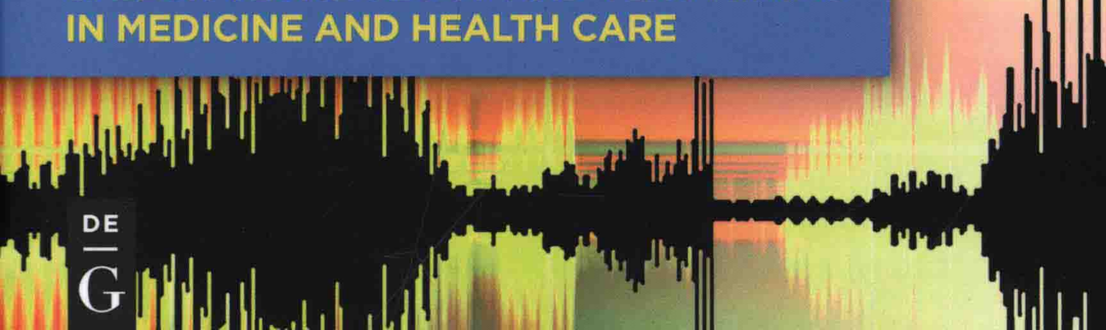
DE GRUYTER

Amy Neustein (Ed.)

TEXT MINING OF WEB-BASED MEDICAL CONTENT

SPEECH TECHNOLOGY AND TEXT MINING IN MEDICINE AND HEALTH CARE



Text Mining of Web-Based Medical Content

Edited by
Amy Neustein

DE GRUYTER

Editor
Amy Neustein
800 Palisade Avenue
Suite 1809
Fort Lee, NJ 07024
USA
amy.neustein@verizon.net

ISBN 978-1-61451-541-8
e-ISBN (PDF) 978-1-61451-390-2
e-ISBN (epub) 978-1-61451-976-8

Library of Congress Cataloging-in-Publication data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Typesetting: Compuscript Ltd., Shannon, Ireland
Printing and binding: CPI Books GmbH, Leck
Cover image: MEHAU KULYK/SCIENCE PHOTO LIBRARY / Agentur Focus

∞ Printed on acid-free paper
Printed in Germany
www.degruyter.com

FSC
www.fsc.org
MIX
Papier aus verantwortungsvollen Quellen
FSC® C003147

Amy Neustein (Ed.)

Text Mining of Web-Based Medical Content

Speech Technology and Text Mining in Medicine and Health Care

Series Editor
Amy Neustein

Additional Titles in the Series

Neustein (Ed.), *Speech and Automata in Health Care*
(forthcoming, November 2014), ISBN: 978-1-61451-709-2

Patil and Kulshreshtha (Eds.), *Signal and Acoustic Modeling for Speech and Communication Disorders* (forthcoming, May 2015), ISBN: 978-1-61451-759-7

Ganchev, *Computational Bioacoustics*
(forthcoming, May 2015), ISBN: 978-1-61451-729-0

Beals, Dahl, and Linebarger, *Speech and Language Technology for Language Disorders* (forthcoming, August 2015), ISBN: 978-1-61451-758-0

Preface

Text Mining of Web-Based Medical Content brings together a talented group of researchers devoted to the study of how to derive high quality information from online data sources, ranging from biomedical literature, electronic health records, query search terms, social media posts and tweets, to general health information found on the web. Using some of the latest empirical methods of knowledge extraction, the authors show how online content, generated by both professionals and laypersons, can be mined for valuable information about disease processes, adverse drug reactions not captured during clinical trials, and tropical fever outbreaks. In this anthology the authors show how to perform information extraction on a hospital intranet and how to build a social media search engine to glean information about patients' own experiences in interacting with healthcare professionals. In addition, some of the authors have studied ways to improve access to online health information for those who suffer from visual impairments, while others have studied the use of information extraction techniques in sifting through YouTube video descriptions and radiographic image data.

This book is divided into four sections:

The first section closely examines methods and techniques for mining biomedical literature and electronic health records.

The section opens with a comprehensive overview of the application of text mining to biomedical knowledge extraction, analyzing both clinical narratives and medical literature. The authors demonstrate how the four main phases of biomedical knowledge extraction using text mining (text gathering, text preprocessing, text analysis, and presentation) may be used to obtain relevant information from vast online databases of health science literature and patients' electronic health records. They present various text mining tools that have been developed in both university and commercial settings, as well as an in-depth analysis of the differences between clinical text found in electronic health records and biomedical text found in online journals, books, and conference papers.

