AUTOMATIC INFORMATION COVERAGE ASSESSMENT OF DIABETES WEBSITES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

GÜLİZ BULUT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

SEPTEMBER 2018

AUTOMATIC INFORMATION COVERAGE ASSESSMENT OF

DIABETES WEBSITES

Submitted by Güliz BULUT in partial fulfillment of the requirements for the degree of **Master of Science in the Department of Information Systems, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

Prof. Dr. Yasemin Yardımcı Çetin
Head of Department, **Information Systems** _____

Assoc. Prof. Dr. Tuğba Taşkaya Temizel
Supervisor, **Information Systems, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Banu Günel Kılıç
Information Systems, METU _____

Assoc. Prof. Dr. Tuğba Taşkaya Temizel
Information Systems, METU _____

Assoc Prof. Dr. Cengiz Acartürk
Cognitive Sciences, METU _____

Assoc. Prof. Dr. Altan Koçyiğit
Information Systems, METU _____

Asst. Prof. Dr. Rahime Belen Sağlam
Computer Engineering, Yıldırım Beyazıt University _____

Date: ***03.09.2018***

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Güliz, Bulut

Signature :

ABSTRACT

AUTOMATIC INFORMATION COVERAGE ASSESSMENT OF DIABETES WEBSITES

BULUT, Güliz

MSc., Department of Information Systems

Supervisor: Assoc. Prof. Dr. Tuğba Taşkaya Temizel

September 2018, 108 pages

People frequently access Internet to look up health information. However, as the quality of websites may vary significantly, the treatment recommendations and guidelines provided by some of these web sites may be fallacious. Consequently, patients may unfollow their current treatments suggested by their doctors or start following unfounded treatments. In this thesis, an automated approach is presented to estimate information coverage of websites. The approach is based on a domain-dependent standard knowledge base (KB) and enhanced by open source resources. Elastic net regularized regression is used to construct a model for estimation. As a case study, data set consisting of type 2 diabetes related web pages is utilized. "Standards of Medical Care in Diabetes" published by American Diabetes Association is processed to obtain factual data about treatment of type 2 diabetes. This standard serves as a detailed KB on type 2 diabetes treatment and enables to produce a trustworthy input for evaluation. In light of this KB, the data set of type 2 diabetes related web pages is processed to retrieve their coverage of factual information. It is observed that, extracting significant terms from a domain-dependent knowledge base provide a basis to measure information coverage of a source.

Keywords: Natural Language Processing, Information Retrieval, Term Extraction, Elastic Net Regularized Regression, Diabetes

ÖZ

DİYABET WEB SİTELERİNİN BİLGİ KAPSAMININ OTOMATİK OLARAK DEĞERLENDİRİLMESİ

BULUT, Güliz

Yüksek Lisans, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Doç. Dr. Tuğba Taşkaya Temizel

Eylül 2018, 108 sayfa

İnsanlar sıklıkla sağlık bilgisi araştırmak için interneti kullanmaktadırlar. Ancak, web sitelerinin kalitesinin önemli ölçüde değişken olabilmesi sebebiyle bazı web sitelerinde yer alan tedavi tavsiyeleri ve prensipleri yanıltıcı olabilmektedir. Sonuç olarak, hastalar, doktorları tarafından tavsiye edilen tedavilerini bırakabilmekte ya da sağlam temellere dayanmayan tedavileri uygulamaya başlayabilmektedir. Bu tezde, web sitelerinde yer alan içeriklerin kapsamını kestirmek üzere otomatikleştirilmiş bir yaklaşım sunulmaktadır. Bu yaklaşım, ilgi alanına bağımlı standart bir bilgi tabanına dayanmakta ve açık kaynaklarla zenginleştirilmektedir. Elastik ağ düzenlenmiş regresyon kullanılarak kestirim modeli oluşturulmaktadır. Durum çalışması olarak, tip 2 diyabet ile ilgili web sayfalarından oluşan bir veri seti kullanılmıştır. Amerikan Diyabet Derneği tarafından yayınlanan "Standards of Medical Care in Diabetes", tip 2 diyabet tedavisi hakkında gerçekçi verileri elde etmek için işlenmektedir. Bu standart, tip 2 diyabet tedavisi ile ilgili detaylı bir bilgi tabanı olarak, değerlendirme için güvenilir bir girdi oluşturulmasını sağlamaktadır. Bu bilgi tabanı ışığında, tip 2 diyabet web sayfalarından oluşan veri seti, içerik kapsamlarının saptanması için işlenmektedir. Önemli terimlerin ilgi alanına bağımlı bir bilgi tabanından çekilmesinin, bir kaynağın bilgi kapsamını ölçmek için temel oluşturabileceği görülmektedir.

Anahtar Sözcükler: Doğal Dil İşleme, Bilgi Çıkarımı, Terim Çıkarımı, Elastik Ağ Düzenlenmiş Regresyon, Diyabet

To My Parents

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my supervisor, Assoc. Prof. Dr. Tuğba Taşkaya Temizel, for her invaluable guidance and knowledge, never-ending motivation and deep understanding especially during the hard times throughout this study. I would not have completed my thesis research without her continuous support.

I am grateful to Assoc. Prof. Dr. Banu Günel Kılıç, Assoc. Prof. Dr. Cengiz Acartürk, Assoc. Prof. Dr. Altan Koçyiğit and Asst. Prof. Dr. Rahime Belen Sağlam for their valuable recommendations.

I would like to thank Didem Tokmak. She not only provided me with her data set, but also supported me at all times that I needed. I would not have finished this thesis without her cooperation. I would also like to thank Şeyma Küçüközer Çavdar for her valuable input on evaluation of models.

I would like to express my deepest gratitude and thanks to my parents, Y. İzzettin Bulut and Gülden Bulut, for their never-ending love, support, understanding and encouragement throughout the process of this research and writing this thesis and also throughout my life. This accomplishment would not have been possible without them.

My sincere thanks also go to my wonderful family that always considers my well-being and supports and encourages me at all times. I am very lucky to have them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADA	American Diabetes Association
AFC	Automated Fact-Checking
AQA	Automatic Quality Assessment
ATR	Automatic Term Recognition
BMI	Body Mass Index
DF	Document Frequency
EBM	Evidence-Based Medicine
ECOSOC	Economic and Social Council of the United Nations
EQIP	Ensuring Quality Information for Patients
FK	Flesch-Kincaid
FRE	Flesch Reading Ease
FS	Feature Set
HON	Health on the Net
HSWG	Health Summit Working Group
ICU	Intensive Care Unit
IDF	Inverse Document Frequency
IPDAS	International Patient Decision Aids Standards
JJ	Adjectives PoS tag
KB	Knowledge Base
LASSO	Least Absolute Shrinkage and Selection Operator
METU	Middle East Technical University
MSE	Mean Squared Error
NHS	National Health Service
NLP	Natural Language Processing
NN	Nouns PoS tag
NO	None
ODPHP	Office of Disease Prevention and Health Promotion

OLS	Ordinary Least Squares
PoS	Part-of-Speech
RAKE	Rapid Keyword Extraction
TF	Term Frequency
TIS	The Information Standard
TTF	Total Term Frequency
TW	Total Word Count
URL	Uniform Resource Locator
US	United States
UW	Unique Word Count
VB	Verbs PoS tag
VM	Vectorization Method

LIST OF SYMBOLS

$\boldsymbol{\alpha}$	Elastic net coefficient to adjust l_1 and l_2 norms
$\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$	i^{th} coefficient (i=1,…,p)
$\boldsymbol{d}$	Denotes any document
$\boldsymbol{DF(t)}$	Document frequency of a word *t*
$\boldsymbol{IDF(t)}$	Inverse document frequency of a word t
$\boldsymbol{\lambda}$	Tuning parameter of penalized regression models
$\boldsymbol{n}$	Number of observations
$\boldsymbol{p}$	Number of predictors
$\boldsymbol{t}$	Denotes any word
$\boldsymbol{TF(t,d)}$	Term frequency of a word *t* for document *d*
$\boldsymbol{TFIDF(t,d)}$	TF-IDF score of a word *t* for document *d*
$\boldsymbol{x}_{\boldsymbol{i}}$	i^{th} predictor (i=1,…,p)
$\boldsymbol{y}$	Response
$\widehat{\boldsymbol{y}}$	Estimation of response *y*

CHAPTER 1

INTRODUCTION

1.1.Problem Statement

More people search for health online however it is difficult to find correct and relevant information from diverse range of web sites. According to a recent survey performed in the United States (Fox & Duggan, 2013), thirty-five percent (35%) of U.S. adults have gone online to figure out a medical condition. In the same survey, among the internet users, seventy-two percent (72%) are reported to have looked online for health information. This group is referred as "online health seekers". Among these online health seekers, seventy-seven percent (77%) state that they refer to a web search engine such as Google, Bing or Yahoo. The quality of websites differs significantly, and some webpages may be misleading for health information seekers. Searching for information on search engines may result in landing on websites that are not prepared by medical domain experts. This may cause patients to unfollow their current treatments suggested by their doctors or start following unfounded treatments. In this thesis, an automated approach is presented to estimate information coverage of diabetes websites related to treatment of type 2 diabetes.

Various studies to assess the quality of websites have been performed. Eysenbach, Powell, Kuss, and Sa (2002) carried out an extensive search on the literature to synthesize the methods used to evaluate quality of health information on the Web. According to this study, the most frequently used quality criteria include accuracy, completeness, readability, design, disclosures, and references provided. This paper also summarizes completeness (used interchangeably with "comprehensiveness", "coverage" or "scope") evaluation methods used in the literature. Some studies are reported to use a 5-point scale, while others use "balance", e.g. whether the disadvantages of a topic are presented besides its advantages, or coverage of topic areas defined a priori are measured. In an effort to measure the quality of Swedish breast cancer websites, Nilsson-Ihrfelt et al. (2004) characterizes "coverage" as "None", "Minimal" and "More than minimal". This methodology for coverage is also employed in Khazaal, Fernandez, Cochand, Reboh, & Zullino (2008) and Morel, Chatton, Cochand, Zullino, & Khazaal (2008) by assigning zero point for "None", one

point for “Minimal” and two points for “More than minimal” or “Sufficient”. In this thesis, this approach is employed to obtain coverage scores.

Coverage can be defined as the extent of addressing the necessary topics in a context. To assign a coverage score manually, one needs to have a guideline to define the context. Moreover, for an automatic coverage assessment, a baseline to extract representative data which will reflect factual information on the topic of interest is required.

Evidence-Based Medicine Working Group (1992) presents evidence-based medicine (EBM) as a new paradigm for medical practice in 1992. In the paper, EBM is reported to de-emphasize intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision-making and stresses the examination of evidence from clinical research. L Sackett, Rosenberg, Gray, Haynes, & Richardson (1996) defines EBM as the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. It is stated that EBM is about integrating individual clinical expertise and the best external evidence. For diabetes treatment, “Standards of Medical Care in Diabetes” released by American Diabetes Association (ADA) (American Diabetes Association, 2016) serves as a gold standard for EBM that intends to provide individuals with components of diabetes care, general treatment goals, and tools to evaluate the quality of care. ADA’s Professional Practice Committee performs an extensive literature search and updates the Standards annually based on the quality of new evidence. In this thesis, it is decided that ADA’s “Standards of Medical Care in Diabetes” can be used as a knowledge base for both manual and automatic information coverage assessments. It will be referred as ADA throughout this report. Automatic information coverage assessment of webpages is also performed using Wikipedia as the knowledge base. These two sources serve as factual information providers in the scope of this thesis.

