internet, comprehensible patterns exist and can be leveraged for ADR discovery. In this study, two experts annotated a random sample of 3600 posting and found 1260 mentions of adverse effects. With this tagged data, they were able to build a classifier that had 70% accuracy, 66% recall, and 68% F1. What is remarkable here is that the incidence of ADRs in the data set was high (35%), substantially more than the estimated 10-25% rate observed elsewhere (T. Nguyen et al., 2017) or the 3.5% rate observed in our own data set.

Sarker & Gonzalez (2015) introduced two important innovations: advanced NLP methods and portable multi-corpus training. They assembled a two-corpus data set consisting of 2972 clinical reports from Medline's Adverse Drug Event database and social media

[2] Leaman et al. measured their classifier at a single cut-off threshold only, and so reported only precision and recall at that threshold, and did not report Area Under Curve (AUC) which is a more comprehensive indicator of precision and recall at multiple thresholds.

postings from Twitter and DailyStrength. They then compiled a lexicon of search terms similar to (Leaman et al., 2010), computed sentiment polarity using a lexicon provided by (Guerini, Gatti, & Turchi, 2013), and extracted topics. The result was an ADR classifier that achieved high levels of precision and accuracy (ADR F1 up to 81%).

This approach was extended in (J. Liu et al., 2016), who added Information Gain feature selection and an ensemble of classifiers using a variety of voting methods and were able to achieve AUC values of up to 78%. They used three data sets, two consisting of postings from health-related forums and another from Twitter. Their feature set included lexical, syntactic, and semantic features, including part of speech tags and syntax trees. They later extended this work to incorporate semi-supervised methods in concert with ensemble learning in (J. Liu et al., 2017). These methods raised the highest AUC to 81%.

A separate thread of research that we contend is relevant to ADR detection is the recent investigation into consumer product safety surveillance in social media. Research into finding safety concerns in online reviews has found that a model consisting of a filter n-gram term list performs well in both precision and recall (Abrahams et al., 2012; Winkler et al., 2016). These models consist of n-grams (unigrams, bigrams, and trigrams) called "smoke terms" weighted according to a prevalence metric (Fan et al., 2005). Successful models have been constructed for finding unsafe defects in automobiles (Abrahams et al., 2012), toys (Winkler et al., 2016), dishwashers (Law et al., 2017), and joint pain treatments (Adams et al., 2017). We therefore tested the smoke word generation method for ADR detection, and test edits performance against the manual term curation seen in other studies.

Our approach is distinct in that we allow individual non-experts to devise their own system for creating lexicons based on exposure to narratives of typical drug reactions. Also, by using Amazon.com, in which reviews are grouped by product, creating drug-reaction

association rules was unnecessary: we had to simply identify them. To our knowledge, no text mining studies of ADRs has specifically investigated Amazon.com reviews.

2.3. Data and Methods

Our methodology consisted of "training" the both human participants and the computer on a balanced data set of 264 Amazon.com reviews, and then testing them on an unseen holdout set of 6000 reviews. Detailed descriptions of the two data sets follow. Steps are pictured in Figure 2.3.1.

Training Data

43,893 reviews were collected from the "Pain Relief", "Allergy & Sinus", and "Digestion & Nausea" categories (see Appendix 2.A for a detailed list of subcategories) in the Amazon.com data set compiled in (McAuley, Targett, Shi, & Van Den Hengel, 2015). A random subset of 5000 was collected for tagging for ADRs by 8 undergraduate university students and 1 authority researcher. A median of 500 reviews were tagged by each tagger. In this round, 4565 tags were collected on 3165 unique reviews. There were 411 cases in which the authority and a tagger tagged the same review, and 371 agreements and 40 disagreements (90.3% agreement), for a Cohen's kappa=.805 (J. Cohen, 1968), indicating "substantial" agreement (Landis & Koch, 1977), almost "near perfect" agreement. A total of 92 ADRs (2.9%) were identified.

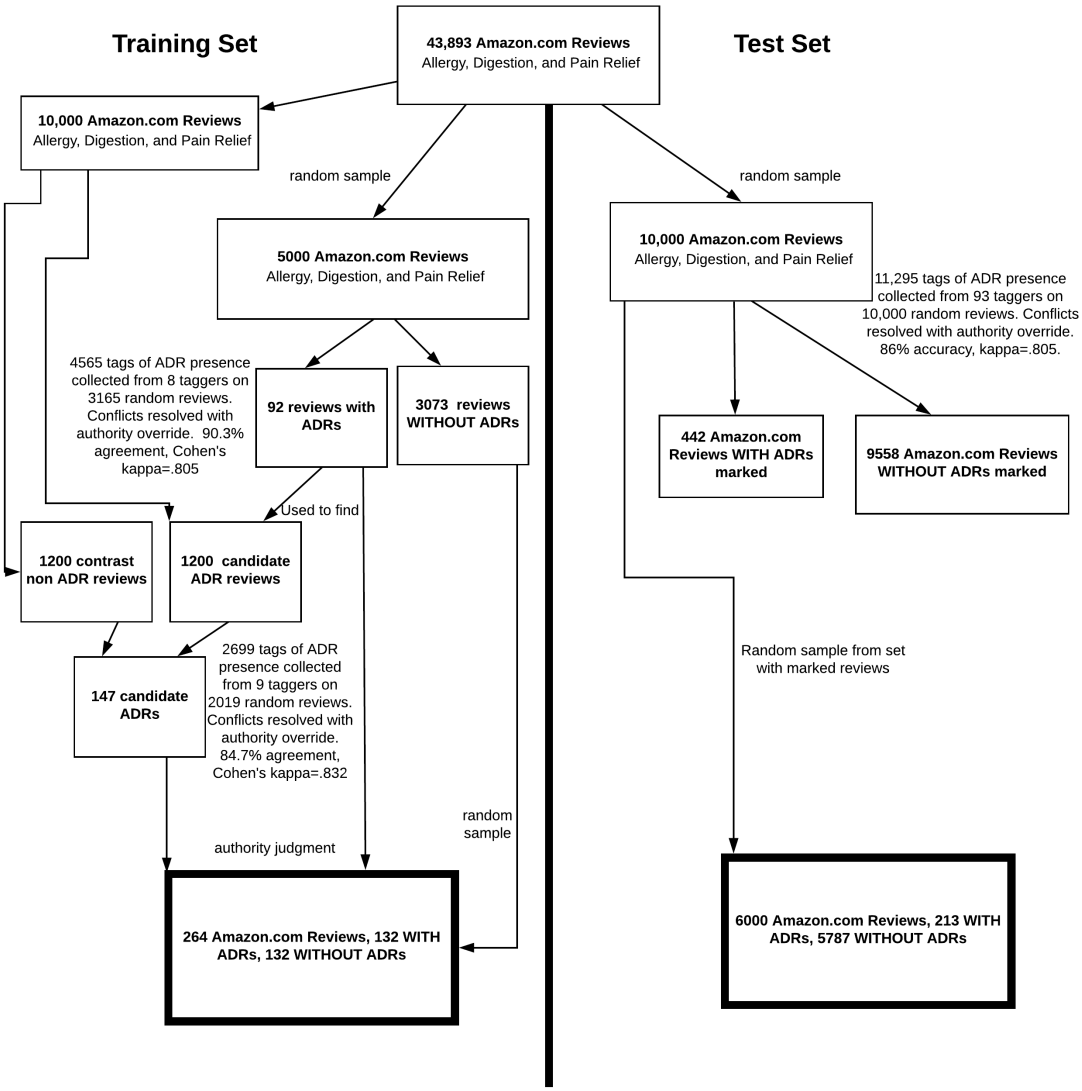
Using the method described in (Law et al., 2017), the 10,000 unseen reviews were scored using each of six methods (Harvard GI, AFINN, ANEW, smoke unigrams, smoke bigrams, smoke trigrams). The top-scoring 200 reviews from each method were collected, making a set of 1200 ADR candidate reviews. A contrast set of 1200 was created using the bottom-scoring 200 reviews for each method. This set of 2400 reviews was then tagged by 9 undergraduate university students, and accuracy was checked as before. This round

produced 2699 tags of 2019 distinct reviews. There were 333 inter-rater agreements and 60 disagreements, meaning 84.7% agreement, or Cohen's kappa(J. Cohen, 1968) .832, indicating "near perfect" agreement(Landis & Koch, 1977). This tagging round identified an additional 147 suspected ADRs. Of the 92+147 ADRs identified altogether, 132 were verified by the authority tagger as ADRs. 132 random reviews were sampled from remainder of initial set of 5,000. This created a balanced set of 264 reviews.

Drugs have known and harmless side effects, and simply finding instances of these in social media is not valuable. However, participants were encouraged to mark any complaint that seemed out of the range of normal side effects. It was important for taggers to be highly sensitive, because any unusual health incident could provide an initial warning of a serious ADR, and false positives are easy to drop in follow-up investigation.

To create the holdout test set, 10,000 reviews that were *not* in the training set were collected from the "Pain Relief", "Allergy & Sinus", and "Digestion & Nausea" categories (see Appendix 2.A for a detailed list of subcategories) in the Amazon.com data set compiled in (McAuley et al., 2015). 11,295 tags across 10,00 OTC medicine reviews were collected from 93 undergraduate university student taggers and 1 researcher authority tagger. Each tagger completed a mean of 126 (M=120) tags. Accuracy was verified against 200 authority tags created by the researcher. In all, there 225 overlaps between authority and non-authority tags, of which 193 were agreements (86% accuracy). Cohen's kappa (J. Cohen, 1968) was .805, indicating "substantial" agreement (Landis & Koch, 1977), almost "near perfect" agreement.

Figure 2.3.1. Construction of training and test data sets.

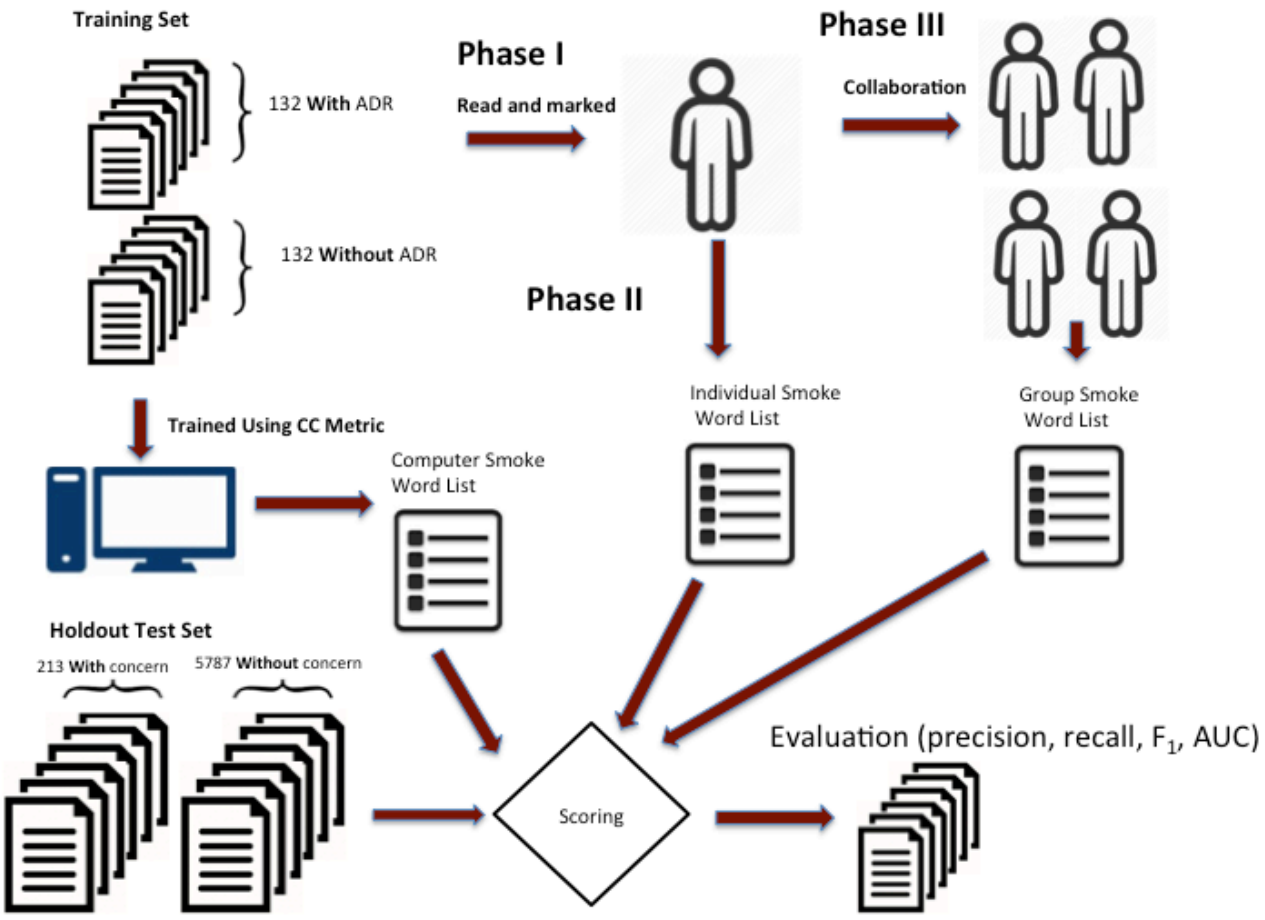


Experiment

Our experimental procedure is detailed in Figure 2.3.2 below. In Phase I of our experiment, 72 undergraduate business students read and tagged each review in the training data set according to whether it appeared to contain an ADR mention. This data set was already a "gold standard", and so the tags were discarded. The rationale for this phase was to provide participants with a rudimentary knowledge of some of the dangers associated with these particular OTC medicines, and thus provide a basic level of domain-specific knowledge. Drugs mentioned in the collection include Allegra, Benadryl, Excedrin, Tylenol, Advil, Aleve, acetaminophen, Claritin, ibuprofen, and Miralax. Negative reactions mentioned include: nausea, anxiety, malaise, stomach ulcer, liver damage, choking on pills, stomach pain, acid reflux, dizziness, burning in the nostrils (nasal spray), burning in the chest, vomiting, skin odor, swollen eyes, hypertension, insomnia, environmental damage, poisonous coating, addiction, nervousness, excessive bleeding, heartburn, scratches in the nose, skin blisters, numbness, watery eyes, cramps, coughing, bleeding gums, ear infection, heart problems, diarrhea, melanosis coli, and rectal bleeding. Reviews also mentioned problems that are not ADRs, but are nevertheless dangerous: e.g., unclear dosage instructions, easily-damaged packaging, and plastic components that cause lacerations. Our hypothesis is that ordinary readers can condense these effects into a small number of key words that would be more descriptive than a list generated by a computer.

Examples of reviews indicative of safety concerns appear in 2.B.

Figure 2.3.2. Experiment Process



In Phase 2, students compiled a list of words and phrases they thought might function as effective filter terms in identifying reviews with safety concerns in a distinct hold-out test set. Students were instructed to use whatever methods they like, but were told that they would be in competition with a computer, which would calculate probabilities based on term frequencies, so they were encouraged to be as creative as they could be. One simple approach might be to find a manufacturer's list of ordinary reactions vs. dangerous reactions (e.g., https://www.rxlist.com/allegra-d-side-effects-drug-center.htm) and use that

as a guide. Each student submitted three separate lists: one for unigrams, one for bigrams, and one for trigrams. Next to each term, participants added a weight between 0 and 1 indicating how important they thought that term was. An example unigram list appears in listing 2.3.1.

Listing 2.3.1. Example unigram list.

burn	0.90
red	0.76
bleed	0.54
hurt	0.54
problem	0.51
caution	0.38

Table 2.3.1 summarizes descriptive statistics of the individual lists. The most common unigrams, bigrams, and trigrams are listed in Appendix 2.C.

Table 2.3.1. Descriptive statistics of human-generated word lists, created by individuals.

		List Length (Count of Words)			
List	**N (# of lists)**	**Mean**	**Median**	**Min**	**Max**
unigrams	72	231	200	16	1392
bigrams	71*	261	202	34	1222
trigrams	72	292	201	29	1001

* One participant forgot to submit a bigram list

Computer smoke word lists were generated using the Correlation Coefficient approach described in Fan (Fan et al., 2005), which is an adaptation of the Chi-squared feature selection method. This method is described in Listing 2.3.2. When comparing the computer lists to the individual lists, the computer term lists consisted of the 200 terms with the highest CC. We chose to cap the number of terms at 200 because this was the median among the individual lists and we did not want to introduce any effects due to list length.

Listing 2.3.2. Correlation Coefficient metric used for computer n-gram lists.

$$\mathcal{C} = \frac{\sqrt{N} \times (AD - CB)}{\sqrt{(A + B) \times (C + D)}}$$

N = number of docs in the collection
A = number of relevant docs with term
B = number of non-relevant docs with term
C = number of relevant docs without term
D = number of non-relevant docs without term

In Phase 3, students collaborated in groups of 2, 3 or 4 to assemble a group list. The intent of this phase was to assess how well collective human intelligence would perform in this task. Group lists were scored on the same metrics as the individual lists. Our experiment included 8 groups of 2, 7 groups of 3, and 8 groups of 4. Table 2.3.2 summarizes descriptive statistics of the resulting group lists.

Table 2.3.2. Descriptive statistics of group word lists

		List Length (Count of Words)			
	N (# of lists)	**Mean**	**Median**	**Min**	**Max**
unigrams	23	492	419	163	1673
bigrams	23	646	537	264	2118
trigrams	23	681	590	327	1270

2.4. Results

We evaluated the performance of each smoke word list using the holdout test data set of 6000 reviews. This set contained 5787 were non-safety concerns and 213 marked as ADRs (3.5%), which made good recall scores difficult. For each smoke word list (individual, computer, and group), we scored each review using this procedure:

1. Each test review was tokenized (words were not stemmed and stop words were not removed) and split into unigram, bigram, and trigram terms.
2. For each individual, group, and computer list of unigrams, bigrams, trigrams, and all-n-grams:
    a. Each review's terms were iterated, and each time a term appeared that was on the term list, we added the value that the list assigned to that term.
    b. The final score for each review-term list pair was this accumulated sum. See a sample of the resulting table in Table 2.4.1 below.
    c. The reviews were sorted from highest to lowest, based on that list's score for each review.
    d. We stepped down this sorted list, checking whether each review included an ADR, and stored counts of true positives and false positives. For each list, we calculated AUC by summing the rectangle area under the resulting ROC curve. We also stored precision, recall, and F1 at 1000-review increments.

Table 2.4.1. Example of data table for review-term list scores.

Review ID	ADR?	computer unigram_score	person_1 unigram_score	person_2 unigram_score	person_2 unigram_score
35384137	No	9.785	1.500	1.350	0.900
35384230	No	1.896	0.000	0.000	0.000
35430503	No	8.086	1.200	0.800	0.300
35444364	Yes	46.420	2.000	2.000	1.700
35484997	Yes	5.667	4.800	2.600	5.800
35489877	No	4.804	0.700	0.000	0.000

In the sections that follow, we first evaluate the classification performance of lists generated by individual humans. Then we evaluate the group lists.

Distribution of Review Scores for Individual Lists

Distributions of review scores from individual lists are summarized in Table 2.4.2. Although the score distributions for Non-ADR and ADR classes overlapped, scores were significantly higher in the ADR class for all lists. Review scores were not normally distributed across participants, so to derive the "human" score for each review, we took the median of the scores assigned to that review by the human lists. In Table 2.4.2, the "Prob >|t|" columns tests against the null that the scores are equals between the ADR and non-ADR classes.

Table 2.4.2. Distributions of review scores.

		Computer			Human(Median)		
		Non-ADR	**ADR**	**Prob > |t|**	**Non-ADR**	**ADR**	**Prob > |t|**
Unigrams	**Mean**	8.954	15.967	p < .0001	.949	2.207	p < .0001
	Median	5.772	11.253		.600	1.550	
	St Dev	10.084	14.241		1.268	2.043	
	Min	0	.383		0	0	
	Max	145.084	81.194		19.700	12.100	
Bigrams	**Mean**	2.470	4.460	p < .0001	.053	.224	p < .0001
	Median	1.540	3.146		0	0	
	St Dev	3.220	4.143		.182	.350	
	Min	0	0		0	0	
	Max	50.580	24.392		3.200	1.80	
Trigrams	**Mean**	.422	.804	p < .0001	.006	.007	NS
	Median	0	.573		0	0	
	St Dev	.760	1.084		.055	.061	
	Min	0	0		0	0	
	Max	11.519	5.35		.9	.62	
Union	**Mean**	11.855	21.231	p < .0001	1.172	2.752	p < .0001
	Median	7.563	14.844		.75	1.9	
	St Dev	13.543	18.573		1.515	2.443	
	Min	0	0		0	0	
	Max	207.184	100.417		1.172	13.8	

T tests indicated that for both computer and human lists, scores were significantly higher in the ADR than the non-ADR class, which confirms that the smoke word list procedure is effective in discriminating the two classes.

We conclude that both the computer and human lists differentiate well between non-safety and safety classes. The human lists had consistently lower scores but similarly significant t-tests; they retrieved fewer true positives, but they also had fewer false negatives.

Classification Performance of Individual Lists

Table 2.4.3 summarizes the classification performance for the best, median, and worst human lists, along with the computer lists, grouped by type of n-grams. AUC appears in the first column, followed by precision (denoted by "prec") and recall (denoted by "rec") at 1000-review increments (denoted by "k"). Across all n-grams, the best humans consistently outperformed the computer, while the median human was not significantly different. The "rec 1k" column shows that the best human list found 47% of the ADRs after looking at the first 1000 reviews, whereas the computer found 40%; by the 2000th review the best human found 70%, whereas the computer found 60%. The best human found 87% of the ADRs by looking at the first half of reviews (3000), whereas the computer found only 78%.

Table 2.4.3. ADR classification performance for individual human lists vs. the computer. Precision and recall are reported for 1000-review increments.

Unigrams													
Participant	**auc**	**prec 1k**	**rec 1k**	**prec 2k**	**rec 2k**	**prec 3k**	**rec 3k**	**prec 4k**	**rec 4k**	**prec 5k**	**rec 5k**	**prec 6k**	**rec 6k**
best human	0.737	0.101	0.474	0.075	0.700	0.062	0.869	0.050	0.939	0.042	0.991	0.036	1.000
median human	0.677	0.088	0.413	0.067	0.629	0.056	0.784	0.045	0.845	0.040	0.930	0.036	1.000
computer	0.677	0.085	0.399	0.064	0.601	0.055	0.775	0.047	0.887	0.041	0.962	0.036	1.000
worst human	0.520	0.053	0.249	0.040	0.376	0.039	0.549	0.036	0.676	0.035	0.822	0.036	1.000
Bigrams													
Participant	**auc**	**prec 1k**	**rec 1k**	**prec 2k**	**rec 2k**	**prec 3k**	**rec 3k**	**prec 4k**	**rec 4k**	**prec 5k**	**rec 5k**	**prec 6k**	**rec 6k**
best human	0.685	0.086	0.404	0.065	0.606	0.055	0.779	0.048	0.892	0.041	0.958	0.036	1.000
computer	0.649	0.081	0.380	0.065	0.606	0.053	0.742	0.044	0.826	0.039	0.906	0.036	1.000
median human	0.580	0.077	0.362	0.053	0.498	0.044	0.620	0.038	0.709	0.037	0.873	0.036	1.000
worst human	0.472	0.034	0.160	0.033	0.305	0.035	0.488	0.035	0.648	0.036	0.850	0.036	1.000
Trigrams													
Participant	**auc**	**prec 1k**	**rec 1k**	**prec 2k**	**rec 2k**	**prec 3k**	**rec 3k**	**prec 4k**	**rec 4k**	**prec 5k**	**rec 5k**	**prec 6k**	**rec 6k**
best human	0.740	0.104	0.488	0.078	0.728	0.060	0.850	0.050	0.934	0.042	0.991	0.036	1.000
median human	0.683	0.090	0.423	0.066	0.620	0.056	0.784	0.047	0.878	0.041	0.953	0.036	1.000
computer	0.677	0.088	0.413	0.062	0.582	0.055	0.779	0.048	0.892	0.041	0.953	0.036	1.000
worst human	0.501	0.049	0.230	0.041	0.380	0.035	0.498	0.036	0.667	0.035	0.812	0.036	1.000
All n-grams													
Participant	**auc**	**prec 1k**	**rec 1k**	**prec 2k**	**rec 2k**	**prec 3k**	**rec 3k**	**prec 4k**	**rec 4k**	**prec 5k**	**rec 5k**	**prec 6k**	**rec 6k**
best human	0.740	0.104	0.488	0.078	0.728	0.060	0.850	0.050	0.934	0.042	0.991	0.036	1.000
median human	0.683	0.090	0.423	0.066	0.620	0.056	0.784	0.047	0.878	0.041	0.953	0.036	1.000
computer	0.677	0.088	0.413	0.062	0.582	0.055	0.779	0.048	0.892	0.041	0.953	0.036	1.000
worst human	0.501	0.049	0.230	0.041	0.380	0.035	0.498	0.036	0.667	0.035	0.812	0.036	1.000

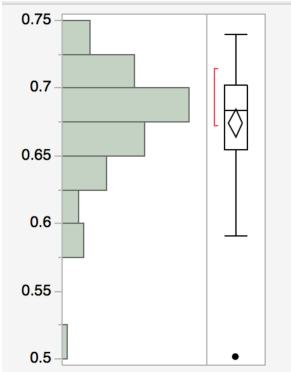
Figure 2.4.1 shows a histogram of the AUC scores for the combined

(unigram+bigram+trigrams) human-generated lists. Most of the lists scored in the high .6's.