In the following chapter, the authors focus exclusively on patients' electronic health records, showing how clinical natural language processing (NLP) can effectively unlock detailed patient information from clinical narratives stored in such records. This chapter introduces the state-of-the-art work in clinical NLP. Using medication information extraction as a use case, the authors describe different methods to build clinical NLP systems, including rule-based, machine learning-based, and hybrid approaches. Applications of medication information extraction systems, such as *pharmacovigilance* (post-market surveillance of drugs) research, are also discussed in this chapter.

The section is rounded out by a fascinating report on two prototypes for performing information extraction on both a hospital intranet and on the World Wide Web. The authors show how they apply ontology-based information extraction to unstructured natural language sources to help enable semantic search of health information. They propose a general architecture capable of handling both private and public data. Two of their novel systems that are based on this architecture are presented in their chapter. The first system, MedInX, is a Medical Information eXtraction system, which processes textual clinical discharge records, performing automatic and accurate mapping of free text reports onto a structured representation. MedInX is designed to be used by health professionals, and by hospital administrators and managers, allowing a search of the contents of its automatically populated ontologies. The second system, SPHInX, attempts to perform semantic search on health information publicly available on the web in Portuguese. The authors provide usage examples and evaluation results that show the potential of their proposed approach to performing information extraction on unstructured text found in hospital records and on the Internet.

The second section explores machine learning techniques for mining medical search queries and health-related social media posts and tweets. In so doing, the authors demonstrate a keen grasp of how laypersons use the web for seeking health information and reassurance. They focus both on search query data entered in the Google search engine and on the health-related user-generated content found on social media sites and on Twitter.

The section begins with a chapter titled "Predicting Dengue Incidence in Thailand from Online Search Queries that Include Weather and Climatic Variables." The chapter presents machine learning techniques to help public health agencies mitigate vector borne disease, in particular dengue outbreaks. Search queries from digital sources are used to forecast the number of dengue cases prior to officially reported cases. This is achieved by processing query terms related to vector-borne dengue disease. Since climate has been correlated to the vector's dynamics, query terms related to weather are utilized for the forecasting of dengue cases.

All in all, one can certainly see the value of mining search query data to predict the number of dengue cases so that public health authorities may devise adequate interventions to address dengue outbreaks before they reach catastrophic proportion.

The chapter on monitoring users' query search terms in predicting a disease outbreak is followed by a fascinating chapter that addresses the other side of the coin. That is, in this subsequent chapter the authors provide a detailed study of how users sometimes divulge *too much* personal health information on line.

The authors opine that with the increasing amount of personal information that is shared on social networks, it is possible that the users might inadvertently reveal some personal health information that may have untoward consequences for the user. They show that personal health information can be detected and, if necessary, protected. They present empirical support for this hypothesis, by showing how two existing well-known electronic medical resources MedDRA and SNOMED help to detect personal health information (PHI) in messages retrieved from a social network site, MySpace. To do so, they introduce a new measure – Risk Factor of Personal Information – that assesses the likelihood that a term would reveal personal health information. They synthesize a profile of a potential PHI leak in a social network, and demonstrate that this task benefits from the emphasis on the MedDRA and SNOMED terms. Using machine learning techniques to validate the importance of terms detected by these two medical dictionaries, they show that their study findings are robust in detecting sentences and phrases that contain users' personal health information.

The section concludes with a thought-provoking analysis of the expanding role of social media for those who seek health information and for those who study social trends based on patient blog postings. Yet, the authors wisely point out that this new medium of communication has its limitations too. Namely, the current inability to access and curate relevant information in the ever-increasing gamut of messages. In their chapter, the authors demonstrate how they seek to understand and curate laypersons' personal experiences on Twitter. To do so, they propose some solutions to improve search, summarization, and visualization capabilities for Twitter (or social media in general), in both real time and retrospectively. In essence, they provide a basic recipe for building a search engine for social media and then make it increasingly more intelligent through smarter processing and personalization of search queries, tweet messages, and search results. In addition, they address the summarization aspect by visualizing topical clusters in tweets and further classifying the retrieval results into topical categories that serve professionals in their work. Finally, they discuss information curation by automating the classification of the information sources as well as combining, comparing, and correlating tweets with other sources of health information. In discussing all these important features of social media search engines they present systems, which they themselves have developed to help identify useful information in social media.