1.2.Research Questions

This thesis presents an automated approach to estimate information coverage of websites related to treatment of type 2 diabetes. Monolingual corpus and data set in English are used in this thesis.

This thesis aims to answer the following research questions:

1) How can we automatically assess coverage of type 2 diabetes websites? Can a knowledge base be used for this purpose?
2) Which knowledge bases – Wikipedia or ADA can be used effectively to measure coverage? What are their benefits and limitations over each other?
3) How can we build a model to estimate coverage of diabetes websites? What insights can the model provide us about the information on websites?
4) How does knowledge base size affect the model performance?
5) Which type of linguistic features are more efficient in estimating coverage of diabetes websites?

To answer the research questions, firstly, a set of type 2 diabetes related websites are collected. Two medical domain experts are asked to evaluate these websites in terms of coverage. An assessment guideline to be provided to the experts for scoring process is prepared using ADA. This guideline included summaries and key points of selected treatment related chapters. Medical domain experts assessed the websites in light of this guideline, detailed chapters of the gold standard and their expertise.

Two knowledge bases are used to generate two corpora that will be used to extract significant terms in order to measure information coverage. The first knowledge base is assembled using the aforementioned "Standards of Medical Care in Diabetes" of ADA. The second corpus is constructed by using Wikipedia content. To attain this construction, all the visible internal links in "Type 2 diabetes mellitus" page of Wikipedia are obtained, and their main content is extracted and processed. Annotation by part-of-speech (PoS) tagging on the two corpora is performed to differentiate effects of linguistic properties of features on model performance.

Finally, occurrences of extracted significant terms are utilized to generate features. Elastic net regularized regression is used to construct a model on these features to estimate coverage of diabetes websites.

1.3.Contributions of the Thesis

The contributions of this thesis are summarized below:

- An automated approach to estimate information coverage of type 2 diabetes websites is provided.
- The knowledge bases that are used to extract factual information and their limitations and advantages are presented.
- The most significant features to estimate coverage of Type 2 diabetes websites are identified.
- The effects of knowledge base size and linguistic properties of features on coverage of websites were identified.

1.4.Organization of the Thesis

The rest of the thesis is divided into five chapters. Chapter 2 introduces the literature review on online information quality assessment, automatic term recognition and elastic net regression model. Chapter 3 explains the methodology employed in thesis studies. This chapter includes details of data set and corpus construction and their preprocessing. It continues with explanation of methodology to construct features. Finally, it gives information about the model used and its construction. Chapter 4 presents the results obtained and discussion on results. The results are covered in two subsections: Candidate Term Extraction and Model Construction to Assess Information Coverage of Diabetes Websites. Model construction section firstly discusses model construction using different features, secondly evaluates higher performing models and their significance in more detail and finally analyzes the significant terms. Chapter 5 includes overall conclusions and discussions on future work.

CHAPTER 2

LITERATURE REVIEW

Internet users progressively search for health information on websites. In a recent survey (Fox & Duggan, 2013), it is reported that seventy-two percent (72%) of Internet users have looked online for health information. US Department of Health and Human Services' Office of Disease Prevention and Health Promotion (ODPHP) sets two objectives in an effort to improve the health of the nation. In short, it aims increasing the proportion of health-related websites that meet certain quality evaluation criteria and follow established usability principles (Devine, Broderick, Harris, Wu, & Hilfiker, 2016). This recent governmental study is a formal indication of the growing significance of online health information quality.

As more people are engaging with online content, online rumour dissemination and fact-checking to detect misinformation through Internet capture growing interest in various domains. Websites that do not provide high quality information may mislead online health seekers. As a result, patients may unfollow their current treatments suggested by their doctors or start following unfounded treatments. Automatic evaluation of health website quality can prove beneficial to overcome this problem and eliminate drawbacks of manual assessment.

This chapter presents the background of the study by outlining research performed on quality assessment of online health information as well as misinformation, rumours and deception on websites. Moreover, background of the techniques employed for automatic evaluation of information coverage throughout this research are presented in “Automatic Term Recognition” and “Elastic Net Regression Model” sections.

2.1.Quality Assessment of Online Health Information

Increasing usage of online content raises concerns about the reliability and quality of health information on websites. Many studies have been performed to assess online health information quality. The tools developed for this purpose include DISCERN; Health on the Net (HON) code; Ensuring Quality Information for Patients (EQIP);

International Patient Decision Aids Standards (IPDAS) checklist; The Information Standard (TIS) and Centers for Disease Control and Prevention (CDC) Clear Communication Index (CCI).

DISCERN is widely recommended and used by authoritative sources for the evaluation of websites (Griffiths & Christensen, 2005). It is defined as "a reliable and valid instrument for judging the quality of written consumer health information" (Charnock, Shepperd, Needham, & Gann, 1999). DISCERN can be used to judge the reliability and quality of a publication as a source of information about treatment choices without the need for specialist knowledge and without reference to other publications or advisers (Charnock, 1998). "It is developed as a national project to establish quality thresholds for written information on treatment choices provided by National Health Service (NHS) organizations, charities and self-help groups, the pharmaceutical industry and other sources of consumer health information." (Charnock, 1998). It measures the quality based on fifteen questions to be answered on a five-point scale about the content. Questions 1-8 are related to the reliability and questions 9-15 concentrate on specific details about treatment choices. Khazaal, et al. (2009) introduces brief DISCERN that includes six items extracted from the original DISCERN.

HON Foundation was born in May 1996 from a collective decision by health specialists. HON was created to promote useful and trustworthy online health information. HON has been granted consultative status with the Economic and Social Council of the United Nations (ECOSOC) and it accredits websites that distribute health-related information. HONcode and its principles aim to improve the quality of health information on the Internet (Health on the Net). Eight HONcode principles are authority, complementarity, privacy policy, attribution and date, justifiability, transparency, financial disclosure and advertising policy.

EQIP is developed to especially assess quality of written patient information (Moult, Franck, & Brady, 2004). In this study, the authors limit the quality assessment tool to completeness, appearance, understandability and usefulness, and 20 criteria that addressed the key concepts were identified from the literature. EQIP is stated to be developed for use by patient information managers and health care professionals and requires at least some knowledge of the topics.

"IPDAS Collaboration is a group of researchers, practitioners and stakeholders from around the world that was established in 2003." (The International Patient Decision Aid Standards (IPDAS) Collaboration). Their aim was "to achieve an international consensus-based framework of quality criteria for patient decision aids that would act as a checklist for developers and users." (Elwyn, et al., 2006).

"TIS was developed in response to the large amount of health and care information available to the public and patients." (National Health Service England). The standard, released by NHS, is made up of six principles: information production, evidence sources, user understanding and involvement, end product, feedback, review. Member organizations have demonstrated their compliance with these requirements with supporting evidence. TIS quality mark on a material indicates that the organization has undergone a comprehensive assessment and the information they provide is high quality.

CDC CCI is a new evidence-based tool to prepare and review health information (Baur & Prue, 2014). There are four open-ended introductory questions and 20 scored items that affect information clarity and audience comprehension in the Index.

Studies have been performed to investigate the use and validity of these tools. Rees, Ford, & Sheard (2002) evaluates the trustworthiness of DISCERN using leaflets on treatment options for prostate cancer. The results showed the instrument could be used by both patients and healthcare professionals to differentiate between low and high-quality publications on prostate cancer. Griffiths & Christensen (2005) performed a study for depression information websites. They sought to determine whether DISCERN is a valid indicator of quality for consumers that do not have specific mental health training. They also investigated whether Google PageRank is a content quality indicator. The findings report that DISCERN is an indicator of evidence-based quality when used by both consumers and health professionals. Moreover, for consumers, Google PageRank is an indicator of evidence-based quality as strong as DISCERN. Finally, sites that are observed as more useful, trustworthy, and relevant by consumers are sites of higher evidence-based quality. The study also states that identifying optimal combinations of multiple indicators of quality may be valuable. A study to evaluate HON label and DISCERN as content quality indicators is performed on the following domains: gambling, alcohol, cocaine, cannabis, bipolar disorder and social phobia (Khazaal, Chatton, Zullino, & Khan, 2012). It is found that content quality was not associated with origin of sites (commercial, university, government, etc.) neither with the HON label, but it was positively correlated with DISCERN. A more recent study compares the reliability of DISCERN and EQIP on written patient information for eczema in German (McCool, Wahl, Schlecht, & Apfelbacher, 2015). The results showed that both tools are reliable and DISCERN is more precise for patient information on the treatment and care of eczema. In another study, Baur & Prue (2014) explores the validity of CDC CCI criteria by comparing original health material with the ones redesigned with CCI.

There are various studies that use DISCERN as a quality index. Some of these studies include comparisons with other tools. For pain-related health information, Kaicker, Debono, Dang, Buckley, & Thabane (2010) investigates the quality of websites and explain the variability in quality and readability. The study uses DISCERN instrument

as a quality index and Flesch-Kincaid (FK) Readability Algorithm. It is reported that overall quality of pain websites is moderate. Moreover, it is found that "websites which contain health related seals of approval and offer information for alternative commercial solutions to pain related conditions have higher DISCERN scores. Lower readability levels were again found in websites with health-related seals of approval and also interactive multimedia options." (Kaicker, Debono, Dang, Buckley, & Thabane, 2010). Another study researches the quality of websites related to stress urinary incontinence and pelvic organ prolapse using DISCERN (Dueñas-Garcia, et al., 2015). It is understood that some of the significant elements for high-quality treatment information are omitted on English-language websites of related professional, governmental and consumer organizations.

Other studies that utilize information quality instruments include evaluating the quality of Internet health sources in pediatric urology using DISCERN and HONcode (Fast, Deibert, Hruby, & Glassberg, 2013); evaluating web searches in childhood epilepsy using DISCERN (Cerminara, Santarone, Casarelli, Curatolo, & El Malhany, 2014); assessing the reliability of websites on the thumb sucking habit with DISCERN and HONcode (Kiran, et al., 2015); evaluating asthma websites using Brief DISCERN, HONcode and FK Readability Algorithm (Banasiak & Meadows-Oliver, 2017); analyzing the quality and readability of online information related to meningiomas by incorporating Flesch Reading Ease (FRE) score, FK grade score, DISCERN, CDC CCI criteria, HON code and TIS certification (Saeed & Anderson, 2017).

Apart from the aforementioned tools, some other measures exist in order to assess online health information. A set of criteria for evaluating online health information quality is defined by the Health Summit Working Group (HSWG) (Association, 2001). Criteria and the factors they incorporate are given below:

- Credibility: the foundation, currency, relevance/utility, editorial review process, and financial disclosure,
- Content: must be accurate and complete, and must provide applicable disclaimer,
- Disclosure: informing the user of the purpose of the site and any profiling or collection of information associated with using the site,
- Links: appraised according to selection, architecture, content, and back linkages,
- Design: accessibility, navigability, and internal search capability,

- Interactivity: feedback mechanisms and means for exchange of information among users,
- Caveats: clarification of the site's main function as marketing products and services or serving as a primary information content provider.

Thakurdesai, Kole, & Pareek (2004) uses HON code and HSWG criteria to evaluate web-based diabetes patient education material.