The worst combined human list (AUC=.501) was an outlier (4 Huber Spreads from the center)

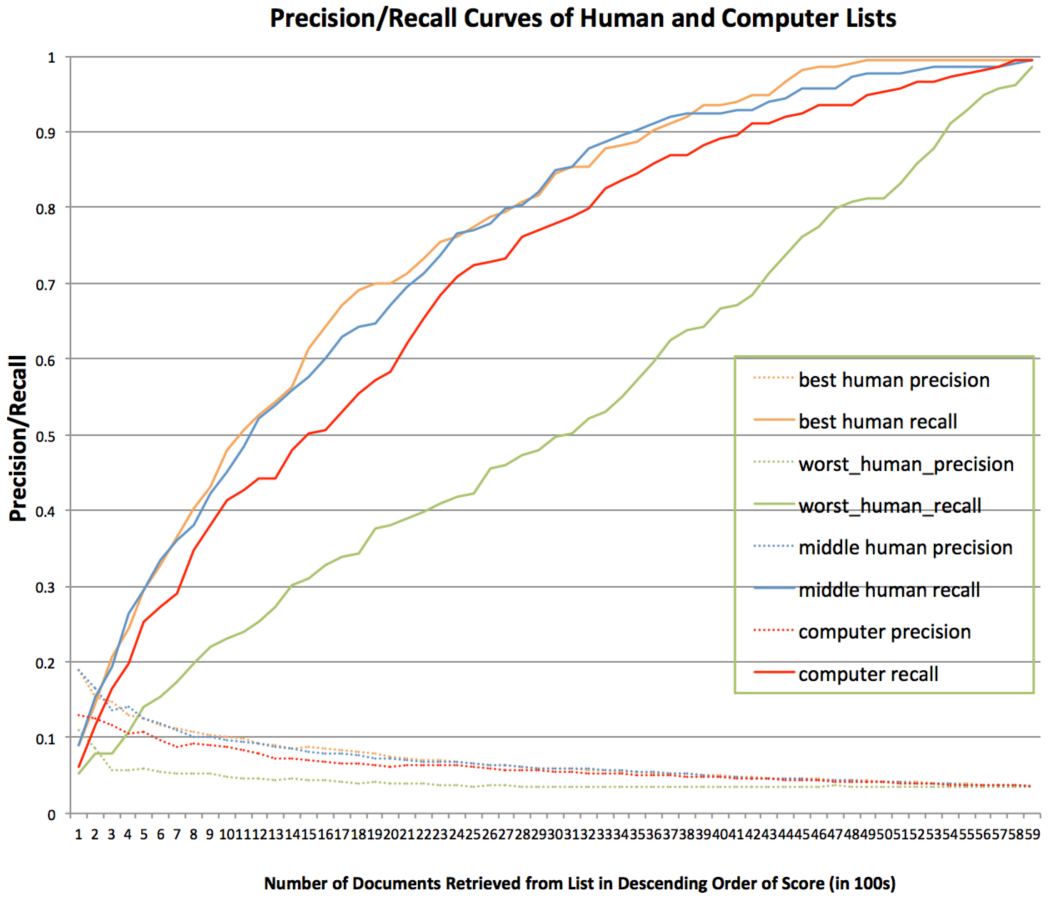
and contained 431 terms, where the average combined list contained 784 terms. Although list length was not significantly correlated with classification performance in general, the short length may have been an indication of this participant's effort. Even without this observation, the mean human AUC was .677, which was not significantly different from the computer AUC of .683 (p=.144).

Figure 2.4.1. Distribution of individual human AUC scores. Longer bars indicate more observations.



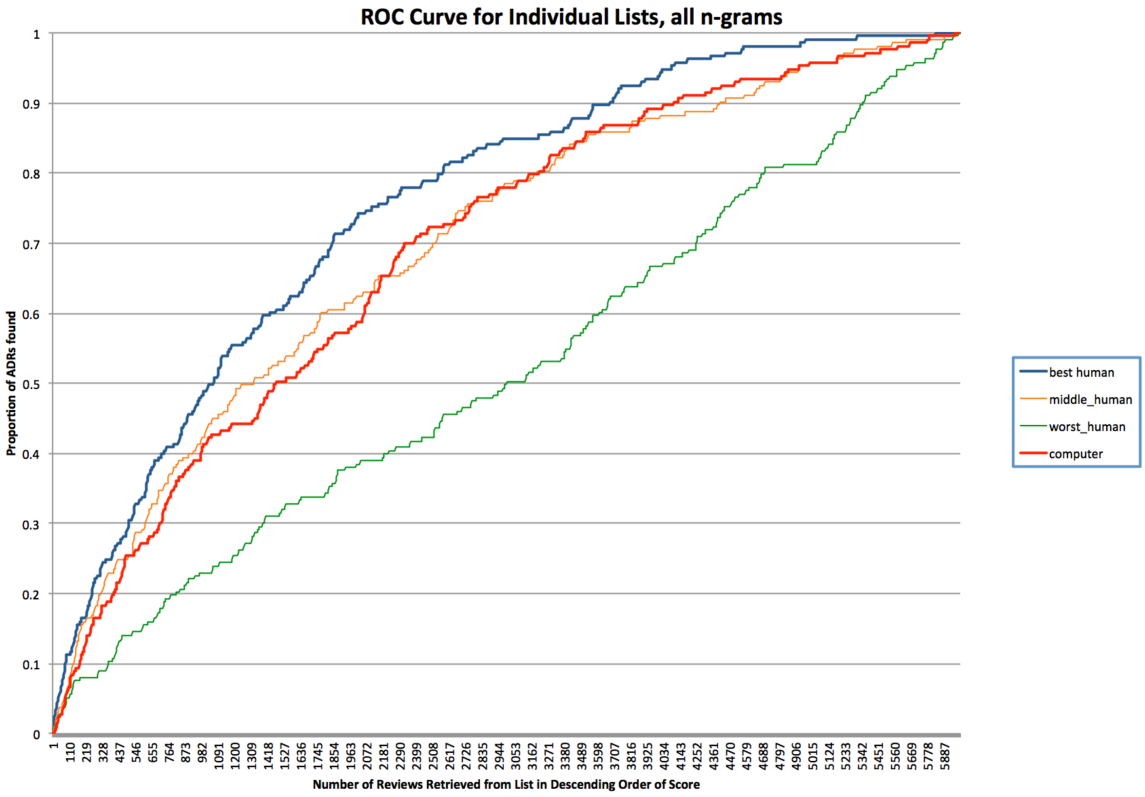
Precision and recall curves across all 6000 reviews appear in Figure 2.4.2. The computer lists start with lower precision than all of the human lists, but the human lists lose precision more quickly. We also note that the human lists pick up more true positives early on, but that recall decreases off more quickly than the computer lists. The best and median human lists stay well above the computer list on recall, and even the worst human list is competitive.

Figure 2.4.2. Precision and recall curves for n-gram lists.



The ROC curves are shown in Figure 2.4.3. The best and median human lists had higher true positive rates earlier in the list, but the computer list catches up, so that AUC across all 6000 reviews is not significantly different

Figure 2.4.3. ROC curve for lists for best human, median human, worst human, and computer (all n-grams).



Although the middle human list did not outperform the computer, as a practical matter, we were interested in whether humans can outperform the computer in the best case. The top quartile of human lists (18 lists) scored an average AUC of .713, significantly higher than the computer list ($p < .0001$). We therefore conclude that human generated lists can outperform the computer's, but only in cases where special effort or skill is applied. On average, there is no difference.

Distribution of Review Scores for Group Lists

Table 2.4.4 reports the computer list vs. group list scores for the test data set. The group scores refer to the median review scores across all groups. We report best, median, and worst group scores later. The group lists, like the individual lists, have significantly higher scores for the ADR reviews than the non-ADR reviews for all lists. We conclude therefore that the

term list generation process works well for differentiating ADR and non-ADR reviews, whether we are using individual, computer, or group lists.

Table 2.4.4. Distributions of review scores, long computer lists vs. group lists.

		Computer			Group(Median)		
		Non-ADR	**ADR**	**Prob > |t|**	**Non ADR**	**ADR**	**Prob > |t|**
Unigrams	**Mean**	11.257	20.080	p<.0001	2.238	4.771	p<.0001
	Median	7.242	14.263		1.450	3.263	
	St Dev	12.413	17.768		2.540	4.218	
	Min	0	.664		0	0	
	Max	162.293	100.315		25.175	23.150	
Bigrams	**Mean**	3.473	6.348	p<.0001	.320	.959	p<.0001
	Median	2.218	4.528		0	.700	
	St Dev	4.034	5.644		.561	1.009	
	Min	0	0		0	0	
	Max	42.223	31.347		6.901	4.7	
Trigrams	**Mean**	.713	1.370	p<.0001	.034	.142	p<.0001
	Median	.547	.953		0	0	
	St Dev	1.069	1.563		.167	.346	
	Min	0	0		0	0	
	Max	10.397	7.479		2.1	2.000	
Union	**Mean**	15.545	27.797	p<.0001	2.812	6.326	p<.0001
	Median	9.943	18.920		1.800	4.500	
	St Dev	17.586	24.090		3.186	5.398	
	Min	0	.664		0	0	
	Max	275.245	132.870		30.250	28.500	

Classification Performance of Group Lists

Table 2.4.5 reports AUC along with precision and recall scores at 1000-review increments. Although the highest-scoring group did not significantly outscore the highest-scoring individual, the median group significantly outperformed both the median human and the computer (Kruskal-Wallis $p < .0001$). The longer computer list only increased AUC from .677 to .678, whereas the median group list had an AUC of .708 vs. individual .683.

Table 2.4.5. Classification performance of human lists generated by groups.

Unigrams													
Participant	**auc**	**prec 1k**	**rec 1k**	**prec 2k**	**rec 2k**	**prec 3k**	**rec 3k**	**prec 4k**	**rec 4k**	**prec 5k**	**rec 5k**	**prec 6k**	**rec 6k**
best group	0.743	0.102	0.479	0.078	0.732	0.061	0.864	0.050	0.944	0.042	0.986	0.036	1.000
median group	0.701	0.088	0.413	0.068	0.634	0.057	0.808	0.049	0.925	0.041	0.972	0.036	1.000
computer	0.678	0.085	0.399	0.065	0.610	0.055	0.775	0.047	0.887	0.041	0.953	0.036	1.000
worst group	0.636	0.075	0.352	0.060	0.559	0.050	0.709	0.045	0.840	0.039	0.925	0.036	1.000
Bigrams													
Participant	**auc**	**prec 1k**	**rec 1k**	**prec 2k**	**rec 2k**	**prec 3k**	**rec 3k**	**prec 4k**	**rec 4k**	**prec 5k**	**rec 5k**	**prec 6k**	**rec 6k**
best group	0.700	0.092	0.432	0.068	0.638	0.058	0.817	0.049	0.911	0.042	0.977	0.036	1.000
computer	0.658	0.080	0.376	0.066	0.615	0.054	0.756	0.045	0.845	0.040	0.930	0.036	1.000
median group	0.653	0.080	0.376	0.059	0.554	0.053	0.746	0.045	0.845	0.040	0.939	0.036	1.000
worst group	0.495	0.036	0.169	0.038	0.357	0.036	0.507	0.037	0.690	0.036	0.854	0.035	1.000
Trigrams													
Participant	**auc**	**prec 1k**	**rec 1k**	**prec 2k**	**rec 2k**	**prec 3k**	**rec 3k**	**prec 4k**	**rec 4k**	**prec 5k**	**rec 5k**	**prec 6k**	**rec 6k**
computer	0.611	0.077	0.362	0.059	0.549	0.045	0.634	0.042	0.789	0.038	0.901	0.036	1.000
best group	0.603	0.070	0.329	0.055	0.512	0.046	0.648	0.042	0.793	0.038	0.887	0.036	1.000
median group	0.566	0.064	0.300	0.047	0.441	0.042	0.596	0.040	0.751	0.037	0.864	0.036	1.000
worst group	0.486	0.036	0.169	0.033	0.310	0.036	0.507	0.036	0.671	0.037	0.864	0.036	1.000
Union													

Participant	auc	prec 1k	rec 1k	prec 2k	rec 2k	prec 3k	rec 3k	prec 4k	rec 4k	prec 5k	rec 5k	prec 6k	rec 6k
best group	0.748	0.104	0.488	0.075	0.704	0.062	0.869	0.051	0.948	0.042	0.986	0.036	1.000
median group	0.708	0.101	0.474	0.069	0.648	0.057	0.808	0.048	0.897	0.041	0.958	0.036	1.000
computer	0.678	0.085	0.399	0.064	0.596	0.054	0.765	0.047	0.887	0.041	0.958	0.036	1.000
worst group	0.644	0.078	0.366	0.061	0.568	0.053	0.751	0.045	0.840	0.039	0.911	0.036	1.000

Precision and recall curves for the best, median, and worst group follow in Figure 2.4.4, and the ROC curves appear in Figure 2.4.5.

Figure 2.4.4. Precision and recall curves for group n-gram lists (all n-grams).

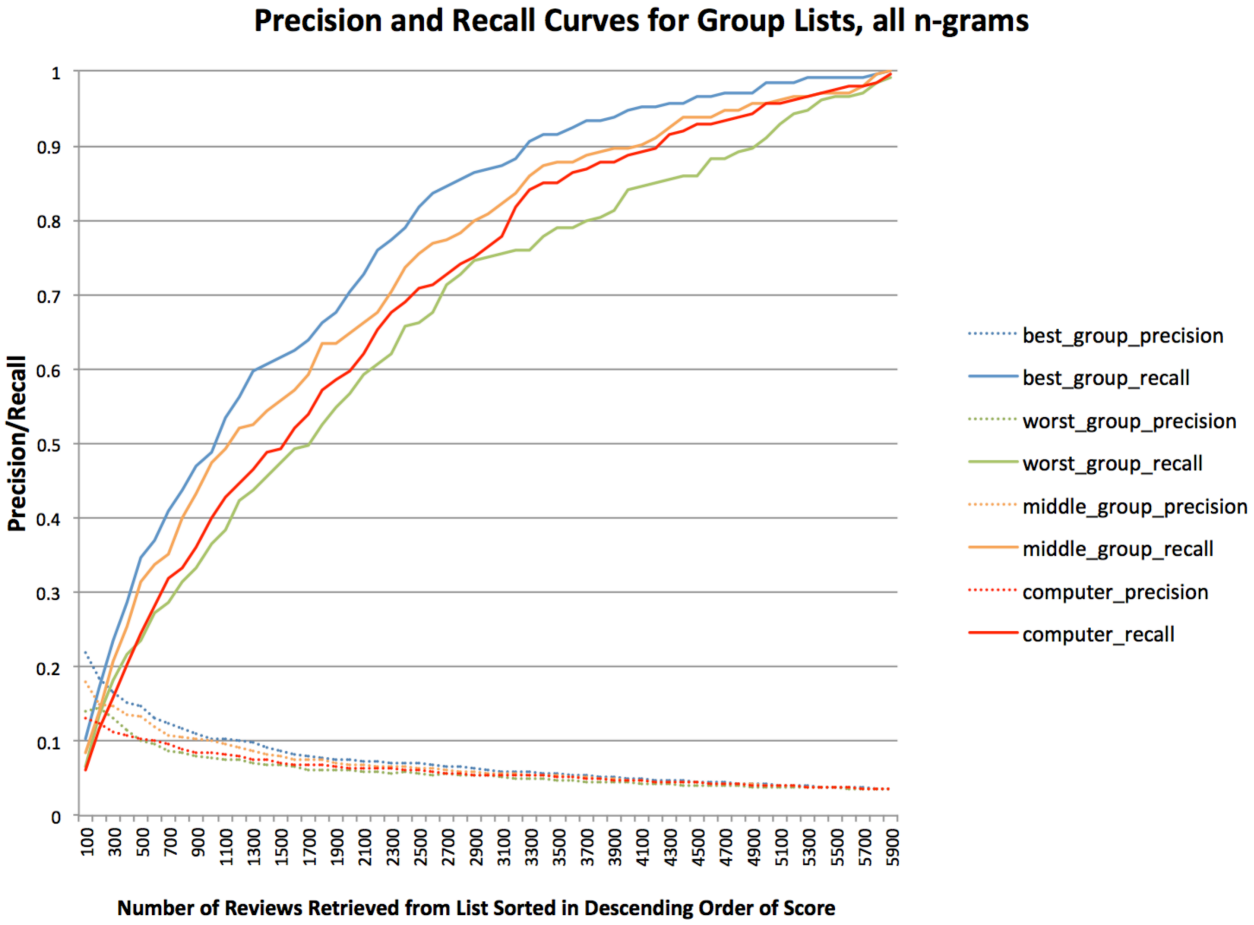
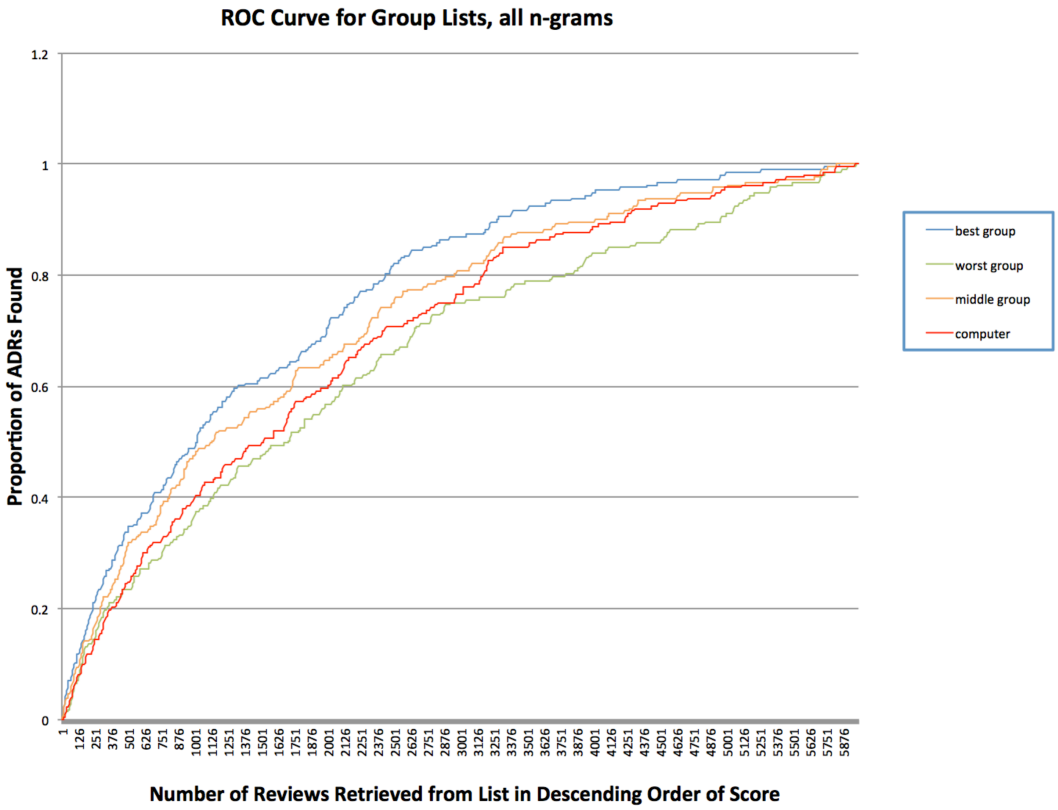


Figure 2.4.5. ROC curve for group n-gram lists (all n-grams).



We were also interested in whether group size made a difference. Our experiment included 8 groups of 2, 7 groups of 3, and 8 groups of 4. A Welch's test for variance indicated that the variance in AUC was not equal across group size, so we used a Welch's ANOVA and concluded that the AUC was not significantly different across group size (p=.30). However, the fact that 5 of the 6 highest-scoring lists were created by groups of 4 is suggestive.

2.5. Discussion

Our results show that when individuals with a basic level of training brainstorm a list of query terms for their information need, their list will outperform a list constructed through correlational analysis of a hand-tagged training set, provided they have a reasonable level of

skill and effort. When groups brainstorm together, the difference is more marked, and even a list of medium quality will outperform a computer list.

In follow-up analyses, we examined two questions: 1) How strong is the relationship between list length and classification performance? We avoided this question in our experiments by constraining the computer list to be the median length of the human lists. Logically, longer lists will have better recall but worse precision, but we wanted to test this intuition. 2) How much were the human lists abetted by synonym-dumping? In a post-hoc reflection, participants were asked to report on their approach in constructing the individual lists and on how their group collaborated. The most common strategies for individuals were unaided synonym generation (14), using a thesaurus (10), and consulting medical dictionaries (7). For the groups, the most common strategies were appending the lists and removing duplicates (37), and discussion/collaboration (18).

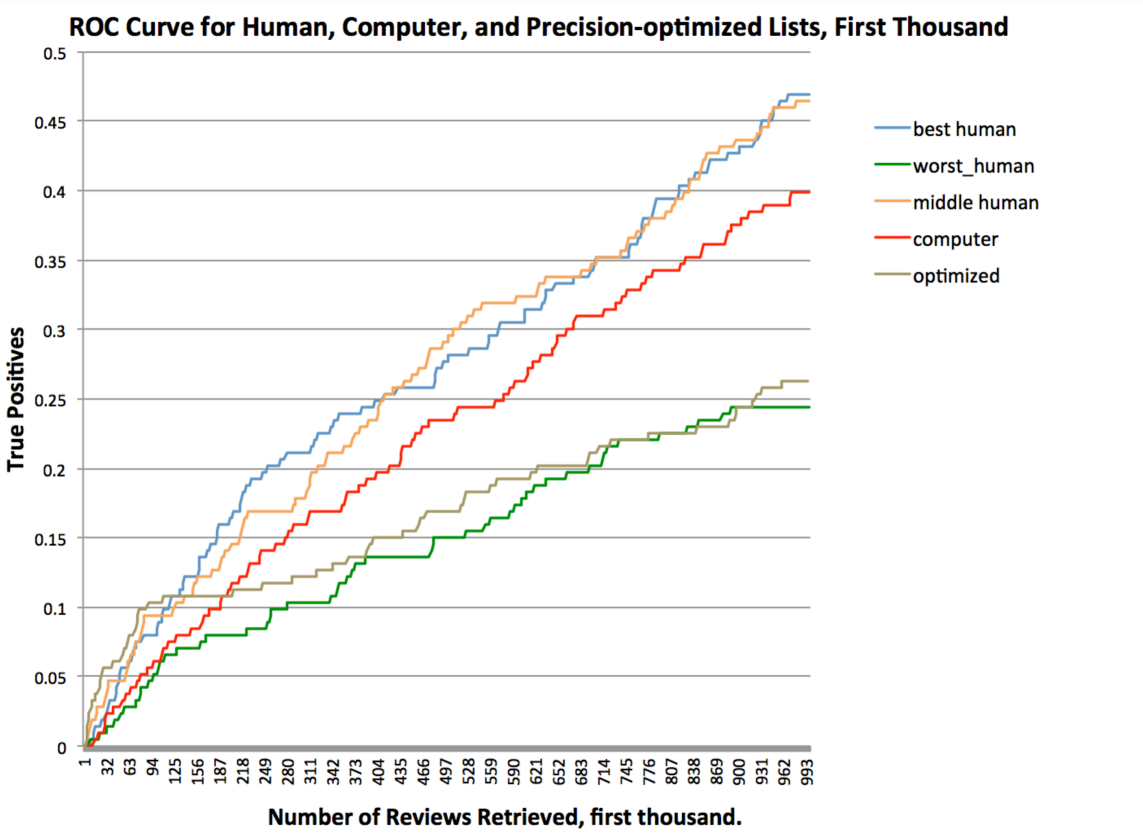
List Length and Classification Performance

We examined the correlation between list length and all classification metrics. Contrary to intuition, list length is not consistently related either precision or recall (p values ranged from .2 to .8). We therefore conclude that attempting to gain performance by blindly heaping terms onto the list is not effective.

However, an interesting observation emerges if we compare the performance of human and computer lists with one optimized for high precision in a small number of retrieved documents, as is the one described in (Goldberg & Abrahams, 2017). In Figure 2.5.1 below, we show the ROC curve over the top 1000 reviews for the best human, the median human, the worst human, the computer, and the short list (11 terms only) optimized for high precision at 100 reviews. We see that the optimized list takes an early lead, picking up true positives before any other list, but at some point, the lead fades and performance

deteriorates. The best human overtakes the optimal at review 134, and the worst human at 153. The computer list does not overtake until review 202. At this resolution, we note a similarity between the short optimal list and the human lists: humans appear to "optimize" for precision, picking up true positives quickly, but the curves level off , while the computer list continues to gain ground.

Figure 2.5.1. ROC curves for the best human, worst human, median human, and computer, alongside that of a short list (11 unigrams) optimized for high precision at 100 documents.



The Impact of Synonyms on Classification Performance

Since a large proportion of the subjects cited synonym-generation as their primary methodology, we decided to test whether the computer-generated lists would benefit from augmentation with basic synonyms. Using the Python Natural Language Toolkit (http://www.nltk.org/), we generated synonyms from WordNet (G. A. Miller, 1995) for all the terms on the unigrams list created by the computer, assigning the weight from the original term to each of the synonyms. A sample of the augmented lists appears in Listing 2.5.1. The addition of synonyms did not significantly improve classification performance on any metric.

Listing 2.5.1. Excerpt from the synonym-augmented computer list. Original terms are in **bold**.

bleed	**0.53**
shed blood	0.53
hemorrhage	0.53
injury	0.53
phlebotomize	0.53
hurt	**0.52**
scathe	0.52
harm	0.52
distress	0.52
bruise	0.52
injure	0.52
ache	0.52
suffer	0.52
pain	0.52
later	**.494**
after	.494
afterwards	.494
subsequently	.494
tardy	.494
belated	.494

2.6. Conclusions

Most text mining approaches to pharmacovigilance entail the creation of a lexicon of term features that help a classifier identify ADRs, whether using EHR, biomedical literature, or social media. The selection of strongly differentiating terms has an element of human judgment, and it remains an open challenge to determine where in the process that judgment should be exercised (Gaines, 1989). In this experimental study, we tested how well non-experts can create lexicons from scratch using creativity and any public sources available. We compared the classification performance of several lexicons: one generated using feature selection on annotated data (supervised machine learning), one made by non-experts on the basis of rudimentary training, and one created by groups of 2, 3, and 4 people.

We found that lists from the top quartile of humans (18 lists) significantly outperform those of the computer on all classification metrics. Group lists outperform the computer even in the average case. Larger group sizes appear to be more effective (5 of the top 6 performing lists came from groups of 4) but we were unable to validate this statistically, possibly due to small sample size (N=23).

Upon closer inspection of how the lists perform on the first 1000 reviews, we observe that humans appear to be optimized for precision, soundly outperforming the computer on shorter lists. We also found that na ve synonym-generation adds nothing to the computer list's performance.

