On Trust, Privacy, and Misinformation in Health Social Media

by

Hamman Waqar Samuel

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

University of Alberta

Abstract

It has gotten increasingly harder for laypersons to determine the veracity of online health information. This is because of the explosion of content in health social media, allowing anyone with an Internet connection to create and propagate health-related content. This includes both innocuous and malignant pseudo-medical advice. On the other hand, medical professionals are able to discern medical facts from fiction using systematic methodologies.

This thesis develops and evaluates pragmatic computational models for evaluating the veracity of health-related online content. Firstly, medical knowledge and evidence-based practices are incorporated into health social media through the *MedFact* algorithm and *Veracity Score*. Secondly, privacy and anonymity requirements of social media are taken into account using the *Iron Mask* algorithm and *Trust-Preserving Pseudonyms*. Thirdly, these solutions are incorporated into a health portal for patients and medics, code-named *Cardea*, that models various types of interactions occurring on health social media, including real-time chat rooms, blogs, question-answering, and support groups. Cardea allows users to share experiences, ask questions, and get answers in three streamlined environments: Patient to Patient, Patient to Medic, and Medic to Medic. Patients are able to chat with other patients, create support communities, and also ask questions specifically to medical experts. Medics can respond to patient questions and also have private and secure discussions with other medics.

The road map for this manuscript is organized by chapters into three predominant groups. Chapters 1 and 2 provide the background to the thesis. In Chapter 1, the motivation, background, and thesis statement are provided. In Chapter 2, a survey of literature related to trust, privacy, and evidence-based medicine is covered.

Chapters 3, 4, and 5 expand the key concepts of the hypotheses on trust, privacy, and health social media. In Chapter 3, the MedFact algorithm is explicitly defined as an objective metric for computational estimation of trust. In Chapter 4, the Iron Mask algorithm is explained as a mechanism towards preventing social stigma while preserving reputation. In Chapter 5, Cardea is described in detail, including its embedded frameworks and components of trust, privacy, and security. Chapters 6, 7, and 8 cover additional artifacts developed as part of this research for use within Cardea for data collection, content recommendation, and duplicate content detection.

Chapter 9 concludes this manuscript with an outlook on future research potential. *Takeaway* boxes are used throughout the manuscript to highlight and summarize key concepts, results, and contributions. Ultimately, this thesis gives new perspectives on a computational definition of trust with an awareness of privacy in health social media. The proposed methods have the potential for assisting users to sift through large volumes of online information and make informed decisions about their health using trustworthy information sources without compromising privacy.

Preface

This thesis and the related publications are original works by Hamman Waqar Samuel, including problem formulation, implementation, evaluation, and writing. A majority of the research described in this thesis is drawn from published peer-reviewed papers authored or co-authored by Hamman Waqar Samuel. The publications are enumerated below, including related thesis chapters. Empirical studies that are part of this thesis received research ethics approval from the University of Alberta Research Ethics Board, under the project title "User Perspectives on Trust in Health Social Media" and project identifier Pro00079019.

The following peer-reviewed publications have been incorporated into the corresponding chapters.

- Hamman Samuel, Osmar Zaïane. MedFact: Towards Improving Veracity of Medical Information in Social Media using Applied Machine Learning. Canadian Conference on Artificial Intelligence (CAI). Toronto, Canada, May 2018 (**Chapter 3**).

- Hamman Samuel, Osmar Zaïane. Iron Mask: Trust-Preserving Anonymity on the Face of Stigmatization in Social Networking Sites. International Conference on Trust, Privacy & Security In Digital Business (TrustBus). Lyon, France, Aug. 2017 (**Chapter 4**).

- Hamman Samuel, Fahim Hassan, Osmar Zaïane. The Need for Medical Professionals to Join Patients in the Online Health Social Media Discourse. International Conference on Health Informatics (HEALTHINF). Feb. 2021 (**Chapter 5**).

- Hamman Samuel, Benyamin Noori, Sara Farazi, Osmar Zaïane. Context Prediction in the Social Web using Applied Machine Learning: A Study of Canadian Tweeters. IEEE/WIC/ACM International Conference on Web Intelligence (WI). Santiago, Chile, Dec. 2018 (**Chapter 6**).

- Hamman Samuel, Osmar Zaïane. PubMedReco: A PubMed Citations Recommender System for Real-Time Chat (*Best Paper - 2nd Place*). IMIA World Congress on Medical and Health Informatics (MEDINFO). Hangzhou, China, Aug. 2017 (**Chapter 7**).

- Mohomed Shazan Mohomed Jabbar, Luke Kumar, Hamman Samuel, Mi-Young Kim, Sankalp Prabhakar, Randy Goebel, Osmar Zaïane. On Generality and Knowledge Transferability in Cross-Domain Duplicate Question Detection for Heterogeneous Community Question Answering. *Preprint* arXiv: 1811.06596 [cs.CL]. Nov. 2018. (**Chapter 8**).

Other research publications that were completed during the development of this thesis are listed below.

- Abhishek Dhankar, Nawshad Farruque, Hamman Samuel, Fahim Hassan, Osmar Zaïane. Detecting COVID-19 Misinformation in Social Media using Augmented Translational Learning. *Under Review* at the Florida AI Research Society (FLAIRS) International Conference - Special Track on AI in Healthcare Informatics. May 2021.

- Hamman Samuel, Mi-Young Kim, Sankalp Prabhakar, Mohomed Shazan Mohomed Jabbar, Osmar Zaïane. Community Question Retrieval in Health Forums. IEEE-EMBS International Conference on Biomedical & Health Informatics (ICBHI). Orlando, USA, Feb. 2017.

- Hamman Samuel, Osmar Zaïane, Patricia Martz. Supporting Digital Epidemiology in Alberta via Twitter Tracking. Presented at the IEEE-EMBS International Conference on Biomedical & Health Informatics (ICBHI). Orlando, USA, Feb. 2017.

- Saeed Mohajeri, Hamman Samuel, Osmar R. Zaïane, Davood Rafiei. BubbleNet: An Innovative Exploratory Search and Summarization Interface with Applicability in Health Social Media (*Best Paper*). International Conference on Digital Economy (ICDEc). Carthage, Tunisia, Apr. 2016.

- Hamman Samuel, Mi-Young Kim, Sankalp Prabhakar, Mohomed Shazan Mohomed Jabbar. Golden Retriever: Question Retrieval System. IEEE Intl. Conference on Healthcare Informatics (ICHI). Dallas, USA, Oct. 2015.

- Hamman Samuel, Osmar R. Zaïane. A Repository of Codes of Ethics and Technical Standards in Health Informatics. Online Journal of Public Health Informatics (OJPHI). Vol. 6, No. 2, Oct. 2014.

- Hamman Samuel, Osmar R. Zaïane. On Management of the Health Content Lifecycle. Public Health Frontiers (PHF). Vol. 2, No. 2, June 2013.

- Hamman Samuel. Evidence-Based Trust Metrics in Web Services. International Conference on Computer Science and Software Engineering (ICSE). Markham, Canada, Nov. 2013.

Finally, the following publications were part of my master's thesis titled "Content Management for Online Health Advice Sharing" and contain research works that were precursors to the development of this thesis.

- Hamman Samuel, Osmar R. Zaïane, Jane Robertson Zaïane. Findability in Health Information Websites. IEEE-EMBS International Conference on Biomedical & Health Informatics (ICBHI). Hong Kong / Shenzhen, China, Jan 2012.

- Hamman Samuel, Osmar R. Zaïane. PSST...Privacy, Safety, Security, and Trust in Health Information Websites. IEEE-EMBS International Conference on Biomedical & Health Informatics (ICBHI). Hong Kong / Shenzhen, China, Jan 2012.

- Hamman Samuel, Osmar R. Zaïane. HCMS: Conceptual Description of a Health Content Management System. International Workshop on Software Engineering in Health Care (SEHC). Honolulu, USA, May 2011.

- Metanat HooshSadat, Hamman Samuel, Sonal Patel, Osmar R. Zaïane. Fastest Association Rule Mining Algorithm Predictor (FARM-AP). Canadian Conference on Computer Science and Software Engineering (C2S3E). Montreal, Canada, May 2011.

- Hamman W. Samuel, Osmar R. Zaïane, Dick Sobsey. Towards a Definition of Health Informatics Ethics. ACM International Health Informatics Symposium (IHI). Arlington, USA, Nov. 2010.

Takeaway

- Peer-Reviewed Publications: **17**
- Under Review Manuscripts: **1**
- Pre-Print Manuscripts: **1**
- Publications During PhD: **12**

For my father,

Naimat Samuel

Little by little, wean yourself.
This is the gist of what I have to say.
From an embryo, whose nourishment comes in the blood,
move to an infant drinking milk,
to a child on solid food,
to a searcher after wisdom,
to a hunter of more invisible game.

Think how it is to have a conversation with an embryo.
You might say, “The world outside is vast and intricate.
There are wheatfields and mountain passes,
and orchards in bloom.
At night there are millions of galaxies, and in sunlight
the beauty of friends dancing at a wedding.”
You ask the embryo why he, or she, stays cooped up
in the dark with eyes closed.

Listen to the answer.

“There is no ‘other world’.
I only know what I’ve experienced.
You must be hallucinating.”

– Rumi

Acknowledgements

My wholehearted appreciation to my supervisor, Dr. Osmar Zaïane, for his absolute support throughout my degree. I am truly grateful to him for encouraging and mentoring me through the challenging and prosperous periods of my doctoral program.

I am very thankful to my supervisory committee, Dr. Lisa Hartling and Dr. Ken Wong for their constructive feedback on my research work. I would also like to acknowledge my examining committee, Dr. Abram Hindle, Dr. Tami Oliphant, and Dr. Sandra Bringay, for their invaluable time.

I also appreciate my co-authors and others that I have done collaborative research with: Mi-Young Kim, Mohomed Shazan Mohomed Jabbar, Sankalp Prabhakar, Patricia Martz, Luke Kumar, Benyamin Noori, Sara Farazi, Abhishek Dhankar, Nawshad Farruque, Fahim Hassan, Saeed Mohajeri, Dr. François Bolduc, Dr. Randy Goebel, and Dr. Davood Rafiei.

I am very grateful to my course professors for expanding my views on various diverse areas of computing science, including software systems, heuristic search, information retrieval, two-player game theory, and formal software verification: Dr. Eleni Stroulia, Dr. Rob Holte, Dr. Denilson Barbosa, Dr. Martin Müller, and late Dr. Piotr Rudnicki.

I would also like to acknowledge the Alberta Machine Intelligence Institute (AMII) for funding various aspects of my research work, and to Cybera for providing free infrastructure and resources towards this research work.

I am always grateful to my family for their support and inspiration: my father Naimat Samuel, my late mother Shahida F. Samuel, my wife Dr. Yanwu Gu, and my sisters Nousheen Samuel, Nayyara Samuel, Naima Samuel, and Saima Samuel.

Contents

Acronyms xv

Glossary xvii

1 Introduction 1
1.1 Motivation . . . 1
1.2 Problem Statement . . . 3
1.3 Research Considerations . . . 4
1.4 Thesis Statement . . . 6
1.5 Key Contributions . . . 7

2 Background 11
2.1 Current State of Research in Trust and Social Media . . . 11
2.2 Psychological Viewpoint on Health Misinformation . . . 15
2.3 Identity Privacy, Stigmatization, and Anonymity . . . 16
2.4 Incorporating Trust in Search and Recommender Systems . . . 19

3 MedFact 22
3.1 Motivating Objective Trust Metrics . . . 22
3.2 Evidence-Based Practice . . . 23
3.3 Combining Evidence with Trust using MedFact . . . 23
3.4 Feedback on Usage of Veracity Score . . . 28
3.5 Survey of Medical Professionals on Controversial Topics in Pediatrics 29
3.6 Veracity Score on Unproven Cancer Treatments . . . 33
3.7 Online Medic Discussions Evaluated with Veracity Scoring . . . 34
3.8 Veracity of COVID-19 Medical Claims and Bootstrapping MedFact 35

4 Iron Mask 37
4.1 Motivation for Anonymity and Drawbacks . . . 37
4.2 Relating Anonymity and Trust . . . 39
4.3 Iron Mask Algorithm Overview . . . 39
4.4 Whiteprint Identification using Probabilistic Classification . . . 42
4.5 Linking Pseudonyms with Level of Expertise . . . 43
4.6 Evaluation of Iron Mask . . . 44

5 Cardea 50
5.1 State of Health Social Media . . . 50
5.2 Functionality Specifications . . . 52
5.3 Prototype with Data . . . 57
5.4 Promoting Credible Content . . . 59

6 Grebe **61**
6.1 Motivation for Social Data Aggregation 62
6.2 Analysis Methodology on Grebe Dataset 64
6.3 Results from Grebe's Analytics 71
6.4 Discussion and Notes on Grebe's Performance 75

7 PubMedReco **78**
7.1 Motivation for PubMed Recommendations 78
7.2 Development Workflow of PubMedReco 80
7.3 Performance Analysis of PubMedReco 83
7.4 Discussion and Appraisal of Results 85

8 DeepDup **88**
8.1 Motivation and Challenges for Duplicate Question Detection 88
8.2 DeepDup Architecture and Design 90
8.3 Performance of Duplicate Detection Methods 93
8.4 Review and Interpretation of Results 94

9 Conclusion **96**

Bibliography **99**

List of Tables

3.1 Survey Statements by Rubric and Topic 31
3.2 Comparison of Responses by Medical Professionals versus MedFact 33

4.1 Quora Dataset for Evaluation Filtered by Topics 45

6.1 Grebe Dataset Statistics 71

7.1 PubMedReco Neural Network Keyword Features 81

8.1 Dataset Properties 90

List of Figures

3.1 Overview of MedFact Algorithm 24
3.2 Summary of Veracity Score Survey Responses 29
3.3 Demographic Information of Participant Medical Professionals . . . 32
3.4 Veracity Score on Random Articles from QuackWatch (all are identified with low veracity score as expected) 34
3.5 Veracity Score Comparisons between DocCheck and QuackWatch (clear separation between true and false health information) 35

4.1 Overview of Anonymization and Pseudonyms 40
4.2 Average Recall for Top-n Configurations 46
4.3 Average Recall for τ Configurations 47
4.4 Iron Mask Survey Step 1 Results 48
4.5 Iron Mask Survey Step 2 Results 49

5.1 Cardea Homepage Screenshot 53
5.2 Cardea's Content Filtering Features 57
5.3 PubMedReco Real-Time Chat Recommendation Demonstration . . 58
5.4 Duplicate Content Detection Examples 58

6.1 Twitter API Geographical Filtering Options 64
6.2 Public Tweet Information Indexed by Grebe 65
6.3 Grebe Visualizations Showcase 72
6.4 Prediction of Canadian Provinces and Cities 73
6.5 Identifying Health-Related Tweets by Wellness Dimension 74
6.6 Average Monthly Polarity of Health Tweets from Canadian Provinces 75
6.7 Emotion Multi-Labels of Health-Related Tweets from Alberta . . . 76

7.1 Sequence diagram depicting interactions between PubMed, MeSH, UMLS and MEDLINE during search 79
7.2 Precision of PubMedReco Neural Network 84
7.3 Kappa Score for Annotations 85
7.4 NDCG for Recommended Citations 85
7.5 Dynamic Window for Discussions D1, D2, D3 86

8.1 Siamese Neural Network for Duplicate Question Detection 91
8.2 Results from Various Machine Learning Approaches on Heterogeneous Datasets 94

Acronyms

ABA Applied Behavior Analysis 30, 31

ADHD Attention Deficit Hyperactivity Disorder 30, 31

AUC Area Under the Curve 93

CHV Consumer Health Vocabulary 25, 72

CNN Convolutional Neural Network 27, 91

CQA Community Question Answering 88, 89, 95

CTR Click-Through Rate 20

DARE Database of Abstracts of Reviews of Effects 7, 26

DCG Discounted Cumulative Gain 20

EBM Evidence-Based Medicine 4, 14, 21, 23, 24, 36

EBP Evidence-Based Practice 2, 6, 7, 11, 14

HSM Health Social Media 11, 50, 51, 52, 53, 57, 59, 60, 61, 77, 96, 97, 98

IR Information Retrieval 7, 35, 89

LDA Latent Dirichlet Allocation 66

LSTM Long Short Term Memory 91

M2M Medic to Medic 8, 53, 59

MAE Mean Absolute Error 20

MAP Mean Average Precision 20, 21

MeSH Medical Subject Headings 79

MLP Multi-Layer Perceptron 66, 73

MSE Mean Squared Error 20

MSSE Medical Sciences Stack Exchange 27, 59

NCBI National Center for Biotechnology Information 78

NDCG Normalized Discounted Cumulative Gain 7, 20, 26, 59, 83, 84

NLM National Library of Medicine 26

NLP Natural Language Processing 7, 25, 35, 42, 89

OEC Owner Engineered Content 1, 51

P2M Patient to Medic 8, 9, 53, 54, 59, 87

P2P Patient to Patient 8, 53, 59

PICO Population, Intervention, Control and Outcomes 14

PII Personally Identifiable Information 4

POS Part-Of-Speech 26

PPE Personal Protective Equipment 2

RMSE Root Mean Squared Error 20, 45

SE Stack Exchange 89, 90, 93, 94

SEW Simple English Wikipedia 25

SNN Siamese Neural Network xiv, 90, 91, 93

SNOMED Systematized NOmenclature of MEDicine 25, 72, 81

SSDLC Secure Software Development Life Cycle 8

SVM Support Vector Machine 66, 73

TCM Traditional Chinese Medicine 2

TL Transfer Learning 92, 93, 94

TPP Trust Preserving Pseudonym 6, 8, 37, 38, 39, 40, 43, 44, 46, 47, 48, 49, 96

TRIP Turning Research Into Practice 26

Triple P Positive Parenting Program 30, 31

UGC User Generated Content 1, 50

UMLS Unifed Medical Language System 25, 79, 82

WOT Web Of Trust 21

Glossary

Asynchronous This form of online communication enables users to send messages and information without the expectation that they would get an instantaneous response 53, 56

Binary Classifier A type of machine learning classifier that predicts only one of two discrete labels, e.g. 0 or 1 25, 92

Bootstrapping Process of loading the bare minimum and necessary data for a system begin operating 3, 35

Cold Start Problem caused by unavailability of data for proper initialization or training of a system 3, 12

Credibility Equivalent to veracity or trust, but usually applies to entities such as users, organizations, or websites with some measure of reputation 3, 4, 5, 8, 13, 39, 48

Cron Job A process or program that will run on regular intervals based on its scheduling, for example every day at noon or weekly 65, 75, 76

Deep Neural Network A type of neural network that typically has many intermediate layers between the initial input and final output layers 25, 73, 81, 89, 90, 92, 95

Deep Learning Deep learning is a category of machine learning that extensively leverages deep neural networks 10, 27, 57, 88

Hidden Markov Models Statistical models working on the assumption that the models themselves are Markov processes, i.e. they are stochastic and the probability of any given event occurring is dependent on prior state 13

Isotonic Regression Statistical and machine learning inference technique for finding best approximation function by fitting a free-form line to observations; the increasing or decreasing trend of the line is maintained while keeping as close as possible to the observations 42

k-Anonymization Method of processing user personal data in order to make it difficult to identify unique individuals with a given dataset; describes the level of anonymity for a given dataset 18, 43

n-Grams A sequence of n-tokens extracted from text, typically by delimiting and splitting the text using spacing or other special characters; unigrams contain only one token, while bigrams contain two tokens in the same sequence as the source text, and so on 42, 43, 66, 68

Naïve Bayes Classifier A type of simplistic probabilistic classifier that leverages Bayes' theorem to compute the probability of membership of a given observation to a given class from a set of classes 18, 42, 68, 69

Neural Network Computing methodology based on the biological structures and abstracted processes of animal brains; a network of nodes and edges connected in a hierarchy where the outputs from one hierarchical level are inputs to the next level 10, 25, 27, 57, 66, 73, 80, 81, 82, 83, 84, 86, 87, 93

Probabilistic Classifier A type of classifier that can predict a probability distribution over a set of classes from observations of inputs 41, 42, 43, 45

Skip-Grams A special case of n-grams where the original sequence of tokens within the text is not preserved 25, 81

Stemming Process to reduce an inflected word to its root form, often by removing suffixes that would occur in variants, e.g. the stem word for both "*wait*ing" and "*wait*ed" would be "*wait*" 25, 42, 68, 80, 82, 90

Stop Words Commonly words within a language that do not add much value to the semantic meaning of phrases or sentences, e.g. "and", "so", "for", etc. 25, 42, 66, 68, 80, 90

Supervised Machine Learning A type of machine learning methodology that uses a known dataset with labels to make predictions on labels for new datasets; given $f(x) = y$, a machine learning model uses training datasets containing x and y to determine the functional relationship between x and y, i.e. f 13

Sybil Attacks Involves an attacker gaining unfair advantage within a social network through the attacker creating multiple profiles; an attacker vote-bombing or flooding comments on social media posts using multiple accounts 18

Synchronous This type of online communication allows users to instantly view messages being sent to them, and respond immediately if needed 9, 54, 56, 86

Veracity Characteristic of information being in conformance with known facts; in the case of this thesis, it applies to how factual is health social media content when compared with known facts from established medical knowledge and literature 3, 4, 7, 8, 9, 10, 12, 14, 17, 21, 23, 24, 26, 28, 29, 30, 32, 33, 34, 35, 36, 52, 55, 59, 60, 96

Whiteprint Unique writing style of an author based on grammatical and lexical patterns observed in their body of writings 39, 42, 44, 45

Whitewashing Situation where a user leaves a social network due to being kicked out or flagged and re-joins using a different identity 3, 12

Chapter 1

Introduction

1.1 Motivation

Health, the focus of this research work, is not merely the presence or absence of disease or infirmity, and can be defined as a combination of physical, intellectual, occupational, spiritual, emotional, and social wellness [1]. Online health information refers to over the Internet content about personal well-being, prevention and management of diseases, and other medical topics related to healthcare [2].

Patients and medical professionals alike have been increasingly using online resources to inform themselves about health topics [3]. Some reasons for this trend include easy accessibility, easy content creation, users' desire to connect and share personal experiences with others, and users' understanding of stories in personal layperson language in contrast with scientific literature. Studies also show users with acute and chronic conditions are more likely to search online for solutions [4]. At the same time, users with possibly trivial medical conditions or symptoms might also search online rather than waiting for prolonged periods at medical facilities [5], [6].

Visitors to health-related websites tend to be confronted with varying degrees of information quality. Most of these websites feature Owner Engineered Content (OEC) that is static and maintained by the website owners. This type of content can be viewed by users but not edited or interacted with through commenting. Social media, on the other hand, provides a new set of online tools and websites that enable users to create their own User Generated Content (UGC) with posts and comments.

Users can interact with existing content through feedback, such as “Likes” or ratings, commenting, conversation threads, and sharing. Using social networks, a subset of social media, users can also connect with each other to chat, share content, target questions, and build virtual support groups. Some popular websites with health information include WebMD, Mayo Clinic, Doctissimo, MedicineNet, MedlinePlus, My Health Alberta, among others.

With the vast amount of information available online, certain information seeking skill sets are needed to locate the right information. Misinformation is getting increasingly prevalent, and health misinformation can have potentially severe consequences on the information consumer’s well-being. For example, viral social media posts were used to falsely associate vaccinations with autism [7]. Articles supposedly written by medical professionals that linked autism and vaccinations were heavily shared on Facebook and other social networks, leading to a persistent perception among many users that vaccinations are harmful. On the other hand, not getting vaccinated can negatively affect public health overall and lead to pandemics. More recently during the COVID-19 pandemic, a large body of misinformation was circulated via social media, leading to the outbreak being termed an “infodemic” [8, 9]. This included conflicting information about symptoms, prevention, and treatments, such as COVID feet as a symptom, debates on the efficacy of Personal Protective Equipment (PPE) such as masks, and ineffective homeopathic, Unani [10], or Traditional Chinese Medicine (TCM) [11] cures. In these cases, matching laypersons to trusted knowledge has become critically important.

There is a clear distinction between how laypersons and medical professionals consume and evaluate health information. Medical experts use the best available evidence coupled with professional experience and expertise to determine trustworthiness of information being presented to them through Evidence-Based Practice (EBP) methodology [12], [13], [14]. The systematic EBP approach is currently underutilized in health social media because the majority of online social interactions are related to user-generated content that is created by non-experts without adequate knowledge of EBP. Research shows that laypersons prefer storytelling and personal reputation rather than systematic methodologies [15].

There are various definitions of trust [16], and it is often used interchangeably with reliability, veracity, and credibility, but the general consensus is that trust involves a willing interaction between two or more entities. There is an implicit belief that the interaction will at least be self-beneficial in the worst case, and mutually beneficial to all entities involved in the best case [17], [18]. There is no guarantee that this belief is correct. However, some level of trust, however minuscule, is fundamentally essential for interactions to happen, even when given limited or nonexistent knowledge about another entity or group of entities. Despite the popularity of health social networks and an implicit sense of trustworthiness between users, untested claims for cures, lack of contraindications, and false claims about disease prevention are common features, having life-threatening potential [19].

Significant work has been done on theoretical foundations of computational trust [20], [21], [22], [23], [24], [25], [26], [27]. However, in practice, common trust metrics found in social networks are scaled unary ratings such as “Likes” or “+1,” binary ratings acting as positive or negative votes, ranked ratings such as Likert scale rankings, as well as reputation systems for measuring user trust using achievement levels, badges, and gamification [28], [29]. Formalized methods have not been adopted in practice because laypersons need a simplified representation of trust. Reputation and ratings-based systems are more accessible to laypersons but are also susceptible to inflation, bootstrapping, whitewashing, and cold start [30].

1.2 Problem Statement

There has been a great deal of research on different approaches for aggregating reputations and ratings, which can be considered *subjective* trust metrics. However, there have been very few studies on *objective* trust metrics that incorporate medical knowledge as ground truth. While subjective opinions about facts may differ, there are certain objective and established facts that are agreed upon by the community. Likewise, in the medical domain, there are established facts and methodologies that ought to be part of the online information consumption process so laypersons can make better decisions regarding fact versus fiction online.

Search and recommender results also need to incorporate an additional dimension of trust with information processing. Essentially, users seeking online information rely on search algorithms to be matched to their intended information needs. However, search engines are only recently beginning to move towards incorporating veracity in their search result rankings. These early efforts rely on ranking curated trusted sources higher in search and recommender results. This can lead to questions around bias and conflict of interest.

Social interactions online also raise questions about privacy, a basic human right to control one's personal information and be able to decide who has access to it [31]. This also includes the content author's right to keep their identity private if they choose to [32]. On the other hand, the content author's identity is tied with their reputation, which is a key factor in determining veracity. It is sometimes desirable to hide Personally Identifiable Information (PII), including identity, to prevent social stigmatization when engaging in topics that may seem sensitive or controversial, such as reproductive health. Research has been done on mechanisms for allowing anonymity in [33], [34], but the risk of de-anonymization and subsequent stigma has not been considered by present work.

1.3 Research Considerations

Research topics that arise in online health information and health social media research include the following. The thesis addresses most of these questions.

- *How to measure trustworthiness, credibility, or veracity of online health information to distinguish misinformation?* Subjective metrics and reputation systems are identified as existing solutions, while objective metrics are proposed as a solution in Chapter 3.

- *How to leverage methods used by medical experts to identify trustworthy content and make them accessible to laypersons?* Evidence-Based Medicine (EBM) is investigated and MedFact is proposed as a solution in Chapter 3.