Eysenbach et al. (2002) carried out an extensive search on the literature to synthesize the methods used to evaluate quality of health information on the Web. According to this study, the most frequently used quality criteria include:

- Accuracy
- Completeness
- Readability
- Design
- Disclosures
- References provided

This paper also summarizes completeness (used interchangeably with "comprehensiveness", "coverage" or "scope") evaluation methods used in the literature.

Khazaal, Fernandez, Cochand, Reboh, & Zullino (2008) uses accountability, interactivity, aesthetic issues, readability and content quality measures. Khazaal, Chatton, Zullino, & Khan (2012) these measures together with HON code and DISCERN. Content quality used in these studies has two aspects: coverage and correctness. In an effort to measure the quality of Swedish breast cancer websites, Nilsson-Ihrfelt, et al. (2004) characterizes "correctness" of information as "Mostly not", "Mostly", and "Completely. Moreover, "information coverage" is characterized as "None", "Minimal" and "More than minimal". This methodology for coverage is also employed in Khazaal, Fernandez, Cochand, Reboh, & Zullino (2008) on social phobia, in Morel et al. (2008) on bipolar disorder, in Khazaal, Chatton, Cochand, & Zullino (2008) on cocaine addiction and Khazaal, Chatton, Cochand, & Jermann (2008) on pathological gambling.

The measures covered in HSWG criteria, literature survey of Eysenbach et al. (2002) and the studies mentioned in the previous paragraph all include information coverage (also "completeness", "comprehensiveness" and "scope") as a primary indicator of

online health information quality. Dutta-Bergman (2004) states that in the context of health information, patients are more equipped in decision-making based on information that is complete. Missing necessary and relevant information misleads the consumer.

2.2.Automated Evaluation of Website Information Quality

All of the studies investigated in the previous section perform manual assessments of online health content. Although the developed tools standardize the measurement of website quality, manual evaluation is subjective and time-consuming. The results may differ if people performing the assessment change. Moreover, it takes time to perform assignment of quality scores to websites manually. However, considering the fast pace that information on websites change and increase, more rapid and objective evaluation of online content would be beneficial for online health seekers. Automated approaches can fulfill these requirements. Some automated approaches employed in the literature are given below.

Griffiths, Tang, Hawking, & Christensen (2005) propose an automated process to assess quality of depression websites. The Automatic Quality Assessment (AQA) procedure calculates scores of depression websites by applying pre-constructed relevance and quality queries. These queries are generated as a collection of representative words and two-word phrases obtained by using a data set that contains websites that are manually judged on relevance. A search engine is used to evaluate queries. A collection of relevant websites is manually judged to be high or low quality in light of Oxford University Centre for Evidence-Based Mental Health's guidelines. The evidence-based scores, AQA scores and Google PageRank scores are compared.

A supervised binary classification using support vector machines for reliability of a webpage based on HONcode principles is presented in Sondhi, Vydiswaran, & Zhai (2012). The features used are:

- Link-based features: normalized counts of internal links, external links, total links as well as presence of Contact Us and Privacy Policy links
- Commercial features: normalized count of commercial links and normalized frequency of commercial keywords
- PageRank features: normalized internal and external PageRank
- Presentation features: percentage of coherent text and percentage of spread-out text
- Word features: normalized frequency of each unique word

Boyer & Dolamic (2014) evaluates machine learning algorithms and different feature types to detect the trustworthiness of a website according to HON code. The following features are used if their document frequency is over a predefined threshold:

- Single word (bag-of-words)
- Two conjunct words (bigrams)
- Word co-occurrence (co-occurring words independent of word order)

A method to rank diabetes websites with respect to their quality, relevance and evidence-based medicine is presented in Belen Saglam & Taskaya Temizel (2015). The study uses evidence-based medicine to rank websites with respect to quality. Indicators of bias are also considered in the evaluation. Bias information is detected using sentiment analysis. Training for bias detection is performed to determine the terms and their weights that are relevant to bias. Average polarity scores for adjectives, nouns and verbs; frequency of positive and negative words; frequency of nouns, adjectives and verbs; the number of sentences, question marks and exclamation marks are used as features. Relevance feedback is used to generate a quality query. Candidate terms are selected among words and phrases from the list of quality criteria for diabetes websites and top search terms for diabetes. The yielding queries are run using a search engine. Okapi BM25 is used to assign a score for each website according to query results.

2.3.Evidence-Based Medicine

Some of the aforementioned sources that aim to measure information quality of websites support the idea of using knowledge bases as factual information source. For health domain, Evidence-Based Medicine Working Group (1992) presents evidence-based medicine (EBM) as “a new paradigm that de-emphasizes intuition, unsystematic clinical experience, and pathophysiologic rationale as sufficient grounds for clinical decision-making and stresses the examination of evidence from clinical research”. ADA provides a gold standard for EBM that intends to provide individuals with components of diabetes care, general treatment goals, and tools to evaluate the quality of care (American Diabetes Association, 2016).

2.4.Misinformation, Rumours and Deception on Websites

Rumour is defined as an item of circulating information with unverified accuracy status at the time of posting (Zubiaga, Aker, Bontcheva, Liakata, & Procter, 2018). Rumours can be short-term, such as those arise during breaking news, or long-term rumours that are debated for long periods of time. As more people engage with online

content, both the quantity and dissemination rate of online rumours increase. Social media even accelerates this with large number of users and ease of sharing information. The veracity of these rumours need to be assessed to distinguish false rumours and prevent dissemination of those. Some of the content may be deliberately generated fake information to deceive people. This leads to misinformation of people and fact-checking needs to be performed to prevent undesired consequences. For these reasons, online rumour dissemination and fact-checking to detect misinformation through Internet capture growing interest in various domains. In the case of healthcare, websites with low quality information may mislead online health seekers causing them to unfollow their current treatments or start unsupported treatments. Zubiaga et al. (2018) presents results of a detailed survey of research on social media rumours. Some examples of the research performed on verification of online rumours are presented in the following paragraphs.

It is emphasized in Ciampaglia, et al. (2015) that traditional fact-checking by expert journalists is insufficient due to large volume of online information and that computational fact checking may enhance accuracy evaluation of content generated online. An approximation to human fact checking by exploiting knowledge graphs is evaluated. Many claims on different domains are examined leveraging a public knowledge graph extracted from Wikipedia, specifically DBpedia database, which consists of factual statements extracted from Wikipedia infoboxes. The findings include that much of the accurate assessment of true statements relies on indirect paths and that an undirected graph produces the best results.

Zhang, Zhang, & Li (2015) explores features that contribute to differentiate true and false health rumours. The findings help health information seekers to assess accuracy of health rumours on the Internet. The study also advises some guidelines to assist decision-making for online users. Length of statements, presence of names of people or places, presence of numbers, presence of pictures, hyperlinks, cues on the information source, and type of rumour as dread or wish are the features investigated to relate to accuracy of Internet health rumours. It is deduced that presence of numbers, source cues and hyperlinks are positively correlated to the probability that a rumour is true. Moreover, dread health rumours are more likely to be true than wish ones. This study investigates effects of directly quantified structural features of data without relating it to a factual information source.

Another study utilizing supervised machine learning methods with rich features is submitted to task 8 in SemEval 2017 Wang, Lan, & Wu (2017). For a rumored tweet and replied tweets, two subtasks are defined: labeling tweets as support, deny, query or comment and predicting their accuracy. The two subtasks are treated as multi-classification problems. The problem is solved by using sentiment-related features. For the first subtask, two-step classifier is used due to imbalance of training data. The first-tier separates tweets as comment and non-comment and the second tier assigns

non-comment tweets as support, deny or query. Three classification system is used for subtask two to label tweets as true, false or unverified. The features used are:

- Linguistic-informed features: word n-grams, named entities
- Tweet domain features: punctuation, emoticon, event (keywords about the events)
- Tweet metadata features: tweet metadata (favorite count, retweet count, pre-retweet count, time gap, tweet level), user metadata (list count, followers count, user favorites count, friends count, verified, protected, default profile, profile use background image, geo enabled)
- Word vector features
- Word cluster feature

Various learning algorithms are investigated: Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, AdaBoost, Gradient Tree Boosting. For submission, an ensemble models are used with top learning algorithms. Logistic regression and support vector machines are reported to perform well consistently. Features closely related to tweets such as tweet domain and metadata performs better than linguistic informed or word vector features.

A recent study embraces an overview of efforts to automatically regulate false or misleading online content and their main challenges Graves (2018). Findings, which are based on a review of efforts and interviews with fact-checkers and computer scientists working in this field, emphasize the need for human judgement for developing generalized and large-scale automated systems. It is stated that although progress is made for a narrow range of simple factual claims, automated fact-checking (AFC) will require human supervision for the foreseeable future. It is agreed by fact-checkers and computer scientists that AFC technologies are promising to help fact-checkers to detect and analyze claims. It is concluded that support from foundations, universities and companies is needed in order to implement large-scale systems and enhance capabilities.

2.5.Automatic Term Recognition

Automatic term recognition (ATR) (or extraction) is a field that has been studied for a long time. It has various application areas and it is widely used for information retrieval applications. As the number of online documents rapidly increase, automated approaches to process large amounts of text data for the purpose of extracting terms has gained importance. For automatic term recognition purposes, depending on the nature of text data and application, one can concentrate on different characteristics of

text such as certain part-of-speech tags and also apply convenient preprocessing techniques such as removing stop words or punctuation. Independent of these decisions, a ranking methodology is required to quantify words or word phrases and select terms accordingly. Examples of methodologies used for ATR are given below.

In ATR process, after preprocessing operations on text data, statistical measures are used to rank the candidate terms. There are two kinds of measures; unithood and termhood (Zhang, Iria, Brewster, & Ciravegna, 2008). Measures of unithood indicate the colocation strength of units that comprise a term and termhood measures indicate association strength of a term to domain concepts. Unithood measure examples include mutual information, log-likelihood and t-test. Measures of termhood are frequency-based and uses a reference corpus. It includes methods such as TF-IDF and weirdness. There are also hybrid approaches, where the two measures are combined. Example measures are C-value, Glossex and Termex. Zhang et al. (2008) also presents a comparison of term recognition algorithms. Five algorithms are selected to compare, which are TF-IDF, weirdness, C-value, Glossex and Termex.

A neural network-based approach to keyphrase extraction from scientific articles is performed by Sarkar, Nasipuri, & Ghose (2010). TF-IDF, position of features, phrase length, word length in a phrase and links between phrases are the features used for this purpose.

Another study on extracting news keywords for topic tracking employs variants of conventional TF-IDF. Cross-domain filtering is used to discard domain-specific stop words (Lee & Kim, 2008). Conrado, Pardo, & Rezende (2013) presents an ATR approach that uses machine learning with various features of candidate terms. Features include statistical ones such as term frequency, linguistic ones such as PoS and more complex ones such as analysis of term context. According to findings of the study, all the tested attribute selection methods indicated TF-IDF to be a significant feature.

A more recent survey presents existing definition of "term" and its linguistic features, formulates the definition of term recognition task and analyzes available methods in domain-specific text collections (Astrakhantsev, Fedorenko, & Turdakov, 2015). Methods based on various factors as statistics of term occurrences, context of term occurrences, topic models, retrieval engines, ontologies, Wikipedia and feature-based inference are described.

Guan (2016) explores the use of automatic keyword and keyphrase extraction techniques for answering biomedical questions. TF-IDF, neighborhood keywords extraction, noun phrases filter and C-value/NC-value are the evaluation approaches. The common terms between extracted terms and "ideal answers" written by biomedical experts could be the answers to biomedical questions. It is concluded that

TF-IDF approach performs the best among other methods and can be used in answering biomedical questions.