Our work has some limitations that suggest avenues of future research. First, our study concentrates on only one particular product, OCT medications. Different products present vastly different potential threats to safety, and previous research has demonstrated that the language of safety concerns varies among industries (Abrahams et al., 2012; Law et al., 2017; Winkler et al., 2016). How broadly our results generalize across different product categories is one area of future work. Second, this work compares the human-generated query term lists against computer lists created using only one method, the CC metric devised in (Ng, Goh, & Low, 1997). It is possible that other metrics could generate lists comparable to or even better than the human lists, and this is another area of future exploration. Given that the human strategies consisted largely of synonym generation and changes in levels of abstraction, it is possible that more sophisticated techniques such as latent semantic indexing (Marchionini, 1997) might capture these powerful intuitive associations and create more effective computer lists. Although augmenting the computer lists with synonyms failed to improve their performance, it is possible that computer lists could benefit from the addition of words that are related but not necessarily synonymous (e.e.g, *headache*, *brain*, and *eyeball*).

The use of word embeddings to generate these "word families" might be a promising avenue for future research (Nikfarjam & Gonzalez, 2011; W. Wang, 2016; Xia, Wang, & Fan, 2017).

Our work makes several contributions to research. First, we confirm that methods that have been effective for finding safety defects in consumer products can also work in discovering safety concerns in OTC medications. Second, we have added to the research on group decision making by determining that at this particular task, groups do in fact outperform individuals. Third, we have established a gold standard for evaluating future algorithms designed to perform this particular task. We have a baseline human performance data set that can be used to benchmark search tools that apply state-of-the-art query enhancement strategies as they are created.

Our work makes practical contributions as well. Our findings suggest a pragmatic approach for hunting through text documents for any kind of business intelligence. A strategy that works surprisingly well is to have a single person with a moderate degree of introductory expertise brainstorm search terms and assign intuitive weights. The best human-generated lists were able to find significantly more ADRs in a set of unseen online reviews as a computer-generated list. Finally, we found that group collaboration improves the quality of list brainstorming, with larger groups apparently creating better lists.

Appendix 2.A. Detailed list of Amazon.com categories and subcategories.

- Health & Personal Care
    - Baby & Child Care
        - Health Care
            - Allergy Medicine
            - Pain Relievers
                - Non-Aspirin
                    - Acetaminophen

    - Health Care
        - Allergy
            - Sinus & Asthma
                - Allergy Medicine
                - Asthma Medicine
                - Nasal Strips
                - Respiration Flow Meters
                - Sinus Medicine
        - Cough & Cold
            - Coughing & Sore Throats
            - Cough Syrups
        - Digestion & Nausea
            - Antacids
            - Gas Relief
            - Lactose Intolerance
            - Laxatives
            - Motion Sickness & Nausea
            - Diarrhea Relief
        - Pain Relievers
            - Non-Aspirin
                - Acetaminophen

Appendix 2.B. Examples of online reviews indicative of safety concerns.

…I took this products every night for about three months. Normally only taking half a dose because I only had time for about four to five hours of sleep. Then, I was at work one day and got a terrible nose bleed. After I finally got it to stop I had a pressure headache into the next day. Ever since then I kept getting daily, sometimes as many as four times a day, pin point headaches…

… There's something more in there that will make you think that you need it and you will have very vivid dreams. I felt like I was coming off drugs….

…I tried it once and threw the bottle away. This stuff really stings like the dickens even in a dilute solution. It brought tears to my eyes…

…I just drank the tea. It smells and taste nasty. I tried my hardest not to throw up, thinking this is normal. But, I did end up throwing up everything. My head is still spinning, I do not know if I am having an allergic reaction to the tea. But I know that I do not recommend this nor will I ever take it again. I am so sick right now….

…I used this product for the 14 days...it causes acute melanosis coli. It can also cause rectal bleeding, the product is not safe and should not be on the market, extremely harsh on your system…

…When I suddenly had extreme intestinal distress, severe cramping, and uncontrollable diarrhea for three months starting in October, I went through numerous examinations, blood and stool samples, prescriptions, CTscans, colonoscopy, and many ruined clothes and missed days of work, to no avail. Imagine my surprise when I realized that the intestinal distress began at about the same time I had inadvertently purchased the new …flavor…

… I had a severe reaction. Inner hives and severe painful itching from my head to toes. There wasn't a single area that didn't hurt and itch. And worse I had to wait hours for this 24 hour medicine to wear off…

…But I cannot wear it! Where the clip part attaches to the frame of each filter part, it has a very sharp edge on the plastic, where the end of the plastic clip pokes through the filter frame on each side. This sharp edge is right where it contacts the inside middle of your nose. Within 10 minutes, it had actually broken the skin inside my nose, and was just painful!...

Appendix 2.C: Most commonly selected unigrams, bigrams, and trigrams. Numbers is parentheses indicate how many lists the term appeared in. (Groupings are provided for convenience, and were created by a research associate who is a PhD is biology who teaches medically-related courses).

General physical reactions
Sick (57)
Reaction (52)
Swelling (47)
Dizziness (46)
Dizzy (44)
Fever (42)
Death (41)
Drowsy (38)
Drowsiness (36)
Jittery (29)
Infection (29)
Irritated (28)
Side effects (36)
Severe reaction (30)
Withdrawal symptoms (25)
Cause death (25)
Very sick (25)
More sick (17)
Bad reaction (16)
Can cause death (30)
Made me sick (24)
Makes me jittery (18)
A severe reaction (18)
Cause withdrawal symptoms (17)
Unclear and dizzy (17)
Decreased sex drive (15)
Bad side effects (13)
Most unpleasant sensations (12)

Pain
Pain (67)
Painful (54)
Hurt (60)
Hurts (40)
Discomfort (56)
Uncomfortable (48)
Sore (47)
Ache (35)
Worst pain (27)
Excruciating pain (26)
Killing me (21)
Becomes painful (21)
Severe pain (20)
In pain (18)
Very painful (17)
So painful (17)
It hurts (15)
Hurts and burns (18)
Is killing me (17)
It's almost painful (17)
Was so painful (17)
Pain and discomfort (15)
The worst pain (15)

Burning/Stinging
Burn (53)
Burning (53)
Burned (43)
Burns (39)
Burning sensation (39)
Sting (33)
Stings (27)
It burned (25)
On fire (24)
Really stings (19)
Really burns (17)
Burned my nose (24)
Burns your throat (21)
Throat started burning (17)
Burned so much (16)
A burning sensation (16)
Was on fire (14)
Stuff really stings (12)

Head
Headache (42)
Migraine (34)
Ear infection (23)
Monster headache (16)
Eyes swelling (16)
Eyes swelling up (23)

Skin/Immune reactions
Rash (43)
Itching (38)
Blister (32)
Allergic (31)
Allergy (31)
Hives (30)
Allergic reaction (41)
An allergic reaction (19)
Severe painful itching (17)
Broken the skin (16)
Can cause scratches (14)
Severe skin irritation (14)
Irritating my skin (12)

Respiratory system
Choking (31)
Cough (27)
Sore throat (29)
Very congested (15)
Still coughing (15)
Very sore throat (25)
Choking on it (18)
Irritate my nose (16)

Digestive System
Heartburn (56)
Diarrhea (51)
Nausea (49)
Nauseous (30)
Acid (42)
Cramping (41)
Cramps (35)
Cramp (28)
Liver (40)
Vomit (38)
Reflux (34)
Stomach (33)
Drymouth (29)
Dry mouth (15)
Stomach pain (44)
Stomach ache (32)
Liver damage (40)
Throw up (38)
Throwing up (34)
Acid reflux (32)
Horrible heartburn (29)
Worst heartburn (25)
Dry heaving (24)
Stomach bug (23)
Bad nausea (22)
Stomach ulcer (21)
Severe reflux (21)
Awful cramping (21)
Bad cramps (17)
Stomach issues (21)
Stomach upset (17)
Upset stomach (17)
Intestinal distress (20)
Uncontrollable diarrhea (18)
Severe stomach pain (36)
Causes stomach pain (24)
Bad stomach ache (22)
Gave me nausea (23)
Very bad nausea (20)
Had horrible heartburn (20)
The worst heartburn (16)
Was throwing up (19)
Throws it up (16)
Had awful cramping (19)
I have diarrhea (15)
Started getting diarrhea (13)
Cause liver damage (14)
Sick and nauseous (14)
Hurt your internals (14)
Noticed stomach pain (14)
Cause severe reflux (13)
Acute melanosis coli (12)
Extreme intestinal distress (12)
Loss of appetite (12)

Cardiovascular System/Bleeding Bleeding (56) Bleed (35) Blood (38) Heart (31) Nose bleeds (30) Nose bleed (24) Rectal bleeding (29) Bleeding gums (26) Heart attack (17) Heart problems (16) Major bleed (15) Elevated blood pressure (27) Cause rectal bleeding (21) Bad nose bleeds (20) Terrible nose bleed (17) Makes me bleed (16) Ears were bleeding (14) **Related to drug expiration/quality** Spoiled (33) Expiration (30) Expired (28) Expiration date (30) Bad batch (19) Safety seal (18) Tablets were missing (19) Scary inactive ingredients (14) Medication was bad (13) Had been broken (12) **Drug efficacy** Totally ineffective (17) Not work (17) Didn't work (16) Did not work (29) Was totally ineffective (12) Does not work (17) Got worse (16) Made it worse (16) I got worse (13) It didn't stop (13)	**Adjectives** Bad (64) Severe (59) Dangerous (57) Danger (30) Intense (50) Broken (48) Worse (48) Worst (36) Horrible (47) Nasty (45) Terrible (42) Negative (34) Unpleasant (34) Awful (33) Ineffective (29) Abrasive (29) Toxic (28) Sharp (28) Not safe (36) Too much (25) Too intense (22) Life threatening (20) Very bad (20) So bad (17) Not gentle (19) Bad tasting (18) Funky taste (15) Extremely harsh (16) Way too intense (18) Very bad tasting (12) **Negation** Never (34) Not (30) Not use (19) Do not (18) Will not (18) Did not (16) **Purchasing** Not purchase (18) Not buy (16) Don't buy (16) Waste of money (45) Won't be buying (20) Would not buy (18) Don't buy this (17) Returned this product (15) I returned it (13)	**Consumer reactions/ recommendations** Disappointed (37) Very disappointed (37) Be careful (38) Not recommend (33) Not good (33) Stay away (28) Dangerous stuff (20) Be cautious (18) Steer clear (17) Be aware (17) Never take (16) Was very disappointed (33) Would not recommend (33) Do not recommend (30) Do not use (28) Is not safe (27) Not worth it (26) Stay away from (25) Be careful with (24) Be really careful (18) Use something else (21) Threw them away (19) Threw it away (17) Do your research (19) Do not take (18) Stopped taking it (16) Will not use (16) Used it once (14) Use with caution (14) Please be careful (14) Could not use (14) **Other** Waste (47) Damage (46) Problem (46) Problems (38) Issues (41) Issue (33) Careful (40) Upset (39) Suffer (37) Concern (34) Warning (34) Threw (33) Tear (32) Distress (32) Chemicals (32) Pressure (32)

Acknowledgements

Special thanks to Professor Laura Gruss for her help in providing term groupings in Appendix 2.C. The authors are also grateful to Siriporn Srisawas for her assistance managing the team of student taggers for the over- the-counter (OTC) medicine dataset. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CHAPTER 3

Text Analytics for C2C Interactions: Numeric Information Extraction

Abstract

There is a growing recognition that social media provide a valuable resource for product development and improvement. Locating relevant and useful information among the voluminous postings remains a challenge, however, and automated methods based on text mining have made significant progress in recent years. This study extends the literature on product defect discovery by contributing a novel set of features based on domain-specific numerical expressions. We show that these numerical expressions can be reliably extracted and classified, and that they can offer improvements in the accurate identification of social media postings that indicate a potential product defect. We further contribute a recommended decision support system for managing product quality using both textual and numerical attributes. This work helps to promote improvements in product quality and consumer safety.

3.1. Introduction

The recent surge of academic interest in the use of social media for quality surveillance reflects the growing role of customer to customer (C2C) communications in the Product Development Lifecycle. Researchers are discovering efficient methods for businesses to find, among the vast quantities of social media postings, opportunities for innovation (Qiao, Wang, Zhou, & Fan, 2018) , manufacturing defects (Abrahams et al., 2015; Abrahams et al., 2012; Law et al., 2017; Qiao et al., 2018; X. Zhang et al., 2016), and unanticipated consumer safety issues (Winkler et al., 2016).

The surveillance of social media for product safety has been an especially rich area of investigation, due to the challenges presented to regulators such as the National Highway

Traffic Safety Administration (NHTSA), Food and Drug Administration (FDA) or the Consumer Product Safety Commission (CPSC). Unlike businesses, which might monitor social media activity for themselves and a few key competitors, regulators have a mandate to oversee several industries, often with extremely limited resources. Automated methods for filtering out noise are especially needed in this case.

Although several approaches for locating high quality content have been explored using social network analysis (SNA), most of the information in social media is expressed as unstructured text, and therefore the most promising techniques are contributed by text mining (TM) and natural language processing (NLP). Some text-based methods that have proven effective in locating defects include term prevalence metrics (Abrahams et al., 2015; Law et al., 2017; Winkler et al., 2016), latent Dirichlet allocation (LDA) (Qiao, Zhang, Zhou, Wang, & Fan, 2017), ensemble methods (Yao Liu, Jiang, & Zhao, 2017), and heuristics (Goldberg & Abrahams, 2017).

One token type that is largely ignored in these examinations is the numerical type. A number by itself conveys very little information, because the lexical token alone does not provide context for understanding its referent or the significance of its magnitude. For example in the two snippets "I was traveling at 25mph …" and "I was going at 15 miles per hour …", a typical token-based approach would extract only unigrams like "25", "15", "mph", and trigrams like "miles per hour"", and these disjoint tokens hold little information value, as they "mph" is not associated with "miles per hour" and the relative magnitude of 25 vs 15 vs other speed values is not recognized. In contrast, a numeric information extraction approach would recognize that both snippets are distinct expressions of measures of "vehicle speed", and that the authors were describing travel at "low" speed (e.g. relative to 80 mph travel).

This study aims to fill this gap by proposing a procedure for creating a set of numerical features that communicate information about the semantics of the number.

When key terms are extracted using prevalence metrics as in (Abrahams et al., 2012), numbers often appear in the list of n-grams that are significantly associated with the target class. These numbers are usually removed from the curated term lists on the grounds that they communicate little information in isolation. Likewise, important numerical terms might not make it onto prevalent term lists because that particular value appears to infrequently when regarded solely as lexical tokens. Also, and of major importance, because the token alone is recognized, and not similar values, new examples would be incorrectly classified if the token value in the new observation is even slightly different from past observations. Consider, "15" and "25" in the example above: these would typically have insufficient prevalence to make it onto the prevalent terms list, as the distinct values are peculiar to each observation. However, once recognized as examples of "low speed", the semantic concept of "low speed" may indeed have unusually high prevalence (e.g. in defects vs. non-defects), and be a candidate for a "most prevalent semantic concepts" list. Furthermore, a new observation where the user is travelling at 10 mph or 20 mph would be correctly recognized as falling into the same category of low speed travel (compared to "10" and "20" being distinct and hitherto unseen tokens, in a conventional token-recognition approach, which would fail to recognize the significance of these values.

This study is an attempt to enrich the smoke word feature set by assigning categories to numbers based on their function in the text. While our case study specifically involves the automotive industry, we maintain that this procedure is generalizable to any domains with a large variety of meaningful numerical attributes.

Tools that make use of numerical expressions in NLP tasks typically follow a procedure in which entities are first located using Named Entity Recognition (NER), and then numerical attributes are attached based on textual proximity. The goal in this methodology is to build a database of entities with a structured representation. For example, in the sentence, "We tested an all-wheel-drive XLE model as well, which also delivered more than its 24-mpg promise. " ("Fuel Economy: 2017 Toyota Sienna Fuel Economy Review," 2017), "XLE" would be identified as a car model, and the algorithm would need to add "mpg=24" to that entity. An example of this approach can be found in (Bakalov, Fuxman, Talukdar, & Chakrabarti, 2011). Our approach differs in that we first find the numbers, learn their magnitude and units, and create indicators on the level of the social media posting. For example, instead of storing the fact that the XLE has an mpg of 24, we store the fact that a high mpg was mentioned in the posting. Our representation of the posting retains the original extracted value, but in addition, we discretize the value also, as that is informative for defect classification and for slice-and-dice drill-down of the textual dataset by numeric bands (e.g. rapidly finding all postings mentioning travel at "low speed").

This paper address three main research questions. *First*, can a reasonably-sized set of domain-specific numerical attributes be identified for an industry? *Second*, can these numbers be processed and interpreted automatically? *Third*, are these numbers useful in other information mining tasks, such as defect isolation? We demonstrate in this paper that the answer to all of these question is yes.

Our primary contribution is methodological. We propose a procedure for identifying and classifying a set of domain-relevant numerical attributes with a high level of precision and recall. Furthermore, we demonstrate how these numerical attributes can be

combined with key term extraction to achieve improved performance in defect isolation tasks.

The rest of this paper is organized as follows. In Section 3.2, we provide our theoretical motivation and make the case for generalizability. In Section 3.3, we review two relevant areas: quality management using social media, and processing numerical attributes. In Section 3.4, we detail the procedure for numerical attribute extraction and classification. In Section 3.5, we evaluate the effectiveness of the numerical attributes in locating defects in social media postings. In Section 3.6, we demonstrate another use case for numerical attributes by proposing a Post Market Defect Surveillance System that offers a dynamic interface for exploring a social media data set using numerical attributes as filters and facets. In Section 3.7, we discuss limitations and future work. Finally, in Section 3.8, we discuss our conclusions and implications for research and practice.

3.2. Theoretical Background

We contend that by advancing from a strictly lexical treatment of a numerical token to an interpretation of the number's function and magnitude, we are augmenting the classifier with some degree of semantic understanding. Many attempts to add semantic comprehension to text classifiers have been inspired by findings from cognitive science about how humans process meaning. The methods of artificial intelligence, in many cases, were derived by defining the processes of human intelligence (however narrowly) and approximating them computationally. Examples from natural language processing (NLP) include word sense disambiguation, topic analysis, named entity recognition, and recognizing textual entailment. We argue that our procedure of extracting and binning of numbers would add numerical intelligence to a variety of tasks in several domains, and we make our case from three perspectives: numeracy, specificity, and semantic richness.

Numeracy, which normally develops in childhood (Doig, McRae, & Rowe, 2003), is a basic understanding of numbers and their magnitudes. A variety of mental competencies are associated with numeracy, including estimating, ranking, understanding probabilities and making comparisons. People who are numerate are less likely fall prey to the biases that lead to poor decisions (Peters et al., 2006). A lack of numeracy has been associated with making poor health decisions (Reyna, Nelson, Han, & Dieckmann, 2009) and defaulting on a mortgage (Gerardi, Goette, & Meier, 2013). An estimation of magnitudes is a fundamental part of human intelligence, and our procedure is an attempt to add this competency to NLP systems. For example, our system can compare the age of cars, comprehend when a speed of outside of the normal range, or recognize a voltage that not commonly associated with a model year. It can also combine numbers to make sophisticated observations, so for example it can deduce that in cars with a high mileage, a certain component lasts an unusually long time.

Another reason why numerical intelligence can aid a variety of NLP tasks is that numbers represent an enhancement of specificity. Human language can be abstract and ambiguous, and so utterances that are specific and concrete provide an opportunity for natural language parsers. When people make reference to numbers, they are producing evidence about a specific case. For example, compare "my car doesn't start on cold mornings" to my "2002 model 56x doesn't start when it falls below 32 degrees." In addition to establishing their own credibility and competence, the speakers are acknowledging the particularity of their case, and this particularity should be leveraged for its information content.

Our third theoretical basis for generalizability has to do with semantic richness. In cognitive science, "semantic richness" refers to the amount of information associated with a concept (Kounios et al., 2009) and is a function of the variability of that concept's usage and

contexts. Three measures of semantic richness are commonly used: 1) number of semantic neighbors (NSN), 2) number of features (NOF), and 3) contextual dispersion (CD). NSN refers to the number of words that are used in a similar context to the focal word, NOF refers to how many different attributes of the concept are available in memory, and CD refers to the number of different contexts in which the word is commonly used.

Several experiments have confirmed that words that are semantically rich are understood more quickly and accurately than those that are semantically impoverished. People are able to perform lexical decision and categorization tasks faster and more accurately for more semantically rich words (Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008; Pexman, Lupker, & Hino, 2002; Pexman, Siakaluk, & Yap, 2014; Yap, Tan, Pexman, & Hargreaves, 2011). When a word with few semantic neighbors, few features, and few contexts is presented to a person, more effort is required to arrive at an understanding of the word's meaning.

The aspect of semantic richness most applicable to machine learning is the number of features (NOF). The "features" of a concept are simply attributes that are associated with it, such as, for "grapefruit", <is a fruit> and <is healthy>. (McRae, Cree, Seidenberg, & McNorgan, 2005) found that people are able to list more features of some words than for others. High-NOF words are comprehended more quickly and accurately than low-NOF words (Grondin, Lupker, & McRae, 2006; Yap et al., 2011). An advance in the semantic information supplied to a machine, therefore, is to add features to concepts. In our case study, we take numerical tokens from automobile postings and add features to them. For example, "200" is the lexical token, but we add features "<is a number>", "<is a speed>", and "<is a high speed>." The addition of these features aids the semantic richness of the text as it is presented to the machine, and aids in the machine's "understanding."

The perspectives of numeracy, specificity, and semantic richness provide theoretical justification for our procedure's broad applicability.

Toward a Taxonomy of Consumer Numerical Attributes

To demonstrate the centrality of numerical intelligence to several domains, we provide examples of numerical attributes from other industries and suggest relevant analytical projects. Table 3.2.1 below shows a sample of snippets extracted from a collection of Amazon.com reviews. Any industry whose online text tends to be dense with numbers could find a promising application of this method.

Table 3.2.1. Other number-intensive industries, with potential use cases for number extraction projects.

Industry	**Example Number Snippets**	**Frequent Type**	**Project**
Food	French Vanilla Pump Bottle , **1.5L** (Pack of 2 Licorice Laces - Red , **6lbs** : **18-Count** Pods (Pack of4) 28 Individually Wrapped - **28Oz** Total Food , **650** mg ,150Vegetarian Capsules is approximately140-**150** calories drawback is **18g** of sugar per bottle	package size serving size calorie count nutrition info	Track problematic packaging, what nutrition information people are concerned with Understand competitors' or consumers' actual or preferred serving sizes or bulk quantity sizing; Understand threshold limits for consumer's nutrition concerns (excess sugar or calories)
Higher Education	Class Is Going To Be **750** Points Out Of 1000 Class Averages Were 42 **67** 72 And 52 Our Final Grades On Dec. **25Th** Had Him For Both Hist **171** And Hist 410 an Drop 2 Quizzes And **2** Hw Scores Midterm Final = 2 (**3-4** Page) Essays	grades points grades percent deadline dates course number assessments page length	Explore student grade priorities, roots of student dissatisfaction, identify focal courses. Understand timing of issues, particularly for issue date earlier than reporting

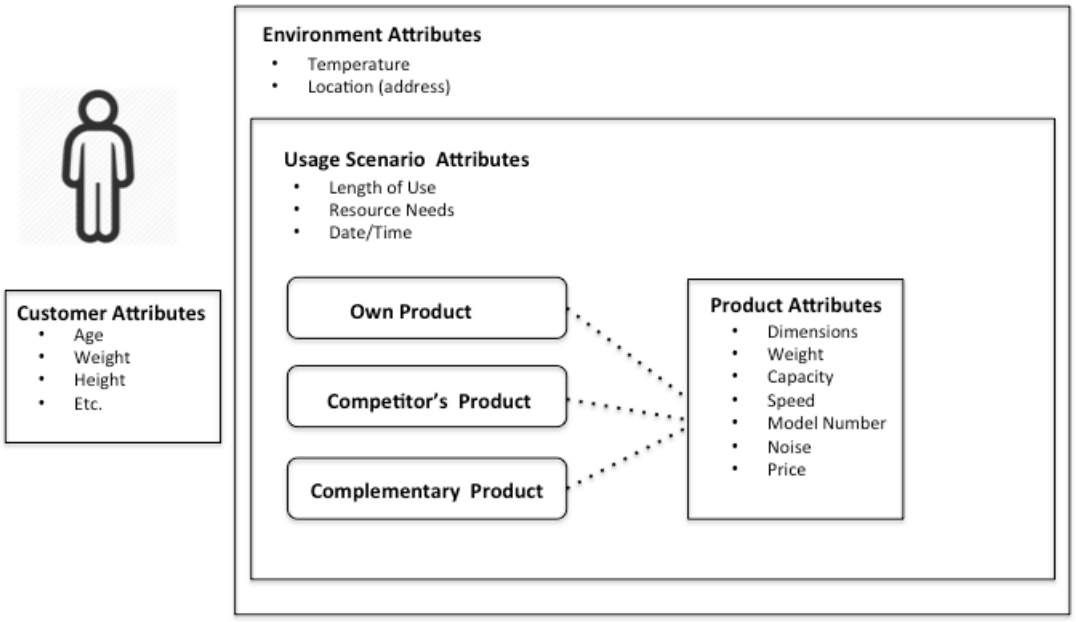
			date.
Banking	ad my bank account for **4** years through me for a $ **3600** line of credit my fico goes down by **80** points women on the order of **98%** that he said made **10%** non compounding yearly mortgage late to the **3** credit bureaus reference to our submission # **218675** . I am **FHA-2291** but also a $ **35** phone fee and then	account duration dollar amount interest percent form number	Better understand consumer choices and needs Understand duration of consumer history, magnitude of competitor fees, correlations between fees and dissatisfaction; credit capacity requirements and credit score magnitudes or changes
Power Tools	lls the Porter-Cable brand **4.5** inch X 10 yard great deal for almost $ **100** I purchased the Dremel 709-11 **110pc** accessory kit this is a **3rd** rate product be careful MANY USES OVER THE PASED **15** YEARS hat is a **3** cutter head hrough were about 6 to **10** inches wide s needed or is the **20** amp breaker in torage shed about **50** feet from my shop	part size price package quant. rating customer loyalty time electrical	Adjust warranty coverage based on time to failure Understand consumer's size (fit), compatibility (e.g. voltage), and capacity (e.g. length) needs Understand assortment variety (e.g. different kit pieces; different cutter heads) Identify focal product models, particularly for competitor comparison and compatibility concerns.
Electronics	Compaq Presario with **2** USB I have a **60**MB zip file on my I purchased this **15ft** . Mediabridge HDMI cable we were getting almost **100m** stream to the tv s switched to the **cat5** connectivity cousin owns a Viewsonic **Pro8500** and even he changed in perspective from **16x10** to	memory size network speed model number data storage	Determine which equipment consumers are using in what combinations Understand size, compatibility and capacity (e.g. length or data volume or speed)

	16x9 disks all the way to **2** TB without breaking a sweat I 'm using Windows 7 (**64** bit)		needs; Identify focal models for compatibility fixes. Identify focal competitor models consumers are using for comparison.
Dishwashers	heavy duty setting takes **2.5** hours , the the size of an **8** oz cup I'm buying other brand after **2** years trying out of the truck (**122lbs**) . There **59** decibels it would be s little as **2.4** gallons of water the temperatures vary from **115 F** to 150 F Fast wash cycle taking only **56** minutes	cycle time capacity customer loyalty length weight gallons temperature	Determine consumer choice priorities, possible co-innovation opportunities Understand consumer requirements related to speed (time to clean), noise, size, capacity, resource (water) utilization, weight.
Credit Cards	to **700** in less than 2 yrs can pay off within **6** months After **7** years of payments have a balance of $ **300** plus dollars without paying off **100** % , you have automatically raised my credit line **4** times i still only have $ **4.00** of credit available store I earn **2** points and I can APR is quite high **19.24** which I do not	credit score payment duration percent dollar amount loyalty points interest rate	Understand duration of issues, or duration of customer history, fees, concerns for consumers within different credit rating segments; understand credit capacity or interest rate utilization or comparison.
Industrial and Scientific Equipment	It is just **24** mm thick features are : *Range : 0.**19** inch - 9.8 feet*Unit ended up picking up about **2** pounds of dust grime and they have a **20** % off coupon	component size	Understand compatibility, dimensions, capacity, and focal specs or focal models for comparison. Understand competitor promotions
Car Seats	seat that accommodates **30** or 35 pounds Addendum : At **9** months he is still s they go out at **90**deg from the back less than **1** inch of movement	child weight child age seat angle price model number	Detecting unsafe usage, improper size for child; Understand buyer or user age, seat fit,

	I am **5 1/2** months pregnant In 2012 , **32** children in the United eplacing a sub- $ **100** seat from a large seat of our BMW **530i** and VW New Jetta		price comparisons; Identify focal vehicle models (e.g. for prioritizing compatibility fixes)

Inspection of the number types commonly found in online reviews, such as those in Table 3.2.1, suggest a taxonomy of the numbers consumers typically reference. Although the numbers will vary across applications, we provide this taxonomy in Figure 3.2.1 as an early step toward a framework for consumer numbers to guide in the construction of future projects.