The third section presents speech and audio technologies for improving access to online content for the computer-illiterate and the visually impaired. The authors report on thoughtfully designed systems that help *democratize* the availability of online health information for those who cannot readily access this information on their own.

The section begins with an empirical study of user satisfaction with a health dialogue system designed for the Nigerian low-literate, computer-illiterate, and visually impaired. The author shows how this health dialogue system provides health information about lassa fever, malaria fever, typhoid fever and yellow fever to those who cannot access this information on line. The author points out that since this information on the Internet is mainly delivered in text format, it is only available to a small percentage of the population due to inadequate Internet access and the low level of literacy in Nigeria. The chapter reports on the development, acceptability, and user satisfaction with this dialogue system, which provides health information about these tropical fevers. The author conducted his cross-sectional study using a questionnaire that gathered demographic data about the study participants and their satisfaction and readiness to accept the health dialogue system. The user satisfaction results showed a mean of 3.98 (approximately 4), which is the recommended average for a good usability study. Dialogue systems of this kind help to provide cost-effective and equitable access to health information that can protect the population from tropical disease outbreaks. They serve the low-literate, the computer-illiterate, and the visually impaired.

The chapter on user satisfaction with a health dialogue system is followed by a fascinating presentation of the Smith-Kettlewell Eye Research Institute's Descriptive Video Exchange (DVX) project, which helps the blind and the visually impaired gain access to the information contained in health-related videos found on the web. The author shows step-by-step how DVX provides a framework that enables a large number of people, both amateur and professional, to create descriptions of video data both quickly and easily. The author shows how DVX distributes those descriptions so that they are available to anyone on the Internet and, in particular, provides a special service for the visually impaired. He points out that DVX when combined with speech recognition can greatly improve video search. In short, this chapter closely examines the use of audio (and text-to-speech) description, created through *crowd sourcing*, to improve video accessibility for the blind and the visually impaired.

The fourth section serves as the coda to this book. The contributors to this section have studied the use of information extraction techniques for accessing both medical images stored in digital libraries and health-related video material found on the web.

The section begins by taking a close look at information extraction from medical images. The authors present in detail their evaluation of a novel automatic image annotation system using semantic-based information retrieval. However, first they show the obstacles for image annotations, namely, the semantic gap problem – it is hard to extract semantically meaningful entities when using low-level image features – and the lack of correspondence between the keywords and image regions in the training data. Then, they show that though content-based visual information retrieval (CBVIR) and image annotation has attracted a lot of interest, namely from the image engineering, computer vision, and database community, current methods of the CBVIR systems only focus on appearance-based similarity, i.e., the appearance of the retrieved images is similar to that of a query image. As a result, there is very little semantic information exploited.

To overcome this obstacle the authors have developed a semantic-based visual information retrieval (SBVIR) system while recognizing that two steps are required: (1) to extract the visual objects from images; and (2) to associate semantic information with each visual object. The authors show that the first step can be achieved by using segmentation methods applied to images, while the second step can be achieved by using semantic annotation methods applied to the visual objects extracted from images. They point out that for testing their annotation module they used a set of 2000 medical images: 1500 of images in the training set and 500 test images. For testing the quality of their segmentation algorithm they used a database consisting of 500 medical images of the digestive system that were captured by an endoscope. Their test results, based on looking at the assigned words to see if they were relevant to the image in question, have proven that their automatic image annotation system augurs well in the diagnostic and treatment process. The authors' novel approach to information extraction from medical images is no doubt a first step toward larger studies of automatic image annotation for indexing, retrieving, and understanding large collections of image data. Moreover, the field of information extraction, which is considered a subtask of text mining, can only benefit from such rigorous studies of multimedia document processing, involving automatic annotation and content extraction from medical images.

The book concludes with a study of video metadata by focusing on the title and description of health-related videos found on the web to see if a lay user in search of medical information can perform a successful online search.

The authors contend that though huge amounts of health-related videos are available on the Internet (and health consumers are increasingly looking for answers to their health problems and health concerns by searching for videos on line), a critical factor in identifying relevant videos based on a textual query is the accuracy of the *metadata* with respect to video content. The authors focus on how reputable health videos providers, such as hospitals and health organizations, describe diabetes-related video content and the frequency with which they use standard terminology found in medical thesauri. Their study compared video title and description to medical terms extracted from the MeSH and ICD-10 vocabularies, respectively. They found that only a small number of videos were described using medical terms (4% of the videos included an exact ICD-10 term; and 7% an exact MeSH term). Furthermore, of all those videos that used medical terms in their title/description, they found an astonishingly low variety of diabetes-related medical terms used. For example, the video titles and descriptions brought up only 2.4% of the ICD-10 terms and 4.3% of MeSH terms, respectively.