Another study proposes machine learning to automatically classify extracted n-grams as term or non-term (Yuan, Gao, & Zhang, 2017). Trials are performed on various domains and languages. Random Forest, Linear Support Vector Machine, Radial Basis Function, Support Vector Machine, Multinomial Naïve Bayes and Linear models of Logistic Regression and SGD Classifier are used as learning algorithms. Features used for training are total term frequency (TTF), average TTF, TTF with inverse document frequency (IDF), residual IDF, C-value, rapid keyword extraction (RAKE), Chi-square, weirdness, glossary extraction and term extraction. Effects of using various classifiers on different corpora are evaluated.

A biomedical corpus of validated terms is created in Sandoval, Diaz, Llanos, & Redondo (2018) for Spanish, Japanese and Arabic corpora. For preselection of terms a morphological tagger; a corpus-based strategy to compare it by a general, large and balanced corpus; and log-likelihood are used. Medical terms are eliminated by the usage of affixes and lemmas. Lastly, each term is evaluated manually. A different methodology is employed for each language to build the term extractor. For Spanish, terms from gold and silver standards are combined with unrecognized items of GRAMPAL lexicon to obtain multi-word terms. The search tool using the result of the studies is developed. It can be used to search for words to also generate information about word distribution, search for medical terms to get the most frequent terms and extract medical terms from an input text.

In order to enhance existing ATR methods, Zhang et al. (2018) proposes to develop generic methods. SemRe-Rank is introduced that includes semantic relatedness into an existing ATR method. Scores of candidate terms calculated by an ATR method are revised using semantic importance scores of SemRe-Rank. It uses a graph of semantically related words and personalized PageRank process. This approach is shown to improve thirteen (13) base ATR methods among four data sets.

2.6.Elastic Net Regression Model

Ordinary Least Squares (OLS) is a linear regression method that aims to minimize the residual sum of squares. Because of the poor performance of OLS, penalization techniques have been proposed to improve it. These techniques include ridge regression and lasso regression. Ridge regression aims to minimize the residual sum of squares subject to a bound on L_2-norm of the coefficients. Lasso regression imposes L_1-penalty on the coefficients (Zou & Hastie, 2005). Both of these methods have some shortcomings. Ridge regression always keeps all predictors in the model. Lasso performs variable selection; however, it selects at most n (number of observations)

variables and tends to select only one of the grouped variables. Zou & Hastie (2005) proposes elastic net regression methodology to overcome these problems. It is reported to perform variable selection and select groups of correlated variables. This model penalizes the coefficients based on both l_1 and l_2 norms. Elastic net encourages grouping effect and performs automatic variable selection for the cases where number of features is much larger than number of observations. When a group of highly correlated features exist, regression coefficients of those parameters tend to be equal.

High dimensionality and correlation between predictors are the common characteristics of text data processing. Because of these properties of data set under observation and the aforementioned characteristics of elastic net regularized regression, elastic net regularized regression method is applied in this thesis both for feature selection and estimation of information quality scores.

Elastic net regression model draws attention for its advantages, especially feature selection when number of predictors is very large. Two examples are given here. Intensive Care Unit (ICU) mortality risk is predicted using two stochastic gradient descent-based classifiers with elastic net regularization based on nursing notes in Marafino, Boscardin, & Dudley (2015). Classifiers used are logistic regression and a linear support vector machine. It is stated that most features selected by both classifiers are relevant and complies with already present predictors of ICU mortality models. Teisseyre (2017) uses elastic net regression to incorporate feature selection property into classifier chains for multi-label classification which is called CCnet.

2.7.Summary

Many studies have been elaborated to evaluate the quality of online health information. However, most of them assess websites manually. This results in subjective and time-consuming evaluations. Automated methodologies can prove beneficial to provide immediate and objective feedback to online health seekers. This methodology does not undervalue the opinions of medical experts, in fact, it provides health seekers with an automated tool to enable shared decision-making of patients and medical experts.

Coverage can be defined as the extent of addressing the necessary topics in a context. It is a convenient measure to match the content of EBM gold standard to websites and pave the way for more detailed further studies. Treated in many studies as an integral part of accuracy, information coverage is a main indicator of website information quality.

In this thesis, an automated approach to estimate information coverage of diabetes websites is presented. The information coverage is based on evidence-based gold standard, ADA. Moreover, Wikipedia is used as an additional knowledge base. Widely

used and reliable ATR method, TF-IDF, is applied to extract candidate terms that are representative of the domain. Features generated using these candidate terms are inputs of elastic net model. Elastic net regularized regression is a method both to estimate information coverage automatically and to detect significant terms for the domain of interest. Model coefficients are evaluated to extract salient terms that are representative of the domain.

To the best of our knowledge, this methodology is the first one to directly incorporate the usage of evidence-based gold standard content to detect important features for the purpose of online health content quality assessment based on information coverage. Other studies were limited in the usage of EBM. They mostly used it as a reference for manual assessments. Moreover, this study is one of the first studies to incorporate analysis of linguistic features for evaluation of website quality. Lastly, linear regression to estimate actual score of websites as opposed to many studies that use logistic regression to classify a website as a quality website or not.

The results represent a first step toward assessing accuracy of health information on diabetes websites related to treatment. Although the correctness of facts is not evaluated in this thesis, the results provide insight on the domain content as well as differentiation of websites according to information coverage. These findings can act as a baseline for further studies such as fact-checking of website content according to EBM.

CHAPTER 3

METHODOLOGY

In this thesis, an automated approach is presented to estimate information coverage of diabetes websites related to treatment of type 2 diabetes.

ADA's "Standards of medical care in diabetes" is an extensive gold standard that explains the treatment process, treatment recommendations and guidelines for diabetes in detail and in a formal format (American Diabetes Association, 2016). However, diabetes related websites are generally written in an informal way and they are not as comprehensive as the ADA guideline. For example; the following sentences of this gold standard, although being informative, include too much formal detail that is generally not included in websites:

1. *"Weight loss can be attained with life-style programs that achieve a 500–750 kcal/day energy deficit or provide approximately 1,200–1,500 kcal/day for women and 1,500–1,800 kcal/day for men, adjusted for the individual's baseline body weight."*

2. *"Consider initiating combination insulin injectable therapy when blood glucose is ≥300– 350 mg/dL (16.7–19.4 mmol/L) and/or A1C is ≥10–12% (86–108 mmol/mol)."*

The related content is written in more informal language and does not include precise metrics even on websites with high coverage scores. The following are examples on two high score websites:

1. Website 3: https://www.diabetesaustralia.com.au/type-2-diabetes (Coverage Score: 10)

    1.1. "Amount of exercise

    For good health, you should be doing about 30 minutes of exercise every day. If this is not possible, then this time can be divided in 3 x 10 minutes sessions. You can break up exercise throughout the day.

    If you need to lose weight, 45-60 minutes every day."

    1.2. a. "When starting insulin, your doctor and Credentialled Diabetes Educator will help you adjust to the new routine and task of giving insulin and find the right dose to reduce your blood glucose levels to acceptable levels."

    b. "However, over time most people with type 2 diabetes will also need tablets and many will also need insulin. It is important to note that this is just the natural progression of the condition, and taking tablets or insulin as soon as they are required can result in fewer complications in the long-term."

2. Website 22: http://www.diabetes.co.uk/type2-diabetes.html (Coverage Score: 10)

    2.1. "To achieve weight loss, a diet should be low calorie and because type 2 diabetes is a lifetime condition, it is important to have a diet you will be able to keep to consistently.

    2.2. "If you are on insulin you may need to regularly test your blood glucose levels to help prevent blood glucose levels from going too low."

Websites with low coverage scores are written with shorter and non-specific statements. Examples from a low score website and how information coverage is assessed manually is given in Appendix A.

Another source of knowledge on type 2 diabetes is the open-source Wikipedia content. It is more detailed and sometimes more precise than websites but still may be written in informal language. Moreover, Wikipedia corpus is unstructured and independent from ADA, therefore, same topics may not be addressed. The following are examples from Wikipedia content related to the above topics:

1. A. Page with the title: "*Insulin resistance*"

    1.1. "Sedentary lifestyle increases the likelihood of development of insulin resistance. [35][36] It has been estimated that each 500 kcal/week increment in physical activity related energy expenditure, reduces the lifetime risk of type 2 diabetes by 9%."

    B. Page with the title*: "Weight loss"*

    1.2. "According to the U.S. Food and Drug Administration (FDA), healthy individuals seeking to maintain their weight should consume 2,000 calories (8.4 MJ) per day."

2. Page with the title*: "Gestational diabetes"*

2.1. "The following are the values which the American Diabetes Association considers to be abnormal during the 100 g of glucose OGTT:

1. *Fasting blood glucose level ≥95 mg/dl (5.33 mmol/L)*
2. *1 hour blood glucose level ≥180 mg/dl (10 mmol/L)*
3. *2 hour blood glucose level ≥155 mg/dl (8.6 mmol/L)*
4. *3 hour blood glucose level ≥140 mg/dl (7.8 mmol/L)"*

Considering this point, in an effort to match the content of ADA's EBM guideline to the content of websites, an approach to identify and analyze salient terms is adopted. Information coverage can be defined as the extent of addressing the necessary topics in a context and it is a convenient measure to determine to what extent information presented in the gold standard is covered on websites. For these reasons, the scope of this thesis is established as mining significant terms using the gold standard as well as open sources such as Wikipedia to generate corpora in order to estimate information coverage of diabetes websites.

The implemented approach enables to identify and analyze terms that are important for diabetes treatment domain. Extracting these terms serves similarly to summarization. It is a way to extract definitive terms of a given corpus instead of extracting definitive sentences in the summarization process. Keywords are a set of significant words in an article that gives high-level description of its contents to readers (Lee & Kim, 2008). A similar approach may be applied to further identify quality related phrases.

The methodology employed in this thesis is summarized in Figure 1.