- *How do medical and information experts match their information needs and query intent with search results?* Research findings related to this question are included in Chapter 2.

- *How to effectively flag health misinformation while avoiding the backfire effect and confirmation bias among laypersons?* Potential solutions to this are covered in Chapters 3 and 7.

- *How to effectively suggest appropriate health content to health information consumers from trusted sources such as PubMed?* PubMedReco is presented as a potential solution and detailed in Chapter 7.

- *How to build a recommender system to show relationships between information, and match information needs with relevant and trusted content?* Potential solutions are discussed in Chapters 5, 7, and 8.

- *How to use active learning to filter information sources and provide only relevant information to a user based on their changing interests?* Some exploration of this question is provided in Chapters 2 and 7.

- *How to determine a new discussion topic is already discussed elsewhere and avoid duplication and information spam?* DeepDup is used within Cardea to address this question, with details provided in Chapters 5 and 8.

- *How to contextualize user credibility ratings according to topics of expertise?* Possible solutions on this question are presented in Chapter 4.

- *How to provide anonymity controls to online users for preventing social stigma while ensuring content from anonymous users can still be trusted?* Iron Mask is developed as a solution, with details provided in Chapter 4.

1.4 Thesis Statement

From the open research challenges in the domain of online health information and health social media, possible are synthesized as hypotheses.

Hypothesis 1: Evidence-based objective trust metrics can distinguish between established medical facts and health misinformation

Objective metrics can leverage medical knowledge established by expert consensus. Ultimately, objective metrics are able to explore agreements and contradictions between layperson opinions and known medical facts. In order to ascertain this hypothesis, a computerized EBP workflow is proposed to incorporate and compare known medical knowledge into social media discourse. The efficacy of this workflow is demonstrated by comparing how objective metrics distinguish between health misinformation and established medical facts.

Hypothesis 2: A Trust Preserving Pseudonym (TPP) can convey an anonymized user's reputation without revealing their identity

While anonymous sharing of health content to avoid stigmatization has been investigated, concurrent prevention of de-anonymization has not been taken into consideration. Identity-based trust from authority figures, mitigating whiteprint identification via differential privacy, and general anonymizaton have been covered separately in literature, but there is little research on the overlap between trust and anonymization. At the same time, this is an important area to cover given the proliferation of health social media into mainstream culture and the sensitivity of personal health topics such as fertility and obesity. To evaluate this hypothesis, the TPP concept is computationally implemented and its effectiveness for anonymization, trust-measurement, and stigma prevention is measured using mixed methods.

Hypothesis 3: Information retrieval metrics in conjunction with trust metrics can provide more credible health information results for search engines and recommender systems

Information retrieval typically focuses on best serving the user's search intent by measuring overlap of query keywords with search results. Matching the user's intent to appropriate content remains an open challenge for search engines and recommender systems. Existing information retrieval metrics such as Normalized Discounted Cumulative Gain (NDCG) rank search and recommender results based solely on semantic details. Trustworthiness of the results was generally not considered until recently. However, current approaches rely on human moderation or hand-curating trusted sources, leading to questions of bias and conflict of interest. Incorporating trust metrics into search and recommender result rankings has the potential to surface credible health content and demote unsafe and health information. This hypothesis will be evaluated by comparing the number of search or recommender results within Cardea that are deemed trustworthy or false.

1.5 Key Contributions

1.5.1 MedFact: Objective Health Trust Assessment

MedFact is a system for objectively computing the veracity of health-related claims contained in text paragraphs. As an example, the text may be taken from layperson social media discussions, and the veracity of some health-related statements may be uncertain. MedFact leverages the EBP systematic approach for appraising health information on the basis of the best current evidence taken from medical knowledge bases. Credible medical knowledge is queried from trusted sources including MEDLINE and PubMed articles, the Cochrane database of systematic reviews, Database of Abstracts of Reviews of Effects (DARE), Cochrane Clinical Answers, WebMD's Medscape, and others. MedFact compares the relationship between claims against related medical facts extracted from the medical knowledge sources through Information Retrieval (IR) and Natural Language Processing (NLP), and quantifies the degree of agreement or disagreement between text using machine learning.

1.5.2 Iron Mask: Trust-Preserving Anonymity

Iron Mask uses the whiteprint or authorship identification approach [35] to take into account the user's historical content, thereby enhancing anonymity by minimizing the risk of re-identification and decreasing the likelihood of online stigma. Iron Mask also provides trust-preservation to balance the social network's needs to generate credible content with the user's need for optional yet reliable anonymity. The naïve approach of explicitly revealing information related to user credibility would constitute a quasi-identifier, and could lead to identity being compromised through correlations [36]. To avoid this, Iron Mask introduces the concept of the TPP, which provides a broader range of pseudonym labels, in addition to the generic "anonymous" pseudonym to mask or cover up the user's actual account name identity while preserving user reputation.

1.5.3 Cardea: Health Social Network

This thesis proposes a model that addresses overarching questions of estimating content veracity against known medical facts, matching content with users' needs, and users' right to privacy. The proposed solutions to these challenges are implemented in *Cardea*, a health portal for medical professionals or medics, and laypersons or patients. Cardea aims to include features from existing social networks [3], and empower patients to determine the veracity of health information. Users can share experiences, ask questions, and get answers in three streamlined environments: Patient to Patient (P2P), Patient to Medic (P2M), and Medic to Medic (M2M). Cardea as an online web application is built using the Secure Software Development Life Cycle (SSDLC) process [37], and incorporates the proposed Med-Fact and Iron Mask algorithms.

1.5.4 Grebe: Health Social Media Data Aggregation

Grebe is a social media data aggregation platform built during this research work to fulfill the need for health data within Cardea for testing, analysis, and training. Grebe has over 28 million tweets indexed since July 2014 to December 2020.

Grebe also enables indexing of health social media data from various social networks and forums, including The BMJ's Doc2Doc, Doctors Lounge, DocCheck, eHealth Forum, and Medical Sciences Stack Exchange. Additionally, data has also been collected from Quora and other Stack Exchange subsidiaries with Grebe. Ultimately, the vision of Grebe is to provide researchers with Canada-specific social web datasets through an open source platform with an accessible RESTful API. Public discourse on health social media websites can also provide valuable opportunities for digital epidemiology, including tracking and predicting disease outbreaks and pandemics such as COVID-19. Grebe is also used in Cardea for populating the prototype with relevant content from external social media sources, including discussions, questions, and blogs. Additional technical details on the Grebe platform are provided in Chapter 6.

1.5.5 PubMedReco: Real-Time Conversation Recommender

Users of health social media websites predominantly use two mechanisms for locating relevant information: search box or recommender system. In order to address the need for incorporating veracity into recommender systems, PubMedReco was developed, a recommender for PubMed citations that can instantaneously recommend medical article citations while users are conversing in a synchoronous communication environment such as a chat room or browsing health-related web pages. PubMedReco integrates into Cardea's P2M chat rooms. It automatically generates the search query and shows relevant citations within the same integrated user interface. The queries are generated from themes and topics across multiple conversations. The citations help users get factual information within the chat room interface. Other trusted sources are also used with PubMedReco, such as layperson-friendly blogs or advisories from Health Canada. PubMedReco is used within Cardea to provide real-time recommendations in P2M chats. Full details about PubMedReco are provided in Chapter 7.

1.5.6 DeepDup: Duplicate Content Reduction

Identification of duplicate questions can reduce the resources required for information retrieval. Reducing duplicate information in social media also helps users find relevant and trusted content faster without information overload. Duplicate information detection is an open research question with the main challenge being how to determine semantic equivalence between text when sentences are lexically dissimilar. This thesis tackles the issue with the development of DeepDup, a deep learning methodology for classifying similar questions as duplicate pairs using a Siamese neural network architecture [38]. DeepDup is utilized in Cardea for ensuring the findability [39] of trusted health information is optimized. The possibility of domain adaptation with transfer learning is also explored to make DeepDup accessible across knowledge domains, including technical medical information and layperson-friendly content. DeepDup is detailed in Chapter 8, and is integrated within Cardea to identify duplicate questions, discussions, and answers.

Visitors to health-related websites tend to be confronted with varying degrees of information veracity. Chapter 1 motivated the need to address the misinformation processing gap that exists between medical experts and laypersons who engage in self-education on personal health and wellness topics. Research questions were outlined, hypotheses enumerated to answer the questions, and contributions of this research highlighted. Chapter 2 will cover related research on these topics, while Chapters 3, 4, and 5 provide details on each of the key contributions. Chapters 6, 7, and 8 contain technical details about auxiliary contributions.

Takeaway

- Laypersons are exposed to potentially harmful health misinformation online
- Medical knowledge ought to be used when evaluating veracity of online health information
- Disclosure of online identity can lead to social stigmatization
- Search and recommender results do not consider veracity
- The thesis proposes MedFact, Iron Mask, and Cardea as key solutions

Chapter 2

Background

In this chapter, other research work related to this research is discussed within three broad categorizations based on the themes of this thesis. Firstly, literature relating to trust is covered, including contextualizing to Health Social Media (HSM), providing a psychological viewpoint on health misinformation, and Evidence-Based Practice (EBP). Secondly, research into privacy and stigmatization is summarized, with emphasis on identity as a privacy dimension, online social stigma, and overlap between anonymity and trust. Thirdly, works on search and recommender systems within existing HSM are covered.

2.1 Current State of Research in Trust and Social Media

Research on trust in social media falls into two categories: empirical analysis and algorithmic contributions. Various studies have been conducted to measure the usefulness of generic trust metrics in forums and online communities. These empirical studies can further be grouped into three categories looking at either the network structure, content, or behavioral signals from users. The network structure and its properties help to iteratively determine trust of a given user based on relationships to other trusted users [40, 41]. Content has also been investigated as an indicator for trustworthiness. However, content assessment in current approaches relies on reputation assessment which is limited by user-based ratings. Collaborative content-based methods have also been investigated for determining user reputation [42].

Other metrics such as frequency and sentiment of follow-up posts in relation to an original post have also been studied.

2.1.1 Crowdsourcing

The popular approach for representing trust within social networks is using ratings. There are various implicit and explicit metrics for trust requiring users to provide subjective feedback. Trust metrics provide an abstracted evaluation of the level of veracity or trust associated with content or users. Common trust metrics found in social networks are scaled unary ratings, such as Facebook "Like" 👍, binary ratings such as up or down votes, ranked ratings such as Likert scale rankings, and reputation systems for measuring user trust using achievement levels, badges, and gamification [29]. Some drawbacks of ratings-based systems include whitewashing and cold start [30].

2.1.2 Truth Discovery

Truth discovery enables identification and selection of the actual true value within different data sources with conflicting information. Truth discovery falls into two categories: single-truth and multi-truth. In the area of single-truth discovery, the aforementioned subjective crowdsourcing methods can also be categorized as a naïve method for resolving truth. Other popular iterative rank computation algorithms also fall under single-truth discovery, such as Hyperlink-Induced Topic Search (HITS) [43]. This and other authority hub-based variants are similar to PageRank, an iterative algorithm links between documents and ratings allocated per document [44]. Specific truth discovery algorithms include TruthFinder, AccuSim, AccuCopy, n-Estimates, among others [45, 46]. Ultimately, these algorithms require initializing trust scores for various sources, and subsequently recalculating the scores based on interlinks and interactions. Additionally, network embeddings have been explored to incorporate link analysis with machine learning to better represent nodes in trust networks [47].

2.1.3 Supervised Machine Learning

Supervised machine learning methods have been applied to misinformation detection. Within text processing, the general strategy is to train machine learning models on untrue and true sources, and the key assumption is that there are implicit linguistic features within the text that can differentiate between true information and misinformation. Commonly used features have included syntax, lexical features, psycholinguistic features, semantics, and subjectivity, including sentiment, emotion, and polarity [48, 49]. Various approaches have been explored in related literature on labelling the training datasets, including binary classes comprising true or false, as well as multi-classification with variants of true, partly true or false, false, and unknown.

2.1.4 Health Misinformation

Research on pragmatic contributions to trust in health information were fewer until the COVID-19 pandemic. The seminal work by [50] on HealthTrust was one of the earlier health information-focused studies on trust. HealthTrust automatically assesses new health information based on a set of health web sites with known credibility. Comparison is based on link analysis and content-based analysis. In link analysis, the assumption is that trustworthy content will point to trustworthy web sites as an appeal for authority. Consequently, TrustRank is used to infer a ranking for new content based on inbound and outbound link analysis. In content-based analysis, topic discovery via the TAGME algorithm [51] is used to classify new content as suspicious or trustworthy based on topic similarity with known content via affinity propagation clustering. Secondly, to improve content matching, Hidden Markov Models are applied to an annotated training set in order to model trustworthy and suspicious sentences. A HealthTrust score is assigned for each web site, which is then iteratively exploited.

Recently, there have been many works published in preprint focusing on detection of health misinformation related to COVID-19. The majority of these methodologies can be grouped as either semi-supervised or supervised machine learning.

These methods require annotated training data to identify misinformation [52, 53]. To support this methodology, various datasets have been annotated independently as well as from fact-checking websites and fact-checked articles covering a broad range of political and medical topics [52, 54, 55].

Veracity of specific health topics such as cancer treatments has also investigated using machine learning techniques such as the study by [56]. Using a bag of words representation as the feature set, web pages with medical advice were labeled as positive or negative based on whether they contained questionable content, and the trained model used to assign new labels to new web pages. This approach relied on keyword co-occurrences and correlations instead of cross-referencing trusted medical knowledge.

2.1.5 Evidence-Based Practice (EBP)

Given uncertain health information, medical experts are able to determine trustworthiness by adhering to EBP, which emphasizes scientific evidence and systematic processes of review. Within the medical field, Evidence-Based Medicine (EBM) focuses on using current best evidence to arrive at decisions about the care of individual patients [57, 58]. Three key pillars of EBM are clinician experience, patient values, and scientific information. The EBM five-step model encourages medical professionals to ask, acquire, appraise, apply, and analyze [59]. Medical experts are able to determine trustworthiness of health information through EBM by systematically organizing pertinent information into a hierarchy of evidence based on methodological quality. From the most reliable Level I up to Level VII, evidence can be grouped into systematic reviews of randomized controlled trials, well-designed randomized controlled trials, quasi-experimental studies, cohort studies, meta-synthesis, single qualitative studies, and reports of expert committees [60]. Additionally, medical experts can clarify clinical questions by investigating research evidence using systematic methodologies such as Population, Intervention, Control and Outcomes (PICO) [61, 62]. However, these methodologies remain largely a manual process requiring medical experts and search tools.

2.2 Psychological Viewpoint on Health Misinformation

Various factors contribute to the present proliferation of unsafe health information online, which need to be taken into consideration when developing any approach for promoting credible information and preventing unsafe viral health campaigns. Apart from the development of technical solutions and useful trust metrics, the psychological biases of users consuming health information also need to be understood, including users' preference for layperson health stories, perceived resistance to medical facts, and the perception of medical expertise among laypersons.

2.2.1 Neural Coupling

The information seeking behavior of laypersons and patients is based on storytelling rather than systematic medical and scientific methodologies. Patients tend to use personal experience and stories as a source of authoritativeness rather than scientific methodology [15]. This behavior is related to neural coupling, an effect observed in neuroscience between storytellers and listeners. Experiments have shown that when a storyteller is communicating with listeners, the listener's brain patterns will eventually mirror the storyteller's patterns. Neural coupling is an evolutionary trait to help human species to learn from each other through emotions [63]. The popularity of story-based narratives on health social media could also be attributed to these primal triggers. In the case of the "anti-vaxxers", even inaccurate stories were effective in convincing people not to vaccinate because of the emotional format of the message [7].

2.2.2 Backfire Effect

Studies related to anti-vaxxers attempted to investigate the efficacy of counter-messages promoting vaccinations for Measles Mumps Rubella (MMR) [64]. In the study, anti-vaxxer parents of children needing MMR vaccinations were presented with various interventions. Firstly, they were presented with information clearly depicting a lack of conclusive evidence associating autism with any vaccinations.

Secondly, they were shown textual information on risks of not getting vaccinated. Thirdly, images of other children who had contracted MMR-related diseases were shown. And finally, parents were told a dramatic story of a child who did not get vaccinated for measles and almost died. Surprisingly, none of the interventions were statistically significant in convincing the parents. In some cases, the parents' belief that vaccinations are harmful was even strengthened, for instance when being shown the imagery of sick children who did not get vaccinated. These counter-intuitive results could be explained by the backfire effect, wherein the presentation of contradictory evidence is not only ineffective in convincing people, but leads people to strengthen their belief [65]. Related to the backfire effect is confirmation bias, where users online tend to seek out and gravitate towards information supporting their beliefs and ignore opposing viewpoints [66].

2.2.3 Dunning-Kruger Effect

The Dunning-Kruger effect is attributed to unskilled persons not realizing their incompetence, resulting in the self-illusion of superior competence [67], a trait that can be readily observed in the online health information communities, where laypersons eagerly and confidently provide medical advice to other laypersons. This phenomenon is clearer in the study of *agnotology*, where inaccurate or misleading scientific information is willfully promoted to induce ignorance about facts [68]. Essentially, online health information is saturated with information that is not credible, yet is being propagated due to users' willingness to look for quick solutions to complex health problems, such as autism [69].

2.3 Identity Privacy, Stigmatization, and Anonymity

A considerable amount of work has been done on the rights of users to privacy and enabling control over content visibility. However, another aspect of privacy is giving users control over how their identity is associated with the content they create. Preventing social stigma is one benefit to users hiding their real identity online, while challenges include reputation preservation and preventing de-anonymization.

2.3.1 Social Media and Social Stigma

The veiled viral marketing approach was suggested by Hansen and Johnson for sending anonymized messages to friends within Facebook [33]. In the study, research on an awareness campaign for Human Papilloma Virus (HPV) showed that people who knew HPV is a sexually transmitted were more likely to feel shame and stigma, and less likely to share or post information about it on their Facebook profiles. Moreover, people were willing to share links to websites about social causes like breast cancer awareness, but were unlikely to do likewise for links to syphilis or gonorrhea websites. The proposed veiled viral marketing approach allowed sending of anonymous "veiled" messages to friends, which essentially substituted the user's identity with the "friend" pseudonym. Users would know that the message came from one of their friends, but would not know which friend actually sent the message. However, this study did not take into account any risk of de-anonymization from the content being a quasi-identifier. In addition, no exploration was made on any relationships between veracity of information and anonymity, although it was implied that users trusted their friends' shared content more than that of strangers.

2.3.2 Anonymity and Trust

The relationship between anonymity and trust has also been explored in peer-to-peer networks for providing ratings and feedback anonymously [70], which also has applications in e-governance and online voting [71]. The proposed approaches focus on the anonymized reporting of aggregated results. This relationship is also important to dematerialized money and cryptocurrencies, where the emphasis is on completing trustworthy transactions while maintaining anonymity of the agents involved [72]. In contrast with these domains, there has not been much direct work done on enhancing the relationship between trust and anonymity in social media. There are various social media websites that have either internal or external anonymity controls. The former, such as Quora, allow users to anonymously post content without revealing their actual registered account's user name. The latter includes websites that let anyone post content without having to register an account.

Pseudo-accounts are a third option in which users register counterfeit accounts to hide their real identities, and subsequently do not require additional anonymity controls or options [73].

2.3.3 De-anonymization Risks

The relationship between historical posted content and user identity was partially investigated by Lebedev and Sukhoparov as a side effect of their study [74]. They looked into the situation where the same person had several different accounts on the same web portal, potentially for manipulation of feedback, ratings and Sybil attacks on the web portal [75]. The study proposed a solution to short messages text authorship determination using a naïve Bayes classifier. The classifier was trained using short messages from known users. This classifier was then used to determine if a new post belonged to an existing user. One drawback of the study was the low accuracy of 50%, which could be attributed to the selection of features, size of the training data, or the classifier used. Another similar study was conducted by Keretna et al. on whiteprint identification in Twitter to recognize multiple accounts being created by the same user [35]. Narayanan et al. [76] investigated different de-anonymization attacks on social networks such as Twitter. Their study looked at possible re-identification risks involved with user information available on more than one social network, i.e. Twitter and Flickr, and how intersection of common information could lead to re-identification. A similar study by Beach et al. [77] also looked at anonymity in social networks and the disadvantages of using traditional anonymization methods such as k-anonymizaion on social media websites like Facebook. However, these studies focused on partial anonymization where some properties of the user are hidden, such as name, while others are visible, like gender or location.

2.4 Incorporating Trust in Search and Recommender Systems

2.4.1 Search Engines

Search engines and search boxes are the most popular method for seeking information. It has also been shown that people tend to be strongly influenced by search results bias, implying the first few results from a search engine can affect one's perception of truth and falsehoods around a topic [78]. In information architecture, matching user intent expressed as a search query to search results can be defined in terms of findability of content. Findability expresses the ease with which a user can locate the content they are seeking within a website [79]. For search engines, a popular approach to ranking results and enhancing findability is the PageRank algorithm, developed by Google [80]. PageRank provides a filtering and sorting mechanism for distinction between content that matches the search query. PageRank is computed iteratively, and for a given website, its PageRank score is computed by accumulating the number of inbound links and outbound links, with each link given a weight based on the PageRanks of other links pointing to the website, and including PageRanks of links being pointed to by the website [80]. Google also uses other algorithms for determining spam links, i.e. nepotistic links. The Google Penguin algorithm penalizes websites that do not follow their guidelines for content and would want to inflate their ratings by keyword stuffing. On the other hand, the Google Panda algorithm promotes websites with high quality original content. The newer Google Hummingbird ranking methodology aims to understand the user's intent within the search query and find content that answers the user's intent. Consequently, this type of content can be said to have a higher findability than content that does not match user intent. The Google Hummingbird algorithm describes user intent in terms of the semantic meaning of the search query. Instead of matching keywords between the query and results, the meaning of the query is related to information using a semantic knowledge base called the Knowledge Graph [80]. For instance, the search results of "time" include the current local time, in addition to web pages containing the search term because either context might be of interest.

This is achieved by considering all the possible contexts and senses of the term and providing results with diverse sense and context matches, as well as considering correlated terms from the user's historical search queries and Click-Through Rate (CTR). Google has also recently partnered with Mayo Clinic to rank trusted results higher using the notion of a trusted database [81].

2.4.2 Recommender Results

Recommender systems are another popular set of software tools and algorithms that can give useful suggestions to users and enhance findability of content [82]. The suggestions are given within the context of the user's domain of interest, such as what items to buy or shop for, which new people to connect with, or new movies to watch. Recommendations enable sifting through large amounts of information previously too massive or complicated to practically navigate. There are various methods of generating recommendations: content-based, collaborative, community-based, demographic, knowledge-based [82]. Content-based recommendations use keywords to suggest new items that are historically similar to previous items a user may have liked or bought in the past. In collaborative recommender systems, suggestions for one user are based on what other users with similar profiles have liked. Community-based recommendations leverage the preferences of users' friends, a popular notion in social media. The demographic approach uses age, gender, ethnicity, and other demographic information about a user, as well as the items they like or buy. These properties are matched to demographic stereotypes, such as English-speaking customers being directed to US-based websites. The knowledge-based approach is to recommend items using specific domain knowledge about how certain item features meet users needs and preferences. Various metrics used for ranking recommendations and measuring the performance of recommender systems. These include prediction-based metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), as well as information retrieval-based metrics such as Precision@n, Mean Average Precision (MAP), Discounted Cumulative Gain (DCG), and NDCG [83].

While Precision@n measures recommender performance from a single user perspective, MAP does the same over the whole set of users [83]. There are also metrics for measuring diversity, novelty and coverage of recommender suggestions. Trust has also been explored as a potential metric for ranking suggestions [84] but a drawback is the lack of publicly available datasets for testing trust-based approaches [85].

The notion of trust dissemination is important for search engine and recommender system results. If the results of a search query or a recommendation show trustworthy content, then the user gets better value out of the interaction. There are add-ons for search engines such as Web Of Trust (WOT) that provide additional information via visual indicators regarding the trustworthiness of links, by using community-based rankings of these links [86]. However, these additional indicators do not factor into the rankings, leading to possibly untrusted content being shown at a high position because of popularity.

Takeaway

- Crowdsourcing and supervised machine learning are the currently popular methods for trust estimation in social media
- Psychology literature suggests that users tend to gravitate towards health misinformation due to emotional language and knowledge gaps
- Evidence-Based Medicine methods require curated access to knowledge bases by experts though the methodologies leverage systematic approaches
- Research on anonymity has room for improvement in terms of de-anonymization risks and effects of pseudonyms on reputation-based trust
- Results from search engines and recommender systems are ranked based on semantic and lexical matching but limited have attention to content veracity

Chapter 3

MedFact: Evidence-Based Objective Metrics to Determine Veracity of Health Information

This chapter elaborates on the MedFact algorithm, and provides details on Hypothesis 1 of this thesis by investigating evidence-based objective trust metrics for identifying health misinformation. This chapter are largely based on the peer-reviewed paper "MedFact: Towards Improving Veracity of Medical Information in Social Media using Applied Machine Learning" published and presented at the Canadian Conference on Artificial Intelligence in Toronto, Canada [87].

3.1 Motivating Objective Trust Metrics

Not too long ago, viral social media posts were used to falsely associate vaccinations with autism [7]. Articles supposedly written by medical professionals that linked autism and vaccinations were heavily shared on Facebook and other social networks, leading to a perception among many users that vaccinations are harmful. Needless to say, not getting vaccinated would give rise to more disease outbreaks and negatively affect public health overall. This is even more evident with the current COVID-19 pandemic, which itself has turned into an infodemic as social media discourse has been flooded with misinformation. In these situations, consensus-based methods relying on the "wisdom of the crowds", likes, or votes can be detrimental, and credible information from medical experts is needed.

3.2 Evidence-Based Practice

Medical experts are able to determine trustworthiness of health information through Evidence-Based Medicine (EBM), a systematic approach for appraising health information on the basis of the best current evidence, clinical expertise, and patient needs in order to facilitate decisions about patient care [58]. Medical knowledge is health information verified through the scientific process and evidence. EBM arranges pertinent information into a hierarchy of evidence based on methodological quality. From the most reliable Level I up to Level VII, evidence can be grouped into systematic reviews of randomized controlled trials, well-designed randomized controlled trials, quasi-experimental studies, cohort studies, meta-synthesis, single qualitative studies, and reports of expert committees [60].

3.3 Combining Evidence with Trust using MedFact

Computing automation can be applied in conjunction with EBM to determine the veracity of online health information. To this end, the MedFact algorithm was developed as a contribution of this thesis, based on EBM and trusted medical information sources, in order to empower and educate online users to determine health information veracity. MedFact addresses the challenges of layperson versus technical vocabularies, and issues of effectively presenting veracity of information in simplified and non-technical formats.

The task of determining the veracity of medical content is defined as a five-step process. Given any textual document, such as a social media post, the first step is to extract health-related phrases $\{x_1, x_2, ..., x_m \in X\}$. The veracity of these phrases is unknown. The second step uses automated information retrieval and processing to search trusted scientific and medical knowledge bases for each of the *unlabelled phrases* $x_i \in X$, representing phrases that need their veracity calculated. In this step, each trusted source would yield zero or more relevant articles, providing a collection of trusted articles which are ranked and filtered by relevance. The trusted articles have various related credible phrases that are identified in the third step to generate a collection of trusted phrases $\{t_1, t_2, ..., t_n \in T\}$.