The authors make the point that these figures certainly give one pause to think as to how many useful health videos are haplessly eluding online patient search because of the sparse use of appropriate terms in video titles and descriptions. Though no one would deny that including medical terms in video title and description is useful to patients who are searching for relevant health information, the authors point out that by adopting good practices for titling and describing health-related videos it may serve another purpose as well. That is, they can help producers of YouTube videos to identify and address the gaps in the delivery of informational resources that patients need to be able to monitor their own health. The authors conclude that sadly, as the situation is now, neither patients nor producers of health videos are able to explore the collection of online materials in the same systematic manner as the medical professional explores medical domains using MEDLINE. The authors pose the question: Why can't we have the same level of rigorous and systematic curating of patient-related health videos as we have for other medical content on the web?

Perhaps this book will provide the answer.

Amy Neustein
Fort Lee, NJ
September, 2014

List of authors

Johan Gustav Bellika
Norwegian Centre for Integrated Care and Telemedicine
Tromsø
Norway

Angel Bravo-Salgado
Center for Computational Epidemiology and Response Analysis (CeCERA)
University of North Texas
Denton

Marius Brezovan
Faculty of Automation
Computers and Electronics
University of Craiova
Romania

Dumitru Dan Burdescu
Faculty of Automation
Computers and Electronics
University of Craiova
Romania

Jedsada Chartree
Center for Computational Epidemiology and Response Analysis (CeCERA)
University of North Texas
Sisaket Rajabhat University
Thailand

Joshua C. Denny
Department of Biomedical Informatics and Medicine
Vanderbilt University
Nashville
Tennessee

Liliana Ferreira
Department of Electronics
Telecommunications and Informatics/ IEETA
University of Aveiro
Portugal

Kambiz Ghazinour
School of Electrical Engineering and Computer Science
University of Ottawa
Canada

Leif Hanlen
NICTA
Faculty of Health
University of Canberra
College of Engineering and Computer Science
Australian National University
Australia

S. Sagar Imambi
T.J.P.S. College
Guntur
India

Tamara Jimenez
Center for Computational Epidemiology and Response Analysis (CeCERA)
University of North Texas
Denton

Randi Karlsen
Department of Computer Science
UiT The Arctic University of Norway
Tromsø
Norway

Stan Matwin
School of Electrical Engineering and
Computer Science
University of Ottawa
Canada
Institute for Big Data Analytics
Dalhousie University
Faculty of Computer Science
Dalhousie University
Halifax
Nova Scotia

Armin R. Mikler
Center for Computational
Epidemiology and Response
Analysis (CeCERA)
University of North Texas
Denton

Jose Enrique Borrás Morell
Department of Computer Science
UiT The Arctic University of Norway
Tromsø
Norway

Amy Neustein
Editor-in-Chief
International Journal of Speech
Technology
Series Editor of Speech Technology
and Text Mining in Medicine and
Health Care (De Gruyter)
Founder and CEO
Linguistic Technology Systems
Fort Lee
New Jersey

Olufemi Oyelami
Department of Computer and
Information Sciences
Covenant University
Ota
Nigeria

Cécile Paris
CSIRO
Computational Informatics
Macquarie University
Sydney
Australian National University
Canberra
Australia

Mário Rodrigues
ESTGA/IEETA
University of Aveiro
Portugal

Vicente Traver Salcedo
ITACA – Health and Wellbeing
Technologies
Universidad Politécnica de Valencia
Spain

Marina Sokolova
School of Electrical Engineering and
Computer Science
University of Ottawa
Canada
Faculty of Medicine
University of Ottawa
Institute for Big Data Analytics
Dalhousie University
Halifax
Nova Scotia

Liana Stanescu
Faculty of Automation
Computers and Electronics
University of Craiova
Romania

Hanna Suominen
NICTA
Faculty of Health
University of Canberra
College of Engineering and Computer
Science
Australian National University
Australia

António Teixeira
Department of Electronics
Telecommunications and Informatics/
IEETA
University of Aveiro
Portugal

Keith M. Williams
Senior Programmer Analyst
Smith-Kettlewell Eye Research Institute
San Francisco

Hua Xu
School of Biomedical Informatics
University of Texas Health Science Center
Houston

Contents