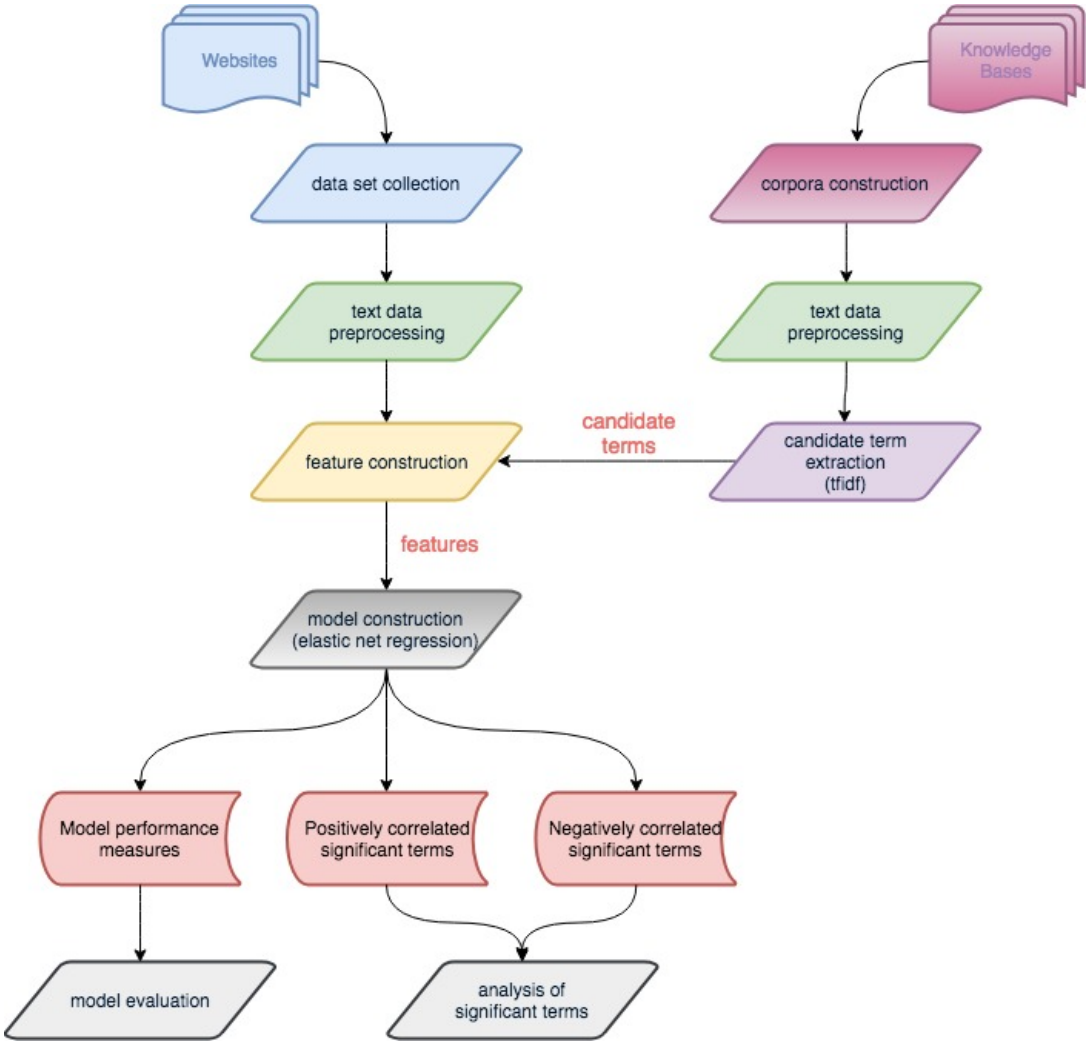


Figure 1: Flow Diagram of the Applied Methodology

Data set collection, corpora construction, text data preprocessing steps, candidate term extraction methodology, feature and model construction procedures, model evaluation approach and analysis of significant terms are explained in detail in the following sections. Monolingual corpus and data set in English are used in this thesis.

3.1.Data Set Collection

Data set of type 2 diabetes related websites is collected by Ölçer (2018). For this purpose, keyword searches are performed on the Internet between June 2016 and August 2016. Two different search engines are used. The keywords used in searches are: “Type 2 diabetes”, “Type 2 diabetes treatments”, “Type 2 diabetes treatments risks”, “Type 2 diabetes treatments benefits”, “Type 2 diabetes no treatment” and “Life with type 2 diabetes”. For each search query; the first thirty websites in English language are collected. Sites were excluded if they were irrelevant, not accessible, require access fee, not English, no info on the domain, journal articles, news, video, personal experiment sites, sponsored links, advertisements, academic press, abstracts, and forum items were excluded. Besides the results of these queries, twelve websites which were assessed as quality websites about type 2 diabetes from earlier studies were included. This process resulted in a data set of 60 observations.

The websites are investigated and scored by two independent medical domain experts in terms of coverage. These experts are six-years general practitioners. Scoring process is defined with the approach employed in Nilsson-Ihrfelt et al. (2004), Khazaal, Fernandez, Cochand, Reboh, & Zullino (2008) and Morel et al. (2008). The scoring table is given in Table 1.

Table 1: Information Coverage Scores

Evaluation	Score
None: Information does not exist	0
Minimal: Some information exists but not sufficient	1
Sufficient: Sufficient information exists	2

Coverage scores are addressed under six topics:

1. Diet, Physical Activity, and Behavioral Therapy
2. Pharmacotherapy
3. Bariatric Surgery
4. Initial Therapy
5. Combination Therapy
6. Insulin Therapy

These topics are generated using subsections of ADA’s gold standard (American Diabetes Association, 2016) chapters: The first three topics are subsections of Chapter 6: “Obesity Management for the Treatment of Type 2 Diabetes” and the final three topics are subsections of Chapter7: “Approaches to Glycemic Treatment”. Experts are

instructed to assign a score according to Table 1 to each website for each of these topics. This results in a maximum score of 12 for the information coverage of each website.

A guideline is provided to domain experts to inform them about the issues that need to be considered for assigning a score. This enabled them to have a common perspective with evidence from EBM. This guideline is prepared by using ADA's gold standard and included summaries and key points of the aforementioned subsections. Medical domain experts assessed the websites in light of this guideline, detailed chapters of the gold standard and their expertise. Ölçer (2018) checked the degree of agreement among experts by quantifying the ratings by kappa. Scores of experts on eleven websites failed to comply after the analysis. For these websites, the experts agreed on the same score for each non-compliant website.

The original data set included 60 websites. In this thesis, since Wikipedia is used as a corpus source, the website with a URL of "https://en.wikipedia.org/wiki/Diabetes_ mellitus_type_2" is eliminated from the data set. Including this page in the data set would be misleading when interpreting the results. This results in a reduced data set of 59 websites.

The html pages of these websites are preprocessed before using them in the automated process. After manually inspecting the content of websites; sidebars, advertisement related fields, navigation sections, references sections and links sections as well as html fields that include unrelated information such as script, meta and style are removed to obtain a structured data set. For getting the main body text out of html files, "BeautifulSoup" package of Python is used.

Each website has a certain number of web pages. The websites used, their web page count, total number of words in each website and information coverage scores assigned to each website by experts are given in Appendix B. The distribution of websites with respect to information coverage score is given in Figure 2.

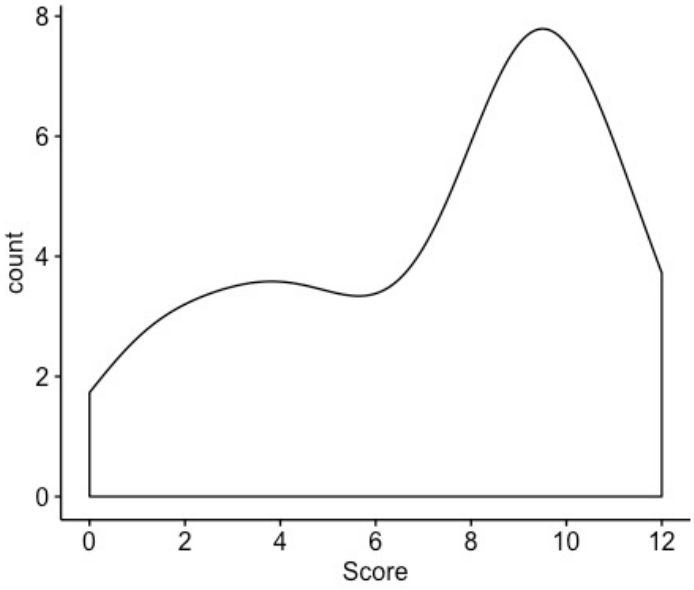


Figure 2: Distribution of Information Coverage Scores

3.2.Corpora Construction

Two knowledge bases are used in this thesis to gather domain specific information and generate two corpora that will be used to extract significant terms. For this purpose, type 2 diabetes related knowledge bases are used because test data of websites are collected only on this subject. The websites all include general terms about type 2 diabetes and significant term extraction enables to identify special terms that yield a difference in information coverage assessment.

The first corpus is assembled using the aforementioned ADA gold standard. ADA corpus is constructed by appending the main text of each topic which are mentioned in Section 3.1 and used as coverage scoring sections. Hence, six documents are present in this corpus. There are 2988 words in total in ADA corpus after preprocessing.

The second corpus is constructed by using Wikipedia content as the knowledge base. To attain this construction, all the visible internal links in “Type 2 diabetes mellitus” page of Wikipedia are obtained, and their main content is extracted and processed to generate the second corpus. Uniqueness of pages are checked according to their “pageid” features. For this purpose, the content of Wikipedia pages is investigated. Fields in the content of Wikipedia pages such as “see also”, “references” and “further reading” are eliminated to retrieve the main body of each page. This corpus has 155 documents. For retrieving Wikipedia information, “Wikipedia” package of Python is

used. Wikipedia pages are fetched on June 2018. The list of titles for retrieved Wikipedia pages is given in Appendix C. There are 255,761 words in total in Wikipedia corpus after preprocessing.

3.3.Text Data Preprocessing

Text data preprocessing steps are applied before using texts in another process. Flow diagram of text data preprocessing employed in this thesis is given in Figure 3. Similar steps are applied both to corpora and data set of web pages.

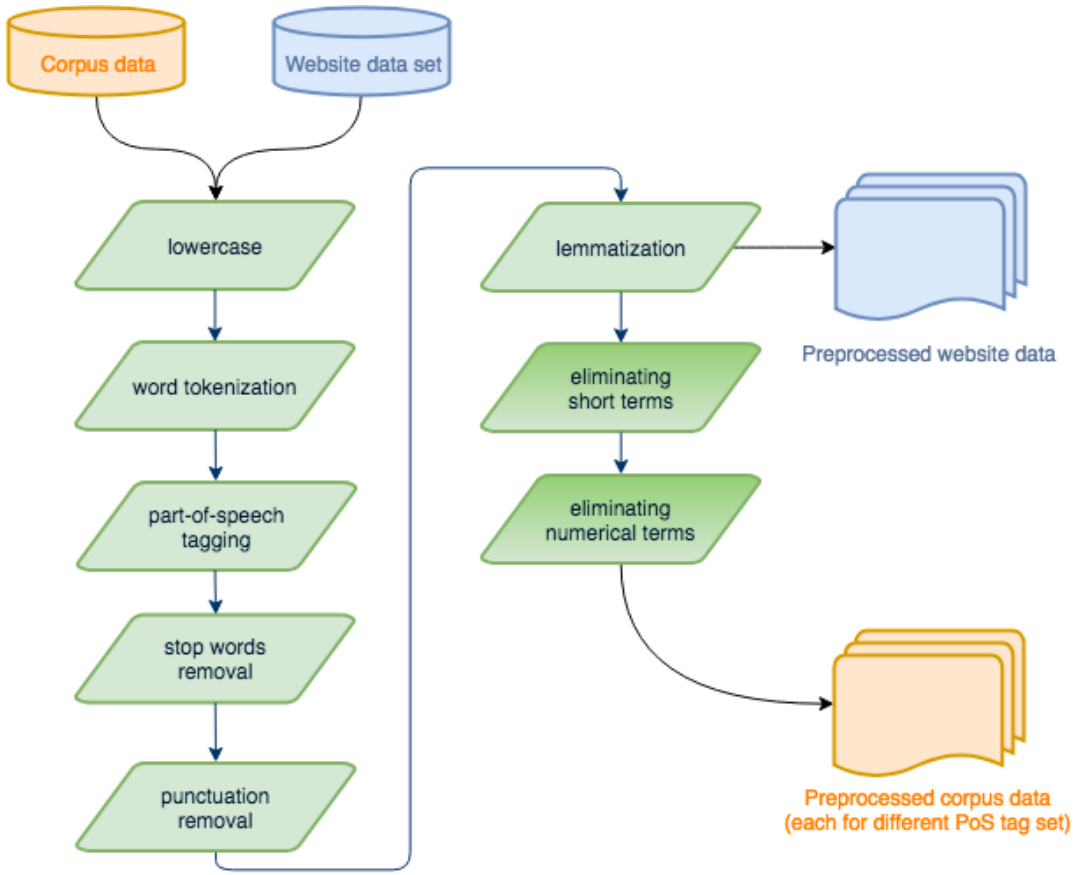


Figure 3 Text Data Preprocessing Flow Diagram

1. Lowercase: All characters of the text are changed to lowercase by using Python's built-in "lower()" function.

2. Word tokenization: The text is tokenized by words using “word_tokenize()” function of Python’s “nltk” package. The steps after this are carried out for each word.