The semantic similarity between a given trusted phrase, $t_j \in T$, and x_i is used for inferring an *agreement score*, $\Upsilon(x_i, t_j)$ between the two phrases. In the fourth step, an aggregated agreement score for a given unlabelled phrase is computed by comparing it with all trusted phrases and averaging the agreement score as formulated in Equation 3.1. In the fifth step, an overall *veracity score* ϑ is computed for the social media post from the aggregated agreement scores of all unlabelled phrases as shown in Equation 3.2.

$$\Upsilon(x_i) = \left(\sum_{p=1}^{n} \Upsilon(x_i, t_p) \right) / n \tag{3.1}$$

$$\vartheta = \left(\sum_{q=1}^{m} \Upsilon(x_q) \right) / m \tag{3.2}$$

This methodology overlaps with the EBM five-step model: ask, acquire, appraise, apply, and analyze [59]. Asking a question entails seeking to investigate the veracity of a social media post, while acquiring involves computationally gathering the available evidence related to the question. The overall pipeline for MedFact is depicted in Figure 3.1. Additionally, the MedFact implementation is open source [1].

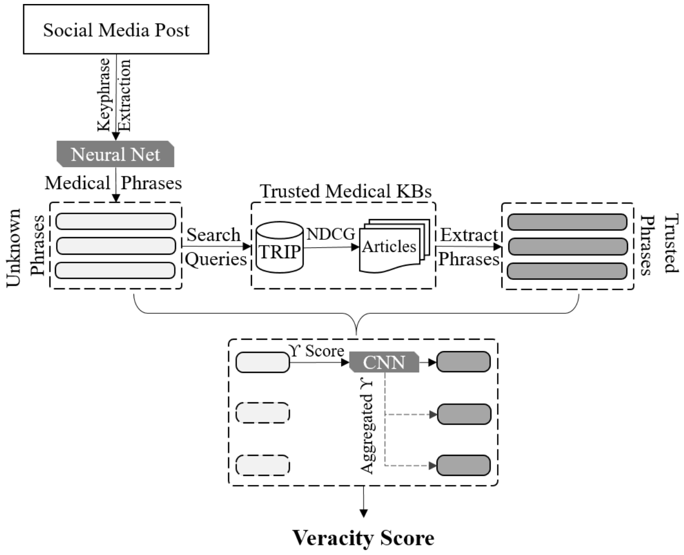


Figure 3.1: Overview of MedFact Algorithm

[1] MedFact GitHub repository https://github.com/hwsamuel/MedFact

Step 1 To extract relevant health phrases from a given social media posting, candidate phrases are extracted using key phrase extraction. The next stage identifies health-related phrases from among the candidate phrases. Extraction of key phrases is done using the TextRank algorithm[2]. The NLP pipeline for this involves tokenization of phrases, stop words removal with Glasgow [88], stemming with Porter [89], and choosing relevant keywords using word frequencies [90]. In the next stage, a supervised learning approach is used to build a binary classifier that for classifying a given phrase as medical or non-medical. The classifier is implemented as an artificial neural network, and medical phrases are input as word embeddings, with output of 0 if the phrase is non-medical or 1 if medical. In order to train the classifier, two categories of datasets were used. The first category corresponds to the "medical" label, including medical phrases from the Systematized NOmenclature of MEDicine (SNOMED) database and layperson health terms from the Consumer Health Vocabulary (CHV) dataset. SNOMED [3] is a digital collection of medical terms provided by the U.S. National Library of Medicine [91]. The CHV dataset[4] provides mappings of common layperson medical terms to technical terms in Unifed Medical Language System (UMLS) [92]. The second category corresponds to the "non-medical" label and contains known non-medical corpora from the Simple English Wikipedia (SEW) dataset[5] [93]. From these datasets, a training sample is created by arbitrary selection of approximately 80% of the phrases from each dataset. A test sample of 20% is kept for internal scoring purposes. The phrases (hyphenated) are converted to word embeddings using the Word2Vec deep neural network model trained on medical corpora with skip-grams [94]. The phrases and their corresponding labels from the training sample are used to train the neural network. The arbitrary selection process is repeated a number of times to achieve non-exhaustive cross-validation and the best trained model is used.

[2] The GenSim Python API includes the TextRank algorithm implementation https://radimrehurek.com/gensim/summarization/keywords.html

[3] SNOMED dataset available from the U.S. National Library of Medicine (NLM) https://nlm.nih.gov/healthit/snomedct

[4] CHV dataset available from the Consumer Health Vocabulary Initiative http://consumerhealthvocab.org

[5] SEW historical dataset available via PIKES home page http://pikes.fbk.eu/eval-sew.html

Step 2 Credible medical knowledge can be searched on the Turning Research Into Practice (TRIP) database[6]. TRIP focuses on evidence-based medical literature from various trusted sources including the National Library of Medicine (NLM) MEDLINE and PubMed articles, the Cochrane database of systematic reviews, the DARE, among others. Moreover, the TRIP database also searches within patient-friendly resources such as Cochrane Clinical Answers and WebMD's Medscape [95]. Results are categorized into the levels of evidence and can be sorted by quality, relevance, or date. A publication score is used to assess and rank quality of the results by incorporating the levels of evidence, Level I receiving the highest weight and subsequent levels receiving progressively lower weights. TRIP's quality metric is used to sort articles and incorporate strength of the evidence. Additional ranking of the articles is performed in order to evaluate the usefulness of the top-*n* articles based on their position in the results using NDCG [96].

Step 3 In order to compare unlabelled phrases with trusted phrases, phrases are extracted from the ranked medical articles via phrase chunking. Firstly, each article's text is split using sentence and word tokenization. Next, Part-Of-Speech (POS) tagging is performed on the tokens, followed by phrase chunking[7] which segments the sentences into noun phrases. After that, each chunked phrase extracted from the medical articles is compared with the set of unlabelled phrases, and trusted phrases that do not correlate with unlabelled phrases are discarded because they will not be useful in the next steps.

Step 4 Given a phrase whose veracity needs to be ascertained, a corresponding set of phrases from a trusted source can be used as evidence for supporting or rejecting the unlabelled phrase as credible. This problem is modelled as that of predicting a class label for a given pair of phrases, where two labels are available: *Yes* or *No*.

[6] The TRIP database is accessible programmatically via as a web service https://www.tripdatabase.com/addtrip

[7] POS tagging is done using the Penn Treebank tags set, all steps in this particular pipeline are programmed with the NLTK Python library http://nltk.org

The former label implies that the two phrases have the same meaning, while the latter label means the phrases could contain incompatible propositions such as contradictions. Given two phrases, their agreement is determined using deep learning, incorporating semantic similarity and sentiment analysis of the two phrases. The feature set consists of the word embeddings of the two phrases, and sentiment information[8] for each phrase, specifically polarity and subjectivity [97]. Polarity for a phrase is in the range [-1.0, 1.0] where -1.0 implies very negative sentiment and 1.0 means very positive sentiment, while subjectivity values are in the range [0.0, 1.0] where 0.0 means very objective and 1.0 implies very subjective. Also, the negation modifier is used from dependency parsing [98] of the related sentence containing the target phrases as an additional binary feature, where 1 implies the presence of the negation modifier and 0 means an absence[9]. For the neural network implementation, a shallow Convolutional Neural Network (CNN) architecture is used[10], which is more suitable for learning from smaller-sized labeled training datasets [99]. The training dataset was built from Medical Sciences Stack Exchange (MSSE)[11], an online question-answering community where users can post health-related questions[12]. Within the MSSE community, moderators can manually flag semantically equivalent posts as *Duplicate*. The training dataset consists of pairs of phrases extracted from the duplicate posts' title and body using phrase chunking. The related medical phrase pairs extracted from these question pairs are assigned the *Yes* label. For question pairs that are not duplicates, the *No* label is assigned to the related phrase pairs in the training dataset. Subsequently, the training dataset was curated for accuracy of the initial labeling in order to verify whether the phrase pairs are in agreement or not. Ultimately, given two phrases, the *agreement score* is defined using the classifier's output label.

[8] Sentiment analysis is performed using the TextBlob Python library http://textblob.readthedocs.io

[9] The spaCy Python library is used for generating dependency trees https://spacy.io

[10] A shallow CNN was implemented with the ConText tool https://github.com/riejohnson/ConText

[11] Medical Science Stack Exchange's beta web site https://medicalsciences.stackexchange.com

[12] Dataset curated from the Internet Archive's Stack Exchange Data Dump https://archive.org/details/stackexchange

Step 5 The veracity score enables aggregation of the agreement scores of many pairs of unlabelled phrases and their respective trusted phrases, and provides a single metric for measuring the veracity of a given social media posting or document. This approach allows for a granular definition of veracity starting from phrase-level agreement to document-level aggregated agreement. Depending on the number of unknown and trusted phrase pairs, the overall veracity score is computed as an average, hence it is within the range [0.0, 1.0], and can be expressed as a simplistic percentage value.

3.4 Feedback on Usage of Veracity Score

To test the usefulness of MedFact, a short survey was designed that was administered to nineteen layperson users. The survey was disseminated through email and the Undergraduate Research Initiative's (URI) Undergraduate Research Portal[13] forums. Hence, the enrolled participants were likely undergraduate students from the University of Alberta. The survey contained polarizing postings on the link between vaccination and autism[14], apricot pits as a cure for cancer[15], and usefulness of flossing for dental care[16]. These three topics were selected because they received attention in news media. Firstly, a posting supporting vaccination and autism was displayed, followed by a post debunking the notion. Similarly, users were then shown a posting supporting apricot pits as a cure for cancer, and then shown an opposing post. Lastly, posts supporting and opposing the need to floss were shown. For each posting shown, the veracity score expressed as a percentage (rounded-off) was visible. The top 3 trusted articles related to the posting were also displayed. After displaying each posting, users were asked three questions about the veracity score and the selected article. Each question required a Yes or No response.

[13] URI Undergraduate Research Portal https://eclass.srv.ualberta.ca/course/view.php?id=759

[14] The Discredited Doctor Hailed by the Anti-Vaccine Movement https://www.nature.com/articles/d41586-020-02989-9

[15] A Man Who Took Apricot Kernels to Beat Cancer Got Cyanide Poisoning https://www.theverge.com/2017/9/11/16288104/apricot-kernels-cancer-cyanide-poisoning-case-study

[16] Feeling Guilty About Not Flossing? Maybe There's No Need https://www.nytimes.com/2016/08/03/health/flossing-teeth-cavities.html

Firstly, they were asked "*Is the veracity score useful in this context?*". Next, they were asked "*Is the veracity score accurate for this post?*". Lastly, they were asked "*Are the links to the medical articles useful?*". At the end of the survey, users were optionally asked to give general feedback in free text form.

A summary of users' responses recorded for the questions is shown in Figure 3.2. Participants provided generally positive feedback to all three questions. However, regarding the accuracy of the veracity score, users gave less than expected positive feedback. Further analysis revealed this was due to differing opinions on apricot pits as a cancer cure, accounting for 70% of the lower positive feedback. The free text feedback was reviewed and investigated to further understand user perspectives. It was discovered that the majority of survey participants viewed apricot pit treatments as a homeopathic remedy that should not be covered by scientific literature. Overall, the survey recorded positive feedback from 68% of the responses regarding the veracity score accuracy.

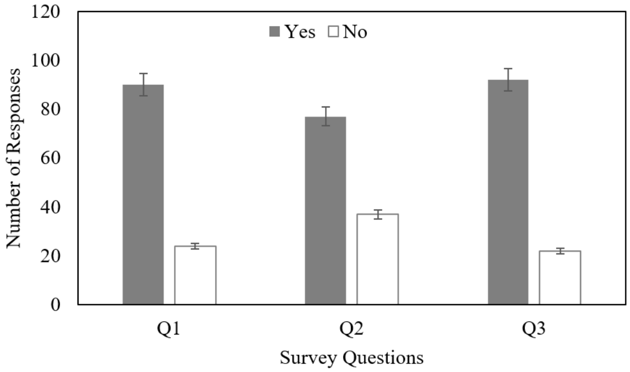


Figure 3.2: Summary of Veracity Score Survey Responses

3.5 Survey of Medical Professionals on Controversial Topics in Pediatrics

This survey was conducted as part of the ethics approval from the University of Alberta Research Ethics Board, under the project title "User Perspectives on Trust in Health Social Media" and project identifier Pro00079019. The survey provided a double-blind comparison on the veracity of medical claims between MedFact's results and medical professionals' responses. Hence, participants were not shown the results of MedFact, but rather were asked to independently evaluate statements related to pediatrics. Also, MedFact's computations for the same statements on pediatrics were computed prior to administering the survey.

A questionnaire was disseminated privately among known medical professionals in pediatrics to avoid layperson opinions. Six statements related to pediatrics were shown to the participant in order to rate each statement based on their professional evaluation of the statement's veracity using a psychometric scale: *Strongly Disagree*, *Disagree*, *Neutral*, *Agree*, *Strongly Agree*, and *Do Not Know*.

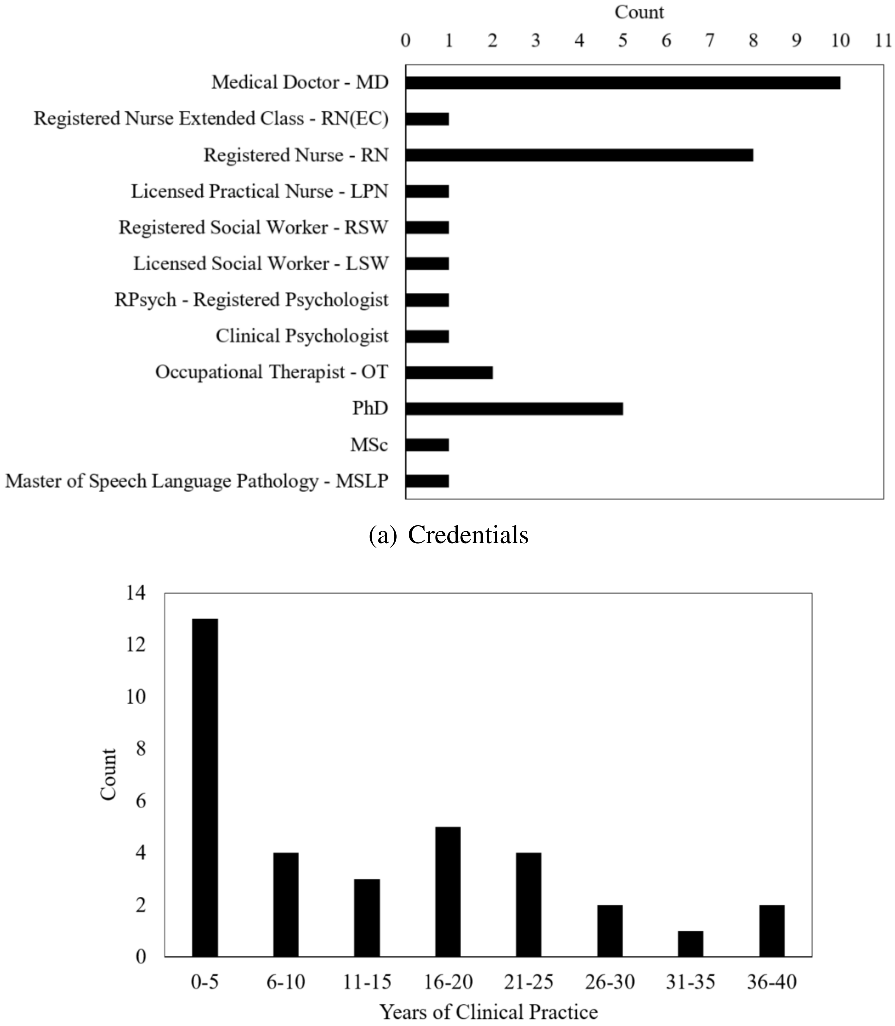
Each of the statements, selected from Facebook, Wikipedia, blogs, and news articles, belonged to one of the following topics: general pediatrics, autism, behavior, Applied Behavior Analysis (ABA), Attention Deficit Hyperactivity Disorder (ADHD), or Positive Parenting Program (Triple P). For each participant, the six statements were selected from three rubrics, A, B, and C, and the statements within the selected rubric were then randomly re-ordered. Hence, each subsequent participant viewed a different set of six statements from each rubric, with the rubric selection being rotated in sequence. A total of 10 respondents viewed rubric A, 11 respondents were shown rubric B, and 13 viewed rubric C. The list of statements and rubrics used to administer the survey are detailed in Table 3.5.

The six statements in each rubric were from varied topics in pediatrics, and a total of 34 participants responded. Aggregated self-reported credentials, years of clinical practice, and areas of practice of the participants are shown in Figure 3.3. The highest years of clinical practice were 36, with mean of 13.44 years and median of 10.50 years.

The statements were evaluated by MedFact and a veracity score computed. Based on the score and confidence, a *MedFact Label* was assigned to each statement. For comparison, the responses of the medical professionals were categorized as either in agreement, disagreement, or uncertain about each of the statements. Based on the majority consensus, a *Medic Label* was assigned to each statement. Ultimately, the two labels were compared to evaluate MedFact's corroboration with medical professionals, with details provided in Table 3.2.

Rubric	ID	Statement	Topic
A	A1	A lot of government-published studies show vaccines cause autism.	Autism
	A2	When dealing with a misbehaving child, intentionally ignore a problem behavior instead of reacting or giving negative attention to the child.	Behavior
	A3	ABA therapy accounts for 45% of pediatric therapies that develop long-lasting and observable results.	ABA
	A4	Parents of children with disabilities should not be allowed to use growth attenuation therapy.	General
	A5	When ADHD is undiagnosed and untreated, ADHD contributes to problems succeeding in school and graduating.	ADHD
	A6	A review of 33 studies published in BMC Medicine found no convincing evidence that Triple P interventions work across the whole population, or that any benefits are long-term.	Triple P
B	B1	Parents can change from using ineffective and coercive discipline such as physical punishment, shouting, and threatening to using effective strategies in specific situations.	Behavior
	B2	Applied Behavioral Analysis (ABA) is based on a cruel premise - of trying to make people with autism 'normal'.	ABA
	B3	Homeopathic treatments for hyperactive children have been generally successful.	ADHD
	B4	The age threshold for using medical intervention for children with gender dysphoria should be lowered.	General
	B5	Environmental factors that could trigger predisposed genes to mutate and cause autism are vast and could include certain drugs, extensive television viewing, or infections during pregnancy.	Autism
	B6	Triple P trials are particularly susceptible to risks of bias and investigator manipulation of apparent results.	Triple P
C	C1	Most scientists agree that genes are one of the risk factors that can make a child more likely to develop autism.	Autism
	C2	The most serious problem with the Triple P literature is the over-reliance on positive but substantially underpowered trials.	Triple P
	C3	Selective Serotonin Reuptake Inhibitors (SSRIs) are an effective treatment for pediatric OCD.	Behavior
	C4	A child with ADHD is accident-prone, likely to make careless mistakes, and take unnecessary risks.	ADHD
	C5	Neurodiversity should be accepted as naturally different rather than abnormal and needing to be fixed.	General
	C6	ABA is just animal training adapted for use with people.	ABA

Table 3.1: Survey Statements by Rubric and Topic



(a) Credentials

(b) Years of Clinical Practice

Figure 3.3: Demographic Information of Participant Medical Professionals

When taking into consideration all the statements, MedFact's automated assessment matched professional opinions of medical personnel by 50%. Even among the professionals, there was no consensus for 50% of the statements, and the statements were marked as uncertain, demonstrating the challenge with determining veracity. Excluding statements where professionals were uncertain, MedFact corroborated even closer by 67%.

ID	**MedFact Veracity Score**	**MedFact Label**	**Medics Label**	**Medics Disagree**	**Medics Uncertain**	**Medics Agree**	**Consensus among Medics**
A1	0.11	**Untrusted**	**Untrusted**	0.90	0.10	0.00	Disagree
A2	0.67	Unknown	Trusted	0.30	0.20	0.50	Agree
A3	0.73	Trusted	Unknown	0.20	0.50	0.30	No Consensus
A4	0.19	Untrusted	Unknown	0.40	0.60	0.00	No Consensus
A5	0.78	**Trusted**	**Trusted**	0.00	0.10	0.90	Agree
A6	0.69	**Unknown**	**Unknown**	0.20	0.50	0.30	No Consensus
B1	0.77	**Trusted**	**Trusted**	0.27	0.18	0.55	Agree
B2	0.66	Unknown	Untrusted	0.55	0.45	0.00	Disagree
B3	0.13	**Untrusted**	**Untrusted**	0.55	0.45	0.00	Disagree
B4	0.12	Untrusted	Unknown	0.27	0.55	0.18	No Consensus
B5	0.80	Trusted	Unknown	0.36	0.45	0.18	No Consensus
B6	0.61	**Unknown**	**Unknown**	0.27	0.64	0.09	No Consensus
C1	0.69	**Trusted**	**Trusted**	0.00	0.08	0.92	Agree
C2	0.66	Unknown	Trusted	0.00	0.46	0.54	Agree
C3	0.04	Untrusted	Unknown	0.31	0.54	0.15	No Consensus
C4	0.82	**Trusted**	**Trusted**	0.23	0.15	0.62	Agree
C5	0.47	**Unknown**	**Unknown**	0.08	0.54	0.38	No Consensus
C6	0.11	Untrusted	Unknown	0.31	0.54	0.15	No Consensus

Table 3.2: Comparison of Responses by Medical Professionals versus MedFact

3.6 Veracity Score on Unproven Cancer Treatments

Thirty articles on cancer were randomly selected from QuackWatch[17], a web site indexing unproven treatments [56]. The selection of random articles was done iteratively, with each iteration retrieving a new set of articles. Hence, each iteration is random sampling without replacement. These selected articles were input to MedFact to compute a veracity score in order to determine whether the score would align with experts' opinions. The veracity score for the selected articles is summarized in Figure 3.4, with low scores below 0.50 for the articles.

[17] QuackWatch web site http://quackwatch.org

The low score is in consensus with the opinions of the experts who identified the unproven claims, showing the usefulness of MedFact in identifying health misinformation. It also aligns with other studies on verifying and evaluating unproven cancer treatments [56]. Variation among samples was significant and all the sample cohorts showed consistent results.

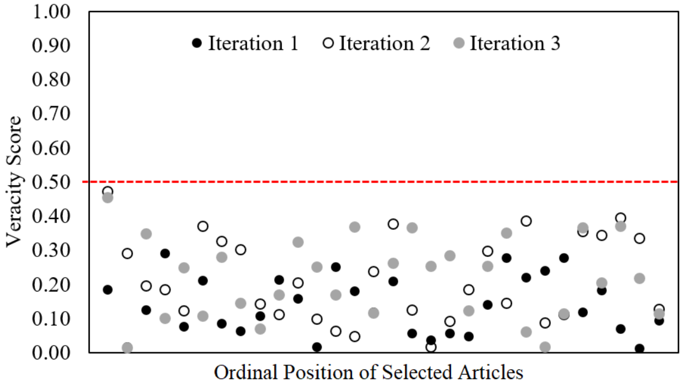


Figure 3.4: Veracity Score on Random Articles from QuackWatch (all are identified with low veracity score as expected)

3.7 Online Medic Discussions Evaluated with Veracity Scoring

To further evaluate the performance and representative accuracy of MedFact's veracity score, a total of thirty answers posted on the DocCheck forums were randomly selected for general health topics[18]. DocCheck allows verified medical professionals to ask questions and post answers. The results showed an average veracity score of 78%. A comparison of the veracity scores for medic posts versus QuackWatch scores averaged across the random selection iterations are presented in Figure 3.5. The results were as expected, with the DocCheck veracity scores being significantly higher than QuackWatch. A clustering effect is observed between credible and untrustworthy posts, showing that MedFact was able to make a clear distinction between trusted and untrusted claims.

[18] DocCheck web site http://doccheck.com

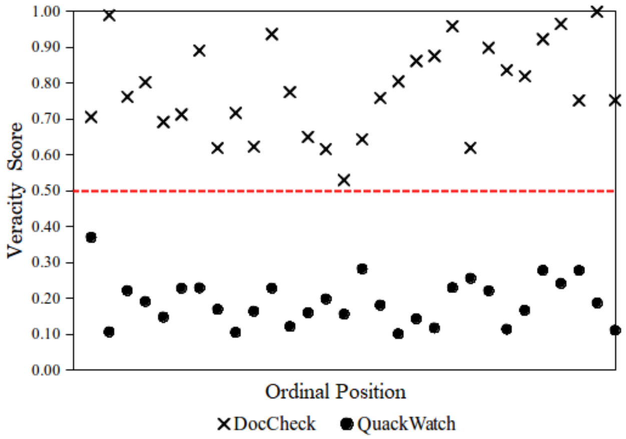


Figure 3.5: Veracity Score Comparisons between DocCheck and QuackWatch (clear separation between true and false health information)

3.8 Veracity of COVID-19 Medical Claims and Bootstrapping MedFact

Ongoing health social media conversations related to the COVID-19 pandemic have led to many controversies and there are a number of topics fueling COVID-related misinformation, from conspiracy theories, misreporting of morbidity and mortality, disease spread mechanisms, recovery experiences, and political controversies [100]. From the medical perspective, controversies surrounding symptoms, treatments, and drugs have been at the forefront. The core approach of MedFact is to use IR and NLP for querying known medical knowledge about topics at hand. When applying MedFact to online social media chatter about COVID-19, a couple of challenges and limitations arose. Firstly, for the majority of COVID-19 controversial topics, there is a lack of strong consensus within medical literature on the related topics. Secondly, medical findings related to COVID-19 are being published and knowledge is being synthesized on an ongoing basis. Consequently, MedFact encountered a cold start scenario leading to two side-effects: low veracity scoring due to a low number of reputed publications on COVID-19, and low confidence score due to current publications being on a lower scale of the evidence hierarchy. Preliminary explorations on overcoming this issue include bootstrapping MedFact with curated reliable sources covering COVID-19.

This chapter provided details on the implementation and usability testing details of MedFact and the veracity score as an objective trust metric. It was demonstrated that MedFact can better identify health misinformation using veracity score where consensus-based subjective rating metrics might be detrimental. The usefulness of MedFact was tested with two surveys of laypersons as well as medical professionals, showing optimistic results. Additionally, the veracity score's usefulness was demonstrated by identification of clear boundaries between true and false claims from DocCheck and QuackWatch.

Takeaway

- Subjective trust metrics do not explicitly incorporate medical knowledge
- MedFact enables medical knowledge in trust metrics for health social media
- Taking inspiration from Evidence-Based Medicine, a veracity score is introduced as an objective trust metric based on trusted knowledge bases
- Surveys of both laypersons and medical professionals provided promising feedback for MedFact and veracity scoring
- Veracity score for online health discussions on topics such as cancer treatments, pediatrics, and general health topics showed favourable results

Chapter 4

Iron Mask: Trust-Preserving Anonymity on the Face of Social Stigmatization

This chapter provides details on the Iron Mask algorithm, and elaborates on Hypothesis 2 of this thesis by looking into the use and evaluation of Trust Preserving Pseudonyms (TPPs) to anonymize users while preserving reputation-based trust. The key motivation for this research topic is grounded in preventing social stigmatization. Details provided in this chapter are largely based on the peer-reviewed paper "Iron Mask: Trust-Preserving Anonymity on the Face of Stigmatization in Social Networking Sites" published and presented at the International Conference on Trust, Privacy & Security In Digital Business in Lyon, France [101].