Figure 3.2.1. A taxonomy of consumer numerical attributes.



3.3. Related Work

3.3.1 Quality Management Using Social Media

Social media provide a rich source of information on customer experience. Studies have confirmed that businesses can gain insights into quality of service in a variety of industries, including hospitality (Berezina, Bilgihan, Cobanoglu, & Okumus, 2016) and healthcare (Greaves, Ramirez-Cano, Millett, Darzi, & Donaldson, 2013).

The most formidable challenge in using social media, especially for regulatory organizations, is the sheer volume of User Generated Content (UGC) being produced every day (Abrahams et al., 2013). Latest numbers indicate that Amazon has over 140 million product reviews (McAuley et al., 2015), TripAdvisor has over 570 million

(TripAdvisor.com), and Yelp has over 142 million (https://www.yelp.com/press). Any system that aims to simplify the discovery of useful information in social media must provide an effective means of filtering out noise and leaving a manageable subset of user postings for further processing by human readers.

Several studies have proposed intelligent means of sifting through social media content for particular information needs. Methods have been proposed to predict stock market activity (L. Liu, Wu, Li, & Li, 2015; T. H. Nguyen, Shirai, & Velcin, 2015), discover socially important locations (Dokuz & Celik, 2017), find important information for crisis management (Pohl, Bouchachia, & Hellwagner, 2018), and gauging customer response to business ads (Jang, Sim, Lee, & Kwon, 2013), to name a few.

One area of active research is the application the text analytic methods to the detection of hazardous product defects in social media. A framework recommending a set of lexical, stylistic, social, sentiment, product, term, and semantic features was introduced in (Abrahams et al., 2015) and confirmed in (Adams et al., 2017; Law et al., 2017; Winkler et al., 2016). A consistent finding among these studies is that distinctive terms, product features, and semantic attributes are helpful to distinguish defects, but stylistic and sentiment features are not. This approach was later extended with a set of contextual features and ensemble methods in (Y. Liu, C. Jiang, & H. Zhao, 2017), and with heuristic methods in (Goldberg & Abrahams, 2017). A procedure for identifying vehicle components to enrich the information set was proposed in (Abrahams et al., 2013). (Bhat & Culotta, 2017) proposed an unsupervised method of defect detection using unlabeled Amazon.com reviews combined with "positive" labeled documents from the CPSC website SaferProducts.gov. Their positive unlabeled learning (PUL) approach provides an improvement in accuracy and a reduction in labor, but their findings may not be extendible to other domains.

3.3.2 Processing Numerical Expressions

The automatic identification and interpretation of numerical quantities is helpful in variety of NLP tasks. It is a critical subtask within question answering (QA) (Yaqing Liu, Wang, Chen, Song, & Cai, 2016). For example, in the 2002 TREC-10 QA contest, Hovy (Hovy, Hermjakob, Lin, & Ravichandran, 2002) showed the importance of constraining answers according to sensible ranges for the target variable; e.g., a nation's population should not be a small number like 10. Akiba et al. (2004) extended this reasoning and argued that numerical expressions can be viewed as a random sample following a Gaussian distribution, and therefore probabilities can be assigned to individual realizations. For example, in our data set, numbers around 60 are likely for a speed (in mph), but numbers greater than 140 are multiple standard deviations too high.

Learning the numerical attributes associated with physical objects can help with certain inferential tasks. For example, knowing the typical dimensions of a strawberry can help a computer vision algorithm to decide whether to label a red object in an image as a strawberry (Takamura & Tsujii, 2015). Davidov & Rappoport (2010) showed that the application of knowledge about object dimensions can help select the appropriate numerical expressions in a sentence when there are several candidates. In this experiment, distributions of sensible size and weight ranges for physical objects were mined from web text by searching for similar objects using WordNet; for example, a "Toyota" and a "Honda" should have approximately the same distribution of widths. The notion of sensible distributions of numerical attributes was termed "numerical common sense" in (Narisawa, Watanabe, Mizuno, Okazaki, & Inui, 2013). In this study, the researchers were able to determine whether a number was uncommonly large, small, or normal, based on context ("the camera weighs *only* 2 pounds"). Banerjee et al. (2009) used web mining to aggregate

numerical references to a particular object and form tight intervals around an answer to a magnitude question.

Domain-specific numerical quantities have been examined in (Mandhan & Niwa, 2016) and (Rubens & Agarwal, 2002) but the emphasis in these works is on accurate extraction, rather than on evaluating applications. The Rubens & Agarwal (2002) study deals specifically with automotive numerical attributes, but the ontology is restricted (year, price, and mileage) and the technique makes heavy use of the structural elements of a classified ad. Our approach employs a greater variety of numerical attributes, and recognition is not dependent on the structural elements of the document.

To our knowledge, no research has investigated the impact of numerical attributes on the performance of text-based defect detection systems. This study aims to fill this gap, using the highly complex automotive domain as an illustrative case.

3.4. Methodology

Extracting numerical attributes can be performed either through rule-based string matching approach or a corpus-based machine-learning approach. A rule-based system would entail writing a set of rules for classifying strings; for example, a regular expression like "/\d[st|rd|th]\sgear/" would detect some (but not all) references to transmission gears. A corpus-based machine learning approach, on the other hand, would involve labeling number snippets and allowing the computer to figure out which features matter and how much. Rule-based systems tend to be overly specific, and difficult to maintain. We propose that a machine-learning methodology is also more generalizable to other domains. We therefore extract and classify numbers by labeling instances and training a classifier.

In Section 3.4.1, we describe our procedure for building the number extractor and classifier. In Section 3.4.2 we describe how we add the numerical attributes as features for defect detection.

3.4.1 *Extracting and Classifying Numerical Attributes*

The classifier training data set is described in (Abrahams et al., 2012), and includes 1500 discussion threads each from Honda-Tech.com, ToyotaNation.com , and ChevroletForum.com, for a total of 4500 threads. 113,355 numbers were extracted using the regular expression [-+]?[A-Za-z]*\d+[A-Za-z]*[\.]?[\,]?\d*, along with a context of 40 previous tokens and 40 following tokens. 35,000 of these "number snippets" were randomly selected for human tagging. Listing 3.4.1 gives examples of number snippets.

Listing 3.4.1. Examples of "Number snippets" extracted from social media postings.

Model Year: [1998] weak ABS question ... Im stumped Well This is not often I need help but this one got me . I know alot of you guys are gods with honda wring so here we go . -Everything Was Fine (===1998===Honda Civic) -Parked car for a week to remove dash -After replacing Dash , Airbags , ECU , and ABS COntrol module , I have ABS Light -Had An SRS light to but I reset that with a

Transmission: [5-speed] are the location of starter relay , clutch start switch and oxygen sensor (bank 1 sensor 1 and sensor 2) . Thank you , Is the clutch in all the way ? I just got a 2004 Matrix ===5-speed===and unless I have the clutch down ALL the way to the floor it will not start either . ^Good point . In most cases , it s simply the button on the floor that is n t depressed

Distance: [650~km] ... the bobber or somethin 1996 Corolla 1.66L - about 400 miles but I travel about 360mi HWY once a week . I went on a few trips over the summer in my 97 AE102 and had consistently gotten ===650~km===or so on 90 /10 % highway/city driving . I still had about 1/8th of the tank left though . I shall start =) 96 DX 1.8L 4 speed auto . I get approx . 325 miles with

A pilot study was conducted to establish a comprehensive list of number types. Undergraduate business students tagged postings from Cars.com using an initial set of 11

number types. Taggers were instructed to write in the number type if the number did not fit into any of the given categories. Based on the results of this study, a list of 45 number types was settled on (see Appendix 3.A for a complete list) for final tagging.

Using the dataset of 35,000 number snippets, 109 Master of Information Technology students were given extra credit to tag numbers using PamTag, a web-based collaborative tagging tool (https://pamtag.pamplin.vt.edu/pamtag/description.html). The instructions and protocol appear in Appendix A. 74 taggers completed the minimum of 200 tags. A gold standard authority set of 927 tags was completed by the lead researcher to check tagger accuracy, and all submissions from taggers with less than 50% agreement with the authority were discarded. After poor performers were dropped, the final set consisted of 20,850 tags (for 19,815 distinct number snippets). Tag comments were inspected for common problems. 104 tags marked "other" contained comments that they were "RPM" (Revolutions Per Minute) readings, so RPM was added as a number type. 61 tags indicated that the number referred to a gear, and 50 tags indicated that the number referred to an oil grade (e.g., 10w40), so Gear and Oil Grade were also added as potentially significant number types. It was also observed that automobile community participants frequently referred to the car model by its generation ("1st gen", "2nd gen", etc.), so Generation was added where appropriate. After these adjustment, average tagger accuracy was 73% (μ=.74, σ=.17).

788 number snippets were tagged by 2 or more taggers (max=4), and inter-rater reliability was strong, with a Cohen's κ of .72, indicating "substantial agreement" (Landis & Koch, 1977). In cases where taggers disagreed, conflicts were resolved using the following rules: 1) If there was an authority tag, it overruled the others. 2) If there was no authority tag, majority won. 3) If there was a tie, the lead researcher (authority tagger) selected the final tag. A total of 19,815 distinct number snippets were tagged from 5,434 threads. Count

of numbers per thread ranged from 1 to 257 with an average of 3.6 number snippets per thread.

An assessment was conducted to see which number types gave taggers the most difficulty. The percentage agreement with the authority for each number type where agreement was greater than 60% appears below in Table 3.4.1.

Table 3.4.1. Percent agreement with authority by field.

Temperature	100%
Calendar Day	100%
Tire Pressure (PSI or kPa)	100%
Rating (Stars)	100%
Word (2 for 'to')	100%
Age of Vehicle	100%
RPM	100%
Oil Grade	100%
Volume	100%
Weight	100%
Calendar Date	100%
Odometer Reading (# Miles)	100%
Count of Doors	100%
Dollar Amount-Price-Currency	96.9%
Model Year	90.8%
Time Duration	90.2%
Speed (mph-kph)	85.7%
Error Code	83.3%
Engine Piston Count (V8-straight 6)	83.3%
Transmission (5 speed-6 speed)	81.3%
Chemical Symbol for Gas (O2-CO2-NO2)	75.0%
Wheel Drive (2WD-4WD-WD-4x4)	75.0%
Gear	75.0%
Length/Height Measure	75.0%
Model Number of Vehicle	68.4%
Listing (1.- 2.)	66.7%
Electrical (watts-amps-volts)	66.7%
Torque	66.7%
Distance Traveled	66.7%
Other	61.0%

Percent	60.0%
Model Number of Component	60.0%
Count Other	60.0%
Tire Model	60.0%
Phone	60.0%
Horsepower	60.0%
Word (2 for"to")	60.0%
Engine Piston Size (liters (L) or cubic centimeters (cc))	60.0%

Numbers with less than 60% agreement were dropped from the analysis. The following categories were dropped, either because the agreement was too low or because too few training examples were available in the training set to ascertain reliability: Calendar Year, Rank (1st-2nd), Age Other, Age of Person, VIN (Vehicle Identification Number), Calendar Month, Passenger Capacity, and Word. The "Word" category comprises inventive respellings of common words with numbers substituted for letters, and examples include b4, st00pid, 2, ub3r, w00t, any1, v3t3c (vetec), 4, y0, inf0, 0well, n00bs, and ph00kin. These are not really numbers, and their inclusion in the tagging was strictly to allow taggers to accurately interpret these numbers.

After problematic categories were dropped, 19,431 distinct number snippets remained.

The lead researcher spot-checked the miscellaneous categories ("Other" and "Count of Other") to check for mis-taggings and potential new number categories. Based on the frequency of octane mentions, another new category was created for "Fuel Octane", consisting of 81 new observations.

The final counts of number type, after all additions and corrections were made, appear in Appendix 3.B. The most common numbers were Model Year (3944), Model Number of Component(2106), Other(1519), Count Other(1358), Dollar Amount (925),

Model Number of Vehicle (873), Time Duration (810), Odometer Reading (765), Listing (699), Engine Piston Count(649), and Engine Cylinder Size(484).

The miscellaneous categories were dropped, as they were too heterogeneous. Due to the extreme imbalance in the representation frequency of the number classes, the classes were either sub-sampled or bootstrapped[3] to create a set of 300 instances of each number category, except for Rating, Address, and Calendar Year, which had very few instances (< 20). Imbalanced classes introduce problems in classification algorithms (Charte, Rivera, del Jesus, & Herrera, 2015), and resampling has been shown empirically to be an effective way of handling class imbalance (López, Fernández, García, Palade, & Herrera, 2013).

Morphological attributes of the number itself play a key role in its classification. Several numbers were discernable based solely on characters. For example, single-characters marked dollar amount ('$'), odometer readings ('k'), oil grade ('W'), percent ('%'), and error codes[4] (which usually start with 'P' or 'U' or 'B'), time of day (':') ; double character sequences marked cylinder configurations ('I4', 'V8'), lengths ('mm', 'cm', 'ft') and gears ('st', 'nd', 'rd'); and triple character sequences marked model generation ('gen'), Pressure ('psi'), speed('mph'), and RPM ('rpm').

In addition to character sequences, the magnitude of the number also revealed something about its class. For example, "Count Other" was rarely larger that 10 ("5 times", "3 wires", "4 lug nuts", "2 engines", etc.), whereas Odometer was rarely smaller than 10,000. Horsepower ranged from 100-400, and Fuel Octane was always 85-93. Therefore, character sequences and number magnitude were critical features in our classification.

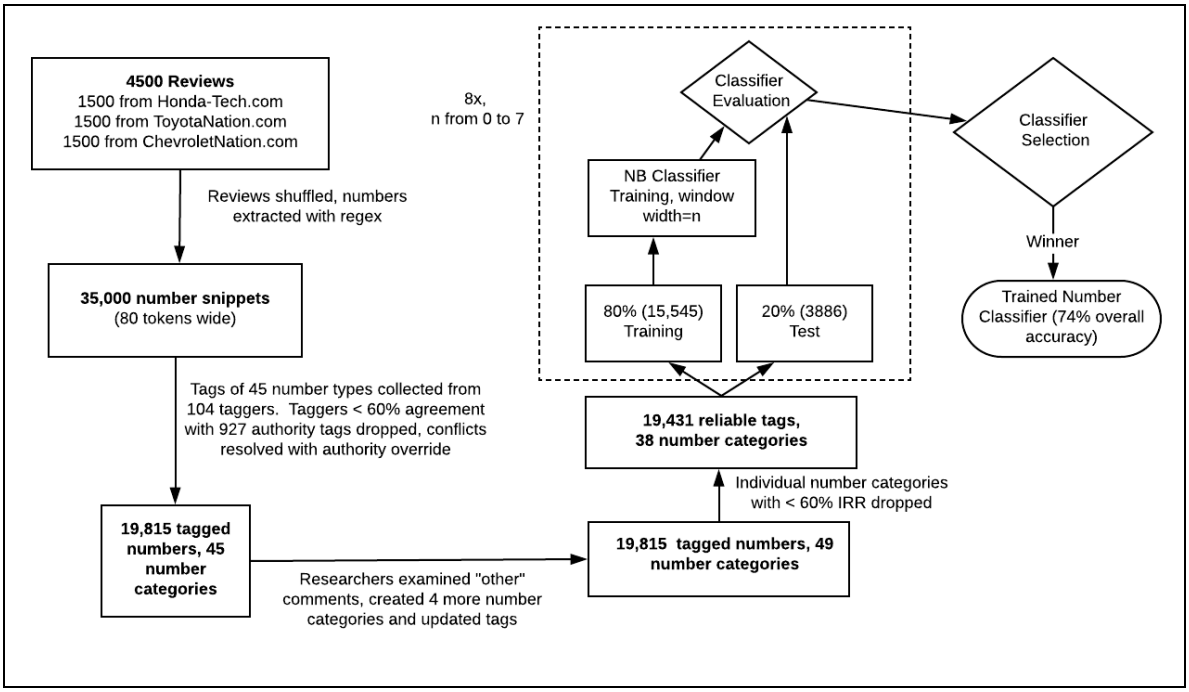
---

[3] In subsampling, a random set of observations of size n is selected (Doucette & Heywood, 2008). In bootstrapping, a set of size n is created through sampling with replacement (Batuwita & Palade, 2013).

[4] Error codes were typically OBD II trouble codes. OBD II (On Board Diagnostics) Diagnostic Trouble Codes (DTCs) are defined in the J2012_201612 "Diagnostic Trouble Code Definitions" from the Vehicle E E System Diagnostic Standards Committee, published by SAE International, and ISO 15031-6:2005.

Eight multinomial Na ve Bayes classifiers were constructed on automatically-extracted features based on widening windows of terms around the focal number. Classifiers were built in Python 2.7.6 using the Natural Language Toolkit (http://www.nltk.org/), and metrics were calculated using scikit-learn (http://scikit-learn.org/). For each classifier, a fresh 80-20 train-test split was created from the tagged data. The first classifier used only morphological elements of the term itself (character sequences of length 1,2, and 3), and the others used window widths of 1, 2, 3, 4, 5, 6, and 7 terms on either side. The complete process for building the number classifier is detailed in Figure 3.4.1.

Figure 3.4.1. Process for building the number classifier.



Number expressions were tokenized on white space and were preserved as strings including all characters, e.g., “15mph”, during classification, to make use of helpful character sequences. Obtaining the number magnitude for binning by stripping out non-numeric characters was delayed until the numbers were extracted and classified from the full set. Feature sets consisted of the character sequences of length 1, 2, and 3 that occurred more than 20 times in the numbers in the training data. Word sequences of varying length (1-7) from a window around the number were also used (see Figure 3.4.1). The vocabulary for each classifier included only terms within the window size around each number in the training data. Vocabulary sizes and examples for each window size appear below in Table 3.4.2.

Figure 3.4.1. Window sizes around the numerical expression.

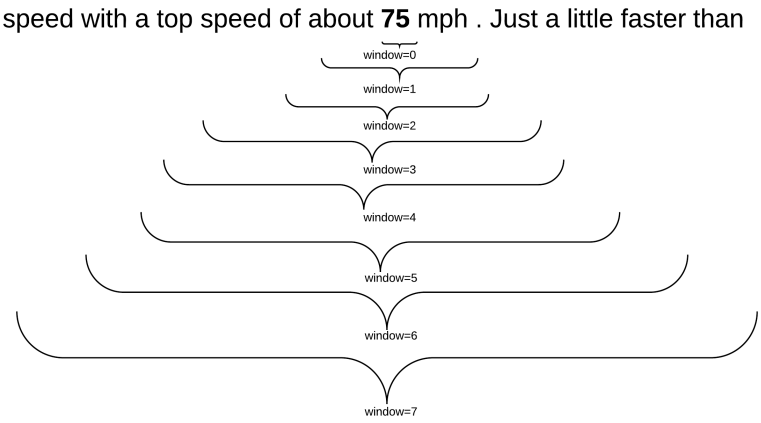


Table 3.4.2. Vocabulary lengths and examples for all window sizes.

Window	Vocabulary Length	Example
0	0	75
1	182	about 75 mph
2	448	of about 75 mph .
3	735	speed of about 75 mph . Just
4	1034	top speed of about 75 mph . Just a
5	1235	a top speed of about 75 mph . Just a little
6	1525	with a top speed of about 75 mph . Just a little faster
7	1786	speed with a top speed of about 75 mph . Just a little faster than

Metrics for recall, precision, and F1 (as defined in (Yiming Yang, 1999)) for the best classifier (window=5) are reported in Table 3.4.3. Table 3.4.3 (column 2) shows illustrative examples of each number type – for additional examples, see Appendix 3.C. The classifier was able to achieve 40% accuracy using only morphological attributes of the number itself. With a window of 1, accuracy jumped to 60%. At window size 5, overall accuracy reached 71%, where it plateaued as the window widened to 6 and 7. We therefore concluded that the 5-window model was the correct tradeoff of accuracy and parsimony. Some additional features were hand-tuned into the 5-window classifier, raising

the overall accuracy to 73%. Specifically, the following were added as features for the number strings, since they were frequently appended the numerals and were longer than 3 characters: 'door', 'amps', 'volts', 'liter', 'litre', 'year', 'mile'.

Table 3.4.3. Best number classifier performance (window=5). See Appendix B for examples of each number type from our data set.

Number Type	Precision	Recall	F1
Address (including streets-zip codes-highways)	0.941	0.955	0.948
Oil Grade	0.935	0.921	0.928
Age of Vehicle	0.853	1.000	0.921
Calendar Date	0.918	0.862	0.889
Phone	0.797	1.000	0.887
Calendar Day	0.893	0.862	0.877
Rating (Stars)	0.882	0.849	0.865
RPM	0.794	0.915	0.850
Fuel Octane	0.794	0.877	0.833
Gear	0.817	0.829	0.823
Temperature	0.806	0.833	0.820
Generation	0.771	0.871	0.818
Fuel Efficiency (MPG-KPG)	0.797	0.839	0.817
Chemical Symbol for Gas (O2-CO2-NO2)	0.727	0.906	0.807
Torque	0.807	0.780	0.793
Time of Day	0.860	0.729	0.789
Wheel Drive (2WD-4WD-WD-4x4)	0.894	0.700	0.785
Percent	0.785	0.761	0.773
Pressure (PSI or kPa)	0.676	0.889	0.768
Model Year	0.764	0.757	0.760
Engine Piston Size (liters (L) or cubic centimeters (cc))	0.677	0.840	0.750
Weight	0.820	0.685	0.746
Count of Vehicles	0.774	0.719	0.745
Speed (mph-kph)	0.764	0.724	0.743
Engine Piston Count (V8-straight 6)	0.679	0.792	0.731
Volume (Non-Engine-gallons)	0.771	0.667	0.715
Dollar Amount-Price-Currency	0.745	0.660	0.700

Distance Traveled	0.733	0.657	0.693
Horsepower	0.685	0.698	0.692
Count of Doors	0.571	0.857	0.686
Transmission (5 speed-6 speed)	0.939	0.525	0.674
Electrical (watts-amps-volts)	0.822	0.569	0.673
Odometer Reading (# Miles)	0.597	0.714	0.650
Error Code	0.661	0.639	0.650
Tire Model	0.635	0.647	0.641
Time Duration	0.618	0.630	0.624
Word (2 for to)	0.492	0.517	0.504
Length/Height Measure	0.325	0.441	0.374
Listing (1.- 2.)	0.352	0.373	0.362
Model Number of Component	0.412	0.241	0.304
Model Number of Vehicle	0.326	0.215	0.259
VIN (Vehicle Identification Number)	None	None	None
Passenger Capacity (4 seats-7-passenger van)	None	None	None
Age Other	None	None	None
Age of Person	None	None	None
Calendar Month	None	None	None
Rank (1st-2nd)	None	None	None

Number categories with an F1 score of less than .65 were considered unreliable and were dropped from consideration.

Some of the challenges in demarcating and classifying the numbers in social media postings can be seen in the examples in Appendix 3.C. There are inconsistent units, abbreviations, spellings, spacing, capitalization, and punctuations. For example, 20 foot-pounds of torque can be written 20 ft.-lbs., 20 foot/pounds, 20 ft/lbs, 20 torque, 20tq, 20 tq, tq. 20, or simply 20. Especially challenging is the disagreement about whether units are part of the number itself (50W, 50watts, 50 watts, 50wts, etc.; or 100lbs, 100 lbs., 100-pounds, 100 pounds, etc.).

Thus, in converting the numerical attributes to features, care had to be taken to derive the correct interpretation of each numerical string. Expressions of the same concept

had different abbreviations, scales, and units. For example, odometer readings, which appear in listing 3.4.1, were written in a variety of different manners: miles vs. kilometers, ranges with hyphens, 'k' for thousand, x for 0's, commas in the wrong place, sometimes '60' intending 60,000, etc. Listing 3.4.1 shows some of the ways of expressing speed. Transmission was always 2,3,4,5, or 6.