3. Part-of-Speech (PoS) tagging: Annotation by PoS tagging is performed by “pos_tag” function of Python’s “nltk” package. PoS tagging labels the words with their part-of-speech according to their placement in context and words adjacent to them. This information is used for lemmatization input and separating words to explore the effects of linguistic characteristics. After this tagging step, words are handled as <word, tag> pairs.

4. Stop words removal: Stop words are removed since they are not representative of a domain and also, they cause noise in text processing. English stop words are obtained from Python’s “nltk” package. The list of stop words are given in Appendix D.

5. Punctuation removal: The remaining punctuation characters at this step are replaced by the blank character, “ ”. A set of punctuation characters is obtained by “punctuation” function of Python’s “string” library.

6. Lemmatization: Lemmatization is a tool of Natural Language Processing. It aims to remove inflectional endings of words and return the dictionary form of a word. This enables us to treat different inflected forms of a word in the same way. Wordnet lemmatizer of “nltk” package is used to lemmatize nouns and verbs. It is based on WordNet’s built-in morphy function. PoS tags are input to this function to reach at accurate lemmas. Adjectives are not lemmatized in order not to lose the comparative and superlative forms.

7. Eliminating short terms: Terms that have two or less alphabetic characters are eliminated since they are not definitive, and they create noise in text processing.

8. Eliminating numerical terms: Terms with a numerical character are removed.

For the corpora, the preprocessing steps 1-8 are applied to each document in the corpus. After obtaining the preprocessed texts of documents, words are extracted according to their PoS tags to generate three feature sets:

- *Feature set 1* consists of nouns,
- *Feature set 2* consists of nouns and adjectives,
- *Feature set 3* consists of nouns, adjectives and verbs.

This incremental approach is implemented to investigate the effects of linguistic features and their relative importance in estimating information coverage.

When the nouns are considered; the following types of words are examples of significant terms:

- medical terms that are directly related to the field of interest such as medicine names or parts of medicine names,
- terms that are related to type 2 diabetes treatment such as “injection”,
- supporting terms that are related to treatment such as “tablet” and “dose”,
- terms that are used to explain alternative treatment methodologies or the treatment process such as “choice” and “combination”
- terms that explain the side effects of medicine such as “vomiting”, “dizziness” and “headache”,
- general terms about medical domain such as “patient” and “condition”

The adjectives may be indications of the following:

- quantities for treatment such as “twice”, “daily”, “weekly” and “monthly”,
- terms that define types of medication such as “rapid”, “short”, “long” and “acting” used for “rapid-acting insulin”, “short-acting insulin” or “long-acting insulin”,
- terms that are used to explain alternative treatment methodologies or the treatment process such as “alternative” and “possible”

Finally, the verbs may be significant in the following ways:

- defining the process of diabetes treatment with words such as “consider”, “add”, “follow”, “advise”, “continue” and “begin”,
- explaining the effects of a treatment or a biological process such as “release”, “activate”, “lower”, “promote” and “effect”

For nouns the PoS tags that start with “NN”, for adjectives the PoS tags that start with “JJ” and for verbs the PoS tags that start with “VB” are retrieved. All PoS tags under these categories are given in Appendix E. After retrieval, the extracted terms are joined with a blank character between them to obtain vectors of three feature sets for each document. These text data are input to candidate term extraction.

For the content of websites, the preprocessing steps 1-6 are applied to each page. Resulting terms are joined with a blank character between them. The resulting text data is input to feature construction together with candidate terms. Steps 7 and 8 are not applied because they would not affect the end result.

3.4.Candidate Term Extraction

Candidate terms are ranked in order to select most relevant ones. To achieve it, Term Frequency-Inverse Document Frequency (TF-IDF) methodology is employed. TF-IDF is a measure of how important a word is for a corpus. It is directly proportional to the frequency of a word in a document and inversely proportional to frequency of a word in the whole corpus. TF-IDF penalizes the words that occur in most of the documents of a corpus and favors the ones that appear in only a few documents. The higher the TF-IDF score, the more significant that word is between the given documents. Only term frequency of words in the corpus could have been used to rank the candidate words, however, they would be related to type 2 diabetes since all text in corpora are type 2 diabetes related or they could be general terms that are not definitive for the domain of interest.

TF-IDF methodology is explained in Jing, Huang, & Shi (2002). The term frequency is the number of times a word *t* occurs in document *d*. It is denoted by *TF(t,d)*. The document frequency *DF(t)* is the number of documents where the word t occurs at least once. The inverse document frequency can be calculated using Equation 1 where |D| is the total number of documents:

$$IDF(t) = \log \frac{|D|}{DF(t)} \tag{1}$$

Then, TF-IDF score of a word *t* for document *d* can be calculated as:

$$TFIDF(t,d) = TF(t,d) * IDF(t) \tag{2}$$

"sklearn" package of Python is used to attain TF-IDF scores for each word in documents. In this package, inverse document frequency is calculated with Equation 3:

$$IDF(t) = \log \frac{|D|+1}{DF(t)+1} + 1 \quad (3)$$

The resulting TF-IDF vectors are then normalized by Euclidean norm:

$$v_{norm} = \frac{v}{\sqrt{{v_1}^2+{v_2}^2+\cdots+{v_n}^2}} \quad (4)$$

For each document of the corpus, Python function returns a vector consisting of TF-IDF scores of all words in all documents. These vectors are merged such that the highest TF-IDF score is taken for each word. The words are then sorted according to decreasing TF-IDF score values. Different percentages of words are selected as candidate terms and used for feature construction to investigate the effect of knowledge base size on estimating information coverage of websites.

3.5.Feature Construction

In order to use words as features, they need to be quantified. This is called "vectorization" or "vector space modelling". For this purpose, occurrences of extracted significant terms in web sites are checked to generate features. When searching for words in a web page, Python's String "count" function is used. In order not to count supersets, the words that are searched for are used with a blank added to the beginning and at the end.

For each website, bag-of-words approach is used to generate features. Each page of a website is treated as a collection of words independent of their order.

Two main approaches are used to quantify the occurrence of candidate words in websites: checking for occurrences of words (binary weighting) and counting total occurrences of words in the websites (term frequency). Term frequencies of words may be affected from the size of documents. Larger documents tend to have higher number of words. Term frequencies are normalized in two-fold: In the first one, the logarithm is calculated. In the second one, frequency of each term is divided by "total word count" and "total unique word count" of websites.

As a result, four vectorization methods are utilized:

- *Vectorization method 1:* The terms are binary coded, where one indicates the presence of a certain term and zero otherwise.
- *Vectorization method 2:* Logarithm of term frequencies is taken for normalization.
- *Vectorization method 3:* Term frequencies are divided by "total word count" for normalization.
- *Vectorization method 4:* Term frequencies are divided by "total unique word count" for normalization.

The resulting feature sets have a size o*f* $p\ x\ n$ where p is the number of predictors and equals the number of words in the feature sets and n is the number of observations, which is the length of data set.

To investigate the effects of "total word count" and "unique word count" of websites on information coverage, these vectors are also used as additional features for some models. In this case, one more predictor is added to the model.

3.6.Model Construction to Assess Information Coverage of Diabetes Websites

Considering the linear regression model, where $x_1, \dots, x_p$ are the p predictors and y is the response;

$$\hat{y} = \hat{\beta}_0 + x_1\hat{\beta}_1 + \cdots + x_p\hat{\beta}_p \tag{5}$$

a model fitting procedure produces the coefficients $\hat{\beta}_i$ where $i = 0, \dots, p$.

Ordinary least squares (OLS) method aims to minimize squared error:

$$\hat{\beta} = argmin_\beta \sum_{i=1}^{p}(y_i - x_i\beta_i)^2 \tag{6}$$

Penalization methods have been proposed to improve OLS. When the data is high-dimensional and the number of observations is low, the number of predictors (p) is much larger than the number of observations (n) (p>>n case). In such a case, a unique combination of p coefficients, such that the model is optimal, cannot be found. "Regularization" aims building a model by reducing the dimensionality of data. Shrinkage is one of the methods used for regularization. It aims to fit a model by using all p predictors but the estimated coefficients are shrunken towards zero (Li & Chen).

Ridge regression is a method that penalizes the coefficients based on their l_2 norm. l_2 norm is defined as the square root of the sum of the squares of components.

$$\hat{\beta} = argmin_{\beta} \sum_{i=1}^{p} (y_i - x_i\beta_i)^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \quad (7)$$

A tuning parameter, λ, adjusts the relative importance of the two terms in the above equation. Ridge regression performs better than ordinary least squares, however, it doesn't set the coefficients exactly to zero, that is, it keeps all the predictors in the model. It cannot perform variable selection.

Lasso is a method that penalizes the coefficients based on their l_1 norm.

$$\hat{\beta} = argmin_{\beta} \sum_{i=1}^{p} (y_i - x_i\beta_i)^2 + \lambda \sum_{i=1}^{p} |\beta_i| \quad (8)$$

A tuning parameter, λ, controls the strength of the penalty. The lasso does continuous shrinkage and it also performs automatic variable selection. However, it has the drawback of instability when the predictors are highly-correlated. It cannot make grouped selection. If some predictors are highly correlated, lasso selects one of them and ignores the rest. Moreover, it selects at most n variables if $p > n$.

For analyzing high-dimensional data, Zou & Hastie (2005) proposed the Elastic Net as a new regularization technique to improve lasso. It penalizes the coefficients based on both their l_1 norm and l_2 norm. The naïve elastic net is defined with Equation 9:

$$\hat{\beta} = argmin_{\beta} \sum_{i=1}^{p} (y_i - x_i\beta_i)^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} \beta_i^2 \quad (9)$$

For $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, the equation becomes;

$$\hat{\beta} = argmin_{\beta} \sum_{i=1}^{p} (y_i - x_i\beta_i)^2 + (1-\alpha) \sum_{i=1}^{p} |\beta_i| + \alpha \sum_{i=1}^{p} \beta_i^2 \quad (10)$$

The elastic net penalization factor in this equation combines lasso and ridge regression methods. The above equation is equivalent to ridge regression if $\alpha = 1$ and to lasso if $\alpha = 0$. For $0 < \alpha < 1$, it is in the form of elastic net. Elastic net yields a model that performs automatic variable selection and that can group correlated variables.

High-dimensionality is a common case for text data processing since several words are converted into a vector space model as features. This results in a large number of predictors and the need for variable selection. Moreover, correlations between predictors can be high which requires grouping effect. For these reasons, elastic net regularized regression models are used in this thesis to estimate coverage of diabetes websites.