4.1 Motivation for Anonymity and Drawbacks

Typically, the author of a posting on a social media website is identifiable by their registered user name or full name being displayed next to the post. However, situations can arise in which the user does not want to be identified. For instance, if a user were to share a link with their friends about sexual dysfunction or infertility, the user may wish to do so anonymously to avoid any potential stigmatization which may result from the assumption that sharer suffers from the condition [33].

Pseudonyms have proven useful within online forum communities for supporting stigmatized issues, and people tend to discuss and learn more openly about stigmatized topics when the perceived risk of being publicly associated with the issue is taken away [102]. On the other hand, the negative effects of users experiencing social stigma can be severe, with outcomes ranging from poorer mental health to increased risk behaviors [103]. It is known that users have been increasingly using the internet for sharing personal experiences and seeking advice about various personal issues, which increases the likelihood of online stigma [104].

Despite the potential severity of online social stigma, options and controls to anonymously post content are not well-supported in most social media websites. Users on these websites may hide their real identity by creating a new account with a non-real name or pseudonym, thereby duplicating the website's user base. This is not ideal and unnecessarily complicates the process of information sharing. From the list of popular social media websites such as Facebook, Twitter, LinkedIn, YouTube, Stack Exchange and Quora, only the latter allows asking questions anonymously without needing to create a new account. Trust Preserving Pseudonyms (TPPs) provide a more convenient and safer alternative.

There are also potential drawbacks with the approach to replace the user's real identity with the generic pseudonym "anonymous". Firstly, despite their name being hidden, users still may be inadvertently revealing their identity because of the similarities between the content they have posted in the past. Phrases, wordings, topics and other nuances about the writing style in the user's past postings may constitute a quasi-identifier that can be associated with a specific user. Quasi-identifiers are not unique by themselves but can be correlated with an user's identity due to frequency of occurrence or other patterns [105]. Secondly, the generic anonymous pseudonym also eliminates the user's associated reputation, motivating the need for trust-preservation during anonymization. Iron Mask addresses both these issues.

4.2 Relating Anonymity and Trust

Information from a known source is easier to identify as being either more or less trustworthy than if it is coming from an unknown source [106]. An unknown person suggesting to take a certain medication will be less credible than a known non-expert. On social media websites, reputation can be a useful indicator of expertise. The reputation of a user in social media websites is often expressed using an aggregate of positive and negative feedback received from other users, such as answers to questions or comments. This mechanism is used by Reddit, Quora, Stack Exchange, and other social media websites where the aggregate points received can be used to determine a user's level of expertise. The assumption is that the higher a user's points, the more knowledgeable they are, given that they have received more positive than negative feedback. However, when users on websites like Quora decide to post anonymously, their reputation history is no longer associated with the content they post. This poses a problem for readers of the content because the author's credibility is hidden. Iron Mask addresses this issue with TPPs.

4.3 Iron Mask Algorithm Overview

The proposed algorithm, Iron Mask provides trust-preservation to balance the social media website's needs to generate credible content with the user's need for optional yet reliable anonymity. To achieve this, Iron Mask introduces the concept of the Trust Preserving Pseudonym (TPP), which provides a broader range of pseudonym labels, in addition to the generic "anonymous" pseudonym to mask or cover up the user's actual account name identity while appropriately contextualizing user credibility information. Even the naïve approach of using the generic "anonymous" label could lead to identity being compromised through grammatical and lexical correlations [36]. To preemptively avoid this, Iron Mask uses the whiteprint or authorship identification approach [35] to take into account the user's historical content. Whiteprint is defined as the unique writing style of an author based on grammatical and lexical patterns [107, 108]. This approach enhances anonymity by minimizing the risk of re-identification or de-anonymization.

The generic workflow of anonymization is summarized in Figure 4.1, where the user is provided a choice of using anonymity. If the user selects in the affirmative, then the author of the posting is reported with the generic label of "anonymous", and no hyperlinks or internal associations to the actual user are maintained. Consequently, anyone viewing the content will see the content's author as anonymous. Otherwise, the actual identity of the user is displayed. These two binary choices are available on social media websites such as Quora. The proposed approach using the Iron Mask algorithm provides an alternative route for anonymity.

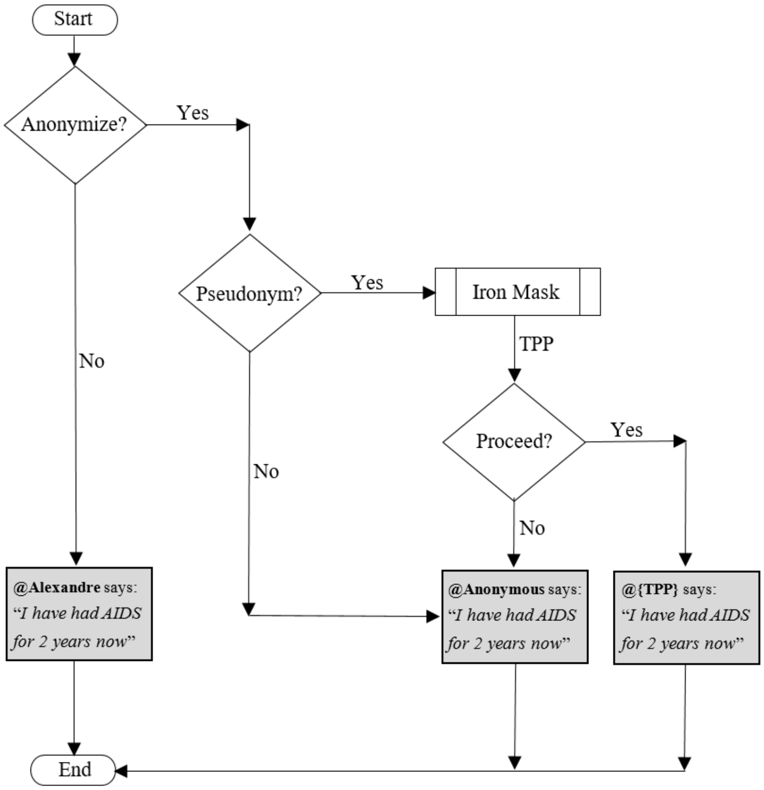


Figure 4.1: Overview of Anonymization and Pseudonyms

Iron Mask assigns a pseudonym using a two-stage approach: firstly, the content to be posted is scrutinized to determine the probability of de-anonymization. Secondly, a TPP is assigned based on the social network's characteristics and the user's profile. For example, the TPP could be "competent" based on a combination of the Dreyfus model of skill acquisition and the user's reputation points on the social media website. This would let the reader know that the user' is knowledgeable or not based on other users' feedback on previous postings.

Programmatically, the procedure for posting new content anonymously is abstracted in Algorithm 1. In the first step, the user can choose to anonymize or not. If they do not want to hide their identity, the content is saved with their actual user ID via the abstracted `SAVE()` function for storing the content to persistent storage. If the user does opt for anonymization through the generic "anonymous" username, this is saved along with the user's posted content.

Finally, if the user chooses to use a pseudonym, the risk of re-identification is assessed using a probabilistic classifier implemented via the `IRONMASK()` algorithm. If there is a risk, then the user is warned of this before proceeding, invoked by the `WARN()` function. It is up to the user to take the risk or not. If the users opts for the risk or in the event there is no risk of re-identification, the user's pseudonym is determined by `TPP()`.

Algorithm 1: ANONYMIZE

```
Input: user, content, anon, pseudo
if anon = False then
    author = user
else
    if pseudo = False then
        author = Anonymous
    else
        proceed = IRONMASK(user, content)
        if proceed = False then
            confirm = WARN()
        else
            confirm = True
        if confirm = False then
            author = Anonymous
        else
            author = TPP(user)
SAVE(author, content)
```

4.4 Whiteprint Identification using Probabilistic Classification

Algorithm 2 outlines the Iron Mask whiteprint identification step-by-step procedure using a probabilistic classifier, which provides the degree of confidence of a sample belonging to a class by predicting the probability distribution over the set of classes, in this case user IDs [109]. Within the workflow of Iron Mask, the probabilistic classifier is used for demonstrating the impact of quasi-identifiers with real-world data, and how Iron Mask uses the classifier outputs to prevent de-anonymization. To initialize, the classifier is trained using existing users and their postings from a social media website, where the user name is the class, and the content is converted to n-grams as features. The NLP pipeline for this involves tokenization of user postings via regular expressions, removing stop words [88], stemming of each token [89], and generation of combinations of stemmed tokens. Training computes a score for how strongly classes and attributes are associated, and the trained model can then be used for making predictions on new data, while probability calibration converts the scores to probabilities [110]. All possible combinations of adjacent words of length n within a posting are referred to as n-grams. For instance, a posting containing words $[w_1, w_2, ..., w_n]$ would yield bigrams as $[w_1w_2, w_1w_3, ..., w_{n-1}w_n]$. For implementation, naïve Bayes classifier with isotonic regression as the probability calibration was used with a combination of uni-, bi- and tri-grams as features.

```
Algorithm 2: IRONMASK
Input: user, content, $\tau$, n
candidates = PROBCLASSIFIER (content)
top_candidates = candidates[:n]
user_prob = candidates.FIND(user).probability
if user in top_candidates or user_prob $\geq \tau$ then
    return False
else
    return True
```

The probabilistic classifier is performing whiteprint identification by associating content and user identity [35], and the content is being used as a quasi-identifier.

More formally, user names and historically posted content can be expressed as the traditional database table defined in k-anonymizaion with n rows and m columns, with the rows representing each user's previously posted content, and the columns representing n-grams from the content, along with the user name. This database table also maps to the classification problem model, where each row comprises a complete tuple, and, in this case, the user name column is the identified class [111].

A new posting is input to the trained probabilistic classifier to get a set of predicted candidate users. On the trained probabilistic classifier, two thresholds are available for making a decision: top-n and τ. The top-n threshold returns the top candidate users based on the sorted (descending) degree of confidence. If the actual author is contained within the top-n candidates, then Iron Mask returns a warning. On the other hand, if the confidence level for predicting the actual author is greater than a given threshold, τ, then Iron Mask also returns a warning. τ is a threshold for the probabilistic classifier's degree of confidence (0 to 1). At higher values, τ enforces more stringent requirements for measuring likelihood of re-identification.

4.5 Linking Pseudonyms with Level of Expertise

The Dreyfus model of skill acquisition is used as a reference for anonymizing a user's online reputation [112]. The Dreyfus model specifies five categories of expertise: novice, advanced beginner, competent, proficient, and expert. Depending on the social media website, there are various reputation attributes available. For instance, Quora and Stack Overflow allow users to "Upvote" or "Downvote" postings based on the voter's perceptions of quality. This feedback, along with general interaction statistics such as number of postings and comments, can be aggregated as a reputation score for each user to determine the user's level of expertise on the Dreyfus hierarchical scale, with pre-configured mappings of scores to each level.

Algorithm 3 outlines this approach as an implementation of TPP using expertise and reputation. The scoring function incorporates the number of upvotes and downvotes received, as well as the total number of postings, while penalizing downvotes. The severity effect of downvotes on reputation is adjusted with a weighting factor.

The reputation score aggregation formulation can be customized to fit the needs of the social media website. Moreover, if there is not enough data available to define level of expertise, the "anonymous" label can be used.

Algorithm 3: TPP

```
Input: user, $[t_1, t_2, ..., t_5]$, w
rep = GETREPUTATION(user)
$score = (rep.upvotes + rep.num_postings)/(w * rep.downvotes + 1)$
if $score \geq t_1$ then
    return Expert
else if $score < t_1$ and $score \geq t_2$ then
    return Proficient
else if $score < t_2$ and $score \geq t_3$ then
    return Competent
else if $score < t_3$ and $score \geq t_4$ then
    return Advanced Beginner
else if $score < t_4$ and $score \geq t_5$ then
    return Novice
else
    return Anonymous
```

4.6 Evaluation of Iron Mask

Two aspects of the Iron Mask algorithm need to be evaluated. Firstly, the whiteprint identification approach is tested using datasets from the Quora question answering community. The evaluation demonstrates the accuracy of predicting the author of a post even when their user name is hidden. Secondly, the trust-preservation approach and TPP are evaluated using a survey-based approach to demonstrate usefulness. Datasets from Quora are used for evaluation of the proposed methodology.

A summary of the number of subsets retrieved is given in Table 4.1, along with the topics used for filtering the postings. The topics were selected in line with the focus of this research on sensitive health content. The retrieval process involved accessing a topic's list of questions, then retrieving the list of followers of the topic.

Subset Retrieved	Men's sexual health	Women's sexual health	Sexua-lity	HIV	Mental health	**Total**
Initial profiles	58	48	110	56	122	394
Questions	179	122	300	151	300	1,052
Answers	895	488	1,500	302	960	4,145
Additional profiles	358	97	750	30	348	1,583

Table 4.1: Quora Dataset for Evaluation Filtered by Topics

Each follower's profile was then programmatically accessed, and questions they have posted were retrieved, as well as upvotes and downvotes on each question. Users are also allowed to post questions anonymously, in which case the questions do not appear on their profile's listing of questions asked. Next, for each question retrieved, the corresponding answers were also enumerated, including the associated upvotes and downvotes, as well as additional profiles of users who authored the related answers.

4.6.1 Accuracy of Content as Quasi-Identifier

In order to determine the accuracy of whiteprint identification, a sample of users were arbitrarily selected from the Quora dataset using random sampling without replacement. Various iterations of this process were performed using different configurations of threshold, while the number of users selected was kept constant for all iterations. The sample dataset was then split into two parts for training and testing. The training set T_r was used for building the probabilistic classifier model. Next, the trained model was used with the other half of the sample dataset, i.e. the test set T_s, to predict user identity. Both the test and training sets were split such that the users in the test dataset were also in the training dataset. However, the content within the test dataset was not in the training counter-part. More formally, if u_i represents users and c represents content of the users, then $u_i \in Tr, c_i \in T_r, uj \in T_s$, and $c_j \in T_s$, but $u_j \subset u_i$ and $c_j \not\subset c_i$. The recall measure was used to determine the effectiveness of the trained model. Probabilistic classifiers are traditionally evaluated using RMSE, but since Iron Mask is being evaluated against various threshold values for top-n and τ, recall is best to measure the classifier's overall correctness.

As an illustrative example of the evaluation strategy, if top-n = 1, that implies that Iron Mask would only detect the user's correct identity and give a warning if the trained model ranked that identity with the highest probability. In other words, if a given user's identity was correctly predicted within the top-n, the recall score was recorded as 1, else it was recorded as 0. An average of the recall was taken for the various users selected for each iteration, shown in Figure 4.2 for different values of top-n. Similarly, for different values of τ, recall was recorded based on whether Iron Mask gave a warning or not, and the results are summarized in Figure 4.3.

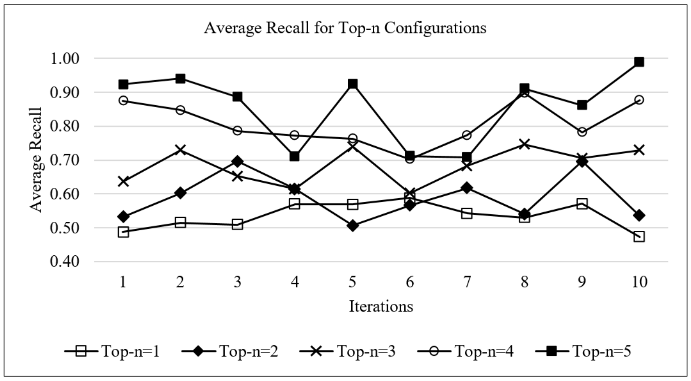


Figure 4.2: Average Recall for Top-n Configurations

For top-n, the recall and prediction of Iron Mask gets better with larger values of n. This is expected, because the larger the options to choose from, the higher the likelihood of discovering the item being searched. Expected results are also observed with τ, where lower values result in a much higher recall. These results demonstrate that Iron Mask is able correlate identity with historical postings to a satisfactory level, even with tight constraints such as $n = 1$ or $\tau = 0.90$.

4.6.2 Testing Trust-Preserving Pseudonyms

For testing the usefulness of TPPs, an online questionnaire-based survey was designed. A total of 46 anonymized responses were recorded for the survey, and there were no specific user profile criteria for participation. The survey started with displaying an arbitrarily selected question from the Quora database. Users were then asked to read the question, and then shown different versions of answers to choose from in two-step stages, with the control group being shown the "anonymous" label.

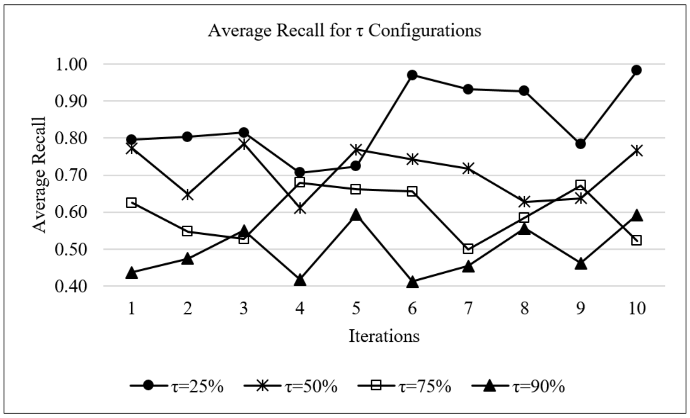


Figure 4.3: Average Recall for τ Configurations

For Step 1, they were shown an answer with two different user labels: one with the generic "anonymous" label, and the other with a TPP from level of expertise. Users were asked to select the answer format that they find more credible from the two choices; a binary comparative choice. In Step 2, users were shown a different answer to the question and asked to select if the answer is trustworthy or not; a binary affirmative yes/no selection. The associated user label in Step 2 was randomly assigned as either "anonymous" or TPP. Hence, some users were shown the "anonymous" label, while others were shown a TPP label computed from the actual user's level of expertise.

Figure 4.4 presents a summary of the results from Step 1, showing the total number of labels presented over the course of the survey, the number of positive selections for each label, and the number of negative selections as well. At first glance, it may look like the "anonymous" label was selected as the majority but this is actually not the case. Relatively, the generic label was selected by 17 out of the 46 users, while 29 users selected one of the TPP labels. In the breakdown shown for TPP labels, negative selections imply the "anonymous" label was preferred. Likewise, for the "anonymous" label, non-selection implies that one of the TPP labels were preferred. Further analysis reveals that out of the 17 selections, 10 were when the "novice" label was presented alongside with "anonymous". This might be due to the surveyors perceiving "novice" and "anonymous" being relatively similar in terms of low level of trustworthiness.

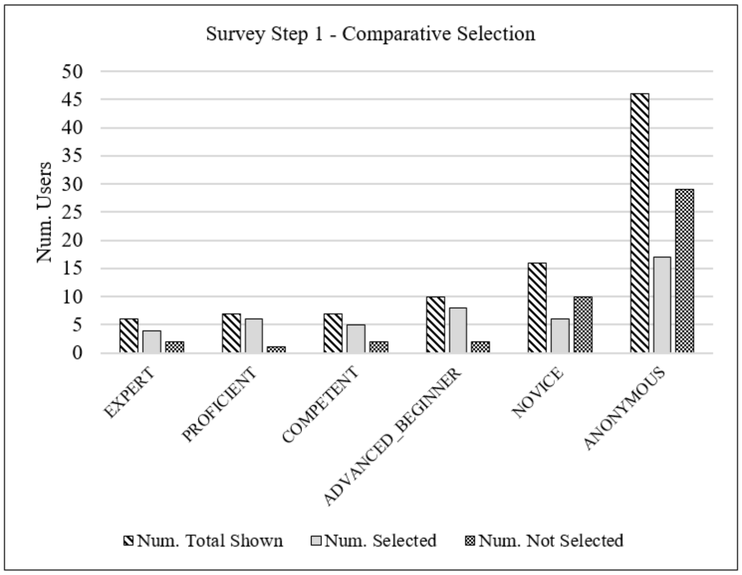


Figure 4.4: Iron Mask Survey Step 1 Results

Figure 4.5 shows the results of Step 2 of the survey, displaying the total number of instances of the labels presented, the number of “yes” selections implying the label was trustworthy, and the number of “no” selections when surveyors disagreed with the labels conveying trustworthiness. The results show that when the “novice” label was used, the users were more likely to disagree with the label conveying trustworthiness. As with Step 1, the users seemed equally likely to select between “anonymous” and “novice”. For the questions showing the higher-level expertise labels, the users agreed in the majority with the label being correlated with trustworthiness. This can be seen in both Steps 1 and 2, implying there was consensus within the sample population about the usefulness of the TPP labels.

This chapter covered implementation and evaluation details on Iron Mask and TPP. It was demonstrated that Iron Mask presents an improvement over existing anonymization options by investigating content as a quasi-identifier while also exploring interdependencies between identity, anonymity, and trust. The results provide a satisfactory baseline for concluding that content created by users can reveal their identity, evaluated via machine learning methods. Moreover, the proposed TPP has shown potential for providing a balance between user credibility and anonymity.

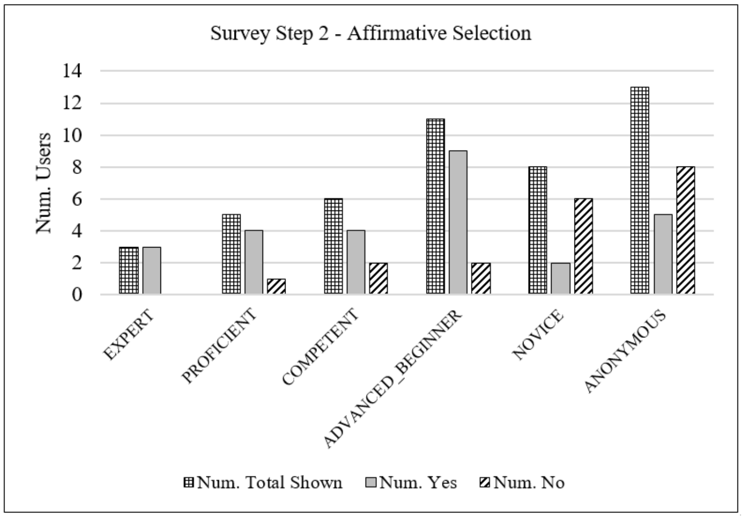


Figure 4.5: Iron Mask Survey Step 2 Results

Takeaway

- Whiteprint identification can leverage content as a quasi-identifier
- Iron Mask successfully predicts authorship to prevent de-anonymization
- Trust Preserving Pseudonyms provide a satisfactory alternative to the generic “anonymous” label
- Users with higher skill-based reputations may be more trusted than users with lower reputations
- The user’s reputation is factored in when generating a Trust Preserving Pseudonym

Chapter 5

Cardea: Health Social Network for Patients and Medical Professionals

This chapter covers details on Cardea, a health social network that consolidates algorithms developed as part of this thesis, including MedFact and Iron Mask. Cardea provides an avenue to explore Hypothesis 3 on the use of information retrieval metrics in conjunction with trust metrics to provide more credible health information results for search engines and recommender systems. The vision of Cardea is to improve social media usage in health care by developing a suite of intelligent tools to provide credible and safe health information to patients and medical practitioners. This chapter is largely based on the paper titled "The Need for Medical Professionals to Join Patients in the Online Health Social Media Discourse" currently *under review* at the International Conference on Health Informatics (HEALTHINF 2021).

5.1 State of Health Social Media

With the advent of social media, anyone with access to the Internet was able to have their say and generate UGC. Laypersons and patients have used this opportunity to interact, collaborate, and share their personal health stories, advise, and opinions on Health Social Media (HSM). However, this has led to varying degree of misinformation being propagated online, with severity ranging from acute to chronic, depending on the nature of the topics being discussed [15]. Early on, there were signs of the severity of this problem when websites promoting harmful cures for cancer using apricot pits were widely accessed, despite the pits containing cyanide.

Also not too long ago, the impact of health misinformation was felt once more as anti-vaccination campaigns started after viral posts on Facebook linked autism and measles vaccines [113]. We are currently living through the COVID-19 pandemic which has turned into a full-blown infodemic with severe consequences for misinformation being spread online about various facets of the disease including origins, causes, symptoms, prevention, and cures [100]. In the face of this new reality, laypersons crucially need trusted information in HSM discourse. Various websites have facilitated patients to seek medical knowledge via Owner Engineered Content (OEC) that is created and maintained by the website owners. In addition, these websites provide forums for patients to have discussions with other laypersons about their personal experiences with treatments and drugs. Some of these websites include WebMD, Doctissimo, Mayo Clinic, MedicineNet, MEDLINEplus, HealthLine, Patients Like Me, to mention a few.

Medical professionals have been using HSM as well, albeit to a more limited extent. A few websites enable medical professionals, such as doctors and nurses, to consult with patients via video conferencing or phone call. For instance, My Health Alberta's 811 HealthLink service[1] allows patients or caregivers to connect with registered nurses via phone call to discuss or ask questions on non-severe health issues [114]. Other paid telemedicine services, such as Dialogue[2], offer virtual healthcare services through web chat and video conferencing. It has also become common for healthcare organizations to have social media presence for brand recognition. Some websites are dedicated to enabling medical professionals exchange information with each other, such as DocCheck[3]. Other forums like Doctors Lounge[4] enable patients to ask questions of medical professionals.

HSM is frequently used for seeking health information online for self-diagnosis, self-treatment, and self-education. HSM also provides various benefits for patients and laypersons, such as allowing users to be part of virtual support groups, having fast and near-ubiquitous access to knowledge and actionable advice via the Internet.

[1] HealthLink https://myhealth.alberta.ca/811
[2] Dialogue https://www.dialogue.co
[3] DocCheck https://www.doccheck.com
[4] Doctors Lounge https://www.doctorslounge.com

At the same time, it raises concerns about misinformation being propagated by laypersons without professional medical expertise, especially during pandemics like COVID-19, leading to an infodemic. There are only a handful of HSM websites that allow medical professionals to have discussions with patients.

It is postulated that the modern face of medicine and healthcare needs more sophisticated algorithms to take veracity of content into consideration, as well as for medical professionals to be included in the online patient discourse so misinformation can be addressed early on and head on. To this end, a new health social network named Cardea is proposed which aims to bring patients, laypersons, and medical professionals together on the world wide web to share experiences, ask questions, and get credible answers. At the core of Cardea[5] is the metaphor of the hospital building, represented in the online environment. A hospital has different rooms, some are public and some are private, some are meant for doctors or nurses only, others for patients, and others for interactions between doctors, nurses and patients. Similarly, Cardea provides secure online pages where users can interact exclusively with their peers or with each other both publicly and privately.

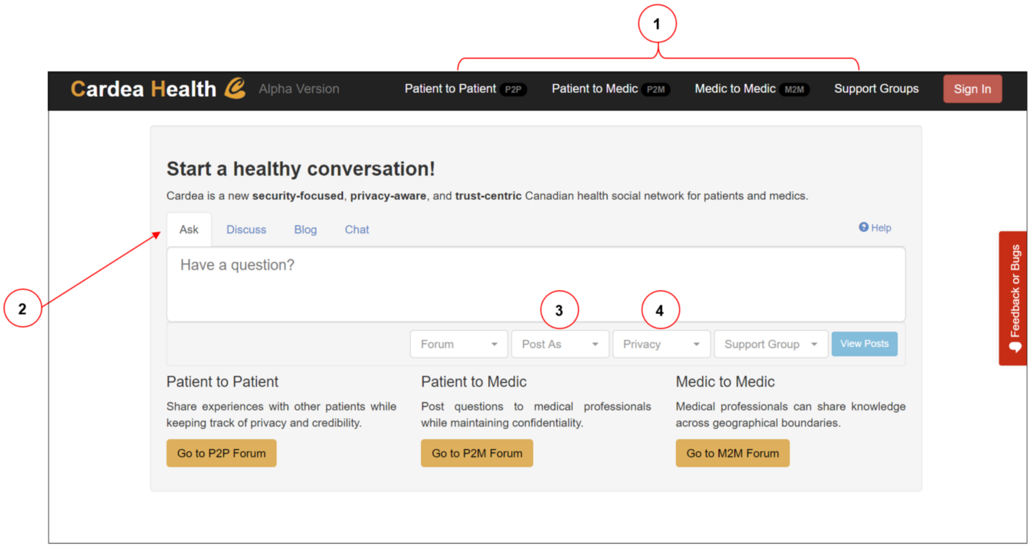
5.2 Functionality Specifications

Cardea aims to address the various challenges outlined with trust, stigmatization, and privacy. From a research perspective, Cardea has been conceptualized as a sandbox for facilitating and understanding HSM interactions. The Cardea homepage is shown in Figure 5.1, and descriptions of its functional aspects follow.