Listing 3.4.1. Different "families" of number expressions from our training data set.

Odometer Reading	**Speed**
'8000km', '50km', '20,000km', ' '109k', '5k', '137k', '37k', '201k' '180000', '15000', '250000' '50xxx', '112,XXX' '57,000', '81,000', '15,000', '103,000', '136,000' '100,000+', '100k+', '200k+' '7900ish' 'w/4k+', 'w/100k' '35-45k', '100-150k', '225,000-250,000'	'0-60mph', '0to60', '60-0mph', '40-45mph' '5mph', '30mph', '=62mph' '0-60', '35-40', '20-25' '.88 km'/h', '55kms', '5km/h' '28', '75', '0', '70' '30+', '60+', '45+'

The count of distinct strings observed for each number type in the data set appear in column 2 of Table 3.B2 in Appendix 3.B.

Discretizing (binning) the Numbers

Since each number type could have an infinite number of values, it was necessary to group (bin) the numbers to derive some kind of interpretable magnitude. To discretize the numbers, we employed the following procedures: In many cases, we found the numeric part of the string, normalized the units, and created a variable indicating whether the number was above or below the median. (Alternative binning approaches are possible – e.g. a domain expert could define low vs medium vs high mileage – however, we chose to employ a simple median-based discretization approach.) In some cases, there were few enough distinct values that we were able to create categorical variables. For example, for "Chemical

Symbol for Gas", our entire training set contained only 6 distinct gases--O2 (oxygen), O3 (ozone), N2O (nitrous oxide), H2 (hydrogen), HO2 (hyroperoxyl, an emission), and NO2 (nitrogen dioxide, also an emission)—so we added a categorical dummy field for the specific gas mentioned. Wheel Drive was always either 2 or 4. Wheel drive, gear, doors, oil grade, and transmission are coded as numerical nominal data, while all others are binary nominal (0-1). For most number categories, we included a dummy variable indicating that it was mentioned, along with two other dummy variables indicating whether the value was above the median or below the median. These are summarized in Table 3.4.4.

Table 3.4.4. Rules for converting numerical strings to social media posting features.

Number Category	Processing Rule	Median	Features
Age of Vehicle	Use context to determine units and normalize to months. Split on median.	72 Months	vehicle_age vehicle_age_low vehicle_age_high
Chemical Symbol for Gas	Remove non-numeric, normalize capitalization, map to categorical variable		gas gas_o2 gas_o3 gas_n2o gas_h2 gas_ho2 gas_no2
Count of Doors	Remove non-numeric characters, discard if < 2 or > 6		doors
Count of Vehicles		4	count_vehicles count_vehicles_low count_vehicles_high
Generation		5	generation_low generation_high
Gear		3	gear
Distance Traveled	Convert k to thousand. Use context to determine units and normalize to miles. Split on median.	305 miles	distance distance_low distance_high
Dollar Amount		253	dollar dollar_high

			dollar_low
Electrical			electrical_present
Engine Cylinder Size		2.7 liters	engine_size engine_size_low engine_size_high
Engine Pistons			engine_pistons
Error Code			error_code
Fuel Octane		91	fuel_octane fuel_octane_low fuel_octaine_high
Fuel Efficiency		30 mpg	fuel_eff fuel_eff_low fuel_eff_high
Horsepower		200	hp hp_low hp_high
Model Year	Split on hyphen and take the first section. Remove any additional punctuation. Convert 2-year strings to 4-years (for <20, prepend '20', otherwise prepend '19'). Discard outliers. Min is 1987, max is 2008, median 1997.	1997	model_year model_year_low model_year_high
Odometer Reading	Convert k to thousand. Use context to determine units and normalize to miles. Split on median. Low is < 10k, middle is 10-50k, high is > 50k.	80,000 miles	odometer odometer_low odometer_middle odometer_high
Speed	If hyphen, it's a range. Otherwise, take numeric part and use context to determine whether mph or kmh. Discard outliers and split on median.	55 mph	speed speed_range speed_low speed_high
Temperature	Use context to determine units (Celsius or Fahrenheit). Split on median.	100 F	temperature temperature_low temperature_high
Tire Model			tire_model
Oil Grade	Expressed as a cold temperature starting		oil oil_low_temp

	performance and a viscosity at 100 degrees, e.g., 10w-40.		oil_viscosity
Pressure	Remove non-numeric and split on median. Convert kilopascals to psi	32 psi	pressure pressure_low pressure_high
RPM	Remove non-numeric and split on median	3000	rpm rpm_low rpm_high
Torque	Remove non-numeric and split on median	195 ft-lbs	torque high_torque low_torque
Transmission	Use number to determine 2,3,4,5, or 6 speed	5 speed	trans_speed
Weight	Use context to determine units. Normalize to pounds.	150 lbs	weight weight_low weight_high
Wheel Drive	Use number to determine whether 2 or 4.	4	wheel_drive

3.4.2. Adding numerical attributes as features of postings for defect detection

We now demonstrate a pragmatic application of numerical attributes as features: enhancing product defect discovery from social media. We will use the data set from (Abrahams et al., 2015), in which 4500 social media postings were labeled as defect = yes/no (1/0). (As in the prior research, we have collapsed performance and safety defects into a single category, titled "defects"). We hypothesize that some of the numerical attributes will make effective features for differentiating automotive social media postings that indicate a manufacturing or design defect.

Although any of the number categories could be useful in some application—say, aspect-based information retrieval—for our particular case study, we select a subset of numbers that might have some bearing on the improper functioning of the automobile. For example, whereas address and phone number might help to locate a dealership

geographically, and a star rating might help to rank cars by customer satisfaction, they are unlikely to help address the functionality of the automobile.

For these reasons, and also to prevent classifier-model over-specification, we concentrate on the numbers listed in 3.4.4. Our theoretical basis for suspecting there is a relationship and the hypothesized direction of the relationship appear in Table 3.4.5.

Table 3.4.5. Hypothesized relevant numerical attributes.

Variable	**Theoretical basis**	**Direction**
vehicle_age_low vehicle_age_high	Defects apply to newer cars, whereas problems with older cars are the result of wear	positive negative
count_vehicles_low count_vehicles_high	Low numbers might indicate repeat purchase from loyal customers, high numbers might indicate vehicles affected by defect.	negative positive
generation_low generation_high	Defects apply to newer cars, whereas problems with older cars are expected	negative positive
gear_present gear	As the most complicated component in a car, apart from advanced electronics, we speculate that transmissions will be prone to defects. Therefore, mentions of gears will be associated with greater defect likelihood.	positive
distance_low distance_high	Defects appear early on during a trip, so low trip distance is more likely to be associated with defects.	positive negative
electrical_present	Due the complexity of electrical components, we suggest that mentions of electrical components will increase likelihood of defect.	positive
engine_size_low engine_size_high	We speculate that larger engines are more prone to problems than smaller ones.	negative positive
error_code	While error codes do not indicate defects, the fact that a consumer is mentioning error codes might suggest a problem with the car.	positive
hp_low hp_high	We speculate that high horsepower vehicles are more prone to defects, due to added performance complexity.	negative positive
model_year_low model_year_high	Newer vehicles are more likely to have undiscovered defects, or owners are more likely to be sensitive to problems experienced shortly after purchase of a recent model-year (owners of older model years may expect and accept defects due to vehicle age), so higher model years will be associated with greater defect likelihood.	negative positive
odometer_low	Newer vehicles are not expected to have	positive

odometer_middle odometer_high	problems, so lower odometer readings will be associated with greater defect likelihood. Middle and higher mileage cars will naturally develop age-related problems over time that are not indicative of manufacturing or design defects.	negative negative
tire_model	Tire defects are entire category of defects according to the NHTSA (https://www-odi.nhtsa.dot.gov/owners/SearchSafetyIssues). We therefore hypothesize that mentions of tires will linked to greater defect likelihood.	positive
rpm_low rpm_high	. High motor revolutions per minute (RPM) puts excessive strain on the vehicle engine, and are more likely to be associated with vehicle defects.	negative positive

Using the number classifier and the rules in Table 3.4.3, we added the numerical attributes to the 4500 postings from Honda-Tech.com, ToyotaNation.com , and ChevroletForum.com. Table 3B3 in Appendix 3B shows the count of each number type identified. Note that the "presence" variables do not always equal the sum of the low and high, due to two facts: 1) a single posting sometimes contained both high and low values for a number (see Table B4 for exact counts) and 2) in 81 cases, the program was unable to determine the magnitude of the number due to unpredictable text formatting, and the program threw an error [5]. Although we retain the presence of dummy variables in the data set as a convenience for other applications, they are not included in the regression because their value is completely determined by the value of the "low" and "high" variables. The three variables together thus function as a single categorical variable, and we exclude the base level.

We then ran a series of logistic regressions to find the best predictive model. All regressions were conducted with JMP Pro Version 13.0, and accuracy measures were computed using the supplied confusion matrices.

[5] One interesting application of these multiple occurrences might be in filtering noisy posts. For example, we had one post with *60* dollar amounts, and it turned out that a user had pasted a large section of a parts catalog.

3.5 Results

Table 3.5.1 summarizes the logistic regression results. To ensure that we were not unjustifiably benefitting from sampling error, we trained all models on the Honda and Toyota sets (n=3000) and tested on the unseen Chevrolet set (n=1500). We report R^2, precision, recall, and AUC for each of 5 models for both the training and test sets.

We begin with the Model 1 described in (Abrahams et al., 2015), which includes context-independent features (word count, Fog index, and average word length), social features (number of views, number of user), and sentiment features (6 principle components). Model 2, also replicated from (Abrahams et al., 2015), includes context-specific features (smoke words, product features, and semantic principle components). Model 3 is the full combined model from (Abrahams et al., 2015). Model 4 consists of only the hypothesized numerical attributes. Model 5 is the full model, with all variables included.

Table 3.5.1. Logistic regression results.

Variable	Model 1 (context independent)	Model 2 (context specific)	Model 3 (full 2015 model)	Model 4 (new numerical attributes only)	Model 5 (full model)
ZwordCount	-0.058***		0.032***		0.012
ZFogIndex	0.101***		0.159***		0.135***
ZAverageWordLength	-0.113***		-0.406***		-0.418***
Zviews	0.041***		-0.100		-0.122***
ZnumOfUsers	-0.106***		0.001***		-0.002
SentiFAC1_2	-0.021		-0.173***		-0.158***
SentiFAC2_2	-0.405***		-0.122***		-0.087***
SentiFAC3_2	0.015		-0.065***		-0.066***
SentiFAC4_2	0.201***		0.211***		0.243***
SentiFAC5_2	-0.150***		-0.189***		-0.191***
SentiFAC6_2	0.004		0.063***		0.089***

ZSmokeWord		1.469***	1.714***		1.736***
AirConditioning[0]		-0.533***	-0.382***		-0.542***
Airbag[0]		-0.257***	-0.160***		-0.180***
Braking[0]		-0.697***	-0.693***		-0.760***
Electricalsystem[0]		-0.353***	-0.316***		-0.409***
Engine[0]		-0.314***	-0.286***		-0.313***
Lights[0]		-0.140**	-0.144**		-0.146**
SeatBelts[0]		-0.261***	-0.276***		-0.312***
Steering[0]		-0.738***	-0.741***		-0.694***
StructureandBody[0]		-0.562***	-0.487***		-0.543***
Transmission[0]		-0.278***	-0.307**		-0.134***
Visibility[0]		-0.093	-0.008		-0.034
WheelsandTires[0]		-0.175***	-0.203***		-0.256***
Other[0]		0.739***	0.706***		0.743
Suspension[0]		-0.141***	-0.124**		-0.099
Acoustics[0]		-0.399***	-0.384***		-0.326***
SemanFAC1_1		-0.096***	-0.233***		-0.208***
SemanFAC2_1		-0.050***	-0.071***		-0.100***
SemanFAC3_1		0.219***	0.311***		0.136***
SemanFAC4_1		-0.130***	-0.159***		-0.114***
SemanFAC5_1		-0.094***	-0.150***		-0.170***
SemanFAC6_1		0.011	0.077***		0.058***
SemanFAC7_1		-0.007	-0.020		-0.006
SemanFAC8_1		0.085***	0.055***		-0.005
vehicle_age_low[0]				0.266***	0.414***
vehicle_age_high[0]				0.494***	0.230***
count_vehicles_low[0]				0.106***	-0.007***
count_vehicles_high[0]				0.018***	-0.089
generation_low[0]				0.161***	0.115***
generation_high[0]				0.291***	0.104***
gear_pres[0]				0.031	-0.144***
gear				-0.002	0.000
distance_low[0]				0.211***	0.097***
distance_high[0]				-0.002	-0.221***
electrical_present[0]				0.004	0.002
error_code_present[0]				-0.226***	-0.114***
hp_low[0]				0.376***	0.274***
hp_high[0]				0.474***	0.177***
model_year_low[0]				-0.015	0.251***

model_year_high[0]				-0.317***	-0.047***
odometer_low[0]				0.191***	-0.345***
odometer_middle[0]				-0.311***	-0.070***
odometer_high[0]				0.037***	-0.440***
tire_model_present[0]				0.274***	-0.252***
rpm_low[0]				0.196***	0.326***
rpm_high[0]				0.088***	0.125
R-squared	0.014	0.219	0.228	0.100	0.250
	Training				
Precision	0.549	0.724	0.726	0.625	0.745
Recall	0.271	0.223	0.224	0.187	0.211
F1	0.363	0.341	0.343	0.287	0.323
AUC	**0.576**	**0.807**	**0.811**	**0.710**	**0.823**
	Test				
Precision	0.752	0.830	0.834	0.743	0.846
Recall	0.452	0.240	0.249	0.272	0.266
F1	0.565	0.372	0.384	0.398	0.404
AUC	**0.558**	**0.685**	**0.692**	**0.610**	**0.716**

**: significant at the 1% significance level
***: significant at the .1% level
Cutoff for precision, recall, and f1 was .5

We note first that the addition of the numerical attributes results in a modest improvement in precision, recall, F1, and AUC on both training and holdout data. Most of the numerical attributes contributed significantly to variance in Model 5 (full model). Table 3.5.2 lists the variables in descending order of LogWorth[6]. Six of the numerical attributes were significant (distance_high, hp_low, hp_high, model_year_high, odometer_middle, and odometer_high) and all except model_year_high had directional effects as hypothesized. The results of the hypothesis tests are summarized in Table 3.5.3.

[6] From the SAS Jmp Documentation: "When you have large effects, the associated *p*-values are often very small. Visualizing these small values graphically can be challenging. When transformed to the LogWorth ($-\log_{10}(p\text{-value})$) scale, highly significant *p*-values have large LogWorths and nonsignificant *p*-values have low LogWorths. A LogWorth of zero corresponds to a nonsignificant *p*-value of 1. Any LogWorth above 2 corresponds to a *p*-value below 0.01." (SAS)

The coefficients represent the percent change in log-odds of a posting containing a defect given that the variable has the value in brackets, over the base level. So for example (consult Table 3.5.1, rightmost column, above) the presence of an error code ("error_code_present") increases the chance of posting indicating a defect by 12% (100*(exp(.114)-1) = 12). Note that the coefficient for this variable is negative, meaning that the level 0 (error code absent) decreases the odds, implying that level 1 (error code present) increases the odds. Again, from the rightmost column of Table 3.5.1, above, the presence of a high model year ("model_year_high") increases the odds of defect by 4.8% (100*(exp(.047)-1) = 4.8), because its absence decreases the odds. The absence of a low rpm increases the odds of a defect by 39% ((100*(exp(.326)-1)) , so the presence of a low rpm decreases the odds.

Table 3.5.2. LogWorth of numerical attributes in Model 5 (full model).

Variable	LogWorth	p-value
ZSmokeWord	88.577	<.001
Electricalsystem	11.435	<.001
Braking	9.438	<.001
StructureandBody	9.393	<.001
Engine	6.916	<.001
model_year_high	6.890	<.001
AirConditioning	6.622	<.001
Other	4.848	<.001
SeatBelts	4.734	<.001
ZAverageWordLength	4.491	<.001
Steering	4.483	<.001
SemanFAC2_1	2.773	0.002
hp_high	2.620	0.002
SemanFAC1_1	2.619	0.002
Visibility	2.613	0.002
odometer_middle	2.080	0.008
hp_low	1.986	0.010
SemanFAC4_1	1.800	0.016
odometer_high	1.736	0.018
SemanFAC3_1	1.381	0.042
distance_high	1.348	0.045
SentiFAC6_2	1.153	0.070
Suspension	1.043	0.091
SemanFAC8_1	1.035	0.092

SentiFAC1_2	0.785	0.164
Zviews	0.784	0.164
SentiFAC4_2	0.716	0.192
Transmission	0.710	0.195
SemanFAC7_1	0.708	0.196
tire_model_present	0.688	0.205
WheelsandTires	0.663	0.218
rpm_high	0.623	0.238
SentiFAC3_2	0.616	0.242
vehicle_age_low	0.608	0.247
gear_present	0.605	0.248
Acoustics	0.584	0.261
odometer_low	0.547	0.284
Lights	0.493	0.321
SemanFAC5_1	0.459	0.348
SemanFAC6_1	0.410	0.389
distance_low	0.400	0.398
elecrical_present	0.302	0.499
SentiFAC5_2	0.262	0.547
ZnumOfUsers	0.183	0.655
model_year_low	0.170	0.676
generation_high	0.163	0.688
ZFogIndex	0.133	0.736
rpm_low	0.129	0.744
count_vehicles_low	0.098	0.799
Airbag	0.097	0.800

error_code_present	0.096	0.802
generation_low	0.066	0.859
ZwordCount	0.061	0.870
vehicle_age_high	0.034	0.925
SentiFAC2_2	0.015	0.966
count_vehicles_high	0.001	0.998

Table 3.5.3. Results of variable hypothesis tests.

Variable	**Hypothesized Direction**	**Observed Direction**
vehicle_age_low vehicle_age_high	positive negative	negative negative
count_vehicles_low count_vehicles_high	negative positive	NS NS
generation_low generation_high	negative positive	negative negative
gear_present	positive	NS
distance_low distance_high	positive negative	negative negative
electrical_present	positive	NS
error_code	positive	positive
hp_low hp_high	negative positive	positive positive
model_year_low model_year_high	negative positive	negative positive
tire_model	positive	positive
rpm_low rpm_high	positive positive	negative NS
odometer_low odometer_middle odometer_high	positive negative negative	positive positive positive

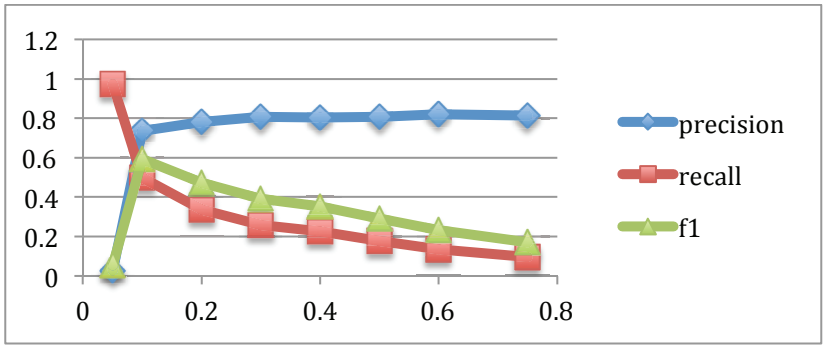
We experimented with different cutoffs in an attempt to find an improved balance between precision and recall: specifically, to see how many available defects we could discover without lowering precision too drastically. Table 3.5.4, and Figure 3.5.1, summarize these results. An optimal tradeoff appears to be a cutoff of .1, which maximizes F1 at .594, increasing our recall to .499 while reducing precision to .735. Any further decreases causes

a rapid degradation in precision. Higher cutoffs (>0.1), provide little improvement in precision but severe drops in recall (Figure 3.5.1).

Table 3.5.4 Full model performance at different cutoffs.

	0.05	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.75**
precision	0.028	0.735	0.783	0.806	0.805	0.808	0.819	0.814
recall	0.972	0.499	0.338	0.26	0.226	0.177	0.134	0.095
f1	0.054	0.594	0.473	0.393	0.353	0.29	0.231	0.17

Figure 3.5.1. Precision, recall, and F1 at different cutoff levels in the logistic regression.



A model with so many variables has the drawback of being complex to build, and such high dimensionality of course runs the risk of multicollinearity. We therefore tested an extremely parsimonious model consisting only of smoke words and our numerical attributes. Such a model can be built automatically once a classifier is trained, without any semantic or sentiment processing (as was previously required in (Abrahams et al., 2015)), and without any expertise in identifying system components being discussed (as was previously necessary in (Abrahams et al., 2013)). Using the same holdout validation sets as above, we ran a logistic regression. Variables sorted by LogWorth are reported in Table 3.5.5, and performance metrics are reported in Table 3.5.6.

Table 3.5.5. Reduced model with only smoke words and numbers.

Variable	LogWorth	p-value
ZSmokeWord	138.685	<.001
model_year_high	5.311	<.001
hp_high	2.702	0.002
hp_low	2.125	0.008
odometer_middle	1.783	0.016
odometer_low	1.083	0.083
odometer_high	0.949	0.113
distance_high	0.844	0.143
tire_model_present	0.829	0.148
vehicle_age_low	0.590	0.257
elecrical_present	0.584	0.261
count_vehicles_high	0.545	0.285
rpm_high	0.429	0.372
generation_high	0.404	0.395
count_vehicles_low	0.362	0.434
model_year_low	0.305	0.496
gear_present	0.228	0.591
rpm_low	0.160	0.692
distance_low	0.138	0.728
vehicle_age_high	0.116	0.765
generation_low	0.023	0.949
error_code_present	0.016	0.963

Table 3.5.6. Simplified model performance metrics

Metric	Training	Test
Precision	.799	.731
Recall	.347	.208
F1	.484	.323
AUC	.79	.65

Although the performance is reduced, the added benefit of a simple model makes it attractive in practice.

We conclude from this case study that the numerical attributes are helpful additions to the construction of models for automotive defect detection in social media. The improvements are modest, but we discuss promising further directions in Section 3.7, Limitations and Future Work.

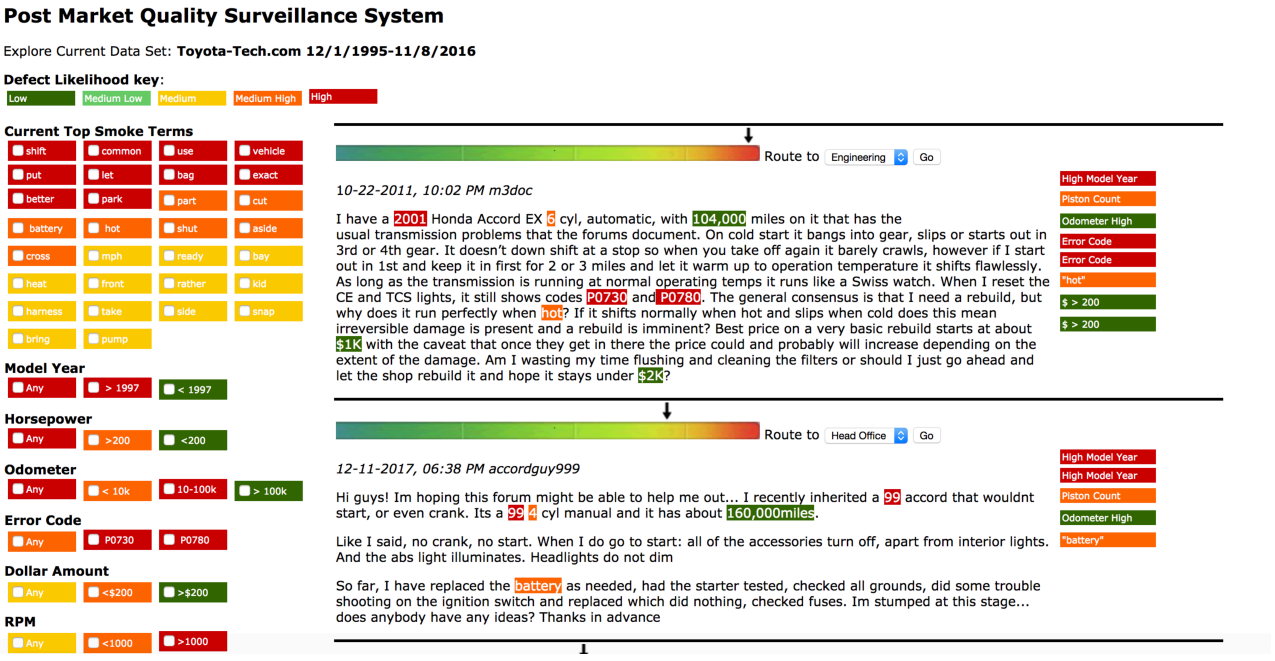
3.6. Numerical Attributes for Information Retrieval Systems

In addition to enabling simple but accurate defect prediction, the numerical attributes provide us with a rich instrument for information retrieval functions for managing defect posts. A dynamic decision support system is ultimately necessary for managing product quality using social media because our predictive models are never perfect, and human interpretation is an important element of the process. Smoke words and numerical attributes provide a powerful means for filtering the data set based on probabilities, but a decision support system that allows humans to explore the data is a way of finding social media content that might cause concern but was nevertheless missed by our predictive models.

Figure 3.6.1 shows a screen shot from a proposed Post Market Quality Surveillance System that makes use of smoke terms and numerical attributes. Social media postings are imported into the system, and smoke term analysis and numerical attribute extraction is completed automatically. In the user view, filtering facets are on the left, and postings

appear on the right. Users can select smoke terms and/or numerical attributes, such as model year, horsepower, odometer, etc. Each facet is colored according to how strongly its presence is tied to defect likelihood (that is, the coefficients on the logistic regression). The identified numerical expressions are highlighted within the posting, and the posting as a whole is assigned a defect likelihood, which is visualized on the slider at the top. Users can route the posting to a relevant organizational unit or to the archive. This interface provides the benefits of the precise predictive model with the flexibility to allow human users to make evaluations and experiment with combinations of numeric attributes and smoke terms. A design science deployment of this system is currently in the planning stages.

Figure 3.6.1. Post Market Quality Surveillance System using smoke terms and numerical attributes.



3.7. Limitations and Future Work

Our study had a number of limitations that can be resolved and explored in future work.