"glmnet" library of R is used to implement elastic net regularized regression. Hastie & Junyang (2016) presents the usage of glmnet library in R. In the implementation of this package, the elastic net penalty is controlled by α and this parameter is used differently than the previous formulations. glmnet solves the problem in Equation 11:

$$min_{\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 - x_i \beta^T)^2 + \lambda[(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1] \quad (11)$$

For this formulation, $\alpha = 1$ corresponds to lasso and $\alpha = 0$ corresponds to ridge regression. The tuning parameter λ controls the strength of the penalty.

The glmnet algorithms use cyclical coordinate descent, which successively optimizes the objective function over each parameter with others fixed, and cycles repeatedly until convergence.

$\alpha = 0.5$ is selected upfront to enable the grouping effect of elastic net. Cross-validation is used to select a value for the tuning parameter, λ. Five-fold cross-validation is performed on all data set. Since the data set size is small ($n = 59$), a test set is not assigned separately. To overcome over-fitting problem, five repetitions are used. The folds are obtained by using "createFolds" function of R's "caret" library. This function takes the response vector as input and the output results in stratified folds.

The features that will be used as predictors in elastic net models are constructed as described in Section 3.5.

"cv.glmnet" function of the glmnet library is used for cross-validation. This function calculates the mean-squared error for a range of λ values. The value of λ for which the mean-squared error is minimum, λ_{min}, is an output of the function.

Initially, average λ_{min} value of different repetitions is calculated for each input data set. Using these values for λ, new models are fitted by "glmnet" function and predicted information coverage scores are obtained for only repetition 1. The predicted scores are compared with actual scores and mean-squared error and correlation between them are calculated. This initial step gives insight about which features are more effective and which models are better performing.

As a second step; for higher performing models and for each repetition, new predicted values are obtained by using "glmnet" function with λ_{min} values of the related repetition. Estimated scores are compared with actual scores and mean-squared error and correlation between actual and estimated values are acquired. Using mean-squared error as the ranking measure, significance tests are performed to investigate effects of using different corpora and different linguistic features.

3.7.Model Evaluation

Friedman's test is used to control joint statistical significance of these models. The Friedman's test is a non-parametric equivalent of the repeated-measures ANOVA. It

ranks the algorithms for each data set separately (Demšar, 2006). As a result, it returns a p-value with the null hypothesis such that all models are equivalent. Friedman's test can be utilized even when the dependent variable is not normally distributed or when it is ordinal (Marshall & Marquier).

Wilcoxon's test is a non-parametric alternative to the paired t-test, which ranks the differences in performances of two classifiers for each data set, ignoring the signs, and compares the ranks for the positive and the negative differences (Demšar, 2006). It is used for assessing pairwise statistically significant difference of models. Wilcoxon's test does not assume normal distributions. The results of this test are adjusted by Bonferroni correction.

3.8.Analysis of Significant Terms

The coefficients of terms provide insight on the importance of words. The direction and significance of the effects of a variable can be inferred directly from the sign and significance of the variables' coefficients (Zhang, Zhang, & Li, 2015). Therefore, as a final step, higher performing models are analyzed in more detail to interpret the important terms that have high coefficients and appear consistently on different data sets. For this purpose, a ranking algorithm is used to collect and investigate all of the significant terms obtained by different models together. In addition to the ranks, average coefficients are calculated for each term.

It is known that terms with a positive coefficient have a positive correlation with the response and terms with a negative coefficient have a negative correlation. In order to rank terms, firstly, for each fold and repetition, the terms are separated as the ones that have positive coefficients and the ones that have negative coefficients. Terms with zero coefficients are treated as insignificant terms and discarded. The positive and negative terms are sorted in decreasing order. Then, union of the terms are generated to yield a unique overall set. For each word in these merged sets, their ranks at each fold and repetition are summed and averaged. The rank of a word is the occurrence order in the sorted sets for the related fold and repetition. If a word does not exist in the related sorted list, its rank for that fold and repetition is taken as $(1 + length\ of\ related\ sorted\ list)$. Finally, the significant terms are ordered with respect to their calculated ranks. This approach enables to prioritize terms that are consistently found to be important with high coefficients among the folds and repetitions of a model.

Rank calculation for a representative case of two repetitions and two folds is given in Figure 4. Calculation of overall rank for "Word 3", "Word 5" and "Word 7" is illustrated. The tables in the figure show a list of words sorted with respect to their coefficients. The position in the list then becomes the rank of the word for that

repetition and fold. If the word does not exist in the list, its rank is assigned by adding one to length of that list. Ranks obtained from each table are averaged to determine the overall rank of a term. In this example, “Word 3”, with an overall rank of 3.25, is more significant than “Word 5”. “Word 7” is the least significant with an overall rank of 6.50.

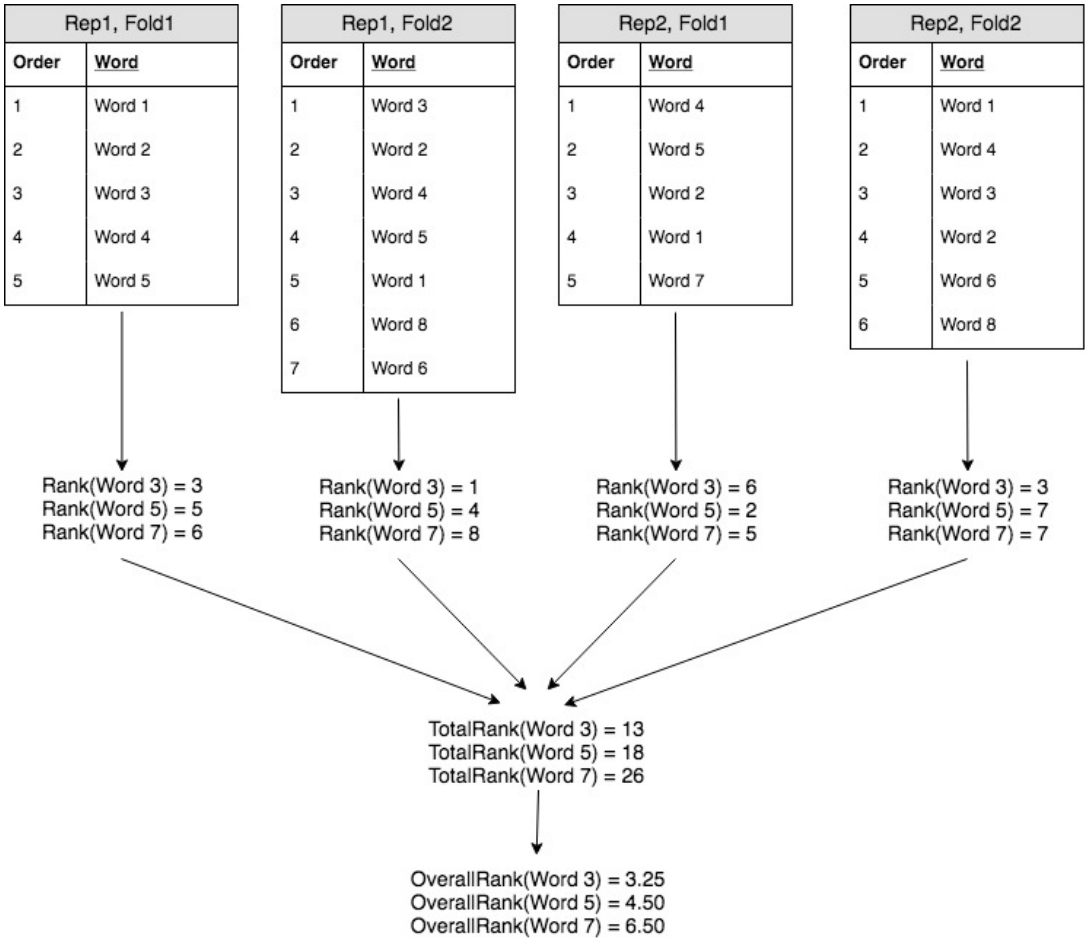


Figure 4: Calculation of Term Ranks

3.9.Summary

This chapter states the problem in more detail with examples both from data set and corpora. Moreover, the methodology applied throughout this study is introduced. Detailed information on data set and corpora is given. Chapter 4 presents the results of applying this methodology and performs discussion on the findings.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents and discusses the results of candidate term extraction and model construction to assess information coverage of diabetes websites applying the methodology described in detail in Chapter 3, Methodology. Candidate term extraction section gives information about TF-IDF scores obtained for ADA and Wikipedia corpora as well as number of candidate terms used in model construction. The results and discussions on Model Construction are divided into three sections: "Model Construction using Different Features", "Detailed Evaluation of Higher Performing Models and Their Significance" and "Evaluation of Significant Terms".

In "Model Construction using Different Features", the features used to construct 360 models that aim providing insights about features and model performance are investigated. Using knowledge gained from this inspection, higher performing models are evaluated in the sections following. In "Detailed Evaluation of Higher Performing Models and Their Significance", the models selected in the previous section and their performance are studied in detail. Inferences about effects of features, corpus size and type are performed. In "Evaluation of Significant Terms" section, examples of positive and negative significant terms are given and evaluated with examples.

4.1.Candidate Term Extraction

ADA and Wikipedia corpora are processed to extract candidate significant terms. Using TF-IDF scores of words in corpora, the words are ranked. It is explained in Chapter 3 that candidate significant term extraction is performed on three feature sets of different PoS tags to investigate the effects of different linguistic characteristics. The TF-IDF scores for different feature sets are depicted Figure 5 and Figure 6 for ADA and Wikipedia respectively.