5.2.1 Folksonomies

A folksonomy is a methodology for allowing users to tag items, whereby the tags can be automatically organized as a classification system based on tag frequencies [115]. In Cardea, content is tagged by *Forums* and *Support Groups*. There are three forums representing hospital buildings having different rooms with varying privacy requirements, some public and others for doctors, nurses, or patients only.

[5] Cardea Alpha is live at https://www.cardeahealth.ca and hosted on Canadian servers by GreenGeeks while code is open source at https://github.com/hwsamuel/Cardea



Start a healthy conversation!

Cardea is a new **security-focused**, **privacy-aware**, and **trust-centric** Canadian health social network for patients and medics.

Patient to Patient

Share experiences with other patients while keeping track of privacy and credibility.

Patient to Medic

Post questions to medical professionals while maintaining confidentiality.

Medic to Medic

Medical professionals can share knowledge across geographical boundaries.

Figure 5.1: Cardea Homepage Screenshot; 1 - Forums and Support Group Folksonomies; 2 - Content Types; 3 - Authorship Settings; 4 - Privacy Controls

Other rooms are for patients, and yet others for interactions between doctors, nurses and patients altogether. Cardea provides Patient to Patient (P2P) secure online pages where laypersons can interact exclusively with other patients. Medical professionals can interact with patients in the Patient to Medic (P2M) online web pages, while medical professionals can engage in public or private discussions on the Medic to Medic (M2M) pages. Support groups constitute specific health-related topics that allow grouping of questions, discussions, and blogs topically. All content is indexed and grouped by health topics, hence patients can join these virtual support groups to get information by health topic.

5.2.2 Content Types

Cardea allows registered users to post text or hyperlinks to websites, images, and videos. Content is created within three categories: *Questions*, *Discussions*, and *Blogs*. Questions are focused on getting answers, while discussions are more open-ended conversations. Blogs provide an avenue for users to curate opinion pieces. Additionally, replies to discussions and blog posts are referred to as *Comments*, while replies to questions constitute *Answers*. The interactions follow the traditional HSM format of asynchoronous textual conversations such as real-time chat rooms.

Additionally, patients and medics are able to chat in a real time synchoronous environment via online chat on the P2M web pages, so patients can ask medics questions and get trusted advise instantly. Hence, a fourth content category of *Chats* between medics and patients is also provided only within the P2M forum.

5.2.3 Authorship Settings

Users are able to control how their identity is displayed and associated with their own content using Iron Mask, which was developed as part of this thesis. Three options are provided to label authorship: *Myself*, *Pseudonym*, or *Anonymous*. For the first option, the user's registered name is shown for authorship. For the second option, a trust-preserving pseudonym will be automatically assigned to the user. The pseudonym is assigned using a two-stage approach: firstly, the content to be posted is scrutinized to determine the probability of de-anonymization using machine learning and the whiteprint identification approach [35] on the user's prior historical content. If there is no risk of de-anonymization, then the user's authorship label shown to other relevant users is "Your Connection". This allows other connections to associate some level of trustworthiness with this user, even though their actual identity is hidden. For all other users, the user's type is displayed, either medic or patient. The final option is the generic label "Anonymous".

5.2.4 Privacy Controls

Users can specify who can view their content using four levels of privacy, from broadest visibility to more limited viewership: *Public*, *Registered*, *Medics / Patients*, and *Connections*. At the first level, content can be shared publicly and viewed by all visitors to the website without needing an account. The next level of privacy restricts content viewing only to users who have registered on the website. Thirdly, the type of user and the forum being browsed dictates viewership, either medic or patient. At the fourth level, users can opt to share the content only with other users that they have explicitly added as *Connections*, which are a two-way virtual relationship between users, similar to Facebook "Friends". Cardea's privacy-aware user recommender suggests similar connections based on topics of mutual interest.

This is especially useful if, for example, a patient is interested in rare diseases and wants to find like-minded users. Connections are suggested depending on two privacy configurations: visibility and discovery. Users can decide to be invisible from connection recommendations.

5.2.5 User Roles

Cardea allows two user roles: *Patient* or *Medic*. The patient is a generic label for anyone who would be seeking health information, while the medic label is for identifying medical professionals. In addition, a select group of users are assigned as *Moderators* for housekeeping and administrative tasks related to the website and content. Medics are manually verified based on their institutional email and correspondence with their organizations to ensure good standing.

5.2.6 Trust Metrics

The veracity of content is established within Cardea using subjective and objective metrics. Cardea allows users to provide subjective feedback on the quality of content by reacting with *Likes* or *Dislikes*. These specific reactions are limited to questions, discussions, blogs, and comments. Answers to questions can be *Up-Voted* or *Down-Voted* to reflect the extent to which the answer addressed the original question, in addition to quality and correctness. Objective metrics are computed via MedFact, developed as part of this thesis to enable credible content to be surfaced using established medical knowledge. This is achieved by processing content containing medical claims through an automated information retrieval process that uses natural language processing and machine learning to compare the claims against known facts extracted from publications in reputed medical journals. At the end, a percentage score is generated to represent the level of *Agreement* of the claims with known medical facts. In addition, votes and likes by verified medics are given more weight when scoring content quality and ranking search results. A reputation system keeps track of positive and negative feedback received by users. Reputations are visible to other registered users and are also contextualized by posts' topics.

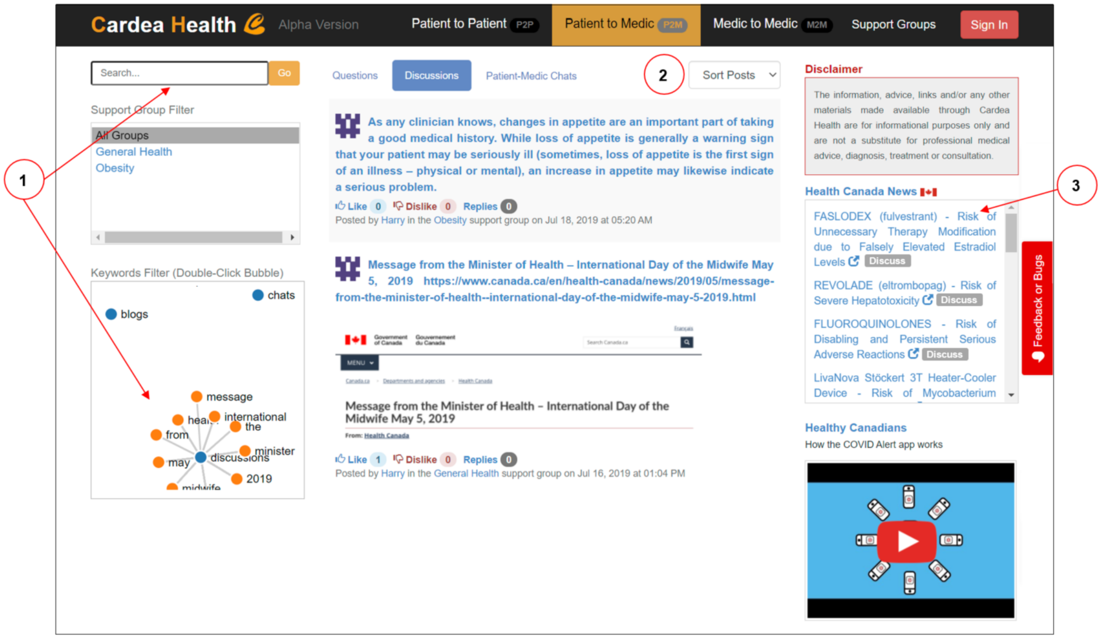
For instance, a user may receive good feedback on posts about pediatrics, but may not be contributing quality content in other topics. User reputation takes into consideration the user's historical postings record regarding the content's topic, as well as the user's overall history.

5.2.7 Faceted and Exploratory Search

Cardea provides standard search mechanisms, while also allowing faceted filtering within results by content topic and type. In addition, exploratory search is available whereby new content can be discovered from search results by displaying related concepts to the user's currently viewed content. For exploratory search, Cardea incorporates BubbleNet [116], which presents an abstract and high-level representation of major concepts, keywords, and topics discussed in a set of conversation threads. The relationships are visualized in the form of a network, showing the topics as well as their inter-connections. This network is built using an estimation of semantic relationships between topics. Having such a network, a user can see the big picture of all major concepts being discussed. The user then can navigate through this network by either refining or expansion. The user can drill down from a given topic to see other related concepts in a lower and more detailed level. The user can also navigate to other related topics and finally find a set of documents talking about their desired topics. This interface is shown in Figure 5.2 alongside other features.

5.2.8 Content Similarity and Real-Time Chat Recommenders

Cardea facilitates discussions by recommending external content that is relevant to existing conversations, including real-time chats between patients and medics using PubMedReco, a real-time recommender developed as part of this thesis to suggest citations from PubMed [117]. Full details on PubMedReco are provided in Chapter 7. PubMedReco analyzes keywords within synchoronous chats, including the relevant time window, and uses the keywords as search query to retrieve citations from PubMed, as shown in Figure 5.3. The same approach is applied to asynchoronous conversations in Cardea for suggesting articles from Health Canada. Users can then



As any clinician knows, changes in appetite are an important part of taking a good medical history. While loss of appetite is generally a warning sign that your patient may be seriously ill (sometimes, loss of appetite is the first sign of an illness – physical or mental), an increase in appetite may likewise indicate a serious problem.

Like 0 Dislike 0 Replies 0
Posted by Harry in the Obesity support group on Jul 18, 2019 at 05:20 AM

Message from the Minister of Health – International Day of the Midwife May 5, 2019 https://www.canada.ca/en/health-canada/news/2019/05/message-from-the-minister-of-health--international-day-of-the-midwife-may-5-2019.html

Message from the Minister of Health – International Day of the Midwife May 5, 2019

From: Health Canada

Like 1 Dislike 0 Replies 0
Posted by Harry in the General Health support group on Jul 16, 2019 at 01:04 PM

Disclaimer

The information, advice, links and/or any other materials made available through Cardea Health are for informational purposes only and are not a substitute for professional medical advice, diagnosis, treatment or consultation.

Health Canada News

FASLODEX (fulvestrant) - Risk of Unnecessary Therapy Modification due to Falsely Elevated Estradiol Levels Discuss

REVOLADE (eltrombopag) - Risk of Severe Hepatotoxicity Discuss

FLUOROQUINOLONES - Risk of Disabling and Persistent Serious Adverse Reactions Discuss

LivaNova Stöckert 3T Heater-Cooler Device - Risk of Mycobacterium

Healthy Canadians

How the COVID Alert app works

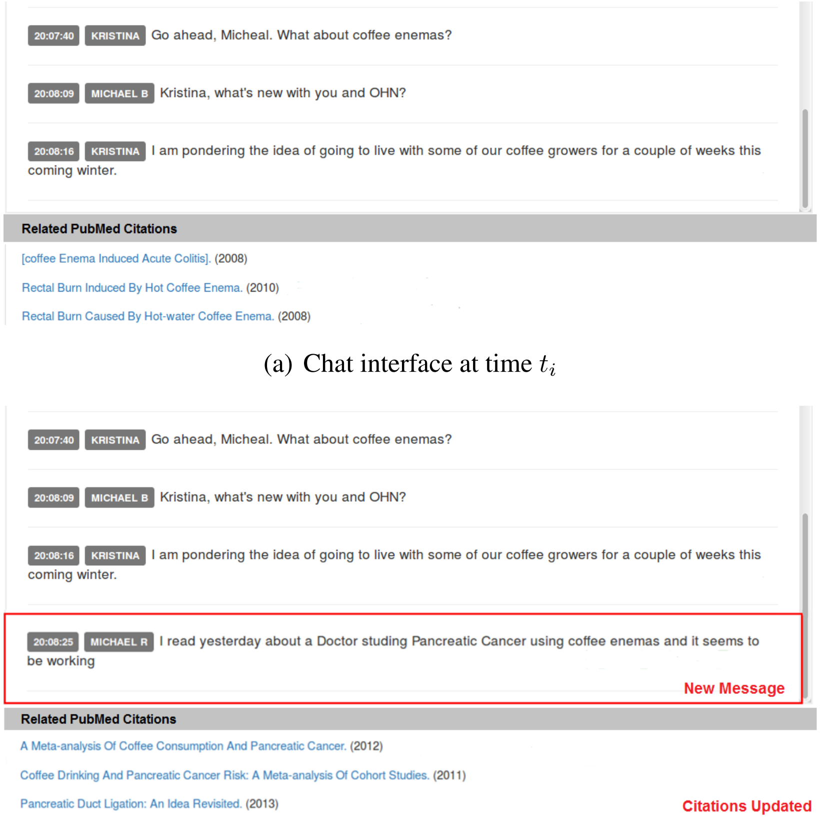
Figure 5.2: Cardea's Search and Recommender Features; 1 - Search Box and BubbleNet; 2 - Content Sorting; 3 - External Content Recommendation

directly discuss these news items within Cardea, as depicted in Figure 5.2. Cardea also enables recommendation of similar content in order to prevent duplication of existing discussions or questions. Figure 5.4 illustrates the practical difficulties of duplicate detection. While users are typing their questions, Cardea matches and displays similar content using DeepDup, a deep learning methodology developed as part of this thesis for surfacing similar questions using a Siamese neural network [38]. Technical details on DeepDup are provided in Chapter 8.

5.3 Prototype with Data

In order to test the HSM functions of Cardea, data was imported from existing websites and populated within Cardea. The data collection was facilitated using Grebe [118], a social data aggregation framework developed as part of this thesis, with details provided in Chapter 6. Grebe[6] contains 28 million tweets at the time of this writing, indexed from Twitter and geo-fenced for provinces within Canada.

[6] Grebe is live and hosted by Cybera at http://199.116.235.207/grebe while the code is open source at https://github.com/hwsamuel/Grebe



20:07:40 KRISTINA Go ahead, Micheal. What about coffee enemas?

20:08:09 MICHAEL B Kristina, what's new with you and OHN?

20:08:16 KRISTINA I am pondering the idea of going to live with some of our coffee growers for a couple of weeks this coming winter.

Related PubMed Citations

[coffee Enema Induced Acute Colitis]. (2008)

Rectal Burn Induced By Hot Coffee Enema. (2010)

Rectal Burn Caused By Hot-water Coffee Enema. (2008)

(a) Chat interface at time t_i

20:07:40 KRISTINA Go ahead, Micheal. What about coffee enemas?

20:08:09 MICHAEL B Kristina, what's new with you and OHN?

20:08:16 KRISTINA I am pondering the idea of going to live with some of our coffee growers for a couple of weeks this coming winter.

20:08:25 MICHAEL R I read yesterday about a Doctor studing Pancreatic Cancer using coffee enemas and it seems to be working

New Message

Related PubMed Citations

A Meta-analysis Of Coffee Consumption And Pancreatic Cancer. (2012)

Coffee Drinking And Pancreatic Cancer Risk: A Meta-analysis Of Cohort Studies. (2011)

Pancreatic Duct Ligation: An Idea Revisited. (2013)

Citations Updated

(b) Chat interface at time t_{i+1}

Figure 5.3: PubMedReco Real-Time Chat Recommendation Demonstration

Query:
Q_1: Sugar free
My 90 year old daddy just got diagnosed, the one thing he loves is ice cream, can he eat sugar free ice cream?

Expected:
Q_2: “Sugar-free” foods have same effect as sugar
I made a mistake last week - I bought and ate about 10 “sugar-free” caramels on my way home. I got home to a blood sugar of 250. I am not sure what artificial sweetener was used in these, or why they had such a profound effect. I would expect this if I ate 10 “regular” caramels. Anyone know?

Not Expected:
Q_3: High blood sugar in the morning
Why does my blood sugar spike so much in the morning, even when I eat a balanced meal for dinner. I takes all day to get it down to a normal level

Figure 5.4: Duplicate Content Detection Examples

In addition, Grebe contains scripts within its MedSpider module for aggregating data specifically needed by Cardea's forums with patient and medic content. For the P2P forum's questions, demo data is scraped from MSSE. Data for P2M discussions, data is retrieved from DoctorsLounge, while the patient-medic chat data is scraped from the eHealth's "Ask a Doctor" forum[7]. Finally, M2M discussions are gathered from Doc2Doc, while medic blogs are aggregated from DocCheck.

5.4 Promoting Credible Content

Cardea provides all the necessary features typical of HSM with the added benefits of being privacy-aware for both content visibility and authorship, as well as trust-centric, with subjective and objective metrics. These metrics for measuring veracity are especially useful for search and recommender results so that only trusted content is being suggested, and misinformation is being demoted, thereby mitigating search bias [78]. To achieve this, Cardea sorts all results in two steps: firstly by ranking results based on relevant keywords and search queries with NDCG, and secondly by re-ranking the top-n results using objective or subjective metrics. As a fallback, in the event that a veracity score is unavailable or inconclusive, subjective metrics from votes and likes are used. However, by default the re-ranking step uses veracity scores from MedFact. Users can also sort content being viewed by recency, subjective, or objective metrics, but results are ranked with objective metrics by default.

This chapter provided an overview of Cardea, a health social network that leverages and consolidates various algorithms developed as part of thesis for objective trust measurement (MedFact), identity privacy preservation (Iron Mask), real-time recommendations (PubMedReco), and duplicate content detection (DeepDup). Cardea is currently in its Alpha stage, and aims to encourage more medical professionals to join the online discourse so health misinformation is dealt with more effectively.

[7] eHealth https://ehealthforum.com/health/ask_a_doctor_forums.html

Takeaway

- Cardea is a Health Social Media portal for patients and medics
- Cardea provides a plethora of features that are useful for both patients and medics while ensuring privacy, security, and trust
- Data within Cardea is populated using Grebe and its MedSpider module
- Cardea incorporates MedFact, Iron Mask, PubMedReco, and DeepDup into a single health portal
- Search and recommender results as well as content filters all rank results using the trust-centric veracity score

Chapter 6

Using the Grebe Social Data Aggregator for Context Prediction in the Social Web

This chapter provides details on Grebe[1], an open source[2] social data aggregation framework developed as part of this thesis for extracting geo-fenced Twitter data and HSM data from various health-related forums. Grebe currently has over 28 million indexed public tweets, and has been leveraged by other researchers for text mining in social media for identifying city-wide events [119], investigating how public, organizations and health care professionals in Alberta, Canada express wellness in relation to children [120] and for exploring the effects of urban design on mental health[3]. Grebe also received news coverage from CBC[4] and Global News[5].

The dataset from Grebe is ideal for investigating the use of applied machine learning to predict three types of contexts: geographical context from user location using supervised classification, topical context via determining health-related tweets, and affective context via sentiment analysis with rule-based approaches.

[1] Grebe Live Demo http://199.116.235.207/grebe

[2] Grebe source code https://github.com/hwsamuel/Grebe

[3] Researchers Tap Twitter to Look at How Urban Design Affects Mental Health https://www.folio.ca/researchers-tap-twitter-to-look-at-how-urban-design-affects-mental-health/

[4] New Tool Engineered in Edmonton Mines Twitter for Health Trends https://www.cbc.ca/news/canada/edmonton/health-twitter-tool-university-alberta-1.5082731

[5] Alberta Scientists Use Machine Learning and Twitter to Better Understand our Health https://globalnews.ca/news/5117599/university-of-alberta-machine-learning-twitter-health/

The combination of these contexts provides useful insights for digital epidemiology. Moreover, various visualization tools can connect with Grebe's API.

6.1 Motivation for Social Data Aggregation

The social web is an ideal source of readily available public conversations on a variety of topics. Various platforms such as Twitter, Facebook, and others provide an avenue for users to publicly express their opinions, advice, and questions on topics such as politics, technology, health, among others. Within the context of health and wellness, this public discourse can provide valuable opportunities for tracking and predicting disease outbreaks, as well as measuring user engagement and opinion towards wellness policies [121]. The keywords used on the social web can enhance understanding about potential health symptoms and risks developing over time. Essentially, public conversations on the social web can be leveraged for epidemiology, involving analysis of public health patterns for disease prevention and promotion of wellness [122].

In order for social web posts to be useful for digital epidemiology, three contexts are identified that need to be available for analysis: location, domain, and sentiment. Firstly, the position of the user allows researchers to know where to look for health-related issues and epidemics. As an example, a post that mentions "Ebola" may not be useful unless users' location is known. Secondly, only health-related posts are relevant for epidemiology, and including posts from other domains would lead to noisy data. For instance, analyzing posts discussing politics or sports would not assist in detecting outbreaks or gauging public wellness. Thirdly, a user may mention a health-related topic, but positive or negative emotions expressed in a post could provide a better indication about the overall health context of the user making the posting. For example, a user may mention "Ebola" in a positive sense of learning about the disease at a seminar, instead of referring to their personal well-being or potentially contracting the disease.

One challenge for researchers is to associate social web posts with location so that the geographical context of opinions and health concerns could be studied. For instance, most users on Twitter disable their location when tweeting, while their profile location is free-text and can refer to fictitious places such as "Narnia". The limited coverage of public tweets with specific and verifiable coordinates limits the usefulness of social datasets for digital epidemiology. The limitations on data gathering also compound this problem. For instance, researchers have very limited access to the Twitter public dataset due to Twitter's rate limits, data collection policies, and high service costs. Full access to Twitter's dataset requires paid enterprise accounts via Gnip. There are rarely any substantive datasets that provide Canada-specific geographical context to digital epidemiology researchers, and most prior studies focus on the United States [123], with some studies on Japan [124]. Recently, Twitter announced their Premium API tiers[6], which are much more cost effective for researchers.

Another challenge is categorization of text-based postings such as tweets. In order to use tweets for epidemiology, they need to be classified as being health-related. However, this is not a trivial task because there are various aspects of health that need to be considered, including physical, intellectual, occupational, spiritual, emotional, and social wellness. Within each category are multiple keywords and variants that need to be detected in a tweet for it to be considered as health-related. In addition, a user may mention health keywords in their tweet without necessarily referring to themselves or talking about their own well-being.

Moreover, once health-related posts are identified, their polarity and sentiment can help further analyze the full status of users' health. In other words, the users feelings about their own health and personal well-being can be understood. For instance, tweets could contain various health-related keywords in a positive context such as feeling healthy and well. On the other hand, tweets could also be referring to health problems, ailments, and symptoms, which would be useful for tracking epidemics and outbreaks.

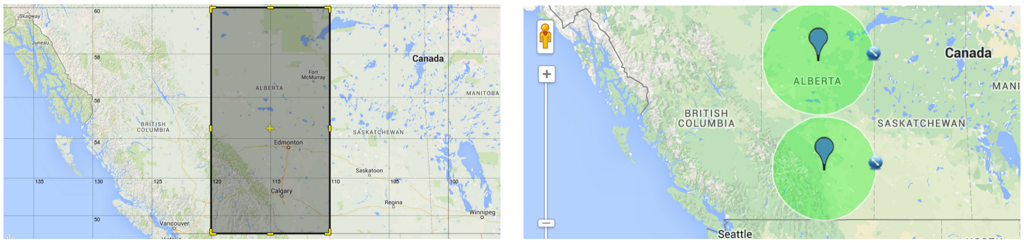
[6] Twitter Premium API https://developer.twitter.com/en/products/twitter-api/premium-apis

6.2 Analysis Methodology on Grebe Dataset

6.2.1 Data Gathering

To gather data from Twitter, the official and freely available Twitter Streaming and Search RESTful Application Programming Interfaces (APIs) were used since the inception of the Grebe project in July 2016. The Streaming API[7] provides access to a limited set of randomized realtime tweets, while the standard Search API[8] allows limited searching of historical tweets retroactively.

Retweets are ignored and incoming tweets are filtered by location using API parameters. The streaming API can return tweets made from within a defined bounding box specified using longitude and latitude of a rectangular region[9]. The search API filters tweets within a specified circular geographical region via a radius around a longitude and latitude point[10]. Figure 6.1 gives a visual illustration of the bounding box and inscribed circle options. Grebe gathers tweets from specific Canadian provinces, so the bounding boxes and inscribed circles are configured according to the geographical coordinates of the specific regions, with multiple bounding boxes and circles per province when necessary for maximum geographical coverage.



(a) Stream API Bounding Box

(b) Search API Inscribed Circle

Figure 6.1: Twitter API Geographical Filtering Options

[7] Twitter Streaming API Developer Documentation https://developer.twitter.com/en/docs/tweets/sample-realtime/api-reference/get-statuses-sample

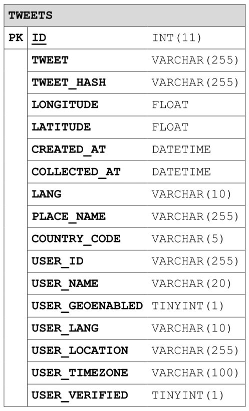
[8] Twitter Standard Search API Developer Documentation https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

[9] BoundingBox tool to configure rectangular geo-fence's edge coordinates and dimensions http://boundingbox.klokantech.com/

[10] FreeMapTools used to configure inscribed circle's centre and radii https://www.freemaptools.com/radius-around-point.htm

Grebe is implemented using Python, Flask, and Tweepy[11] on an Infrastructure-as-a-Service (IaaS) cloud platform running Ubuntu, with the Grebe web application and Grebe's RESTful API served via Web Server Gateway Interface (WSGI). The aggregation of tweets is done via cron job while respecting the Twitter rate limits[12]. Grebe is available as an open source Git project via GitHub for researchers[13].

Ultimately, public tweets retrieved from the Twitter API are indexed and stored in a MariaDB SQL database, with the indexed fields shown in Figure 6.2. The tweets are denormalized and users entities in order to capture snapshots of user information at the time of tweeting. NoSQL databases such as MongoDB were also investigated, but it was found that the disk space and performance metrics for MariaDB were optimal. MongoDB requires significantly more disk space, while performance for Create, Read, Update, Delete (CRUD) operations is similar between MongoDB and MariaDB.



TWEETS		
PK	ID	INT(11)
	TWEET	VARCHAR(255)
	TWEET_HASH	VARCHAR(255)
	LONGITUDE	FLOAT
	LATITUDE	FLOAT
	CREATED_AT	DATETIME
	COLLECTED_AT	DATETIME
	LANG	VARCHAR(10)
	PLACE_NAME	VARCHAR(255)
	COUNTRY_CODE	VARCHAR(5)
	USER_ID	VARCHAR(255)
	USER_NAME	VARCHAR(20)
	USER_GEOENABLED	TINYINT(1)
	USER_LANG	VARCHAR(10)
	USER_LOCATION	VARCHAR(255)
	USER_TIMEZONE	VARCHAR(100)
	USER_VERIFIED	TINYINT(1)

Figure 6.2: Public Tweet Information Indexed by Grebe

For optimizing **C**reate operations, a hash of each tweet was stored and checked before new tweets were saved. To optimize **R**ead operations, table indexes for the `tweet`, `tweet_hash`, `created_at`, and `place_name` fields were used. The system does not carry out any **U**pdate or **D**elete operations because saved tweets do not need to be updated or deleted.