Some of the numbers in our data that were the same number category but referred to different functions, so strict continuous comparison against a median did not always make sense: normal pressure in a tire is not the same as normal pressure in a cylinder. Future studies should incorporate some kind of topic extraction to determine exactly what component the number is referring to.

Additionally, there could be splits for each number category that would better separate the defect and non-defect classes, and these optimal splits should be investigated. This should be approached with caution, however, because many splitting methods have an exponential run-time (Elomaa & Rousu, 1999). Discretization of numerical attributes is a key technique of data preprocessing (Ramírez-Gallego, García, Benítez, & Herrera, 2016), and a more granular discretization of numerical attributes is a logical next step. One important question to address is whether an optimal discretization is more useful than one informed by field experts, who have their own pragmatic thresholds for real world applications.

Future studies should experiment with integrating the numerical attributes with proprietary data sources, such as engineering manuals that indicate acceptable tolerances for number ranges. Many businesses are seeking ways to integrate social media with in-house structure data stores, and this provides a promising use case.

The Defect Management System interface proposes a potential Information Retrieval (IR) application for our numerical attributes, and its utility should be assessed in a production context. A follow-up design science project which deploys this system and evaluates its effectiveness in quality control is currently in the planning stages.

3.8. Conclusions

We have presented a novel approach for handling domain-specific numerical attributes within social media text and demonstrated their utility in tasks related to quality management.

Our work has a number of implications for research and practice.

For research, we contribute a novel approach to dealing with numerical data in text. We demonstrate that a corpus-based method in which numbers are extracted and classified using supervised machine learning can be accurate, effective, and straightforward. This is a robust and proven procedure for creating structure from unstructured data, which makes classification easier in any domain.

For management practice, we supply a methodology for producing informed diagnostics from social media postings. Our case study applies directly to the automotive industry, but our method can be expanded to other number-intensive industries as well.

Appendix 3.A. Tagging Protocol

Thank you for participating in The Vehicle Number Extraction Project! The purpose of this tagging round is to collect training data for computer algorithms that extract numbers from automotive postings.

Recent research has demonstrated that online forums contain text that can be processed to identify automotive defects that pose a safety hazard. Automatically identifying defects remains an area of active research. In addition to text, online postings contain a rich variety of numbers, and the goal of this project to test whether numerical features improve the accuracy of the defect classification.

Instructions
Go to https://pamtag.pamplin.vt.edu and select "Forgot password" to retrieve the password assigned to your email. After you log in, click on the project called "Vehicle Number Extraction Project".

On each page, you will see 5 numbers to tag. The target number appears at the beginning of the text entry in brackets [xxx]. The number will also be highlighted in context with "===xxx===" markings. Based on the text snippet given, select a number category on the right.

The categories appear below. Most are self-explanatory, but note some important points:

- The most common numbers are "Model Year"
- "Other" should be used sparingly, and when used, *please write how the number should be categorized in the comment field on the right*.
- A "Calendar Year" is different from a "Model Year" in that "Model Year" refers specifically to a vehicle, whereas "Calendar Year" simply states the year in which something happened.
- "Time of Day" refers to an actual hour, whereas "Time Duration" describes how long some action took, such as "I waited 4 hours."
- In many cases, a posting will refer to a count of something. We only track "Count of Doors" and "Count of Vehicles". Everything else is "Count of Other."

Please familiarize yourself with these categories before starting:

Address (including streets, zip codes, highways)
Age of Vehicle
Age of Person
Age Other
Calendar Date
Calendar Day
Calendar Month
Calendar Year
Chemical Symbol for Gas (O2, CO2, NO2)
Count of Doors
Count of Vehicles
Count Other

Distance Traveled
Dollar Amount, Price, Currency
Electrical (watts, amps, volts)
Engine Piston Count (V8, straight 6)
Engine Piston Size (liters (L) or cubic centimeters (cc))
Error Code
Fuel Efficiency (MPG, KPG)
Horsepower
Length/Height Measure
Listing (1., 2.,..)
Model Number of Component
Model Number of Vehicle
Model Year
Odometer Reading (# Miles)
Passenger Capacity (4 seats, 7-passenger van)
Percent
Phone
Rank (1st, 2nd)
Rating (Stars)
Speed (mph, kph)
Temperature
Time of Day
Time Duration
Tire Model
Tire Pressure (PSI or kPa)
Torque
Transmission (5 speed, 6 speed)
VIN (Vehicle Identification Number)
Volume (Non-Engine, gallons)
Weight
Wheel Drive (2WD, 4WD, AWD, 4x4)
Word (2 for "to")
Other

Please complete **300** tags. You will be saving 5 tags at a time by clicking on the red "Submit" button down at the bottom of the page.

Appendix 3.B. Number classification tables.

Table 3.B1. Count of Number Types before sampling.

Model Year	3944
Model Number of Component	2106
Other	1519
Count Other	1358
Dollar Amount-Price-Currency	925
Model Number of Vehicle	873
Time Duration	810
Odometer Reading (# Miles)	765
Listing (1.- 2.)	699
Engine Piston Count (V8-straight 6)	649
Engine Piston Size (liters (L) or cubic centimeters (cc))	484
Length/Height Measure	436
Transmission (5 speed-6 speed)	429
Count of Doors	375
Error Code	355
Speed (mph-kph)	309
Electrical (watts-amps-volts)	265
Percent	254
Distance Traveled	248
Wheel Drive (2WD-4WD-WD-4x4)	228
RPM	208
Word	173
Generation	195
Gear	173
Tire Model	159
Horsepower	147
Chemical Symbol for Gas (O2-CO2-NO2)	130
Weight	127
Calendar Date	109
Pressure (PSI or kPa)	104
Time of Day	93
Volume (Non-Engine-gallons)	91
Fuel Efficiency (MPG-KPG)	88
Horsepower	83
Count of Vehicles	81

Fuel Octane	81
Phone	55
Temperature	52
Torque	48
Temperature	43
Calendar Day	41
Oil Grade	29
Age of Vehicle	26

Table 3.B2. Count of number types after sampling and dropping miscellaneous "other" categories. These 44 number sets were used to train the classifiers described in the Methodology section.

Number	**Distinct Strings Observed in Training**
Age of Vehicle	20
Calendar Date	79
Calendar Day	26
Chemical Symbol for Gas (O2-CO2-NO2)	12
Count of Doors	34
Count of Vehicles	32
Distance Traveled	98
Dollar Amount-Price-Currency	159
Electrical (watts-amps-volts)	112
Engine Piston Count (V8-straight 6)	55
Engine Piston Size (liters (L) or cubic centimeters (cc))	106
Error Code	216
Fuel Efficiency (MPG-KPG)	68
Fuel Octane	10
Gear	23
Generation	23
Horsepower	100
Length/Height Measure	166
Listing (1.- 2.)	61

Model Number of Component	204
Model Number of Vehicle	195
Model Year	94
Odometer Reading (# Miles)	209
Oil Grade	20
Percent	57
Phone	37
Pressure (PSI or kPa)	70
Rating (Stars)	13
RPM	99
Speed (mph-kph)	128
Temperature	59
Time Duration	84
Time of Day	71
Tire Model	91
Torque	45
Transmission (5 speed-6 speed)	42
Volume (Non-Engine-gallons)	49
Weight	92
Wheel Drive (2WD-4WD-WD-4x4)	31
Word	22

Table 3.B3. Number feature counts by type in the full data set (4500 postings).

Numerical Attribute	**Posts with at Least One Mention**	**Posts with Multiple Mentions**
vehicle_age vehicle_age_low vehicle_age_high	195 138 64	31 18 7
fuel octane	753	256
doors	174	
count_vehicles count_vehicles_low count_vehicles_high	357 276 111	85 47 31
generation generation_low generation_high	322 250 106	118 68 36
gear	493	203

distance	686	226
distance_low	377	82
distance_high	420	116
dollar	1125	506
dollar_high	683	249
dollar_low	702	254
electrical	584	
engine_size	956	330
engine_size_low	484	
engine_size_high	588	
error_code	528	258
fuel_octane	409	87
fuel_octane_low	209	38
fuel_octaine_high	232	28
fuel_eff	262	62
fuel_eff_low	113	32
fuel_eff_high	183	35
hp	290	53
hp_low	102	19
hp_high	170	25
model_year	3554	2043
model_year_low	2181	959
model_year_high	2041	1010
odometer	1338	607
odometer_low	524	147
odometer_middle	353	113
odometer_high	893	313
speed	713	256
speed_range	122	0
speed_low	540	169
speed_high	275	67
temperature	498	
temperature_low	333	81
temperature_high	214	47
tire_model	276	71
oil	99	23
pressure	340	101
pressure_low	180	40
pressure_high	194	54
rpm	581	
rpm_low	447	158
rpm_high	214	48
torque	101	
high_torque	52	13
low_torque	52	13
weight	371	72
weight_low	153	26
weight_high	234	37

*Note: presence variables do not necessarily equal low+high. In some cases, both high and low values were mentioned in the same thread. In other cases, the program was unable to determine the magnitude of the number due to inconsistent structure.

Appendix 3.C. Sample word contexts from training set. Bolded tokens are instances of the type specified in the heading.

Address (including streets-zip codes-highways):
coyotes and other varmints across **i-19** i of course could install; yeah i went to a 6th ave electrons like a circuit; and got into the interstate **i-95** s and the oil light; i live in southern california **90706** area code and came on; remotes2003 inc) remotes2003 inc **945** walnut creek dr lilburn ga; 111 mechanic st lexington ky **40507** usa ph : (606; yeah i went to a **6th** ave electrons like a circuit; on the hov lanes on **i-95/395** you ll see a couple; me up the curves of us26 from portland to beaverton and;

Age of Vehicle:
exotic car but on a **20** year old honda no the; so iv had it for 6 months now just replace the; well as laborthe truck is **21** years old and i dont; exotic car but on a **20** year old honda no the; the truck is now about **6** months old first time it; well as laborthe truck is **21** years old and i dont; this way my miata is **6** years old - never had; gary the radiator is **2 1/2 yrs** old just took a; which i have had for **3** yrs now started with 98000;

Calendar Date:
modified by bigmoose 4:41 pm **2/12/2003**] the msd boxs is; (supersedes 02-031 dated september 17 2002) background electrical contacts; vibration # pit3511 - (**03/29/2005**) models : 2000-05 cadillac; modified by wundedsaint 10:06 pm **3/31/2003**] good stuff shadow way; by caramel turtle 9:48 am **11/27/2002**] i have a 99; modified by razor 2:25 am **5/29/2002**] no this is a

Calendar Day:
good news*** as of january **1st** 2008 there will be no; 98v231000 recall date : sep **22** 1998 component : electrical system; next issue out on september **14** 2005 crazy looks like a; : 3:59 pm edt october **12** 2007 updated : 6:25 pm; module owner notification began august **2** 2002 owners who take their; switch owner notification began june **14** 2002 owners who take their; = tochigi japan **12th** thru 17th position production sequence number this; the garage going on day 6 they need to replace the; the recall began on april **2** 2007 owners may contact honda;

Chemical Symbol for Gas (O2-CO2-NO2):
? are you using oem **o2** sensor or aftermarket ? you; cai or tbs ecu issue **o2** sensor issue associated with the; low voltage DTC P1164 ho2s bank 2 sensor 3 rich or high voltage say anything about cats or **o2** sensors since youre an 87; single wire oxygen sensor (02s) and the heated oxygen; map sensor 48 primary heated **02** sensor 5 map sensor 54; at temperature i disconnect the **n20s** and measured open resistance on; some keyboard bangers i/h/e and no2 yeah pretty much new at; say it s the primary **o2** sensor (front) based; the egr pcv alternator starter **o2** plugs wires and painted the; van , 100 shot of **n2o** and all , and no issues whatsoever

Count of Doors:
door 1~5l eh3 = civic **3** door 16l eh6 = civic; door sedan automatic or cvt 6 = **4** door truck automatic; 9 = civic si w/abs **3**-door 9 = civic ex 4-door; civic **3dr** dx 1992-1995 civic 3dr vx

1992-1995 civic 4dr dx; = accord dx w/abs **2/4-door** 2 = accord dx 2/4-door 2; 2 = accord dx 2/4-door 2 = civic cx **3-door** 2; 5 = civic ex a/abs 2-door 5 = cr-v w/abs 5; door hatchback auto 5 = **4** door sedan manual 6 =; its a 99 accord ex 4 door automatic i got in; spoiler will fit on a 4dr ? or has anybody tried;

Count of Vehicles:
better and i ve owned **2** automatic veichles that i actually; off with insurance with just **2** cars i just love how; has clips toyota to recall 3.8m vehicles over floor mats toyota; when parked i ve seen **8-9** civics in the past month; underhood potential units affected : 967,771 you computer is not recalled; read that i can use **2** accords and a prelude mount; sensor/control module # vehicles affected **81000** defect on certain vehicles the; ever had and we have **7** toyota s in the family; dx 4-door (1998) 4 = accord lx w/abs 2/4-door; seems like **8** of every 10 hondas here are boosted b;

Distance Traveled:
which stopped the racing engine **7000** miles later i now do; really no closer dealers ? 500 miles ? ! wow i; to add a quart every **3000** or so (runnin sysentic; the car 2 days about 20 miles to town and back; to honda 1 qt around 3k miles as a honda owner; mild tire hum i drove **250** miles at 70mph without any; ran a **159** in the 1/4 mile with a bone stock; changed out the oil like **1,000miles** later then 3000miles like normal; i change the oil every **2500** miles w/ 10w30 valvoline maxlife; sometimes after driving for 20 miles or so it still;

Dollar Amount-Price-Currency:
labour about 15 hours so **250**.00cdn regardless if i used the; for about $ 30- **$ 40** i would just get a; it s like a $ **50** part i had my dealer; i believe i paid around 600 for it circuit city is; do you trust a $ **9.00** under drive pulley and do; aquarium tubing new **$ 1.51** honda o-ring torque wrench (; bad discs would be around 160 $ for brembo blanks new;) which cost me $ **35** for 3qts ! the drain; problem today i get $ **10** more of 89oct at murphy; do a $ **500**- $ 600 worth of work in a; wants to dip into the **27k** range tell him to look;

Electrical (watts-amps-volts):
039 as i pulled the **7.5a** under hood fuse the draw; min later we put a 20a in and drove it down; this ? hard wire a **12v** source to the battery terminals; if the voltage is above .8 volts at normal operating temp; if the voltage is above **.8** volts at normal operating temp; high beams it goes to **12.8** is the eld not seeing; 5/40 or 10/40 in both 300v and 6100 **300v** of course; relay one line has a **12v** reading checked another relay on; the voltage drop is above **200** mv replaced the effected cable; it went from **94kw** to 108kw) i have a lead; and hook it to a **12v** and watch that sucker blow; as well just short it **2** ohm is really low just; heater circuit low voltage bank 1 sensor 2 dtc p0038 ho2s; the vehical computer sends a **12** volt signal to it to;

Engine Piston Count (V8-straight 6):
do you have **i4** or v6 ? fluid is less than; speed automatic comparing to the **v6** but if he can afford; [**4cyl**] my 96 4cyl camry has 155k miles fully; n t even have a v6 in accords in 1994 something; and mumbles vw makes a w8 not a **w16** vw doesnt; in super taikyu use spoon n1 coilovers made by showa *smacks; the frame in short a v6 in a civic : there; vs the f-series motor the **v6** on the other hand will; my dad s automatic accord **v6** gets about the same mpg; have a 98 4dr ex **4** cyl accord i have a; the autos it affects all **v6** accords and some acura models; so hello all 86 4runner **4cyl** straight drive 22re fisrt let; thermostat my 08 tacoma 4x4 **6** cylinder has wonderfully hot heat;

Engine Piston Size (liters (L) or cubic centimeters (cc)):
a 2000 malibu with the **3.1** when im driving this car; extended cab 4x4 with a **5.3** liter i am having trouble; have a 1998 chevy cavalier **2.2l** automatic trans history : my; lx automatic carburetter not efi **1.8 l** does the car make; of a **16l** nope no 1.6l s came with 4 speed; had out of my 2007 **6.0l** ltz max 3500 miles old; 1992 honda prelude si 4ws **7.8** 159 1994 honda prelude vtec; in my opinion325hp from a **5.3** that gets 16+ mpgmixed driving; 1995- new v6 models with **2.7** liter v6 from 1987-90 acura; as low as 47:1 early 930 ? got ta love those; 1995 honda accord ex sedan **2.2l** vtec 5spd correct me if; i know squat about the 3.0 v6 but a oil to;

Error Code:
! thank you hervecheck ur **codes0** error code appears nothing at; end sensor 1/left if dtc **b0104** or b0105 is set current; ground/voltage out of range dtc **b0051** deployment commanded dtc **b0053** deployment; codes found and auxillary code p1607 ? ? ? anyone have; do u have a code **20** ? ? electric load ?; codes flashed were 5 and 6 one an oxygen sensor and; -- -- cont : dtc **p0801** reverse inhibit control circuit malfunction;) sensor circuit intermittent dtc **p0340** camshaft position (cmp); p0780 shift malfunction dtc **p0781** 1-2 shift malfunction dtc **p0782** 2-3; codes let me knw plz .1,15,12,16,20,22,**14**,18 1 oxygen sensor 12 exhaust; 58 pu **719** low reference 59-60 not used 61 gy 1884; inspection i throw 2 codes **14** and 19 how do i;

Fuel Efficiency (MPG-KPG):
speed i am currently averaging **40** mpg and do n t; it i could get over 300 that was the only sign; on plus and get around **12** mpg efficient power is fine; it should get at least **17mpg** do i need a tune; ppm hc and only manages **27** mpg or so freeway my; smooth as going gt ; auto and my best is **27** mpg highway the epa for; far i have been getting **36-38mpg** consistantly 75 % highway driving; damn latvia of course city- 12-13 l/**100km** highway 8-10 l/100km depends; mileage on the highway probably 5mpg extra gil ``; auto and my best is **27** mpg highway the epa for; aforementioned si furthermore i averaged **45.5** mpg for my first tankful;

Fuel Octane:
if i started to put **89** or 87 octane in it; stuff so ill stick with **87** just for regular driving thanks; type elevations anything higher than **87** is a waste ans will; gas damn you only get **91** where you re at ?; at ? ? we have **93** for our premium around here; the two (87 and 91) but the point at; time with a fillup of **87** then a half tank of; area we only get to **91** octance and it costs me; do n t need to 87 will do just fine **87**; use either irving or shell **87** octane gas and have throughout;

Gear:
location indicator revised 4th and **5th** gear ratio vsa (vehicle; gears but no reverse or **4th** ah now i understand so; to just keep it in 4th to extend the life of; downshifts but only really from **2nd** to 3rd i will post; cruising at around 60-70 in 5th on the stock gauge my; suspect you re feeling the 4th gear lockout on a cold; 98 v6 would go into **4th** at 35 or higher and; second gear slam it to 2 and when it is ready; car will accellerate quicker in d3 since it does n t; **3rd** lockup a2454e has 1st 2nd 3rd 4th (overdrive); mph before it changes to **2nd** it would rev up between; 91 stang hard and holding **3rd** gear too long and letting; going into **2nd** gear and 1st and 5th are just about;

Generation:
help 4th gen is 4x114 **3rd** gen is 4x100 thx so; its 4th i dont consider **5th** to be early gen but; always more powerful than the **6th** gen av6 i ll give; your car s a gen 6 and your not expected to; for a cb accord the **5th** gen ones are italicized just; `` [ 3 ] gen 3 : my headlights are awesome; auto d or buy a **5th** gen x2 ``; the engine compartment than the 5th gen s (the whole; i heard from other professional 4th gen accord that already done; j32 s (is the **1st** gen tl a c25 ?; and cnc all the crap 6th gen v6s are known to; last thread in the gen 3 section thanks guys damon how; up the hood on my **gen3** the hood came down and;

Horsepower:
sho daily driven up to **400whp** with sc and another sho; as for your supercharger comment 40-**50hp** at the wheels is a; they dont do anything maybe 10 hp at the most and; the acura cl with only **20hp** less than the type-s and; gon na net you maybe 3-10hp at the wheels if you; the new accord of course **260hp** and a 6speed manual i; only a guess civic is **127hp** ``; up an turbo h22a with **400hp**+ for that kind of money; only a guess civic is **127hp** ``; a 1999 toyota camry le 4 cylinder i brought in a; a million it is rated **245** hp ! and the oem; ll be getting up to 100 more horses the way i; mine to push out about **200** crank hp check out http; the new accord of course **260hp** and a 6speed manual i;

Length/Height Measure:
644 height (in) 51.0 track (in front /; unit with ss coil pack **8mm** wires with stock distributor cap; coated) pauter rods eagle **92mm** crank (1599 r/s still; the pivot rivet drop down 1 1/16 and center punch a; are the shorter bolts (**8mm**) 28 lb ft for; them back them to a 13 rotor and bigger pads and; as long as they are 5x114.3 those look pretty nice where; shoots a fat stream like **30** feet i m sure that; even more and jam the 19 s through the hood and; ? second question- how much **0** gauge wire to buy and; and if i remember right 2-**14** gauge green wires the purple; i used **25** feet of 1/2 inch aluminim tubing i bent; dimensions wheelbase (in) **91.3** length (in) 1614; when we were installing my 6x9 s i believe you can;

Listing (1.- 2.):
2 when traveling between the united; turn the ignition switch off **2)** turn the fan switch; narrowed down to **3** or 4 things either the fuel pump; poewer the dispaly is blank 1 can this be the cause; supply voltage c4 not used c5 d-gn 801 retained accessory power; = civic ex a/abs 2-door 5 = cr-v w/abs **5** =; the ignition is switched off 2 connect the 2 pins in;) both batteries were dead **2)** both day time running; changed the thermostat two times 4 changed the coolant water sensor; hardware in the motor step **4.** put the screws back and; on the engine and a/c 4 if the clutch pulls in; get all the steps correctly **1.** remove both fuse panels and;

Model Number of Component:
but id just swap a **b18** into either slam it and; be programmed to operate the **rs3000** the status monitors led turns; from my y8 to a y7 tranny feels no differnt and; kept i think 16000 the os2 should not go bad i; little uneasy it s the **d16y7** pretty much stock i ve; cr and the rl has 11.0:1cr moot points ? the ecu; the short term kep if **ob1** system isnt as sensative as; computer and ecu like and obd1 car : i will pay; vehicles there that carted the **b20b** s around so according to; such as radio shack p/n 33-3013) plugged into the chassis; incorrect 3rd gear ratio dtc p0734 incorrect 4th gear ratio dtc; ton cargo van with a **305ci** has been shifting hard for; control module (bcm) **c3** pin wire color circuit no;

Model Number of Vehicle:

block for example a jdm h22 will say **h22a** while the; my 1996 chev blazer (**s-10**) wo n t start; good luck ! ! 1997 **k1500** 350 intermittent miss 1997 k1500; in close proximity to the **rav4** s receiver the tire pressue; tension sbc **350** problem sbc 350 problem ``;] i have a 1996 **c1500** with a vortec 57l engine; i was cruising in the gen2 in the middle of winter; is a black lt3 package z71 i got a great deal; [**s10**] my 94 **s10** tail lights quit working the; same problem with my 86 s10 blazer and it ended up; a turbo timer to my **a4** and when i ran the;

Model Year:
recall date : may 16 **1995** component : seat belts :; critical my information shows the **97** has a prime connector for; auto question i have a **1996** integra ls with an automatic; 2002] hello i have **2002** avalon xls and it just; recall on these for the 97-**2000** civic i had mine replaced; 96-98 97-98 n/a 96-97 96-98 96-**98** ect sensor (engine coolant; `` [01] **01** crv-shifting issue please help 01; think it was just luck 2003 changed battery gauges dead **2003**; ignition problem on **97** accord **97** accord dx manual transmission non-vtec; have a similar issue ? **94** accord lx if that matters; florida yesterday with my new **2004** mdx i am liking it; behind the car is a **1994** accord lx coupe 5spd what; thing happen started up my **2007** impala but the a/c did; v6 models except for the 03 6spd i noticed that was;

Odometer Reading (# Miles):
1st owner in 1994 with **24000** on it happy camper here; through 4 delco alternators between **110,000** and 195000 on my 94; an 03 tahoe with about **100,300** miles on it this past; **35000** miles and now at 37201 it happened again the previous; gears is awful i have **520** miles on the car and; which is a 99 with **150k** for $ 4395 with some; have a gen 35 with **104k** i m starting to notice; here again mine has about **43,315** on it and 50 psi; **11500** mi and at guessing 700-800 mi in 4wd works like; fully loaded the car has **70,000** miles on it never been; my 94 accord ex has 153xxxx miles blew two trannys and; and the warranty is actually **109k** or 7years+9mos i had a; transmissions that never last over **100-150k** quite supriseing to me !;

Percent:
loading is about 70/30 with **70 %** done by the front; of higher boost levels about **50** % of the car n; ? any one here its **100** % they work for both; might be a little diff 100 % sure about this ?; is a fuel pump relay **100** % - you hear the; and in black i have 15 % on the side windows; dominated by japanese cars therefore **99.9999** % of the mechs only; addition the front dampers generate **14-percent** less high speed compression damping; liek that but im not **100 %** sure but i found;

Phone:
time should contact honda at **1-800-999-1009** or acura at 1-800-382-2238 notes; 40507 usa ph : **(606) 233-1173** for your missing; usa-mn (big-lake) request_quote **1-800-527-4895** request_insurance_quote 1102 another thing comes; here call toyota corporate at **800-331-4331** and speak to a customer;

Rating (Stars):
and i give it a **8** out of 10 b/c of; and they only have a **3** star front crash test rating; them on overall everthing and **10** was the score was 10; they are referred to as **5** star rims and i think; and i give it a 8 out of 10 b/c of; my moms 01 xls **10/10** exceptional quality at 65000mi my; done i would get the **level10** rebuild for it though instead;

Speed (mph-kph):

into reverse and accelerate to **20** mph 7) bring speedo; put me around 3600 at **70** with the ls 5th so; with the cruise on at **70 - 75** and it kicks; had it in cruise at 55 it did the same thing; trips driving most interstate around 77mph no ac with cooler air; idea ? jerks 35mph - 75mph (fastest i ve had; on i slowed down to **55** and the light cut offnow; 5th with 8500 cutoff is **122mph** with a 44 fd it; bumper traffic that was moving **30-35mph** and had a good 40; at all it will reach **65-70mph** and run at about 4; shoots up to **30** to 40 mph almost as if something; mph and is sometimes at 0 mph any help suggestions would; temp and at speeds beyond **50** mph does do it when; application if i accelerate to **45mph** i can feel the car;