[11] Tweepy Library http://www.tweepy.org

[12] Twitter Rate Limits
https://developer.twitter.com/en/docs/basics/rate-limiting

[13] Grebe GitHub Repository https://github.com/hwsamuel/Grebe

6.2.2 Tweet Location Prediction

The location of a tweet is defined as the geographical position from which a tweet was made. Grebe also aggregates tweets with missing longitude and latitude information. From the collected dataset of over 18 million tweets, 14% of the tweets contained verifiable location information. This is a general trend, as few users enable geo-tagging when tweeting [125]. In this case, this was also because the Twitter API often returns tweets using an estimated location based on the free-text user profile location. Hence, while the Twitter API categorizes these tweets from a specific region, the tweets' actual longitude and latitude is missing. Nevertheless, these tweets are useful for expanding the sample size. Moreover, this dataset of unmarked tweets, along with tweets with verified location, is useful for investigating whether a tweet's location could be predicted without geo-coordinates.

Other studies have attempted to identify granular tweet location from text using word collocations [126]. For tweet location prediction, a supervised classifier is used to predict the Canadian province from where the tweet was posted. City-level predictions are also investigated. For the features of the classifer, tweet n-grams are used. Tweets are tokenized using regular expressions, with spacing and variants as separator, and converted to n-grams. Stop words are then removed using the Glasgow list [88], and all possible combinations of adjacent stemmed words of length *n* within a posting are referred to as n-grams. For example, a tweet containing words $[w_1, w_2, ..., w_n]$, would generate the list of bigrams as $[w_1w_2, w_1w_3, ..., w_{n-1}w_n]$. A combination of uni-, bi- and tri-grams as features are used.

The class labels correspond to the city names and postal abbreviations of provinces to be indexed, for example AB, ON, SK, BC, MB, and QC. For evaluation, 10-fold cross validation was used with 80-20% split between training and holdout data respectively at each iteration. Four classification approaches were explored: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) neural network, Latent Dirichlet Allocation (LDA), and Random Forest (RF) [127].

6.2.3 Classifying Health-Related Tweets

In order to determine whether a tweet is referring to the tweeter's personal health status, the six dimensions of wellness model was used [1], which broadens the meaning of health as not merely the absence of disease or infirmity. In essence, health can be defined as a combination of physical, intellectual, occupational, spiritual, emotional, and social wellness.

Physical wellness covers physical activity, healthy eating, use of appropriate drugs and supplements to avoid stress, fatigue, and diseases. Intellectual wellness covers lifelong learning, acquisition of skills, and self-education. Occupational wellness covers use of personal talents and skills to perform paid professional work or unpaid volunteering. Spiritual wellness covers the pursuit of peace and harmony through a value system. Emotional wellness covers mental and psychological stability, enablement of positivity, avoidance of negativity, coping with life challenges, and expressing feelings. Finally, social wellness covers personal and communal relationships with friends or strangers, generally anyone users interact with [1].

The problem of classifying tweets as health-related can be formulated as follows. Firstly, how can tweets be classified as being related to health? And secondly, if a tweet is related to health, which dimension of health is it related to? Four approaches are explored for a potential optimal solution: keyword search (KS), document search (DS), supervised learning (SUP), and semi-supervised learning (SSUP).

Keyword search is simplistic solution, where a list of keywords is created, each related to one dimension of wellness. If a tweet contains any of the keywords in these lists, it is considered as being related to the dimension the list corresponds to. A tweet can be related to multiple dimensions of health with this approach. If a tweet has more than one of the keywords associated with a dimension, a stronger relationship between the tweet and that dimension of health is assumed. For this approach, six lists of keywords per wellness dimension were manually curated by reading documents describing the dimensions. Tweets were then categorized based on keywords.

With the document search method, this problem is approached from the point of view of information retrieval. Firstly, a dataset of ten documents for each dimension of health is curated. Hence, a total of sixty grouped documents describing the various dimensions of health is available. Secondly, every tweet is seen as a query to this database of documents. Based on various measures of similarity, a decision can be made about the label of a tweet.

This dataset is queried with each tweet to fetch the most similar documents. The query keywords are extracted from the tweet by firstly tokenizing, then removing stop words via the Glasgow list [88], and finally stemming each token using the Porter stemmer [89]. Next, each tweet is classified as being related to a health dimension if there is at least one document with a similarity score higher than a pre-defined threshold. The tweet is then classified to the label of its most similar document. Two measures of similarity are used: cosine similarity and set containment. In the experiments, a threshold of 0.2 is used to determine whether a document is related to any of the health dimensions.

For cosine similarity, both the document and the query are transformed, in this case the tweet, into a vector representation. This vector has n items, with n being the number of distinct terms in the query and the document. The value of each item in the vector shows the number of times that term has appeared in the query or the document. Using these vectorized formats, the similarity of the document and the query is defined as the cosine of the angle between the two vectors. For set containment, both the document and the query are considered as a set of terms, and define similarity based on common keywords as $\frac{|D \cap Q|}{|Q|}$, where D are the document keywords and Q are the tweet keywords. Manually labeled tweets are used for evaluation of the results.

In the supervised learning method, the manually labeled set of tweets is sued to train a naïve Bayes classifier for each health dimension. For every data point, tweet n-grams are used as features for training. To evaluate this method, 5-fold cross validation is used. At every iteration, 20% of the data is set aside as holdout, and train the classifier with the remaining 80%. Accuracy is calculated from the holdout set.

Other studies have proposed a semi-supervised binary approach to label a stream of incoming tweets [128]. For this approach, two sets of keywords are curated for the six dimensions of wellness. The first set contains keywords specific to each dimension, while the second set contains general health keywords. If a tweet includes keywords in the first set, they are labeled with the corresponding health dimension. On the other hand, if a tweet has none of the keywords from the first set, but at least one from the second set of general health terms, it is marked as possibly being related to health and set aside. If a tweet has no match in both lists, it is not health-related.

Using these initial labels, a naïve Bayes classifier is trained for each health dimension. The trained model is then used to label the tweets that were set aside. There is now a larger set of labeled data available for training that reveals new collocated keywords, and the newly labeled tweets can be included in the training dataset. The expanded dataset is then used to train a better, hopefully more accurate classifier. By repeating this process as more tweets are received, the classifier can improve iteratively. Accuracy is calculated every time processing a batch of tweets is completed by evaluation on an evenly split set of labeled samples.

6.2.4 Tweet Sentiment Analysis

Sentiment analysis enables evaluation of the emotions expressed in text. There are two potential objectives: continuous metrics or discrete labels. For the former, the polarity of a given text is computed to give an indication about positive, negative, or neutral emotions and the degree or strength of the sentiment. For the latter, emotion-based labels are assigned to the text, such as the eight human emotions of *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust* [129]. For this research, rule-based approaches are applied to determining polarity of the health-related tweets, and also investigate multi-label emotion mining.

To determine the degree of polarity of the health tweets, rule-based lexicon approaches are used via the Liu & Hu[14] lexicon [130] and the VADER (Valence Aware Dictionary and sEntiment Reasoner)[15] lexicon for social media datasets [131]. The rule-based approach was initially applied to customer review datasets, while the VADER tool is specifically targeted towards Twitter datasets. Both lexicons contain positive, negative and neutral words, and the frequency of words in-text is used to compute polarity within the range $[-1, 1]$.

Multi-label emotion mining from text is an interesting area that few studies have explored previously [132]. Since human emotions often tend to co-occur, multi-labels are applied to health tweets using the National Research Council (NRC) Canada's Word-Emotion Association Lexicon (EmoLex) [133], using labels from psychology literature, specifically Plutchik's wheel of 8 human emotions [129]. EmoLex contains multiple labels associated with words, hence a simplistic rule-based approach is applied. Given a tweet with tokenized words as $[w_1, w_2, ...w_n]$, the associated labels for each word are determined, $L(w_i) = [l_1, l_2, ...l_m]$. Hence, the tweet's emotion multi-labels are all labels with aggregated frequency above a given threshold.

Additionally, a self-reference filter is applied by detecting use of personal pronouns within health tweets. With this filter, the final dataset contains tweets mentioning personal health issues rather than user commentaries on general or public health topics. This simplistic heuristic enables more focused digital epidemiology because users' personal health condition might be more useful for predicting outbreaks.

[14] NLTK's Sentiment Analyzer contains Liu & Hu's lexicon http://www.nltk.org/api/nltk.sentiment.html#nltk.sentiment.util.demo_liu_hu_lexicon

[15] VADER sentiment analysis tool is open source and code is available via GitHub https://github.com/cjhutto/vaderSentiment

6.3 Results from Grebe's Analytics

6.3.1 Showcase of Dataset and Interface

Table 6.1 provides statistics on the present size of the dataset collected by Grebe from July 2016 till date. It should be noted that aggregation for some provinces was started later. Also, various provinces are not presently being indexed, but there are future plans to expand Grebe's coverage. There are some non-Canadian tweets captured from the neighbouring United States due to the approximate nature of some of the bounding boxes and inscribed circles used in geo-fencing.

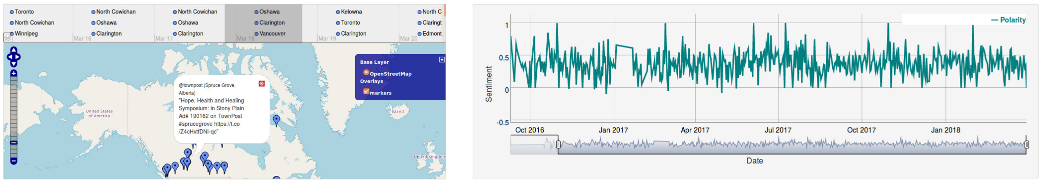
Table 6.1: Grebe Dataset Statistics

Statistics	**Total Amount**	**Quick Glance**
All Tweets	18,250,853	+18M
Tweets with Coordinates	2,555,973	+2M
Tweets from Alberta (AB)	336,228	+300k
Tweets from Ontario (ON)	552,428	+500k
Tweets from Saskatchewan (SK)	56,118	+50k
Tweets from British Columbia (BC)	319,185	+300k
Tweets from Manitoba (MB)	76,026	+70k
Tweets from Quebec (QC)	102,379	+100k

The vision of Grebe is to provide Canadian and worldwide researchers geo-fenced social data specific to Canada. Grebe consists of three main sub-systems: data aggregator, RESTful API, and tools. Firstly, the data aggregator presently focuses on Twitter but future plans include adding other social web sources. All aggregated data is geo-fenced so that longitude-latitude pairs are partially available. In addition, the preliminary results for location prediction are promising. Secondly, the data stored in Grebe is accessible via a RESTful API[16] as JSON output, with various filtering and querying options. Researchers can use the data to develop tools and perform analysis, such as classification of health tweets or sentiment analysis. Thirdly, Grebe provides visualization tools for analysis: time map and trend graph.

[16] Grebe API documentation http://199.116.235.207/static/api_docs.pdf

The time map enables visualization of data on a map, along with a temporal overview of data variations. The trend graph demonstrates summarization of statistics over time, such as top keywords and hash tags being tweeted, or sentiment polarity. Figure 6.3 shows the time map and trend graph visualizations.



@townpost (Spruce Grove, Alberta)
"Hope, Health and Healing Symposium: in Stony Plain Ad# 190162 on TownPost #sprucegrove https://t.co /Z4cHxtfDNI-qc"

(a) Time Map

(b) Trend Graph

Figure 6.3: Grebe Visualizations Showcase

Grebe currently provides hash tag search that can allow filtering of data being shown on the visualization tools. The most frequently used general hashtags are also recommended for filtering. Another feature under development and testing is a keyword recommender that can suggest health keywords to use for filtering. An inverted index of all keywords within tweets is used, and the top-*n* medical keywords from the past *m* months are suggested to prevent outdated keywords from appearing, and increase serendipity.

General health keywords are identified by their occurrence in SNOMED, a digital collection of medical terms provided by the U.S. National Library of Medicine [91]. In order to properly identify layperson health terms, the CHV mapping is also used [134].

6.3.2 Location Prediction Performance

For province predictions, class labels are balanced by selecting 50,000 tweets per province based on the minimum number of indexed tweets from Saskatchewan, giving a total of 300,000 tweets used with 80-20% training-holdout split.

For city predictions, out of 3,843 Canadian cities indexed in Grebe, 19 cities (those with over 10,000 tweets) were selected, with class labels balanced at 10,000 tweets per label, giving a total of 190,000 tweets in the training and holdout datasets. The accuracy metrics for province and city predictions are shown in Figure 6.4.

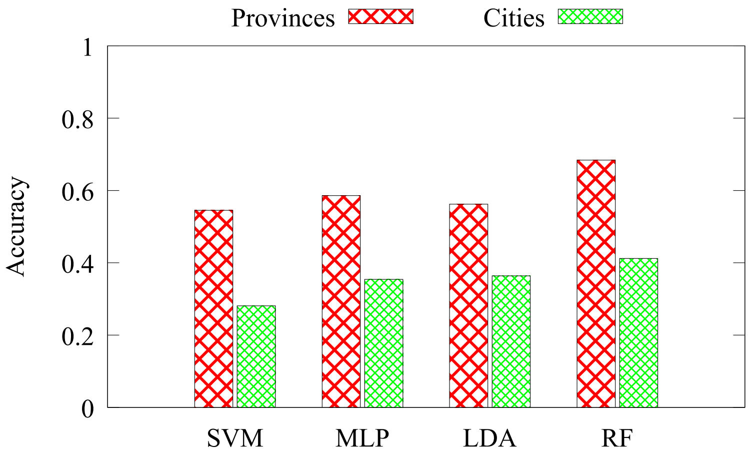


Figure 6.4: Prediction of Canadian Provinces and Cities

For the MLP neural network, the logistic sigmoid activation function is used, and Adam solver with adaptive learning. A deep neural network approach was also explored with up to five hidden layers, but the overall performance decreased. For the SVM classifier, the sigmoid kernel was used, while for LDA the best performance was obtained with a learning decay set at 0.65. Overall, the Random Forest classifier recorded the best performance with 68% accuracy in predicting provinces and 41% prediction accuracy for cities. The low predictability of cities could be attributed to the large size of class labels. Accuracy@k was also evaluated, where k represents the top-k labels predicted. For accuracy@3, provinces are predicted with 78% accuracy, while cities can be predicted with 53% accuracy using the Random Forest classifier.

6.3.3 Health Context Classification

Over 118,000 tweets from the province of Alberta were used to evaluate the health-related tweet classification strategies. For the evaluation dataset, 100 tweets related to each of the six dimensions of wellness were manually labeled, resulting in 600 tweets. 221 tweets were also labeled as not being related to any of the dimensions of health. 821 tweets with 221 negative and 600 positive examples were manually labeled.

The process of manual labeling was done by reading through the list of available tweets sequentially, to a point where Enough labeled samples were available for physical and emotional dimensions of wellness. Next, keyword search was used to find possibly relevant tweets for other dimensions with sparse tweets. A summary of the accuracy for different classification strategies is shown in Figure 6.5.

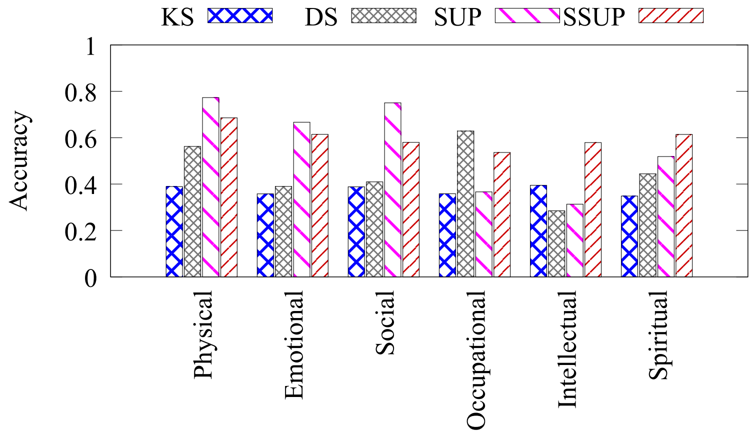


Figure 6.5: Identifying Health-Related Tweets by Wellness Dimension

Supervised learning provided best results for the physical, emotional and social wellness labels, while the semi-supervised learning approach provided best results on average for the occupational, intellectual, and spiritual wellness labels. Generally, the performance of the classifiers was lower for intellectual, spiritual and occupational dimensions of wellness across all methods. This can be attributed to the fact that these dimensions are more difficult to define and to find keywords for. The semi-supervised approach gave better performance for some categories because finding descriptive and well-defined documents for these dimensions proved to be more difficult. Hence, supervised methods could not readily detect patterns and correlations within these documents and required semi-supervised human moderation. Generally, there was less agreement over the label of tweets in these categories compared to other dimensions. For the document search approach, cosine similarity outperformed set containment, while for semi-supervised learning, accuracy increased proportionally with the iteratively increasing size of the training set. The keyword search approach performed poorly in general.

6.3.4 Sentiment Analysis Insights

Firstly, trends in polarity across provinces were analyzed by aggregating average polarity scores per month using both the Liu & Hu lexicon and the VADER lexicon for the classified health-related tweets. An overview of polarity of various Canadian provinces over the period of July 2016 to March 2018 is shown in Figure 6.6. Data collection for some provinces was started later than July 2016. There is some suggestion from the Alberta statistics that users tended to be more negative seasonally around November, but more data is needed to confirm any patterns.

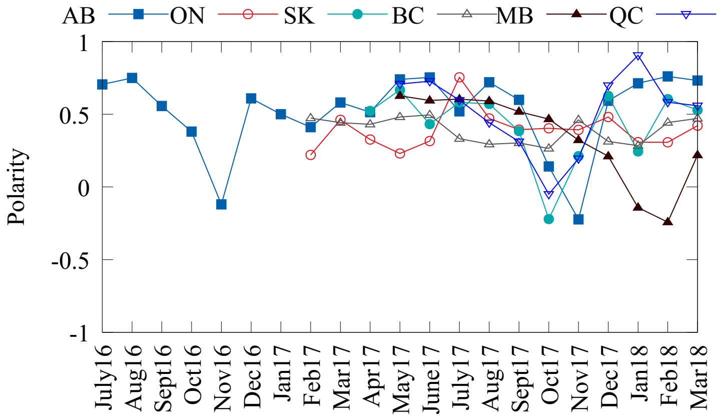


Figure 6.6: Average Monthly Polarity of Health Tweets from Canadian Provinces

Secondly, motion multi-labels of health tweets were investigated from the province of Alberta over time using the NRC-Canada's EmoLex lexicon. The frequency of emotion labels is summarized in Figure 6.7. For Alberta, *joy* was the most frequent emotion expressed, while the *sadness-anger* labels occurred frequently together.

6.4 Discussion and Notes on Grebe's Performance

Only publicly available tweets are indexed by Grebe to respect users' right to privacy [135]. Moreover, minimal personal information about users is collected, as the focus is on analysis of textual conversations on the social web rather than user profiling. Technical considerations for Grebe include up-time of the RESTful API, stoppages affecting the cloud architecture, Twitter rate limits, and cron job errors.

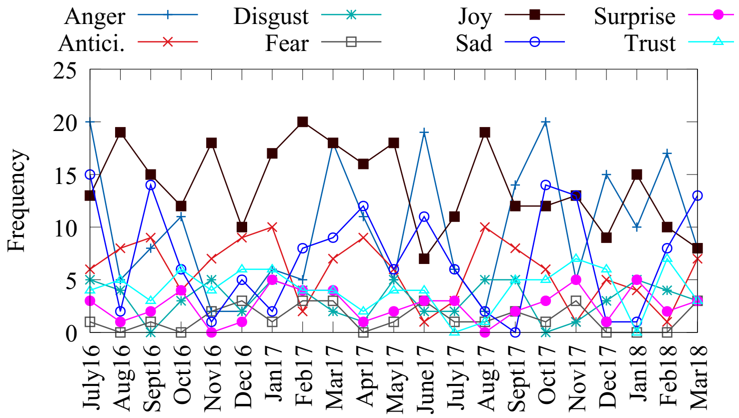


Figure 6.7: Emotion Multi-Labels of Health-Related Tweets from Alberta

The RESTful API is available on-demand as long as the Ubuntu cloud virtual machine does not experience outages from the service provider.

Grebe respects Twitter's rate limits for retrieving tweet streams by suspending the cron job when the rate limit is reached, and resuming after the timeout periods recommended by Twitter. There have been instances when the cron job do not resume correctly, and this has led to some missing data for certain periods. Over time, with continuous testing, the Python scripts and cron job have gotten more stable, leading to a higher volume of data collected.

Limitations of the system include sample size, collection period, and health tweets. The sample size, though large, is still a small representative of the overall population. This is a general limitation for social web data because not everyone in the population uses social networks. Secondly, the data covers just under two years, and has not reached full maturity for identifying patterns and predicting future incidents. Grebe has the potential to provide these services in the near future, as the dataset is continuously expanding. Thirdly, the size of users who talk about their health online is a small representative of the overall population. More research and data is needed to determine whether this is a representative sample.

This chapter showcased the open source Grebe social data aggregator that was developed as part of this theis, and has been used for extracting geo-fenced Twitter data, in addition to datasets from other HSM sources. From this research, there are promising results for predicting a tweet's province and city using supervised learning, even when it is not geo-tagged. It was also demonstrated that health-related tweets can be identified and used for further analysis such as determining tweet polarity and emotion labels. For this thesis, Grebe is utilized for data collection for Cardea, but it also has broader uses for social data analysis and digital epidemiology, as demonstrated by its usage by other researchers, and news media attention.

Takeaway

- Grebe is a social media data aggregator for data collection from Twitter
- Grebe has over 28 million geo-fenced tweets made from within Canada and available to researchers via a RESTful API
- Grebe is utilized for data collection of Cardea for testing and demonstration
- Three contexts have been explored on the Grebe dataset: geographical, topical, and affective
- The dataset from Grebe is ideal for digital epidemiology and Health Social Media analysis

Chapter 7

Using PubMedReco for Suggesting Trusted Content in Real-Time Chat Rooms

This chapter gives details on PubMedReco, a recommender system developed as part of this thesis which can analyze discussion threads to extract relevant medical terms, and then query trusted knowledge sources like PubMed to suggest related content related to the ongoing discussion. PubMedReco is utilized within Cardea to provide recommendations from credible sources. The contents of this chapter are largely based on the paper "PubMedReco: A PubMed Citations Recommender System for Real-Time Chat" published and presented at the IMIA World Congress on Medical and Health Informatics (MedInfo). The publication was runner up (second place) for the conference's *Best Student Paper* award.

7.1 Motivation for PubMed Recommendations

PubMed is the de facto tool for searching biomedical and life sciences literature, comprising of the MEDLINE bibliographic database, which covers academic journals on medicine, pharmacy, nursing, dentistry, and medical care, among others. PubMed is the web interface to the MEDLINE database, with over 23 million trustworthy and peer-reviewed articles [136]. It is typically used from its home page at the website of the National Center for Biotechnology Information (NCBI). Based on the keywords entered in the search engine, matching citations are returned.

The MEDLINE database is indexed using the Medical Subject Headings (MeSH), an indexing medical vocabulary [137]. To facilitate natural language usage in the PubMed search engine, PubMed uses the process of Automatic Term Mapping, which matches non-MeSH keywords to the MeSH index. A MeSH translation table is used for each query issued by the user, which maps the natural language keywords to equivalent MeSH keywords. In order to broaden search results, PubMed also leverages mappings of the search query keywords derived from the UMLS.

UMLS is a collection of biomedical vocabularies and standards, and PubMed can leverage these vocabularies to expand search queries using hypernymns, hyponyms, synonyms, and other semantic relationships [138]. The interactions between PubMed, MeSH, UMLS, and MEDLINE are depicted in Figure 7.1.

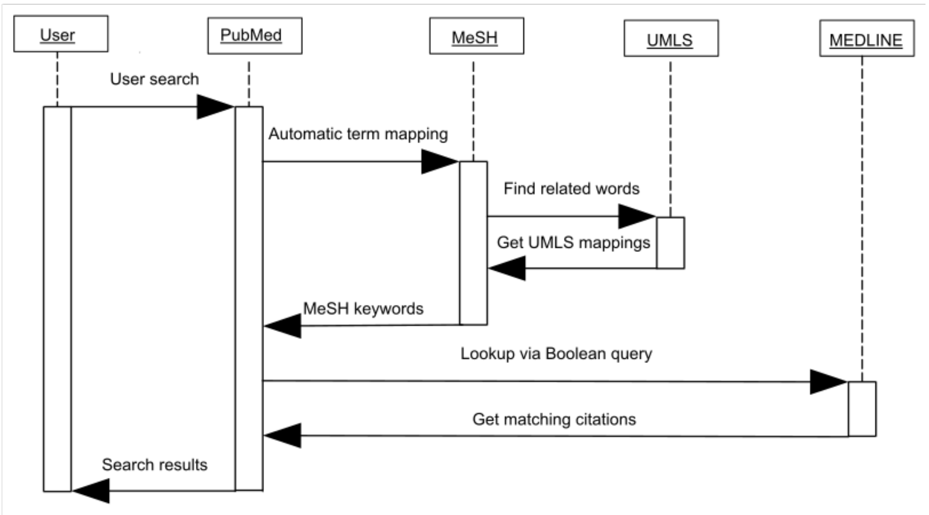


Figure 7.1: Sequence diagram depicting interactions between PubMed, MeSH, UMLS and MEDLINE during search

However, using the PubMed web interface for searching is not ideal when users are in a contained environment, such as a chat room or a forum discussion thread. Users would have to leave their current web page in order to go to the PubMed website and retrieve the information they need to look up. Furthermore, users need to have a sense of the context of the overall conversation and topics being discussed.

The proposed recommender system integrates PubMed citations into a unified user experience so that medical citations relevant to the on-going conversation are conveniently accessible for the users. From a survey of existing literature, the proposed system is the first of its kind in the medical domain. It should be noted that access to reading an article depends on the individual user's subscription to PubMed; the system displays only the citation, and a clickable hyperlink to the article conveniently within a chat environment interface.

7.2 Development Workflow of PubMedReco

To demonstrate the feasibility of PubMedReco, a prototype is developed and populated with conversations taken from a health forum, thereby simulating a chat-like environment where new messages are being added over time. For testing, three aspects of accuracy are considered: selection of relevant keywords, agreement about what constitutes keyword relevance, and correlation between recommended citations and selected keywords. For each new message, medical keywords are extracted. A subset of all the keywords extracted are then used to query PubMed and get related citations. In this way, users can view citations related to the overall conversation without needing to go to the PubMed website, or having to determine which specific keywords to use for querying PubMed.

7.2.1 System Design

As a new message arrives, it is tokenized via regular expressions into individual keywords. Next, stop words are discarded, residual keywords are retained along with relevant details such as message timestamp. For execution time optimization, no stemming or lemmatization was used. In the next step, the incoming keywords from the new message are looked up in an index of all retrieved keywords, initially empty. If the incoming keyword exists, only its related information is updated, including the latest message number containing the keyword. Otherwise, the new keyword is added to the index. The keywords index is then fed to an artificial neural network that selects keywords that are both temporally and contextually relevant.

Artificial neural networks are computational models that are based on the structure of the biological brain, where a large collection of neurons and axons can enforce or inhibit signals, and in turn activate different states [139]. Neural networks have been used for modelling binary classification problems, where elements of a given set need to be categorized into two groups based on specific rules. The task of determining relevant keywords is modelled by inputting various features for each keyword and then outputting 0 if the keyword is irrelevant, or 1 if it is relevant. The keyword features are listed in Table 7.2.1.

Property	**Description**
keyword	Word embeddings representation of keyword
isMed1	Is keyword in the Merriam-Webster's Medical Dictionary [1]
isMed2	Is keyword in SNOMED database [2]
firstPos	Message number/reference where keyword first appears
lastPos	Message number/reference where keyword last appeared
firstTime	Seconds passed since epoch till first occurrence of keyword
lastTime	Seconds passed since epoch till last occurrence of keyword
frequency	Number of times the keyword has appeared in the chat
numMsgs	Number of messages in the entire conversation

Table 7.1: PubMedReco Neural Network Keyword Features

7.2.2 Neural Network Design

The keyword features are converted to binary form for input into the neural network's neurons, as neuron states are more readily represented as 0 or 1. This conversion includes the keywords themselves, which are converted to word embeddings using the Word2Vec deep neural network model trained with skip-grams on the entire Doc2Doc dataset [94]. The word embeddings enable each unique keyword to be assigned a corresponding binary vector in the space. Also, keywords with common contexts, such as synonyms, are positioned close each other in the vector space, and have similar binary vector values.

[2] Merriam-Webster's Medical Dictionary with Audio API available at http://www.dictionaryapi.com

[2] U.S. National Library of Medicine (NLM) SNOMED International edition

7.2.3 Training Dataset

Initially, the neural network needs to learn how to associate these features to the groupings by using a training dataset with keywords already classified as relevant or irrelevant. The neural network model is trained by manual annotation of a subset of The BMJ's Doc2Doc forum discussions dataset[3]. The annotation process involves manually inspecting the keywords for each new incoming message, and marking their relevancy based on the current context. The Doc2Doc forums allow doctors to have online discussions with other doctors on various health-related topics. The temporal nature of forum conversations is ideal for testing PubMedReco, as forum discussions progress over time like online chats. Also, the technical nature of doctors' conversations makes the dataset suitable for querying PubMed. It should be noted that a forum discussion or chat containing n messages yields n training sets because annotations are made for each new incoming message. The trained model can then be used to predict the relevance of new forum discussions or chats that do not have any manual annotations.

7.2.4 Citations Retrieval

Once the relevant keywords are selected, they are then used to query the PubMed database programmatically using Entrez Programming Utilities (E-utilities), a RESTful programming interface [140]. E-utilities accepts natural language queries and converts them into Boolean queries by inserting Boolean operators and using the words as operands. Stemming and lemmatization is also performed on the keywords by the E-utilities API which can infer synonyms and other relationships to the query words via UMLS. E-utilities has options for specifying what citation fields to search within, such as title, abstract, full text (where available), author and others. The proposed system restricts search to the citation title because the recommendations are ultimately presented as full citation titles. Consequently, the user would decide initial interest or disinterest in the recommendation from the article's title.

[3] This dataset is no longer available, but is still accessible via the Internet Archive's Wayback Machine at http://web.archive.org/web/20160615110024/doc2doc.bmj.com

The citations returned can also be sorted using various options available in E-utilities, and sorted by relevancy, which takes into account the frequency of matched keywords within the title. Hence, citations containing more of the searched keywords would be ranked higher.

7.2.5 Testing Criteria

Three aspects of PubMedReco need to be tested for accuracy: neural network, annotations, and recommended citations. The neural network needs to be appraised for accuracy, in order to determine how it would perform when given datasets that have no annotations. The evaluation is done via the standard precision metric. A sampling out of the total number of training sets generated is selected, and the precision measured. This process is iteratively done in order to see which random sampling provides the best precision value. The annotations are related to the performance of the neural network. For evaluating the quality of annotations about the relevance of keywords, the Kappa score is used to compare multiple annotations made on the same dataset. As baseline, arbitrary annotations are used so that there is no planned correlation between a keyword and its relevance. Finally, for evaluating the quality of the recommended citations, NDCG [96] is used to quantify if the recommended citations indeed contain the keywords that were used to query PubMed. NDCG measures both the number of matching keywords in a search hit, as well as the usefulness of the hit based on its position in the results. As baseline, the "Title" is used as the sorting option within E-utilities, so that the top-n citations are sorted alphabetically. This sorting option will not rank citations based on number of keyword hits, but rather based on the title's alphabetic ordering.

7.3 Performance Analysis of PubMedReco

Firstly, the results of evaluating the neural network are presented. The Doc2Doc dataset contained a total of 1,400 discussions. From these, 10 were used as the training dataset with varying number of message threads, averaging 9.8 threads per discussion. Hence, a total of 98 training points were generated from the discussions.

The neural network was trained by iteratively and randomly selecting 50 samples out of the 98 training datasets and choosing the model with the highest accuracy. Figure 7.2 shows the precision for 20 iterations of the sampling. The average precision was 56%, while the highest precision was 61%.

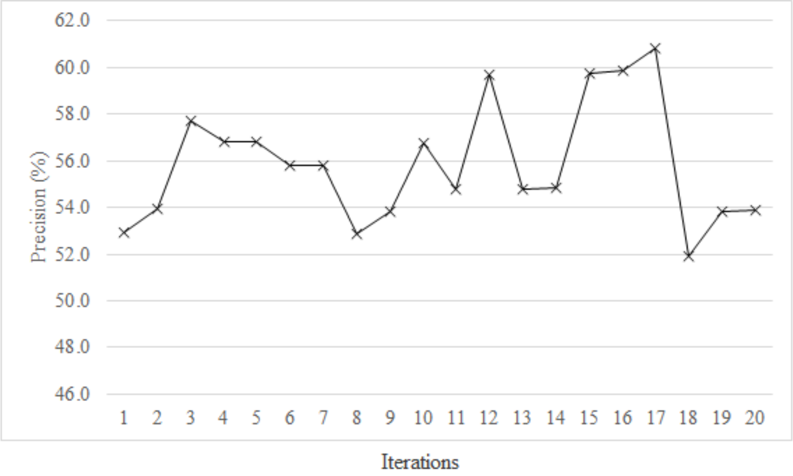


Figure 7.2: Precision of PubMedReco Neural Network

Secondly, testing of the manual annotations is presented. Ten discussions with annotations were selected and re-annotated without cross-referencing the previous annotations. An average Kappa score of 52% was achieved, showing borderline acceptable agreement with the annotation method. Figure 7.3 shows the Kappa scores for the 10 annotated discussions (M), their counter-part annotations (N), and baseline arbitrary annotations (Base).

Thirdly, statistics are presented on the quality of the citations using NDCG. Five discussions were arbitrarily selected and NDCG computed for each query to E-utilities, resulting in 32 data points, and an average number of 6.4 messages in the selected discussions. Figure 7.4 shows the NDCG values with E-utilities optimal sorting (Relevancy), which averaged 80%, and also the baseline using E-utilities alphabetic ordering of the citation (Title). Similar results were also achieved for other iterations of arbitrarily selecting 5 discussions.

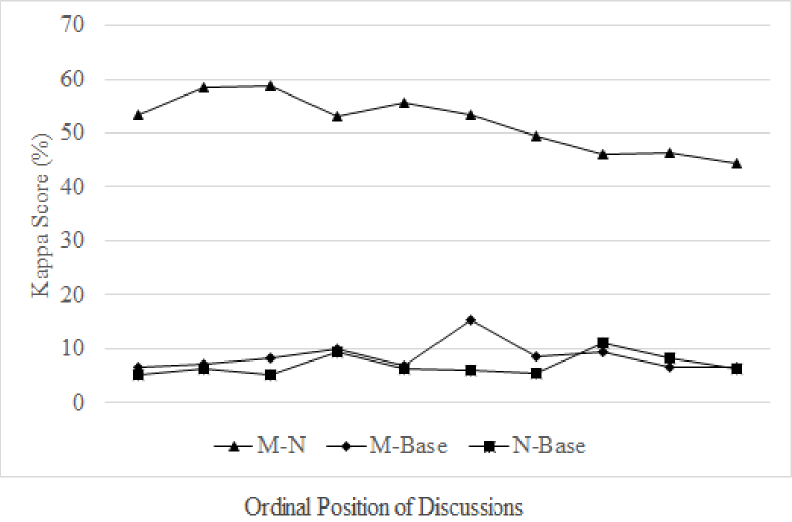


Figure 7.3: Kappa Score for Annotations

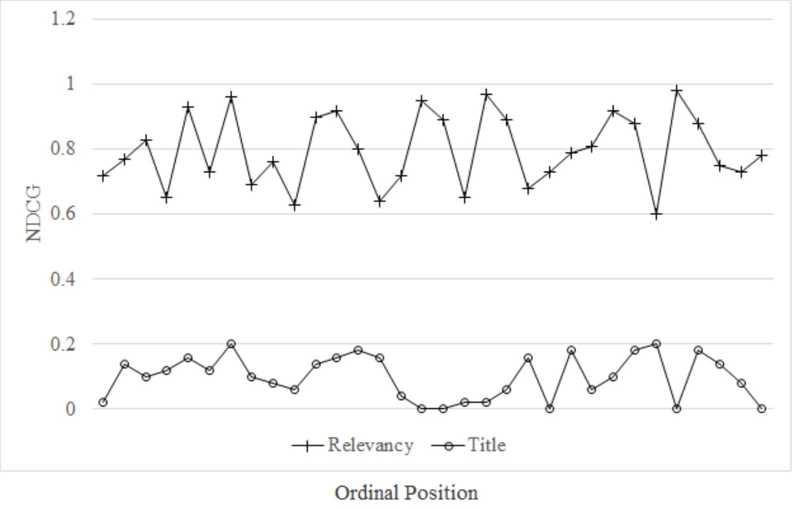


Figure 7.4: NDCG for Recommended Citations

7.4 Discussion and Appraisal of Results

The results show that the proposed approach was able to retrieve citations based on forum discussion, while taking into account the relevance of keywords. Other methods for retrieving keywords in chat rooms rely on static heuristics, such as a fixed time window [141]. As an example, with static heuristics only the last 5 messages might be used to determine the keywords for retrieving recommendations.

In contrast, the proposed method moves away from static heuristics and applies machine learning for dynamic retrieval of relevant keywords. Figure 7.5 shows how the neural network's keywords selection relates to a dynamic window metric for 3 randomly selected Doc2Doc forum discussions over 10 incoming messages. The window is set to the number of messages from the latest to the oldest keyword selected by the neural network.

Moreover, text summarization methods such as TextRank [90], and ensemble keyword extraction systems such as AlchemyAPI[4] are not adequate for extracting keywords to summarize a forum or chat room discussion because they do not take into consideration the decay in relevance of the keywords over time.

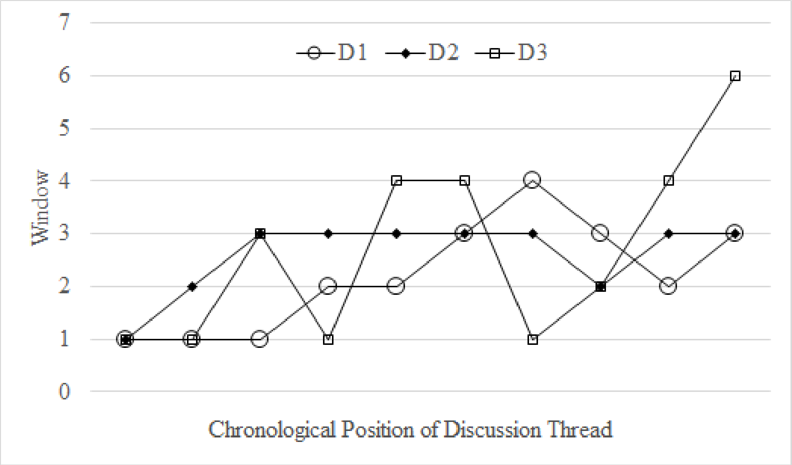


Figure 7.5: Dynamic Window for Discussions D1, D2, D3

A limitation of this study is reliance on forum discussions to simulate chats, instead of actual real-time synchoronous conversations. This impacts the trained neural network because feature properties such as the `firstTime` and `lastTime` features from Table 7.2.1 will have relatively much smaller values for real-time chats. Another area of improvement is increasing the training dataset size, which could improve precision. Having multiple annotations would strengthen the neural network and provide insights on user preferences for dynamic time window values.

[4] AlchemyAPI available at http://www.alchemyapi.com

This chapter presented the PubMedReco recommender system which can analyze a forum or chat discussion to extract the relevant medical terms, and then query PubMed to suggest citations that are related to the ongoing discussion. PubMedReco overcomes the limitation imposed on users of online discussion environments whereby users would have to leave their chat interface to search for medical articles in a new web browser window, and manually formulate an appropriate search query from arbitrary keywords to get results from general search engines. The feasibility of PubMedReco was demonstrated using The BMJ's Doc2Doc forum datasets. The system was also tested to determine the quality of its neural network's relevance predictions, the training annotations used, and the recommended citations. To the best of our knowledge, the proposed system is the first of its kind in the medical domain. Unlike other real-time chat recommender systems surveyed that use static time window heuristics, the proposed system presents a novel and dynamic machine learning approach for determining keyword relevance within health forums and chat rooms.

Takeaway

- PubMedReco is a recommender system for suggesting PubMed citations based on discussion threads
- In real-time chats, the relevant context of the discussion frequently changes
- PubMedReco selects keywords that are temporally and contextually relevant
- PubMedReco leverages the Entrez Programming Utilities (E-Utilities) API
- PubMedReco is utilized within Cardea to provide recommendations from credible sources within Patient to Medic chat rooms

Chapter 8

DeepDup and Cross-Domain Duplicate Question Detection for Heterogeneous Community Question Answering

This chapter covers details about DeepDup, a duplicate question detection system using deep learning developed as part of this thesis. DeepDup won a bronze medal at the Kaggle Quora Question Pairs competition[1], and is integrated into Cardea for detecting duplicate questions, discussions, and answers.

8.1 Motivation and Challenges for Duplicate Question Detection

To efficiently exploit Community Question Answering (CQA) forums, users need to know if their question has already been asked, to avoid re-posting a duplicate question. The identification of duplicate questions in CQA forums can provide at least three main advantages. Firstly, finding duplicate questions saves users' time because they do not have to wait for responses. Secondly, users searching for questions will be presented with better results with duplicates pruned. Thirdly, the overall retrievability of information for the CQA forum will be enhanced by reducing duplication.

[1] Kaggle Quora Question Pairs Prediction Competition https://www.kaggle.com/c/quora-question-pairs

Identifying two questions as duplicates can be challenging because the choice of words, structure of sentences, and even context, can vary significantly between questions, even if the intended semantics are near identical. In addition, questions with similar verbiage are not necessarily duplicates. Traditional IR and NLP methods have achieved limited success in detecting semantically identical text-pairs. Popular CQA forums like Quora and Stack Exchange (SE) have many new questions posted daily, some of which have been previously asked but have variations in wordings, synonyms, phrases, or sentence structure.

When comparing state-of-the-art machine learning methods for this task, an interesting observation is that a classification model trained on a dataset from one domain cannot achieve the same performance to predict text-pair duplicates in another domain. For instance, the similarity between two question pairs can be completely different depending on the domain of the dataset which was used to train the classification model. The question pair "Where can I find a place to eat pizza?" and "What's the closest Italian restaurant?" can be classified as duplicate or not, depending on the domain of the dataset used in model training. With Quora, the similarity was 6%, while with Stack Exchange, it was 46% [142].

The results of an empirical analysis of popular state-of-the-art machine learning methods for text-pair duplicate classification are presented: deep neural network and gradient tree boosting. Also, the possibility of domain adaptation is explored to increase the performance of under performing domains using transfer learning.

Three research questions are targeted. Firstly, the best approaches for text-pair duplicate detection is investigated. Secondly, the possibility of a general-purpose cross-domain duplicate detection approach for heterogeneous datasets is explored. Ultimately, this research aims to determine whether dataset domain affects the outcomes of the trained model, and consequently evaluate the null hypothesis that the meaning of a "duplicate" is universal.

8.2 DeepDup Architecture and Design

Three publicly available datasets were used from Quora[2], and Stack Exchange's Ask Ubuntu and English forums[3]. The datasets contain forum moderator annotated labels for duplicate and non-duplicate question pairs. The study used only the question's title; the question's full body, tags and other meta data were not used for the SE datasets in order to be fair for the Quora dataset which only had succinct questions. Properties of the datasets are summarized in Table 8.1, including total number of question pairs and Words Per Question (WPQ).

Property	Quora	AskUbuntu	English SE
Question Pairs	404,303	131,271	33,661
Max WPQ	237	33	32
Mean WPQ	11.0	8.7	8.9

Table 8.1: Dataset Properties

8.2.1 Data Preprocessing

Each question was tokenized, and question pairs whose data types do not match were filtered. Non-English questions were removed by checking for non-English vowels. Stop words removal, lemmatization, and stemming were also performed. Finally, abbreviated forms such as "what's" and "i'm" were transformed to their unabbreviated forms of "what is" and "i am" respectively.

8.2.2 Deep Neural Network Models

Underlying semantic similarity between questions can be learned with a better numerical representation of the texts, such as the ones learned through deep neural network models. The datasets used have sufficient attributes to be used with a variety of deep neural network models. Siamese Neural Networks (SNNs) have been popularly used to compare two objects and find similarity relationships between them [143].

[2] Quora dataset https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

[3] Stack Exchange dataset https://archive.org/details/stackexchange

A salient feature of these Siamese networks is that they employ two sub-networks, which share parameters, thus reducing the number of parameters to learn, and give a consistent representation for the two objects being compared. A similar architecture is adadpted in this research to compare question pairs, and to determine whether they are duplicates. In Figure 8.1 illustrates an abstracted view of this architecture, which features three major modules: i) Representation module (R), ii) Aggregation module (A), and iii) Decision module (D).

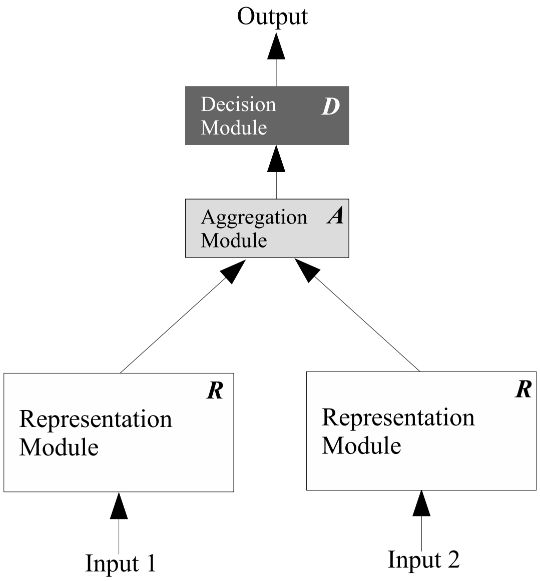


Figure 8.1: Siamese Neural Network for Duplicate Question Detection

The representation module learns the representation of a question. This typically consists of an Embedding layer (E), and either Long Short Term Memory (LSTM) layers or Convolutional Neural Networks (CNNs), and, optionally a few fully connected layers to flatten and summarize the output as a concise vector representation. The embedding layer converts the question words to vectors in the embedding space; GloVe [144] pre-trained word embeddings were used for this.

The aggregation module takes the representations of the question pairs and performs an aggregation operation to prepare them for the decision module. $e^{-|Q_1-Q_2|}$ (negative absolute exponential distance) and the simple vector concatenation are two such successful aggregation methods experimented with.

The output of the aggregation module is fed to the decision module, which consists of one or more fully connected layers and a decision node with `sigmoid` activation at the end. `nadam` was used as the optimization function and binary cross entropy as the loss function. The datasets were split into 60% for training, 20% for validation, and 20% for testing. Training and validation subsets were used to tune for the optimal hyper-parameter combinations, which included the optimal number of iterations to train, types and number of layers in the representation module, type of aggregation, number of nodes in each layer, and the best input length.

8.2.3 Gradient Tree Boosting Models

Gradient Tree Boosting (GTB) is a popular machine learning method which uses an ensemble of weak prediction models (typically decision trees) to build a strong predictor. It has shown strong results in various real-world applications [145]. Efficient gradient boosted tree implementations such as XGBoost have demonstrated very good performance in large datasets. Given the robustness of it, a GTB-based binary classifier was used to address the duplicate question detection problem.

As input features to the above GTB model, more than 40 hand-crafted features were used, reflecting the semantic and structural similarities between two questions. These features included many traditional and non-traditional distance metrics such as TF-IDF distance, word movers distance, graph based structural question similarity distances, Word2Vec-based distances [94], and Doc2Vec-based distances [146].

8.2.4 Transferability of Neural Networks

Transfer Learning (TL) aims to utilize the knowledge learned from a better performing source domain to increase the performance of an under-performing target domain with insufficient or sparsely labeled examples [147]. Prior work in deep neural network-based computer vision models indicates that transfer learning can be successfully utilized [148]. Recently, NLP applications have used similar ideas to improve performance in certain target domains [147, 149].

This work explores the possibility of transferring and utilizing knowledge learned from large datasets such as Quora to improve the performance in other target domains such as Ask Ubuntu or English. The intent is that this will lead to generally improved duplicate question detection across domains. Specifically, the INIT TL approach is adopted [149], which uses parameters trained on a source domain to initialize parameters of the target domain's model.

With INIT [149], a neural network model was first trained on the source dataset and experimented with three initialization states, I_i, on the target model: i) initialize the target parameters using source parameters but freeze further training, denoted I_1, ii) initialize the target parameters using source parameters and fine tune it further on target dataset, denoted I_2, and iii) random initialization, denoted I_3. Combinations of these initialization states were experimented with on each module of the SNN (Representation [without the Embedding layer], Aggregation, and Decision) and the Embedding layer, and reported the best results obtained. Some example configurations are [$E(I_2)$, $R(I_3)$, $A(I_3)$, $D(I_3)$], [$E(I_2)$, $R(I_2)$, $A(I_3)$, $D(I_3)$], and [$E(I_2)$, $R(I_2)$, $A(I_2)$, $D(I_2)$].

8.3 Performance of Duplicate Detection Methods

For performance evaluation, the Area Under the Curve (AUC) metric is used. The results are presented in Figure 8.2, including an additional naïve approach, in which a model is trained by combining training data from all three of the datasets. The trained model is then used to make predictions on the individual hold out test sets. The model achieved state-of-the-art performance with Quora dataset using XGBoost, with 94% AUC. This XGBoost approach was also best for the AskUbuntu dataset with 66% AUC. For TL, Quora was selected as the source domain with neural networks as the preferred model, and models based on AskUbuntu and English SE as the targets. The TL approach gave the best performance for the English SE dataset at 58% AUC, but only slightly better than XGBoost at 56%. On the other hand Ask Ubuntu did not gain any improvement through transfer learning indicating the context specific difference in the duplicate question detection task.

For all the approaches, there were significant differences between performance across the datasets. While the approach performed considerably well on the Quora dataset, the AskUbuntu and English SE datasets did not give comparatively good results even with the TL approach. This indicates that, across domains, the knowledge which can be positively transferred is low and the meaning of duplicates is vastly different.

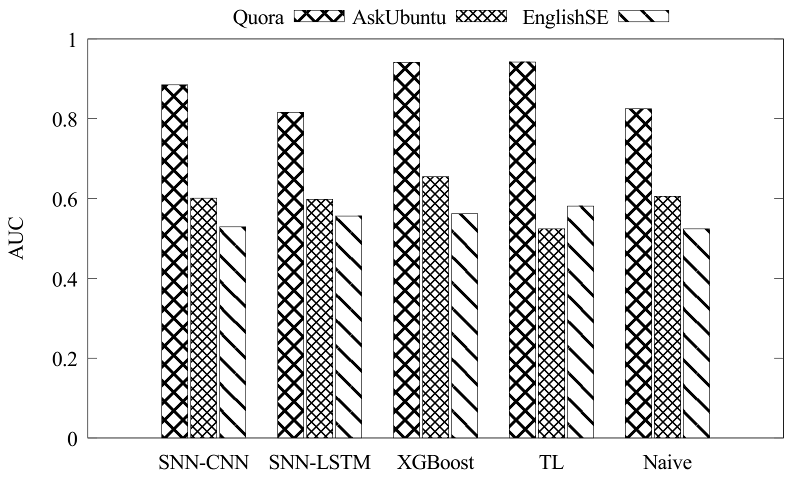


Figure 8.2: Results from Various Machine Learning Approaches on Heterogeneous Datasets

8.4 Review and Interpretation of Results

Hence, the results provide some support for the alternative hypothesis that semantic representation of duplicates is significantly affected by specific domains. It is postulated that the nature of the domain's language makes it difficult to predict duplicate text-pairs across domains. For example, Quora has simplified layperson English words, AskUbuntu has technical jargon and acronyms, while English SE has academic phrases.

Duplicate question detection is an ongoing challenge in CQA because semantically equivalent questions can have significantly different words and structures. In addition, the identification of duplicate questions can reduce the resources required for retrieval, when the same questions are not repeated. This chapter compared the performance of deep neural network and gradient tree boosting, and explored the possibility of domain adaptation with transfer learning to improve the under-performing target domains for the text-pair duplicates classification task, using three heterogeneous datasets: general-purpose Quora, technical Ask Ubuntu, and academic English Stack Exchange. Ultimately, this study shed more light on the alternative hypothesis that the meaning of a "duplicate" is not inherently general-purpose, but rather is dependent on the domain of learning, hence reducing the chance of transfer learning through adapting to the domain.

Takeaway

- Duplicate question detection improves the quality of social media content by increasing findability and reducing redundancy
- DeepDup is a Siamese deep neural network for detecting duplicate questions
- DeepDup utilizes three main modules: representation, aggregation, decision
- Semantic representation of duplicates can vary significantly across domains
- DeepDup leverages domain characteristics of text to optimize results

Chapter 9

Conclusion

In this thesis, an important area of research is covered that has gotten even more relevance during the global COVID-19 pandemic. Health Social Media has been use variously by laypersons for sharing opinions, discussing about health topics, self-educating, self-diagnosis, and self-treatment. However, it has become increasingly obvious that there are severe consequences when health misinformation gets disseminated widely. Patients and laypersons aim to seek credible advise on various issues including sensitive topics where they wish to do so anonymously to avoid social stigma. This thesis proposed solutions to these research challenges via objective metrics for measuring the veracity of health information, trust-preserving anonymization, and a platform for enabling online discourse between patients and medics in a privacy-aware, trust-centric, and security-focused health portal.

Firstly, this thesis contributed to research on veracity of health content by developing and evaluating pragmatic computational models to measure the veracity of health-related online content. Medical knowledge and evidence-based practices were incorporated through the MedFact algorithm. The effectiveness of MedFact was evaluated against surveys by both laypersons and medical professionals, as well as datasets containing true and false claims.

Secondly, this thesis contributed to research on anonymity, trust, privacy, and stigmatization by development of the Iron Mask algorithm. Evaluation by survey demonstrated the usefulness and effectiveness of Iron Mask and Trust Preserving Pseudonym, especially in avoidance of whiteprint identification and de-anonymization by using structural and content-based features of social networks.