Temperature:
set to automatic and about **68** degrees i noticed that it; @ # $ % g **95** degrees outside warmer air doesnt; cat if the temp is ~**100** deg hotter after the cat; outside whenever it goes over **210** degrees i shutoff the ac; day my wifes suburban blew **54** degrees at the same time; genuine product may breakdown at **160** which might be what you; out on track on a **90** degree day placement of the; i think it will be **115** today we need coolant mixed; like **200** but never hitting 210 i also noted the volt; out on track on a **90** degree day placement of the; emratures reach an excess of **2500** degrees farenheit and pressure is; runs in the **220** to 230 range and it is normal;

Time of Day:
addition [modified by 4doorh22 **4:40** pm 1/24/2002] ``; ? [modified by ideal **1:15** am 7/25/2002] you probably; constitute a violation of section 27151 vc ? no section 27151; it [modified by losrocket_ **4:04** am 3/3/2003] hey calm; the rpms to like a 1500 and then pops it back;) [modified by tinkerbell **1:30** pm 4/14/2002] [modified; the tire and wheel at **9:00** and 3:00 and see if; it [modified by b16astard **8:53** am 4/1/2002] [modified; [modified by jdm girl **10:09** pm 3/6/2003] oh really; addition [modified by 4doorh22 4:40 pm 1/24/2002] ``;

Time Duration:
and we got paid the **4** hours to this day all; 96 22l v-tec accord about **2** months ago it has almost; like i had said about **3** months ago on tsxclubcom rdx; car is off for about **10** minutes the fan do n; switch ? mine broke about **3** weeks after my air conditioner; it had one previous owner **3** months ago the supercharger caused; it and sold it not **2** weeks later the kid he; will come out 1 or **2** months after the regular coupe; of course i never did **3** weeks ago i started to; dealer says it s a 1-hour job for a mechanic w/; keep your car for another **6** months or so who cares; believe it is supposed to 30 minutes you disconnect it for; i start the car after 5 -**10** minutes the temperature gauge;

Tire Model:
wheels 16-inch alloy tires **205/50 r16** 87v exterior dimensions wheelbase (; kill relay and reconnect the 12-gauge starter wires together in order; dunlops website i ve got **205-50** 15 s dz-101 s going; i have a trunk with 15 s in it and the; for a 2000 civic si **195/55-15** tires read the outter wall; dunlops website i ve got **205-50 15** s dz-101 s going; sedan front and rear tires **p175/70r13** available tires 1996 honda civic; perfect match ? my oe **195/65/15** tires have a diameter of; i would n t a **225/40/16** is 113 smaller in overall; the same set of toyo **205/60x15** spectrum s for the last; perform better **195 65 15= 205/60/15** would be closer to your

Oil Grade:

your year shoud n t **10w30** be fine though ? ``; can run stuff up to 50 weight if the conditions are; always good ! i run **10w30** in my 60 2500 4x4; 2002 civic si runs on 91 (or at least thats; the car i d use **10w-30** ! four quarts fill it; always good ! i run 10w30 in my 60 2500 4x4; it down like **10w-40** or 20w-50 other than that keep the; performance/race use 8100 e-tech lite 0w30 gasoline and diesel engine oil; every 15000 miles i use 50w synthetic ; it only takes; performance/race use 8100 e-tech lite **0w30** gasoline and diesel engine oil; just changed the oil to 10w40 b/c i know the car; can run stuff up to **50** weight if the conditions are; every 15000 miles i use **50w** synthetic ;

Pressure (PSI or kPa):
make ur car ready for **10psi** thats silly ``; the front dropped them to **50** psi back in action for; have 720cc injectors and at **10psi** you think you need a; psi higher for the street 32-**40psi** the vehicle manufacturer determines the; see values ranging anywhere from 130 to **160** psi i was; on it and drive it 45/**50** psi all the time fuse; my guage says it has **47** psi while running the needle; does have to be at 60psi to start iircdo n t; pressure gauge is still showing 42psi the fuel pump itself is;) again it depends on tires..33-**35** imho is too low for; : **220** kpa rr : 220 kpa when you re saying; regulator will not cause a **0**-psi reading that i have ever; be an under-inflated 29-**30psi** (32psi recommended for me) yep;

RPM:
brake then rev it to **2800** rpm (which is the; oh and dont drive over **3000** rpm so what are you; my car wont go past **5000** rpm and supposly i have; 193 n m) @ 4000 rpm o f22a6 : 142; net) : 185lbs-ft @ 3900rpm redline : **6800rpm** fuel cutoff; the automatic my tach reads **8200**+ by the time my rev;) the tachometer goes to **7100** rpm speedometer to 50 mph; heard they pull hard to **9000** rpm ! check out hmotorsonline; s worse between 1200 - **2200** rpms not saying that it; revs to a little over 2000 and a few seconds later; poorly and slowly (about **800** rpm) four about one; engine just revved up to **3000** rpm and it would hardly; tachs are accurate to about **4000rpm** then they are pointless webcam;

Torque:
to **75lb**-ft for aluminum wheels 80lb-ft for steel wheels 13 do; 200-hp @ 5500rpm engine torque **195** lb-ft @ 4700rpm engine bore; rebuild i think theres only **1** tq converter upgrade i ve; dohc) ~ 150hp / **137lb**-ft; : 150 lb/ft 150 x 3.42 x 427 = 21905 lb/ft; acura tl $ 35395 270hp **240lbs** torque 148 @ 966 infiniti; them and it does say **11ft** lbs for an existing head; 160 @ 5800 190 @ 6800 torque (**lbft**) 142; to be torqued correctly to **85** ft/lbs as well http :;

Transmission (5 speed-6 speed):
the gs has the optional **6** speed but what else does; got a 95 ex accord **5spd**.. the car srs light stays; rebuild would have rebuilt the **5** speed if i had one; 2 auto swap will rsx **6** speed transmission fit and can; think its a 91-92 legend **6** spd call i 800 79; about swapping over to a **5** speed ? its easier to; 4 = **4** speed manual 5 = **5** speed manual 6; crazy markup on it and **6sp** only comes in ex and; save your money for your **5-spd** the only way you can; have less hp then the **5** speed the 5 speed is; unless you find another inline 5 from a volvo or a; standpoint to put in a **5**-speed ?
Volume (Non-Engine-gallons):
looking for ? it had **3/4** tank of gas when this; oil change i already put **3** more quarts of oil what; 11) fill transmission with **2 1/2** quarts of atf start; or can i just use 2 quarts ? any input would; capacity (l) - 60 the highest output (kw; was alittle bit less than 1.5 quarts it s only 15qts; you are using put exactly **10** gallons in & gt ;; b20b/z high compression block ie

9.6:1 cr ? thanks t i; i believe it holds about **4.5** quarts so those letter sg; car my max vol is **35** and it seems like 25-35; you are using put exactly **10** gallons in & gt ; suck up about 1/3 to **1/2** the can turn the vehicle; that because it s a 4 barrel carb and will be; do this there is a **12** gallon fuel cell in the;

Weight:
are adding an additional like **40** pounds between the pumps compressors; and that door felt like **50lbs** ! i placed 2 donuts; and the seatbelt retractors yesterday 12lb savings the belts were 6lbs; see how can i take **100lbs** out my car that i; my civic is lighter than 2326lbs that the us model weighs; are adding an additional like **40** pounds between the pumps compressors; we are running 10:1 and 28lbs on boost on a fully; height 272 inches dry weight **769.6** pounds fuel capacity 62 gallons; piece u see outside the car-**7 lbs** what i want to; n t be much thought 2-3 lbs) and will never; pully ? thats what maybe **10** pounds weight savings ? ahaha; x distance here such as 50 lbs of force applied to; 200hp/ 200tq and weighed about **3300** or so but on the;

Wheel Drive (2WD-4WD-WD-4x4):
a 1988 cheyenne and my **4x4** light does not work i; start the truck and no **4wd** when i bought the truck; km (~50000 miles) **4wdi** last morning when i arrived; workingwhat s happening with the **4wd** ? glad you got the; matter what the salesman saysthe **4wd** does not work the truck; trace the wiring on your **4wd** unit to this `` switch; steering problem 05 chevy 1500 **4x4** steering problem; 6 = rti civic wagon **4wd** manual seat belt 8 =; an 01 tacoma 27 auto 4x4.122000 on the od had a; 4:10 gears it s a **4x4** crew cab i normaly get; not change from **2wd** to 4wd or vice versa just a; t know the `` service **4wd** indicator bulb was missing should;

CONCLUSION

The three studies in this dissertation delve into various aspects of interacting with customers in social media. An enormous increase in non-transactional but value-adding interactions with customers has been enabled by social media, and exploring the various dimensions of "customer engagement" has been a growing interest in business research. These interactions take place online largely through text, so we adopted a text analytic methodology in each study.

In Study 1 – "Text Analytics for B2C Interactions: Engaging Customers in Online Communities" – we explored the drivers of customer interest in interacting with business in online communities. We conducted an empirical test of the theory that belonging to some form of community motivates online activities such as likes, comments, and shares on Facebook. We defined "community building" and compiled a set of textual phrases that constitute appeals to community. We then tested whether these messages prompt greater response from online community members, controlling for various factors such as audience size, length of message, readability, time of day, media type, etc. We discovered they are consistently associated with higher likes, but not necessarily with comments or shares. This suggests that different kinds of textual content will lead to different kinds of responses. Low-effort interactions such as likes, which still have value because they keep organizational content in a user's news feed, can be garnered through community appeals, but comments and shares have different triggers.

Finding the attributes of content that provoke comments or shares is a topic for future studies. One potential study might test whether content unpredictability drives comments. There are text analytic ways of quantifying how dissimilar a message is from a corpus of messages from the same organization using topics and word distributions, and it would be useful to know whether surprising content in fact results in more comments. It would also be

useful to know whether there are topics that consistently drive comments, or whether there is a constantly evolving and unpredictable range of interests that cannot easily be accurately predicted.

More generally, more research is needed into how customers respond to a variety of B2C messages. Lacking better options, businesses have frequently adopted the same tone and style in online communication as they have always used for person-to-person customer messages. When a B2C message is no longer read only by the recipient, the nature of the communication changes, and more research needs to be dedicated to finding the best approaches. For example, what is the best way for a business to respond to negative criticism online? More investigation is needed into not only content, but style and tone.

In Study 2 – "Text Analytics for C2C Interactions: Application in Pharmacovigilance" – we investigated how businesses can use online reviews for product quality management. The volume of online reviews makes it impossible to thoroughly read all of them, especially for industry regulators who monitor thousands of businesses, so automated methods for locating information of interest are critical. In this study, we adapted a text analytic methodology that has elsewhere proven effective in finding safety defects in consumer products to the task of automatically locating adverse drug reactions among thousands of online reviews for over-the-counter medications. We found that this method is effective at finding these reactions. We also found that individuals are astonishingly adept at constructing lists of search terms, often outperforming machine-learning methods, and that groups perform even better than individuals.

Text analytic research into locating adverse drug reactions in social media is still nascent, and there are many useful avenues for future research. One confounding issue in social media text is that there are many usage scenarios for individual words or phrases, so bag-of-words approaches may result in misleading findings. For example, a review for a

medicated cream can have the word "burn" appear in any of the following contexts: "I had a burn and this helped." "This product caused a burn on my skin so I threw it away." "Other products burn but this one doesn't." "Don't put this on a sun burn or it'll blister." "This was such a rip-off I felt like I was burned." "I would rather burn down my house than use this product again." More research is needed into not only word sense disambiguation and part of speech (syntax), but a word's semantic function within its sentence. Current studies into word embeddings appear to be a promising way of addressing these problems.

In Study 3 - "Text Analytics for C2C Interactions: Numeric Information Extraction" – we conducted a further inquiry into automated ways of using customer communications for product quality management. We proposed and tested a procedure for extracting, identifying, and binning numerical attributes. We then demonstrated that these numerical features were useful for classification tasks by showing that they improved defect discovery.

This work proposed a new decision support system that uses numerical intelligence for searching social media postings for product defect information. This system allows quality managers to slice and dice information along a variety of dimensions, allowing them to employ both the automated numerical intelligence created in Study 3, and the combined human- and machine-driven judgments that proved so valuable in Study 2. A promising next step would be to implement this system in a real world setting and conduct a design science study to evaluate its effectiveness.

A variety of other research projects are suggested by the numerical approach. For example, a food producer could use the procedure to organize and index online reviews to better understand consumer preferences regarding package sizes or consumer thresholds for nutrition content (sugar, fat, etc.). Electronics companies can examine how various components (memory capacities, CPU speeds, model numbers) are combined to identify

compatibility problems faced by consumers. A car seat manufacturer could use the number approach to investigate regular usage or unsafe usage by rapidly reviewing consumer postings specifying different groups of ages, sizes, weights, and angles.

The framework in Figure 1 can be used as a broad outline for future studies of how organizations derive benefit from interacting with stakeholders on social media. The boxes represent the stakeholders, and the arrows represent the messages. In this dissertation, we dealt with the interactions between businesses and customers over social media, but future work can investigate the dynamics of interacting with other entities, such as competitors, partners, regulators, suppliers, and employees.

Social media present an immense opportunity for businesses to engage with customers, and both parties benefit from the enhanced relationship. Customers have the chance to voice their opinions, express dissatisfaction, request new features or services, and receive information. Businesses benefit as customers act as brand advocate, product champion, troubleshooter, and co-innovator. In this dissertation, we investigated various dimensions of these interactions, and arrived at several pragmatic recommendations for posting in a social community (Study 1) or finding useful feedback in online reviews (Study 2 and 3). Text analytic methods will continue to provide a valuable set of methodologies for these exploring these interactions, and further developments in natural language processing will yield valuable insights for perfecting customer engagement practices.

REFERENCES

Abbasi, A., Adjeroh, D., Dredze, M., Paul, M. J., Zahedi, F. M., Zhao, H., . . . Shaker, R. (2014). Social media analytics for smart health. *IEEE Intelligent Systems, 29*(2), 60-80.

Abbasi, A., & Chen, H. (2008). CyberGate: a design framework and system for text analysis of computer-mediated communication. *Mis Quarterly*, 811-837.

Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z. J., & Jiao, J. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management, 24*(6), 975-990.

Abrahams, A. S., Jiao, J., Fan, W., Wang, G. A., & Zhang, Z. (2013). What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings. *Decision Support Systems, 55*(4), 871-882.

Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems, 54*(1), 87-97.

Abualigah, L. M., Khader, A. T., Al-Betar, M. A., Alyasseri, Z. A. A., Alomari, O. A., & Hanandeh, E. S. (2017). *Feature selection with β-hill climbing search for text clustering application.* Paper presented at the Information and Communication Technology (PICICT), 2017 Palestinian International Conference on.

Academic Ranking of World Universities. Shanghai Ranking Consultancy. Retrieved from http://www.shanghairanking.com/ARWU2014.html

Adams, D. Z., Gruss, R., & Abrahams, A. S. (2017). Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews. *International Journal of Medical Informatics, 100*, 108-120.

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*: Springer Science & Business Media.

Akiba, T., Itou, K., & Fujii, A. (2004). *Question Answering Using" Common Sense" and Utility Maximization Principle.* Paper presented at the National Institute of Informatics Testbeds and Community for Information Access Research (NTCIR-4), Tokyo, Japan.

Al-Daihani, S. M., & Abrahams, A. (2016). A Text Mining Analysis of Academic Libraries' Tweets. *The Journal of Academic Librarianship*.

Allison, P. (2012). When can you safely ignore multicollinearity. *Statistical Horizons, 5*(1).

Alvarez, C., & Fournier, S. (2016). Consumers' relationships with brands. *Current Opinion in Psychology, 10*, 129-135.

Alvarez-Requejo, A., Carvajal, A., Begaud, B., Moride, Y., Vega, T., & Arias, L. M. (1998). Under-reporting of adverse drug reactions Estimate based on a spontaneous reporting scheme and a sentinel system. *European journal of clinical pharmacology, 54*(6), 483-488.

Angela Hausman, D., Kabadayi, S., & Price, K. (2014). Consumer–brand engagement on Facebook: liking and commenting behaviors. *Journal of Research in Interactive Marketing, 8*(3), 203-223.

Aral, S., Dellarocas, C., & Godes, D. (2013). Introduction to the special issue—social media and business transformation: a framework for research. *Information Systems Research, 24*(1), 3-13.

Aramaki, E., Miura, Y., Tonoike, M., Ohkuma, T., Masuichi, H., Waki, K., & Ohe, K. (2010). *Extraction of adverse drug effects from clinical records.* Paper presented at the MedInfo.

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media.* Paper presented at the Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01.

Bagozzi, R. P., & Dholakia, U. M. (2006). Antecedents and purchase consequences of customer participation in small group brand communities. *International Journal of research in Marketing, 23*(1), 45-61.
Bakalov, A., Fuxman, A., Talukdar, P. P., & Chakrabarti, S. (2011). *Scad: Collective discovery of attribute values.* Paper presented at the Proceedings of the 20th international conference on World wide web.
Banerjee, S., Chakrabarti, S., & Ramakrishnan, G. (2009). *Learning to rank for quantity consensus queries.* Paper presented at the Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.
Batuwita, R., & Palade, V. (2013). Class imbalance learning methods for support vector machines.
Baumeister, R. F., & Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological bulletin, 117*(3), 497.
Beckers, S. F., van Doorn, J., & Verhoef, P. C. (2017). Good, better, engaged? The effect of company-initiated customer engagement behavior on shareholder value. *Journal of the Academy of Marketing Science*, 1-18.
Benton, A., Ungar, L., Hill, S., Hennessy, S., Mao, J., Chung, A., . . . Holmes, J. H. (2011). Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of biomedical informatics, 44*(6), 989-996.
Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. *Journal of Hospitality Marketing & Management, 25*(1), 1-24.
Bhat, S. K., & Culotta, A. (2017). Identifying leading indicators of product recalls from online reviews using positive unlabeled learning and domain adaptation. *arXiv preprint arXiv:1703.00518*.
Bhattacharya, M., Snyder, S., Malin, M., Truffa, M. M., Marinic, S., Engelmann, R., & Raheja, R. R. (2017). Using Social Media Data in Routine Pharmacovigilance: A Pilot Study to Identify Safety Signals and Patient Perspectives. *Pharmaceutical Medicine*, 1-8.
Bian, J., Topaloglu, U., & Yu, F. (2012). *Towards large-scale twitter mining for drug-related adverse events.* Paper presented at the Proceedings of the 2012 international workshop on Smart health and wellbeing.
Bijmolt, T. H., Leeflang, P. S., Block, F., Eisenbeiss, M., Hardie, B. G., Lemmens, A., & Saffert, P. (2010). Analytics for customer engagement. *Journal of service research, 13*(3), 341-356.
Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition, 30*(7), 1145-1159.
Brandtzaeg, P. B., & Haugstveit, I. M. (2014). Facebook likes: a study of liking practices for humanitarian causes. *International Journal of Web Based Communities, 10*(3), 258-279.
Braojos, J., Benitez, J., & Llorens-Montes, F. J. (2017). Contemporary Micro-IT Capabilities and Organizational Performance: The Role of Online Customer Engagement.
Brehm, S. S. (1987). Social support and clinical practice *Social processes in clinical and counseling psychology* (pp. 26-38): Springer.
Brodie, R. J., Hollebeek, L. D., Jurić, B., & Ilić, A. (2011). Customer engagement: Conceptual domain, fundamental propositions, and implications for research. *Journal of service research, 14*(3), 252-271.
Budd, J. M. (1998). *The Academic Library: Its Context, Its Purpose, and Its Operation*: ERIC.
Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine, 9*(2), 48-57.

Campbell, J. D. (2006). Changing a Cultural Icon: The Academic Library as a Virtual Destination. *EDUCAUSE review, 41*(1), 16.

Cao, Z., Li, W., Li, S., & Wei, F. (2017). *Improving Multi-Document Summarization via Text Classification.* Paper presented at the AAAI.

Castellanos, A., Castillo, A., Lukyanenko, R., & Tremblay, M. C. (2017). *Understanding Benefits and Limitations of Unstructured Data Collection for Repurposing Organizational Data.* Paper presented at the EuroSymposium on Systems Analysis and Design.

Chaganty, A., Paranjape, A., Liang, P., & Manning, C. D. (2017). *Importance sampling for unbiased on-demand evaluation of knowledge base population.* Paper presented at the Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

Charte, F., Rivera, A. J., del Jesus, M. J., & Herrera, F. (2015). Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing, 163*, 3-16.

Chen, G., Ye, D., Xing, Z., Chen, J., & Cambria, E. (2017). *Ensemble application of convolutional and recurrent neural networks for multi-label text categorization.* Paper presented at the Neural Networks (IJCNN), 2017 International Joint Conference on.

Chen, R., & Xu, W. (2017). The determinants of online customer ratings: a combined domain ontology and topic text analytics approach. *Electronic Commerce Research, 17*(1), 31-50.

Chen, Y. H. (2011). Undergraduates' perceptions and use of the university libraries Web portal: Can information literacy instruction make a difference? *Proceedings of the Association for Information Science and Technology, 48*(1), 1-10.

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2015). *Antisocial Behavior in Online Discussion Communities.* Paper presented at the ICWSM.

Chou, W.-Y. S., Hunt, Y. M., Beckjord, E. B., Moser, R. P., & Hesse, B. W. (2009). Social media use in the United States: implications for health communication. *Journal of medical Internet research, 11*(4).

Chu, S. K.-W., & Du, H. S. (2013). Social networking tools for academic libraries. *Journal of librarianship and information science, 45*(1), 64-75.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(4), 213.

Cohen, S., & Wills, T. A. (1985). Stress, social support, and the buffering hypothesis. *Psychological bulletin, 98*(2), 310.

Connors, L., Mudambi, S. M., & Schuff, D. (2011). *Is it the review or the reviewer? A multi-method approach to determine the antecedents of online review helpfulness.* Paper presented at the System Sciences (HICSS), 2011 44th Hawaii International Conference on.

Coulter, K. S., Gummerus, J., Liljander, V., Weman, E., & Pihlström, M. (2012). Customer engagement in a Facebook brand community. *Management Research Review, 35*(9), 857-877.

Cvijikj, I. P., & Michahelles, F. (2013). Online engagement factors on Facebook brand pages. *Social Network Analysis and Mining, 3*(4), 843-861.

Davidov, D., & Rappoport, A. (2010). *Extraction and approximation of numerical attributes from the web.* Paper presented at the Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.

De Vries, N. J., & Carlson, J. (2014). Examining the drivers and brand performance implications of customer engagement with brands in the social media environment. *Journal of Brand Management, 21*(6), 495-515.

Doig, B., McRae, B., & Rowe, K. (2003). A good start to numeracy: Effective numeracy strategies from research and practice in early childhood. *Australian Council for Educational Research (ACER): ACEReSearch*.
Dokuz, A. S., & Celik, M. (2017). Discovering Socially Important Locations of Social Media Users. *Expert Systems with Applications, 86*, 113-124.
Dong, W., Liao, S., Xu, Y., & Feng, X. (2016). Leading Effect of Social Media for Financial Fraud Disclosure: A Text Mining Based Analytics.
Doucette, J., & Heywood, M. (2008). GP classification under imbalanced data sets: Active sub-sampling and AUC approximation. *Genetic Programming*, 266-277.
Edwards, I. R., & Lindquist, M. (2011). Social media and networks in pharmacovigilance: Springer.
Elomaa, T., & Rousu, J. (1999). General and efficient multisplitting of numerical attributes. *Machine learning, 36*(3), 201-244.
Eranti, V., & Lonkila, M. (2015). The social significance of the Facebook Like button. *First Monday, 20*(6).
Fan, W., Gordon, M. D., & Pathak, P. (2005). Effective profiling of consumer information retrieval needs: a unified framework and empirical comparison. *Decision Support Systems, 40*(2), 213-233.
Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research, 19*(3), 291-313.
Fournier, S., & Avery, J. (2011). The uninvited brand. *Business horizons, 54*(3), 193-207.
Friedman, C. (2009). Discovering novel adverse drug events using natural language processing and mining of the electronic health record. *Artificial Intelligence in Medicine*, 1-5.
Fuel Economy: 2017 Toyota Sienna Fuel Economy Review. 2017. Caranddriver.com. Retrieved from https://www.caranddriver.com/reviews/2017-toyota-sienna-in-depth-model-review-2017-toyota-sienna-fuel-economy-review-car-and-driver-page-3
Gaby, S., & Caren, N. (2012). Occupy online: How cute old men and Malcolm X recruited 400,000 US users to OWS on Facebook. *Social Movement Studies, 11*(3-4), 367-374.
Gaines, B. R. (1989). *An ounce of knowledge is worth a ton of data: quantitative studies of the trade-off between expertise and data based on statistically well-founded empirical induction.* Paper presented at the Proceedings of the sixth international workshop on Machine learning.
Gangadharbatla, H. (2008). Facebook me: Collective self-esteem, need to belong, and internet self-efficacy as predictors of the iGeneration's attitudes toward social networking sites. *Journal of interactive advertising, 8*(2), 5-15.
Gardner, W. L., Pickett, C. L., & Brewer, M. B. (2000). Social exclusion and selective memory: How the need to belong influences memory for social events. *Personality and Social Psychology Bulletin, 26*(4), 486-496.
Gerardi, K., Goette, L., & Meier, S. (2013). Numerical ability predicts mortgage default. *Proceedings of the National Academy of Sciences, 110*(28), 11267-11271.
Gerolimos, M. (2011). Academic libraries on Facebook: An analysis of users' comments. *D-Lib Magazine, 17*(11), 4.
Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., O'Connor, K., Sarker, A., . . . Gonzalez, G. (2014). *Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark.* Paper presented at the Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing.