Thirdly, this thesis addressed the need for medical professionals to join the online health discourse, and incorporating MedFact and Iron Mask into the Cardea health portal for patients and medics. Cardea models various types of interactions occurring on Health Social Media, including real-time chat rooms, blogs, question-answering, and support groups. Cardea also enables similar content recommendations for preventing duplication of questions and discussions via the DeepDup algorithm, as well as suggesting new content from reputable sources such as PubMed and Health Canada via the PubMedReco algorithm. Moreover, the search engine and recommender systems within Cardea incorporate standard metrics from information retrieval literature to rank results while also including trust metrics developed as part of this thesis to promote credible content.

For future work, the thesis opens new research questions, three of which are highlighted here. Firstly, ‘*How can the cold start problem be addressed with objective trust metrics?*’. This challenge arises for ground-truth based methods when the knowledge base itself does not yet have established facts from domain experts regarding a topic, such as COVID-19 in the medical domain. Secondly, ‘*How can the psychological effects of misinformation be addressed by computational methods?*’. Social media users gravitate towards content with emotional storytelling more than factual information. Moreover, their personal estimates about their information-seeking skill sets are likely exaggerated. Not all users may be willing to accept veracity estimations that oppose their own personal opinions and beliefs, leading to the Dunning-Kruger effect. The research challenge is on informing the user about content veracity without shunning them away. Thirdly, ‘*How can medical professionals and patients engage more effectively in online discourse?*’. This thesis has proposed the Cardea health portal for bringing medics and patients together. Additionally, the need for medical professionals to actively engage with patients online has been highlighted due to patients’ needs for self-education. While telemedicine solutions such as phone calls and video chats have been available for a while now, the appeal and effectiveness of online chat and text messaging is noticeable in other domains, but remains to be fully evaluated in online health.

Ultimately, this thesis provides new perspectives on a computational definition of trust with an awareness of privacy in Health Social Media. The contributions of this thesis have the potential for assisting users to sift through large volumes of online information and make informed decisions about their health using trustworthy information sources without compromising privacy.

Bibliography

[1] Kelley Ann Strout and Elizabeth P. Howard. The Six Dimensions of Wellness and Cognition in Aging Adults. *Journal of Holistic Nursing*, 30(3):195–204, 2012.

[2] Helga Rippen and Ahmad Risk. e-Health Code of Ethics (May 24). *Journal of Medical Internet Research*, 2(2), May 2000.

[3] Michele P. Hamm, Annabritt Chisholm, Jocelyn Shulhan, Andrea Milne, Shannon D. Scott, Lisa M. Given, and Lisa Hartling. Social Media Use Among Patients and Caregivers: A Scoping Review. *BMJ Open*, 3(5), 2013.

[4] Michele P. Hamm, Jocelyn Shulhan, Gillian Williams, Andrea Milne, Shannon D. Scott, and Lisa Hartling. A Systematic Review of the Use and Effectiveness of Social Media in Child Health. *BMC Pediatrics*, 14(1), 2014.

[5] Melanie Michael, Susan D. Schaffer, Patricia L. Egan, Barbara B. Little, and Patrick Scott Pritchard. Improving Wait Times and Patient Satisfaction in Primary Care. *Journal for Healthcare Quality*, 35(2):50–60, 2013.

[6] Khalid Syed, David Parente, Samuel Johnson, and Joel Davies. Factors Determining Wait-Time and Patient Satisfaction at Post-Operative Orthopaedic Follow-Up. *Open Journal of Medical Psychology*, 2:47–53, 2013.

[7] Anna Kata. Anti-Vaccine Activists, Web 2.0, and the Postmodern Paradigm–An Overview of Tactics and Tropes Used Online by the Anti-Vaccination Movement. *Vaccine*, 30(25):3778–3789, 2012.

[8] Cristina M Pulido, Beatriz Villarejo-Carballido, Gisela Redondo-Sama, and Aitor Gómez. COVID-19 Infodemic: More Retweets for Science-based Information on Coronavirus than for False Information. *International Sociology*, 2020.

[9] Jinling Hua and Rajib Shaw. Corona Virus (COVID-19) “Infodemic” and Emerging Issues through a Data Lens: The Case of China. *International Journal of Environmental Research and Public Health*, 17(7), 2020.

[10] CTV News. Indian Government Slammed for Recommending Homeopathy for Coronavirus Prevention, 2020.

[11] NBC News. China is Encouraging Herbal Remedies to Treat COVID-19. But Scientists Warn Against It, 2020.

[12] Bernadette Melnyk and Ellen Fineout-Overholt. *Evidence-Based Practice in Nursing & Healthcare: A Guide to Best Practice*. In Focus. Wolters Kluwer/Lippincott Williams & Wilkins, 2011.

[13] Bonnie Spring. Evidence-Based Practice in Clinical Psychology: What It Is, Why It Matters; What You Need To Know. *Journal of Clinical Psychology*, 63(7):611–631, 2007.

[14] Marita G. Titler. The Evidence for Evidence-Based Practice Implementation. In *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*. Agency for Healthcare Research and Quality, 2008.

[15] Tami Oliphant. "I Am Making My Decision On The Basis Of My Experience": Constructing Authoritative Knowledge about Treatments for Depression. *Canadian Journal of Information and Library Science*, 33(3-4):215–232, 2009.

[16] Stephen Paul Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling, Scotland, 1994.

[17] John Child. Trust - The Fundamental Bond in Global Collaboration. *Organizational Dynamics*, 29(4):274–288, 2001.

[18] Jennifer Ann Golbeck. *Computing and Applying Trust in Web-based Social Networks*. PhD thesis, University of Maryland at College Park, 2005.

[19] James G Anderson and Kenneth Goodman. *Ethics and Information Technology: A Case-Based Approach to a Health Care System in Transition*. Springer Science & Business Media, 2002.

[20] Daniel W. Manchala. Trust Metrics, Models and Protocols for Electronic Commerce Transactions. In *Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS)*, pages 312–321. IEEE, 1998.

[21] George Theodorakopoulos and John S. Baras. On Trust Models and Trust Evaluation Metrics for Ad Hoc Networks. *IEEE Journal on Selected Areas in Communications*, 24(2):318–328, 2006.

[22] John Erickson. Trust Metrics. In *Proceedings of the International Symposium on Collaborative Technologies and Systems*, volume 0, pages 93–97. IEEE Computer Society, 2009.

[23] Ioan Alfred Letia and Radu Razvan Slavescu. FloodTrust for Improved Trust Transitivity. In *Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 276–279. IEEE Computer Society, 2011.

[24] Donovan Artz and Yolanda Gil. A Survey of Trust in Computer Science and the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):58–71, 2007.

[25] Mirjam Šitum. Analysis of Algorithms for Determining Trust Among Friends on Social Networks. Master's thesis, University of Zagreb, Faculty of Electrical Engineering and Computing, 2014.

[26] Wanita Sherchan, Surya Nepal, and Cecile Paris. A Survey of Trust in Social Networks. *ACM Computing Surveys (CSUR)*, 45(4), 2013.

[27] Audun Jøsang, Roslan Ismail, and Colin Boyd. A Survey of Trust and Reputation Systems for Online Service Provision. *Decision Support Systems*, 43(2):618–644, 2007.

[28] Alexander Halavais, K. Hazel Kwon, Shannon Havener, and Jason Striker. Badges of Friendship: Social Influence and Badge Acquisition on Stack Overflow. In *Proceedings of the Hawaii International Conference on System Sciences*, pages 1607–1615. IEEE, 2014.

[29] Scott Grant and Buddy Betts. Encouraging User Behaviour with Achievements: An Empirical Study. *IEEE International Working Conference on Mining Software Repositories (MSR)*, pages 65–68, 2013.

[30] Zainab Mohammed Aljazzaf. *Trust-Based Service Selection*. PhD thesis, University of Western Ontario, 2011.

[31] Helena Leino-Kilpi, Maritta Välimäki, Theo Dassen, María Gasull, Chryssoula Lemonidou, Anne Scott, and Marianne Arndt. Privacy: A Review of the Literature. *International Journal of Nursing Studies*, 38(6):663–671, 2001.

[32] Hamman W. Samuel and Osmar R. Zaïane. A Repository of Codes of Ethics and Technical Standards in Health Informatics. *Online Journal of Public Health Informatics*, 6(2):1–17, 2014.

[33] Derek L. Hansen and Christianne Johnson. Veiled Viral Marketing: Disseminating Information on Stigmatized Illnesses via Social Networking Sites. In *Proceedings of the ACM SIGHIT International Health Informatics*, pages 247–254. ACM, 2012.

[34] Daniel Cvrček and Vaclav Matyáš Jr. Pseudonymity in the Light of Evidence-Based Trust. *Security Protocols*, pages 267–274, 2006.

[35] Sara Keretna, Ahmad Hossny, and Doug Creighton. Recognising User Identity in Twitter Social Networks via Text Mining. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 3079–3082. IEEE, 2013.

[36] Cynthia Dwork. Differential Privacy: A Survey of Results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.

[37] Karen Mercedes Goertzel and Theodore Winograd. Enhancing the Development Life Cycle to Produce Secure Software. *U.S. Department of Defense*, 2008.

[38] Mohomed Shazan Mohomed Jabbar, Luke Kumar, Hamman Samuel, Mi-Young Kim, Sankalp Prabhakar, Randy Goebel, and Osmar Zaïane. On Generality and Knowledge Transferability in Cross-Domain Duplicate Question Detection for Heterogeneous Community Question Answering. *arXiv*, pages arXiv–1811, 2018.

[39] Hamman W Samuel, Osmar R Zaïane, and Jane Robertson Zaïane. Findability in Health Information Websites. In *Proceedings of IEEE-EMBS International Conference on Biomedical and Health Informatics*, pages 709–712. IEEE, 2012.

[40] Iraklis Varlamis, Magdalini Eirinaki, and Malamati Louta. A Study on Social Network Metrics and their Application in Trust Networks. In *Proceedings of the IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 168–175, 2010.

[41] Sonja Grabner-Kräuter and Sofie Bitter. Trust in Online Social Networks: A Multifaceted Perspective. *Forum for Social Economics*, 44(1):48–68, 2015.

[42] Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. Collaborative Content-Based Method for Estimating User Reputation in Online Forums. In *Web Information Systems Engineering*, pages 292–299. Springer, 2015.

[43] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth Finding on the Deep Web: Is the Problem Solved? *arXiv preprint arXiv:1503.00303*, 2015.

[44] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab, 1999.

[45] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A Survey on Truth Discovery. *ACM Sigkdd Explorations Newsletter*, 17(2):1–16, 2016.

[46] Jeff Pasternack and Dan Roth. Knowing What To Believe (When You Already Know Something). In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 877–885, 2010.

[47] Haochen Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. A Tutorial on Network Embeddings. *arXiv preprint arXiv:1808.02590*, 2018.

[48] Julio CS Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. Supervised Learning for Fake News Detection. *IEEE Intelligent Systems*, 34(2):76–81, 2019.

[49] Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. Fake News Detection Using Machine Learning Ensemble Methods. *Complexity*, 2020.

[50] Meeyoung Park. *HealthTrust: Assessing the Trustworthiness of Healthcare Information on the Internet*. PhD thesis, University of Kansas, 2013.

[51] Paolo Ferragina and Ugo Scaiella. TAGME: On-The-Fly Annotation of Short Text Fragments (By Wikipedia Entities). In *ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1625–1628, 2010.

[52] Gautam Kishore Shahi and Durgesh Nandini. FakeCovid–A Multilingual Cross-domain Fact Check News Dataset for COVID-19. *arXiv preprint arXiv:2006.11343*, 2020.

[53] Debanjana Kar, Mohit Bhardwaj, Suranjana Samanta, and Amar Prakash Azad. No Rumours Please! A Multi-Indic-Lingual Approach for COVID Fake-Tweet Detection. *arXiv preprint arXiv:2010.06906*, 2020.

[54] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, K. Funk, Rodney Michael Kinney, Ziyang Liu, W. Merrill, P. Mooney, D. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, B. Stilson, A. Wade, K. Wang, Christopher Wilhelm, Boya Xie, D. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The Covid-19 Open Research Dataset. *arXiv 2004.05125*, 2020.

[55] Arkin Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. Drink Bleach or Do What Now? Covid-HeRA: A Dataset for Risk-Informed Health Decision Making in the Presence of COVID19 Misinformation. *arXiv preprint arXiv:2010.08743*, 2020.

[56] Yin Aphinyanaphongs, Constantin Aliferis, et al. Text Categorization Models for Identifying Unproven Cancer Treatments on the Web. In *World Congress on Medical Informatics (MedInfo)*, page 968. IOS Press, 2007.

[57] David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. Evidence Based Medicine: What It Is and What It Isn't, 1996.

[58] Trisha Greenhalgh. *How to Read a Paper: The Basics of Evidence-Based Medicine*. John Wiley & Sons, 2010.

[59] Sharon E. Straus, W. Scott Richardson, Paul Glasziou, and R. Brian Haynes. *Evidence-Based Medicine: How to Practice and Teach EBM*. Elsevier/Churchill Livingstone, 2005.

[60] Betty J Ackley. *Evidence-Based Nursing Care Guidelines: Medical-Surgical Interventions*. Elsevier Health Sciences, 2008.

[61] Katy L James, Nicola P Randall, and Neal R Haddaway. A Methodology for Systematic Mapping in Environmental Sciences. *Environmental Evidence*, 5(1):7, 2016.

[62] Caio V Meneses Silva, Raphael Silva Fontes, and Methanias Colaço Júnior. Intelligent Fake News Detection: A Systematic Mapping. *Journal of Applied Security Research*, pages 1–22, 2020.

[63] Greg J Stephens, Lauren J Silbert, and Uri Hasson. Speaker–Listener Neural Coupling Underlies Successful Communication. *Proceedings of the National Academy of Sciences*, 107(32):14425–14430, 2010.

[64] Brendan Nyhan, Jason Reifler, Sean Richey, and Gary L. Freed. Effective Messages in Vaccine Promotion: A Randomized Trial. *Pediatrics*, 2014.

[65] Brendan Nyhan and Jason Reifler. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior*, 32(2):303–330, 2010.

[66] Scott Plous. *The Psychology of Judgment and Decision Making*. McGraw-Hill, 1993.

[67] David Dunning. The Dunning-Kruger Effect: On Being Ignorant of One's Own Ignorance. *Advances in Experimental Social Psychology*, 44:247, 2011.

[68] Robert Proctor and Londa L Schiebinger. *Agnotology: The Making and Unmaking of Ignorance*. Stanford University Press, 2008.

[69] Julie Henderson. Expert and Lay Knowledge: A Sociological Perspective. *Nutrition & Dietetics*, 67(1):4–5, 2010.

[70] Aaron M Johnson, Paul Syverson, Roger Dingledine, and Nick Mathewson. Trust-Based Anonymous Communication: Adversary Models and Routing Algorithms. In *ACM Conference on Computer and Communications Security*, pages 175–186. ACM, 2011.

[71] Vladimiro Sassone, Sardaouna Hamadou, and Mu Yang. Trust in Anonymity Networks. In *International Conference on Concurrency Theory*, pages 48–70. Springer, 2010.

[72] Felix Konstantin Maurer. A Survey on Approaches to Anonymity in Bitcoin and other Cryptocurrencies. *Informatik*, 2016.

[73] Michael Scott Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. 4chan and //b//: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *International AAAI Conference on Weblogs and Social Media*. AAAI, 2011.

[74] Ilya Lebedev and Milhail Sukhoparov. Methodologies of Internet Portals Users' Short Messages Texts Authorship Identification based on the Methods of Mathematical Linguistics. *In the World of Scientific Discoveries*, 54(6):599–622, 2014.

[75] John R Douceur. The Sybil Attack. In *International Workshop on Peer-to-Peer Systems*, pages 251–260. Springer, 2002.

[76] Arvind Narayanan and Vitaly Shmatikov. De-Anonymizing Social Networks. In *IEEE Symposium on Security and Privacy*, pages 173–187. IEEE, 2009.

[77] Aaron Beach, Mike Gartrell, and Richard Han. q-Anon: Rethinking Anonymity for Social Networks. In *IEEE International Conference on Social Computing*, pages 185–192. IEEE, 2010.

[78] Amira Ghenai. *Health Misinformation in Search and Social Media*. PhD thesis, University of Waterloo, 2017.

[79] Peter Morville. *Ambient Findability*. O'Reilly Media, 2005.

[80] Marjut Lievonen. *Understanding Google Algorithms and SEO is Essential for Online Marketer*. PhD thesis, Tampere University of Applied Sciences, 2013.

[81] Ramanathan Guha. Search Result Ranking Based on Trust, 2009. US Patent 7,603,350.

[82] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor. *Recommender Systems Handbook*. Springer-Verlag, New York, NY, USA, 2011.

[83] Denis Parra and Shaghayegh Sahebi. Recommender Systems: Sources of Knowledge and Evaluation Metrics. In *Advanced Techniques in Web Intelligence*, pages 149–175. Springer, 2013.

[84] Paolo Massa and Paolo Avesani. Trust Metrics in Recommender Systems. In *Computing with Social Trust*, pages 259–285. Springer, 2009.

[85] Patricia Victor, Martine De Cock, and Chris Cornelis. Trust and Recommendations. In *Recommender Systems Handbook*, pages 645–675. Springer, 2011.

[86] Sourav Bhattacharya, Otto Huhta, and N Asokan. LookAhead: Augmenting Crowdsourced Website Reputation Systems with Predictive Modeling. In *International Conference on Trust and Trustworthy Computing*, pages 143–162. Springer, 2015.

[87] Hamman Samuel and Osmar Zaiane. MedFact: Towards Improving Veracity of Medical Information in Social Media using Applied Machine Learning. In *Canadian Conference on Artificial Intelligence*, pages 108–120. Springer, 2018.

[88] Joel Nothman, Hanmin Qin, and Roman Yurchak. Stop Word Lists in Free Open-Source Software Packages. In *Workshop for NLP Open Source Software (NLP-OSS)*, pages 7–12, 2018.

[89] Sandeep R Sirsat, Vinay Chavan, and Hemant S Mahalle. Strength and Accuracy Analysis of Affix Removal Stemming Algorithms. *International Journal of Computer Science and Information Technologies*, 4(2):265–269, 2013.

[90] Rada Mihalcea and Paul Tarau. TextRank: Bringing Order into Text. In *EMNLP*, volume 4, pages 404–411, 2004.

[91] Ronald Cornet and Nicolette de Keizer. Forty Years of SNOMED: A Literature Review. *BMC Medical Informatics and Decision Making*, 8(Suppl 1):S2, 2008.

[92] Catherine Smith and P Stavri. Consumer Health Vocabulary. *Consumer Health Informatics*, pages 122–128, 2005.

[93] Francesco Corcoglioniti, Marco Rospocher, and Alessio Palmero Aprosio. Extracting Knowledge from Text with PIKES. In *International Semantic Web Conference*, 2015.

[94] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv*, 2013.

[95] Jon Brassey. TRIP Database: Identifying High Quality Medical Literature from a Range of Sources. *New Review of Information Networking*, 11(2):229–234, 2005.

[96] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[97] Bo Pang, Lillian Lee, et al. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[98] Marie-Catherine De Marneffe and Christopher D Manning. Stanford Typed Dependencies Manual. Technical report, Technical Report, Stanford University, 2008.

[99] Rie Johnson and Tong Zhang. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, 2015.

[100] Emilio Ferrara. What Types of COVID-19 Conspiracies are Populated by Twitter Bots? *First Monday*, 25(6), May 2020.

[101] Hamman Samuel and Osmar Zaïane. Iron Mask: Trust-Preserving Anonymity on the Face of Stigmatization in Social Networking Sites. In *International Conference on Trust and Privacy in Digital Business*, pages 66–80. Springer, 2017.

[102] Marsha White and Steve M Dorman. Receiving Social Support Online: Implications for Health Education. *Health Education Research*, 16(6):693–707, 2001.

[103] David M Frost. Social Stigma and its Consequences for the Socially Stigmatized. *Social and Personality Psychology Compass*, 5(11):824–839, 2011.

[104] Yi Hong, Timothy B Patrick, and Rick Gillis. Protection of Patient's Privacy and Data Security in E-Health Services. In *International Conference on BioMedical Engineering and Informatics*, pages 643–647. IEEE, 2008.

[105] Latanya Sweeney. Simple Demographics Often Identify People Uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.

[106] Piotr Cofta. Confidence, Trust and Identity. *BT Technology Journal*, 25(2):173–178, 2007.

[107] Priyanka Kayarkar and Prashant Ricchariaya. An Enhanced Approach for Digital Forensics using Innovative GSP Algorithm. *International Journal of Computer Applications*, 103(6), 2014.

[108] Faiza Masood, Ahmad Almogren, Assad Abbas, Hasan Ali Khattak, Ikram Ud Din, Mohsen Guizani, and Mansour Zuair. Spammer Detection and Fake User Identification on Social Networks. *IEEE Access*, 7:68140–68152, 2019.

[109] Benjamin Taskar, Eran Segal, and Daphne Koller. Probabilistic Classification and Clustering in Relational Data. *International Joint Conference on Artificial Intelligence*, 17(1):870–878, 2001.

[110] Bianca Zadrozny and Charles Elkan. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699. ACM, 2002.

[111] Yoshitaka Kameya and Kentaro Hayashi. Bottom-Up Cell Suppression that Preserves the Missing-At-Random Condition. In *International Conference on Trust and Privacy in Digital Business*, pages 65–78. Springer, 2016.

[112] Stuart E Dreyfus and Hubert L Dreyfus. A Five-Stage Model of the Mental Activities involved in Directed Skill Acquisition. Technical report, California Univ Berkeley Operations Research Center, 1980.

[113] Brendan Nyhan, Jason Reifler, Sean Richey, and Gary L Freed. Effective Messages in Vaccine Promotion: A Randomized Trial. *Pediatrics*, 133(4):e835–e842, 2014.

[114] Mike Sharun. MyHealth.Alberta.ca: Strategies and Lessons Learned for a Provincial Health Portal. In *Medicine 2.0 Conference*. JMIR Publications Inc, 2012.

[115] P. Jason Morrison. Folksonomies: Why Are They Tagging, and Why Do We Want Them To? *Bulletin of the American Society for Information Science and Technology*, 34(1):12–15, 2007.

[116] Saeed Mohajeri, Hamman W Samuel, Osmar Zaïane, and Davood Rafiei. BubbleNet: An Innovative Exploratory Search and Summarization Interface with Applicability in Health Social Media. In *International Conference on Digital Economy (ICDEc)*, pages 37–44. IEEE, 2016.

[117] Hamman Samuel and Osmar Zaïane. PubMedReco: A PubMed Citations Recommender System for Real-Time Chat. In *16th IMIA World Congress on Medical and Health Informatics (MedInfo)*, 2017.

[118] Hamman Samuel, Benyamin Noori, Sara Farazi, and Osmar Zaïane. Context Prediction in the Social Web Using Applied Machine Learning: A Study of Canadian Tweeters. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 230–237, 2018.

[119] Esha. Harnessing Tweets to get the Pulse of a City. Master's thesis, University of Alberta, Department of Computing Science, 2020.

[120] Patricia Martz. An Observational Study of Social Media Technology Conversations: Exploring how Members of the Alberta Public, Organizations and Health Care Professionals Express Wellness, in Relation to Children. Master's thesis, University of Alberta, Public Health Sciences, 2015.

[121] Sean Brennan, Adam Sadilek, and Henry Kautz. Towards Understanding Global Spread of Disease from Everyday Interpersonal Interactions. In *Proceedings of the AAAI International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2783–2789, 2013.

[122] Marcel Salathé. Digital Epidemiology: What Is It, And Where Is It Going? *Life Sciences, Society and Policy*, 14(1):1, 2018.

[123] James T Schlitt, Bryan Lewis, and Stephen Eubank. ChatterGrabber: A Lightweight Easy to Use Social Media Surveillance Toolkit. *Online Journal of Public Health Informatics (OJPHI)*, 7(1), 2015.

[124] Hassan Zaraket and Reiko Saito. Japanese Surveillance Systems and Treatment for Influenza. *Current Treatment Options in Infectious Diseases*, 8(4):311–328, Dec 2016.

[125] Luke Sloan and Jeffrey Morgan. Who Tweets with their Location? Understanding the Relationship Between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter. *PloS One*, 10(11), 2015.

[126] Bo Han, Paul Cook, and Timothy Baldwin. Text-Based Twitter User Geolocation Prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014.

[127] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[128] Ahmed Ali, Walid Magdy, and Stephan Vogel. A Tool for Monitoring and Analyzing Healthcare Tweets. In *Proceedings of the ACM SIGIR Workshop on Health Search & Discovery*, volume 28, page 23, 2013.

[129] Robert Plutchik. The Nature Of Emotions: Human Emotions Have Deep Evolutionary Roots, A Fact That May Explain Their Complexity And Provide Tools For Clinical Practice. *American scientist*, 89(4):344–350, 2001.

[130] Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. ACM, 2004.

[131] C.J. Clayton Hutto and Eric Gilbert. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Eighth International Conference on Weblogs and Social Media*, 2014.

[132] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaïane. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Computing Surveys*, 50(2):25, 2017.

[133] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the 7th International Workshop on Semantic Evaluation Exercises*, June 2013.

[134] Alla Keselman, Catherine Arnott Smith, Guy Divita, Hyeoneui Kim, Allen C Browne, Gondy Leroy, and Qing Zeng-Treitler. Consumer Health Concepts That Do Not Map To The UMLS: Where Do They Fit? *Journal of the American Medical Informatics Association*, 15(4):496–505, 2008.

[135] Lee Humphreys, Phillipa Gill, and Balachander Krishnamurthy. How Much Is Too Much? Privacy Issues On Twitter. In *Conference of International Communication Association*, 2010.

[136] Zhiyong Lu. PubMed and Beyond: A Survey of Web Tools for Searching Biomedical Literature. *Database*, 2011, 2011.

[137] Stuart J Nelson. Medical Terminologies That Work: The Example of MeSH. In *International Symposium on Pervasive Systems, Algorithms, and Networks*, pages 380–384. IEEE, 2009.

[138] Eric Sayers. A General Introduction to the E-utilities. *National Center for Biotechnology Information*, 2010.

[139] John J Hopfield. Artificial Neural Networks. *IEEE Circuits and Devices Magazine*, 4(5):3–10, 1988.

[140] Eric Sayers. The E-utilities In-Depth: Parameters, Syntax and More. *Entrez Programming Utilities Help [Internet]*, 2009.

[141] Miquel Montaner, Beatriz López, and Josep Lluís De La Rosa. A Taxonomy of Recommender Agents on the Internet. *Artificial Intelligence Review*, 19(4):285–330, 2003.

[142] Matthew Honnibal. Supervised Similarity: Learning Symmetric Relations from Duplicate Question Data. Online: https://explosion.ai/blog/supervised-similarity-siamese-cnn, 2017. Retrieved May 20, 2017.

[143] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 539–546. IEEE, 2005.

[144] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[145] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[146] Jey Han Lau and Timothy Baldwin. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. *arXiv Preprint arXiv:1607.05368*, 2016.

[147] Tushar Semwal, Gaurav Mathur, Promod Yenigalla, and Shivashankar B Nair. A Practitioners' Guide to Transfer Learning for Text Classification using Convolutional Neural Networks. *arXiv Preprint arXiv:1801.06480*, 2018.

[148] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.

[149] Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. How Transferable are Neural Networks in NLP Applications? *arXiv Preprint arXiv:1603.06111*, 2016.