Goh, K.-Y., Heng, C.-S., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Information Systems Research, 24*(1), 88-107.
Goldberg, D. M., & Abrahams, A. S. (2017). A Tabu search heuristic for smoke term curation in safety defect discovery. *Decision Support Systems*.
Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ Qual Saf, 22*(3), 251-255.
Grondin, R., Lupker, S. J., & McRae, K. (2006). *Shared features dominate the number-of-features effect.* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.
Gu, B., & Ye, Q. (2014). First step in social media: Measuring the influence of online management responses on customer satisfaction. *Production and Operations Management, 23*(4), 570-582.
Guerini, M., Gatti, L., & Turchi, M. (2013). Sentiment analysis: How to derive prior polarities from SentiWordNet. *arXiv preprint arXiv:1309.5843*.
Gummerus, J., Liljander, V., Weman, E., & Pihlström, M. (2012). Customer engagement in a Facebook brand community. *Management Research Review, 35*(9), 857-877.
Gurulingappa, H., Fluck, J., Hofmann-Apitius, M., & Toldo, L. (2011). *Identification of adverse drug event assertive sentences in medical case reports.* Paper presented at the First international workshop on knowledge discovery and health care management (KD-HCM), European conference on machine learning and principles and practice of knowledge discovery in databases (ECML PKDD).
Gurulingappa, H., Mateen‐Rajpu, A., & Toldo, L. (2012). Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics, 3*(1), 15.
Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.
Harpaz, R., DuMouchel, W., Shah, N. H., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel Data‐Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clinical Pharmacology & Therapeutics, 91*(6), 1010-1021.
Harrison, A., Burress, R., Velasquez, S., & Schreiner, L. (2017). Social media use in academic libraries: A phenomenological study. *The Journal of Academic Librarianship*.
Hartley, D. M. (2014). Using social media and internet data for public health surveillance: the importance of talking. *The Milbank Quarterly, 92*(1), 34-39.
Hartmann, W. R., Manchanda, P., Nair, H., Bothner, M., Dodds, P., Godes, D., . . . Tucker, C. (2008). Modeling social interactions: Identification, empirical methods and policy implications. *Marketing Letters, 19*(3-4), 287-304.
He, W., Tian, X., Chen, Y., & Chong, D. (2016). Actionable social media competitive analytics for understanding customer experiences. *Journal of Computer Information Systems, 56*(2), 145-155.
He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management, 33*(3), 464-472.
Heinrichs, J. H., Lim, K. S., Lim, J. S., & Spangenberg, M. A. (2007). Determining factors of academic library web site usage. *Journal of the Association for Information Science and Technology, 58*(14), 2325-2334.
Hennig-Thurau, T., Malthouse, E. C., Friege, C., Gensler, S., Lobschat, L., Rangaswamy, A., & Skiera, B. (2010). The impact of new media on customer relationships. *Journal of service research, 13*(3), 311-330.

Hildebrand, C., Häubl, G., Herrmann, A., & Landwehr, J. R. (2013). When social media can be bad for you: Community feedback stifles consumer creativity and reduces satisfaction with self-designed products. *Information Systems Research, 24*(1), 14-29.
Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter, 2*(1), 58-64.
Hoffman, D. L., & Fodor, M. (2010). Can you measure the ROI of your social media marketing? *MIT Sloan Management Review, 52*(1), 41.
Hollebeek, L. D. (2011). Demystifying customer brand engagement: Exploring the loyalty nexus. *Journal of marketing management, 27*(7-8), 785-807.
Hollebeek, L. D., Conduit, J., & Brodie, R. J. (2016). Strategic drivers, anticipated and unanticipated outcomes of customer engagement: Taylor & Francis.
Hong, H., Xu, D., Wang, G. A., & Fan, W. (2017). Understanding the determinants of online review helpfulness: A meta-analytic investigation. *Decision Support Systems*.
Hovy, E., Hermjakob, U., Lin, C.-Y., & Ravichandran, D. (2002). *Using knowledge to facilitate factoid answer pinpointing.* Paper presented at the Proceedings of the 19th international conference on Computational linguistics-Volume 1.
Hu, N., Pavlou, P. A., & Zhang, J. (2006). *Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication.* Paper presented at the Proceedings of the 7th ACM conference on Electronic commerce.
Huizinga, T., Ayanso, A., Smoor, M., & Wronski, T. (2017). Exploring Insurance and Natural Disaster Tweets Using Text Analytics. *International Journal of Business Analytics (IJBAN), 4*(1), 1-17.
Ittoo, A., Nguyen, L. M., & van den Bosch, A. (2016). Text analytics in industry: Challenges, desiderata and trends. *Computers in Industry, 78*, 96-107.
Jaakkola, E., & Alexander, M. (2014). The role of customer engagement behavior in value co-creation: a service system perspective. *Journal of service research, 17*(3), 247-261.
Jang, H.-J., Sim, J., Lee, Y., & Kwon, O. (2013). Deep sentiment analysis: Mining the causality between personality-value-attitude for analyzing business ads in social media. *Expert Systems with Applications, 40*(18), 7492-7503.
Jiang, K., & Zheng, Y. (2013). *Mining Twitter data for potential drug effects.* Paper presented at the International Conference on Advanced Data Mining and Applications.
Jiang, Z., Zhang, Y., & Li, X. (2017). *MOOCon: A Framework for Semi-supervised Concept Extraction from MOOC Content.* Paper presented at the International Conference on Database Systems for Advanced Applications.
Johnson, S. L., Safadi, H., & Faraj, S. (2015). The emergence of online community leadership. *Information Systems Research, 26*(1), 165-187.
Jones, K. S. (1997). *Readings in information retrieval*: Morgan Kaufmann.
Kang, J., Tang, L., & Fiore, A. M. (2014). Enhancing consumer–brand relationships on restaurant Facebook fan pages: Maximizing consumer benefits and increasing active participation. *International Journal of Hospitality Management, 36*, 145-155.
Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons, 53*(1), 59-68.
Khan, Z., Khan, Z., Vorley, T., & Vorley, T. (2017). Big data text analytics: an enabler of knowledge management. *Journal of Knowledge Management, 21*(1), 18-34.
Kilambi, A., Laroche, M., & Richard, M.-O. (2013). Constitutive marketing: Towards understanding brand community formation. *International Journal of Advertising, 32*(1), 45-64.

Kounios, J., Green, D. L., Payne, L., Fleck, J. I., Grondin, R., & McRae, K. (2009). Semantic richness and the activation of concepts in semantic memory: Evidence from event-related potentials. *Brain research, 1282*, 95-102.
Kozinets, R. V. (2002). The field behind the screen: Using netnography for marketing research in online communities. *Journal of marketing research, 39*(1), 61-72.
Kronrod, A., & Danziger, S. (2013). "Wii will rock you!" The use and effect of figurative language in consumer reviews of hedonic and utilitarian consumption. *Journal of Consumer research, 40*(4), 726-739.
Kuh, G. D., & Gonyea, R. M. (2003). The role of the academic library in promoting student engagement in learning. *College & Research Libraries, 64*(4), 256-282.
Kumar, V. (2013). Challenges and future consideration for pharmacovigilance. *Journal of Pharmacovigilance*.
Kumar, V., & Pansari, A. (2016). Competitive advantage through engagement. *Journal of marketing Research, 53*(4), 497-514.
Kwok, L., & Yu, B. (2013). Spreading social media messages on facebook an analysis of restaurant business-to-consumer communications. *Cornell Hospitality Quarterly, 54*(1), 84-94.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.
Law, D., Gruss, R., & Abrahams, A. S. (2017). Automated defect discovery for dishwasher appliances from online consumer reviews. *Expert Systems with Applications, 67*, 84-94.
Lazarou, J., Pomeranz, B. H., & Corey, P. N. (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *Jama, 279*(15), 1200-1205.
Leaman, R., Wojtulewicz, L., Sullivan, R., Skariah, A., Yang, J., & Gonzalez, G. (2010). *Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks.* Paper presented at the Proceedings of the 2010 workshop on biomedical natural language processing.
Leary, M. R., Kelly, K. M., Cottrell, C. A., & Schreindorfer, L. S. (2007). Individual differences in the need to belong: Mapping the nomological network. *Unpublished manuscript, Duke University*.
Leary, T. (1958). Interpersonal diagnosis of personality. *American Journal of Physical Medicine & Rehabilitation, 37*(6), 331.
Lee, D., Hosanagar, K., & Nair, H. (2014). The effect of social media marketing content on consumer engagement: Evidence from facebook. *Available at SSRN*.
Leonardi, P. M., Huysman, M., & Steinfield, C. (2013). Enterprise social media: Definition, history, and prospects for the study of social technologies in organizations. *Journal of Computer‐Mediated Communication, 19*(1), 1-19.
Libai, B., Bolton, R., Bügel, M. S., De Ruyter, K., Götz, O., Risselada, H., & Stephen, A. T. (2010). Customer-to-customer interactions: broadening the scope of word of mouth research. *Journal of service research, 13*(3), 267-282.
Lin, K.-Y., & Lu, H.-P. (2011). Intention to continue using Facebook fan pages from the perspective of social capital theory. *Cyberpsychology, Behavior, and Social Networking, 14*(10), 565-570.
Liu, J., Zhao, S., & Wang, G. (2017). SSEL-ADE: A semi-supervised ensemble learning framework for extracting adverse drug events from social media. *Artificial Intelligence in Medicine*.
Liu, J., Zhao, S., & Zhang, X. (2016). An ensemble method for extracting adverse drug events from social media. *Artificial Intelligence in Medicine, 70*, 62-76.

Liu, L., Wu, J., Li, P., & Li, Q. (2015). A social-media-based approach to predicting stock comovement. *Expert Systems with Applications, 42*(8), 3893-3901.
Liu, X., & Chen, H. (2013). *AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums.* Paper presented at the International Conference on Smart Health.
Liu, Y., Jiang, C., Ding, Y., Wang, Z., Lv, X., & Wang, J. (2017). Identifying helpful quality-related reviews from social media based on attractive quality theory. *Total Quality Management & Business Excellence*, 1-20.
Liu, Y., Jiang, C., & Zhao, H. (2017). Using contextual features and multi-view ensemble learning in product defect identification from online discussion forums. *Decision Support Systems, 105*, 1-12.
Liu, Y., Wang, L., Chen, R., Song, Y., & Cai, Y. (2016). A PUT-Based Approach to Automatically Extracting Quantities and Generating Final Answers for Numerical Attributes. *Entropy, 18*(6), 235.
Loncar, M., Barrett, N. E., & Liu, G.-Z. (2014). Towards the refinement of forum and asynchronous online discussion in educational contexts worldwide: Trends and investigative approaches within a dominant research paradigm. *Computers & Education, 73*, 93-110.
López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113-141.
Luo, X. (2009). Quantifying the long-term impact of negative word of mouth on cash flows and stock prices. *Marketing Science, 28*(1), 148-165.
Lusch, R., Liu, Y., & Chen, Y. (2010). The phase transition of markets and organizations: the new intelligence and entrepreneurial frontier.
Malbon, J. (2013). Taking fake online consumer reviews seriously. *Journal of Consumer Policy, 36*(2), 139-157.
Mandhan, S., & Niwa, Y. (2016). Numerical Atrribute Extraction from Clinical Texts. *arXiv preprint arXiv:1602.00269*.
Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. *Business horizons, 52*(4), 357-365.
Manning, C., & Raghavan, P. S. et-al. 2008. Introduction to Information Retrieval: Cambridge University Press. USA.
Marchionini, G. (1997). *Information seeking in electronic environments*. New York: Cambridge University Press.
McAlexander, J. H., Schouten, J. W., & Koenig, H. F. (2002). Building brand community. *Journal of marketing, 66*(1), 38-54.
McAuley, J., Targett, C., Shi, Q., & Van Den Hengel, A. (2015). *Image-based recommendations on styles and substitutes.* Paper presented at the Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval.
McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods, 42*(2), 381-392.
McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods, 37*(4), 547-559.
Miller, A. R., & Tucker, C. (2013). Active social media management: the case of health care. *Information Systems Research, 24*(1), 52-70.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.
Miller, T., Leroy, G., Chatterjee, S., Fan, J., & Thoms, B. (2007). *A classifier to evaluate language specificity of medical documents.* Paper presented at the System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on.
Min, B., Freedman, M., & Meltzer, T. (2017). *Probabilistic Inference for Cold Start Knowledge Base Population with Prior World Knowledge.* Paper presented at the Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers.
Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*(10), 4241-4251.
Müller, O., Debortoli, S., Junglas, I., & vom Brocke, J. (2016). Using Text Analytics to Derive Customer Service Management Benefits from Unstructured Data. *MIS Quarterly Executive, 15*(4).
Mummalaneni, V., Gruss, R., Goldberg, D., & Abrahams, A. S. (2018). Text Analytics for Safety Hazard Detection in Baby Cribs. *Safety Science, Forthcoming*.
Muniz, A. M., & O'Guinn, T. C. (2001). Brand community. *Journal of Consumer research, 27*(4), 412-432.
Myers, D. G. (1992). *The Pursuit of Happiness: What Makes a Person Happy-And Why*: William Morrow & Company.
Narisawa, K., Watanabe, Y., Mizuno, J., Okazaki, N., & Inui, K. (2013). *Is a 204 cm Man Tall or Small? Acquisition of Numerical Common Sense from the Web.* Paper presented at the ACL (1).
Negi, P. S., Rauthan, M., & Dhami, H. (2010). Sentence Boundary Disambiguation: A User Friendly Approach. *International Journal of Computer Applications, 7*(8), 33-37.
Ng, H. T., Goh, W. B., & Low, K. L. (1997). *Feature selection, perceptron learning, and a usability case study for text categorization.* Paper presented at the ACM SIGIR forum.
Ngai, E. W., Tao, S. S., & Moon, K. K. (2015). Social media research: Theories, constructs, and conceptual frameworks. *International Journal of Information Management, 35*(1), 33-44.
Nguyen, L. T., Wu, P., Chan, W., Peng, W., & Zhang, Y. (2012). *Predicting collective sentiment dynamics from time-series social media.* Paper presented at the Proceedings of the first international workshop on issues of sentiment discovery and opinion mining.
Nguyen, T., Larsen, M. E., O'Dea, B., Phung, D., Venkatesh, S., & Christensen, H. (2017). Estimation of the prevalence of adverse drug reactions from social media. *International Journal of Medical Informatics, 102*, 130-137.
Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42*(24), 9603-9611.
Nikfarjam, A., & Gonzalez, G. H. (2011). *Pattern mining for extraction of mentions of adverse drug reactions from user comments.* Paper presented at the AMIA Annual Symposium Proceedings.
Nirenburg, S., & McShane, M. (2016). Natural language processing. *The Oxford Handbook of Cognitive Science*, 337.
O'Connor, K., Pimpalkhute, P., Nikfarjam, A., Ginn, R., Smith, K. L., & Gonzalez, G. (2014). *Pharmacovigilance on twitter? mining tweets for adverse drug reactions.* Paper presented at the AMIA annual symposium proceedings.
Onah, D. F., Sinclair, J. E., & Boyatt, R. (2014). *Exploring the use of MOOC discussion forums.* Paper presented at the Proceedings of London International Conference on Education.

Packard, G., & Berger, J. (2017). How language shapes word of mouth's impact. *Journal of marketing Research, 54*(4), 572-588.
Patki, A., Sarker, A., Pimpalkhute, P., Nikfarjam, A., Ginn, R., O'Connor, K., . . . Gonzalez, G. (2014). Mining adverse drug reaction signals from social media: going beyond extraction. *Proceedings of BioLinkSig, 2014*, 1-8.
Peacemaker, B., Robinson, S., & Hurst, E. J. (2016). Connecting best practices in public relations to social media strategies for academic libraries. *College & Undergraduate Libraries, 23*(1), 101-108.
Peters, E., Västfjäll, D., Slovic, P., Mertz, C., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological science, 17*(5), 407-413.
Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review, 15*(1), 161-167.
Pexman, P. M., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review, 9*(3), 542-549.
Pexman, P. M., Siakaluk, P. D., & Yap, M. J. (2014). *Meaning in mind: Semantic richness effects in language processing*: Frontiers E-books.
Phillips, N. K. (2011). Academic library use of Facebook: Building relationships with students. *The Journal of Academic Librarianship, 37*(6), 512-522.
Pimpalkhute, P., Patki, A., Nikfarjam, A., & Gonzalez, G. (2014). Phonetic spelling filter for keyword selection in drug mention mining from social media. *AMIA Summits on Translational Science Proceedings, 2014*, 90.
Pohl, D., Bouchachia, A., & Hellwagner, H. (2018). Batch-based active learning: Application to social media data for crisis management. *Expert Systems with Applications, 93*, 232-244.
Poon, E. W. N. J., Lam, S., & Moon, K. K. (2017). Design and Development of Intelligent Decision Support Prototype System for Social Media Competitive Analysis in Fashion Industry. *Fashion and Textiles: Breakthroughs in Research and Practice: Breakthroughs in Research and Practice*, 211.
Qiao, Z., Wang, G. A., Zhou, M., & Fan, W. (2018). The Impact of Customer Reviews on Product Innovation: Empirical Evidence in Mobile Apps *Analytics and Data Science* (pp. 95-110): Springer.
Qiao, Z., Zhang, X., Zhou, M., Wang, G. A., & Fan, W. (2017). *A Domain Oriented LDA Model for Mining Product Defects from Online Customer Reviews.* Paper presented at the The 50th Hawaii International Conference on System Sciences Waikoloa, HI.
Ramírez-Gallego, S., García, S., Benítez, J. M., & Herrera, F. (2016). Multivariate discretization based on evolutionary cut points selection for classification. *IEEE transactions on cybernetics, 46*(3), 595-608.
Relations, U. o. O. I. o. G., & Sherif, M. (1961). *Intergroup conflict and cooperation: The Robbers Cave experiment* (Vol. 10): University Book Exchange Norman, OK.
Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological bulletin, 135*(6), 943.
Ridings, C. M., & Gefen, D. (2004). Virtual community attraction: Why people hang out online. *Journal of Computer‐Mediated Communication, 10*(1), 00-00.
Rosenthal, S., Farra, N., & Nakov, P. (2017). *SemEval-2017 task 4: Sentiment analysis in Twitter.* Paper presented at the Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).

Royo-Vela, M., & Casamassima, P. (2011). The influence of belonging to virtual brand communities on consumers' affective commitment, satisfaction and word-of-mouth advertising: The ZARA case. *Online Information Review, 35*(4), 517-542.

Rubens, M., & Agarwal, P., 2002. Information Extraction from Online Automotive Classifieds. Retrieved from http://www-nlp.stanford.edu/courses/cs224n/2004/cs224n_final_mrubens_agarwalp.pdf

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J, 2*(1), 127-133.

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., . . . Gonzalez, G. (2015). Utilizing social media data for pharmacovigilance: A review. *Journal of biomedical informatics, 54*, 202-212.

Sarker, A., & Gonzalez, G. (2015). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of biomedical informatics, 53*, 196-207.

Sarker, A., O'Connor, K., Ginn, R., Scotch, M., Smith, K., Malone, D., & Gonzalez, G. (2016). Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. *Drug safety, 39*(3), 231-240.

SAS, FDR LogWorth by Effect Size. *Jmp Online Documentation.* Retrieved from https://www.jmp.com/support/help/13-2/FDR_LogWorth_by_Effect_Size.shtml

Sashi, C. (2012). Customer engagement, buyer-seller relationships, and social media. *Management decision, 50*(2), 253-272.

Sawhney, M., Verona, G., & Prandelli, E. (2005). Collaborating to create: The Internet as a platform for customer engagement in product innovation. *Journal of interactive marketing, 19*(4), 4-17.

Senter, R., & Smith, E. (1967). *Automated readability index*. Retrieved from

Shelke, N., Deshpande, S., & Thakare, V. (2017). *Domain independent approach for aspect oriented sentiment analysis for product reviews.* Paper presented at the Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications.

Shi, D., Guan, J., Zurada, J., & Manikas, A. (2017). A Data-Mining Approach to Identification of Risk Factors in Safety Management Systems. *Journal of management information systems, 34*(4), 1054-1081.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of Management Information Systems, 29*(4), 217-248.

Stone, P., Dunphy, D. C., Smith, M. S., & Ogilvie, D. (1968). The general inquirer: A computer approach to content analysis. *Journal of Regional Science, 8*(1), 113-116.

Stvilia, B., & Gibradze, L. (2014). What do academic libraries tweet about, and what makes a library tweet useful? *Library & Information Science Research, 36*(3), 136-141.

Swani, K., Milne, G., & P. Brown, B. (2013). Spreading the word through likes on Facebook: Evaluating the message strategy effectiveness of Fortune 500 companies. *Journal of Research in Interactive Marketing, 7*(4), 269-294.

Tafti, A. P., Badger, J., LaRose, E., Shirzadi, E., Mahnke, A., Mayer, J., . . . Peissig, P. (2017). Adverse drug event discovery using biomedical literature: a big data neural network adventure. *JMIR medical informatics, 5*(4).

Takamura, H., & Tsujii, J. i. (2015). *Estimating Numerical Attributes by Bringing Together Fragmentary Clues.* Paper presented at the Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, Denver, CO.
Turner, J. C. (1975). Social comparison and social identity: Some prospects for intergroup behaviour. *European journal of social psychology, 5*(1), 1-34.
Van Doorn, J. (2011). Comment: customer engagement: Essence, dimensionality, and boundaries. *Journal of service research, 14*(3), 280-282.
Van Doorn, J., Lemon, K. N., Mittal, V., Nass, S., Pick, D., Pirner, P., & Verhoef, P. C. (2010). Customer engagement behavior: Theoretical foundations and research directions. *Journal of service research, 13*(3), 253-266.
Vanden Bergh, B. G., Lee, M., Quilliam, E. T., & Hove, T. (2011). The multidimensional nature and brand impact of user-generated ad parodies in social media. *International Journal of Advertising, 30*(1), 103-131.
Verhoef, P. C., Reinartz, W. J., & Krafft, M. (2010). Customer engagement as a new perspective in customer management. *Journal of service research, 13*(3), 247-252.
Vilar, S., Friedman, C., & Hripcsak, G. (2017). Detection of drug–drug interactions through data mining studies using clinical sources, scientific literature and social media. *Briefings in Bioinformatics*, bbx010.
Vivek, S. D., Beatty, S. E., & Morgan, R. M. (2012). Customer engagement: Exploring customer relationships beyond purchase. *Journal of Marketing Theory and Practice, 20*(2), 122-146.
Vohra, A., & Bhardwaj, N. (2016). A conceptual presentation of customer engagement in the context of social media-An emerging market perspective.
Wallace, E., Buil, I., & de Chernatony, L. (2014). Consumer engagement with self-expressive brands: brand love and WOM outcomes. *Journal of Product & Brand Management, 23*(1), 33-42.
Wallace, E., Buil, I., de Chernatony, L., & Hogan, M. (2014). Who "likes" you… and why? A typology of Facebook fans. *Journal of Advertising Research, 54*(1), 92-109.
Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., & Zhang, Z. (2013). ExpertRank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems, 54*(3), 1442-1451.
Wang, W. (2016). *Mining adverse drug reaction mentions in twitter with word embeddings.* Paper presented at the Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing.
Wang, X., Hripcsak, G., Markatou, M., & Friedman, C. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association, 16*(3), 328-337.
Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems, 105*, 87-95.
WHO, World Health Organization. Pharmacovigilance. *Essential Medicines and Health Products.* Retrieved from http://www.who.int/medicines/areas/quality_safety/safety_efficacy/pharmvigi/en/
WHO, 1972. World Health Organization. International Drug Monitoring: the Role of National Centres. *Technical Report No. 498.* Retrieved from https://www.who-umc.org/media/2680/who-technical-report-498.pdf
Winkler, M., Abrahams, A. S., Gruss, R., & Ehsani, J. P. (2016). Toy safety surveillance from online reviews. *Decision Support Systems, 90*, 23-32.

Wissmann, M., & Toutenburg, H. (2011). Role of Categorical Variables in Multicollinearity in Linear Regression Model. *Journal of Applied Statistical Science, 19*(1), 99.
Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology, 8*(1), 52-69.
Xia, L., Wang, G. A., & Fan, W. (2017). *A Deep Learning Based Named Entity Recognition Approach for Adverse Drug Events Identification and Extraction in Health Social Media.* Paper presented at the International Conference on Smart Health.
Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management, 58*, 51-65.
Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management, 44*, 120-130.
Xiang, Z., Schwartz, Z., & Uysal, M. (2017). Market Intelligence: Social Media Analytics and Hotel Online Reviews *Analytics in Smart Tourism Design* (pp. 281-295): Springer.
Xu, J., Xu, B., Wang, P., Zheng, S., Tian, G., & Zhao, J. (2017). Self-Taught convolutional neural networks for short text clustering. *Neural Networks, 88*, 22-31.
Yang, M., Wang, X., & Kiang, M. Y. (2013). *Identification of Consumer Adverse Drug Reaction Messages on Social Media.* Paper presented at the PACIS.
Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information retrieval, 1*(1), 69-90.
Yang, Y., Bao, F. S., & Nenkova, A. (2017). Detecting (Un) Important Content for Single-Document News Summarization. *arXiv preprint arXiv:1702.07998*.
Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review, 18*(4), 742-750.
Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112*(4), 1036-1040.
Zhang, M., Guo, L., Hu, M., & Liu, W. (2017). Influence of customer engagement with company social networks on stickiness: Mediating effect of customer value creation. *International Journal of Information Management, 37*(3), 229-240.
Zhang, M., Hu, M., Guo, L., & Liu, W. (2017). Understanding relationships among customer experience, engagement, and word-of-mouth intention on online brand communities: The perspective of Service Ecosystem. *Internet Research*(just-accepted), 00-00.
Zhang, S., Jiang, H., & Carroll, J. M. (2010). Social identity in Facebook community life. *IGI Global*, 64-76.
Zhang, X., Qiao, Z., Tang, L., Fan, W., Fox, E., & Wang, G. (2016). Identifying Product Defects from User Complaints: A Probabilistic Defect Model. https://vtechworks.lib.vt.edu/handle/10919/64902.
Zhang, Y., Dai, H., Kozareva, Z., Smola, A. J., & Song, L. (2017). Variational Reasoning for Question Answering with Knowledge Graph. *arXiv preprint arXiv:1709.04071*.
Zhang, Y., Dang, Y., Chen, H., Thurmond, M., & Larson, C. (2009). Automatic online news monitoring and classification for syndromic surveillance. *Decision Support Systems, 47*(4), 508-517.
Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management, 29*(4), 694-700.

Zhao, K., Stylianou, A. C., & Zheng, Y. (2017). Sources and impacts of social influence from online anonymous user reviews. *Information & Management*.
Zheng, W., Tang, D., Zhang, H., & Tang, H. (2017). *Feature Selection with Structural Sparse Mode for Text Categorization.* Paper presented at the Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2017 9th International Conference on.