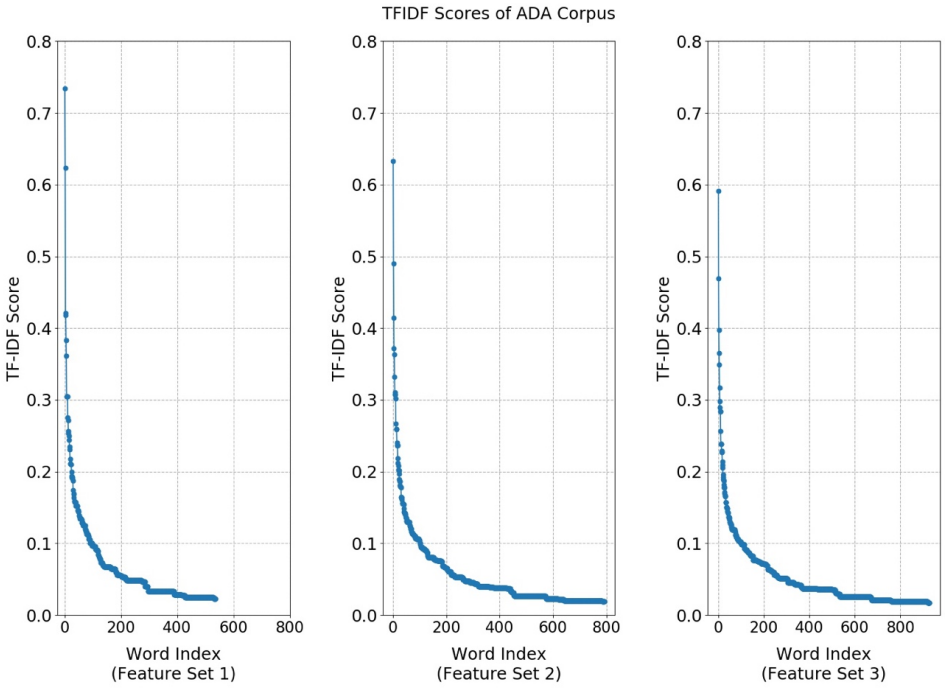


Figure 5: TF-IDF Scores of ADA corpus

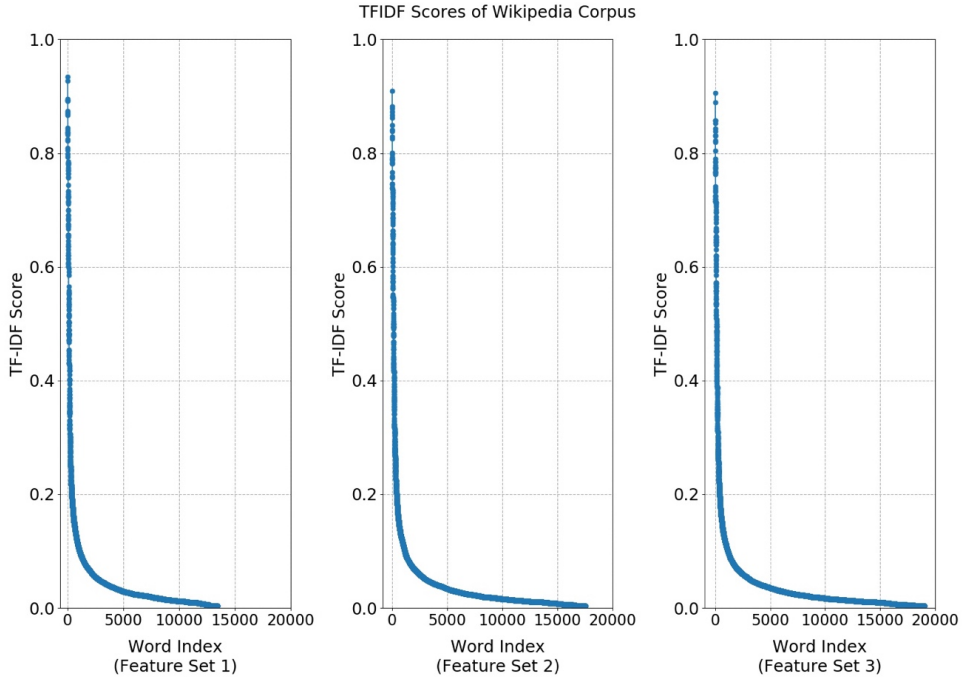


Figure 6: TF-IDF Scores of Wikipedia corpus

It is observed that TF-IDF scores drop significantly after a certain point and they do not change much after this point. For ADA corpus, this low score region covers approximately the second half of candidate words for all PoS tag sets. For Wikipedia corpus, the points where TF-IDF scores start to be lower than scores of initial candidate words correspond to approximately thirty-percent of the total number of candidate words. In light of this information, different percentages of the candidate terms that are sorted with respect to TF-IDF scores are extracted. The percentages used are 10, 20, 30, 40 and 50. This selection enables us to evaluate the effect of corpus size on estimating information coverage of websites.

Table 2 presents the number of candidate terms extracted with respect to feature sets used and percentage of the extracted candidate words. Naturally, the number of candidate terms increase as the percentages increase. Also, as context of feature sets increase, the number of candidate terms increase.

Table 2: Number of Candidate Terms

No	Candidate Term Set	ADA	Wikipedia
1	Feature set 1, Percentage: 10	54	1345
2	Feature set 2, Percentage: 10	80	1752
3	Feature set 3, Percentage: 10	93	1906
4	Feature set 1, Percentage: 20	108	2689
5	Feature set 2, Percentage: 20	159	3503
6	Feature set 3, Percentage: 20	186	3812
7	Feature set 1, Percentage: 30	161	4033
8	Feature set 2, Percentage: 30	238	5255
9	Feature set 3, Percentage: 30	279	5718
10	Feature set 1, Percentage: 40	215	5377
11	Feature set 2, Percentage: 40	317	7006
12	Feature set 3, Percentage: 40	372	7624
13	Feature set 1, Percentage: 50	268	6721
14	Feature set 2, Percentage: 50	396	8758
15	Feature set 3, Percentage: 50	464	9530

4.2.Model Construction to Assess Information Coverage of Diabetes Websites

4.2.1 Model Construction using Different Features

The features used in model construction are described in Chapter 3. The models are named in the format "a.b.c.d.e" according to Table 3.

Table 3: Model Naming

Numbering Element	**a: Corpus**	**b: Feature Set**	**c: Vectorization Method**
Possible Values	ADA: ADA WKP: Wikipedia	FS1: Feature set 1 FS2: Feature set 2 FS3: Feature set 3	VM1: Vectorization Method 1 VM2: Vectorization Method 2 VM3: Vectorization Method 3 VM4: Vectorization Method 4

Numbering Element	**d: Additional Features**	**e: Percentage** *(used as is)*
Possible Values	NO: None TW: "total word count" UW: "unique word count	10 20 30 40 50

In total, 360 models are constructed to acquire insight about which features are more effective and which models are better performing. As performance measures of models, the correlation and mean squared error (MSE) between actual and predicted values of information scores are used. In Table 4, the results for a representative case, which is Feature set 1 and Percentage 10, are tabulated. The results for all models can be seen in Appendix F.

To calculate performance measures in this section, the predicted scores are calculated at average λ_{min} and for the first repetition as explained in Chapter 3.

Table 4: Model Performance Measures

Model	Correlation	MSE
ADA.FS1.VM1.NO.10	0.763	4.982
ADA.FS1.VM1.TW.10	0.763	4.997
ADA.FS1.VM1.UW.10	0.763	4.991
ADA.FS1.VM2.NO.10	0.772	4.901
ADA.FS1.VM2.TW.10	0.772	4.901
ADA.FS1.VM2.UW.10	0.772	4.901
ADA.FS1.VM3.NO.10	0.377	10.309
ADA.FS1.VM3.TW.10	0.404	10.078
ADA.FS1.VM3.UW.10	0.462	9.441
ADA.FS1.VM4.NO.10	0.506	8.929
ADA.FS1.VM4.TW.10	0.504	8.960
ADA.FS1.VM4.UW.10	0.542	8.483
WKP.FS1.VM1.NO.10	0.792	4.617
WKP.FS1.VM1.TW.10	0.792	4.617
WKP.FS1.VM1.UW.10	0.792	4.617
WKP.FS1.VM2.NO.10	0.796	4.400
WKP.FS1.VM2.TW.10	0.796	4.400
WKP.FS1.VM2.UW.10	0.796	4.400
WKP.FS1.VM3.NO.10	0.489	9.327
WKP.FS1.VM3.TW.10	0.493	9.323
WKP.FS1.VM3.UW.10	0.548	8.685
WKP.FS1.VM4.NO.10	0.530	8.639
WKP.FS1.VM4.TW.10	0.529	8.645
WKP.FS1.VM4.UW.10	0.549	8.394

When the model outputs are examined, VM3 and VM4 perform the worst. The highest correlation obtained from VM3 and VM4 vectorization methods is 0.619 for ADA corpus when 50% of Feature set 2 is used with VM3 and “unique word count” is added as a feature. For Wikipedia corpus, the highest correlation value is 0.575 when 30% of Feature set 2 is used with VM3 and “unique word count” is added as a feature. For

these reasons, VM3 and VM4 vectorization methods can be excluded from the models for detailed analysis.

It is also recognized that additional features of "total word count" and "unique word count" are mostly not selected as important when used with VM1 and VM2 methods. Therefore, these additional features did not affect the model performance measures. To be selected, a feature needs to have a coefficient other than zero. Even for the cases where these additional features are selected, they have very little effect on model performance. For these reasons, addition of these features can be omitted in the models selected for detailed evaluation. Moreover, it is observed that VM1 and VM2 perform the best among others. These findings are also seen in Table 5 and Table 6.

Considering the aforementioned findings, detailed analysis is decided to be performed with the following properties in models:

- ADA and Wikipedia corpus
- Feature set 1, 2 and 3
- VM1 and VM2
- Percentages: 10, 20, 30, 40, 50

This reduction results in 60 models.

Table 5: Best Performing Models (ADA Corpus)

Model	Correlation	MSE
ADA.FS2.VM2.NO.30	**0.818**	**3.962**
ADA.FS2.VM2.TW.30	**0.818**	**3.962**
ADA.FS2.VM2.UW.30	**0.818**	**3.962**
ADA.FS3.VM2.NO.30	**0.810**	**4.116**
ADA.FS3.VM2.TW.30	**0.810**	**4.116**
ADA.FS3.VM2.UW.30	**0.810**	**4.116**
ADA.FS3.VM1.NO.30	**0.804**	**4.233**
ADA.FS3.VM1.TW.30	**0.804**	**4.233**
ADA.FS3.VM1.UW.30	**0.804**	**4.230**
ADA.FS2.VM1.NO.30	**0.803**	**4.257**
ADA.FS2.VM1.TW.30	**0.803**	**4.257**
ADA.FS2.VM1.UW.30	**0.803**	**4.257**

Table 6: Best Performing Models (Wikipedia Corpus)

Model	Correlation	MSE
WKP.FS1.VM1.NO.40	**0.849**	**3.639**
WKP.FS1.VM1.TW.40	**0.849**	**3.639**
WKP.FS1.VM1.UW.40	**0.849**	**3.639**
WKP.FS1.VM1.NO.50	**0.844**	**3.689**
WKP.FS1.VM1.TW.50	**0.844**	**3.689**
WKP.FS1.VM1.UW.50	**0.844**	**3.689**
WKP.FS2.VM1.NO.40	**0.835**	**3.844**
WKP.FS2.VM1.TW.40	**0.835**	**3.844**
WKP.FS2.VM1.UW.40	**0.835**	**3.844**
WKP.FS2.VM1.NO.10	**0.830**	**3.924**
WKP.FS2.VM1.TW.10	**0.830**	**3.924**
WKP.FS2.VM1.UW.10	**0.830**	**3.924**

When the model outputs are observed, it is seen that Feature set 2 (nouns and adjectives) is more significant for the ADA corpus. The second-best performing model has a lower correlation and higher MSE than the best performing model and the only change is addition of verbs in Feature set 3. Also, the third and fourth best models have almost the same measures and the only difference between these models is presence of verbs in the feature set. Therefore, we cannot conclude that verbs are significant at first glance. The best performing two models use VM2 as vectorization method and VM1 is used in third and fourth best performing models. All high-performance models of ADA corpus use 30% of the candidate terms.

For the Wikipedia corpus, nouns are perceived to be the best performing feature set and when the adjectives are added, high performance measures are still observed. These deductions and difference from ADA corpus can be due to the corpus size. Wikipedia corpus is much larger than ADA corpus and also Wikipedia corpus is unstructured compared to ADA corpus which is directly paired with scoring guideline chapters. Therefore, in Wikipedia corpus, verbs and to some extent adjectives may be too general terms and may not have a distinctive effect on output. For this larger corpus, nouns may be more determinant. All the best performing models have VM1 vectorization method. The extracted percentage of candidate terms is not observed to have a consistent effect on model outputs.

In conclusion, nouns and adjectives are more dominant than verbs. In addition, VM2 and VM1 vectorization methods are more effective in matching the models with actual scores. The extracted percentage of candidate terms does not have a consistent influence on these models. These effects are studied in more detail in the following section.

4.2.2 Detailed Evaluation of Higher Performing Models and Their Significance

In this section, the higher performing models are analyzed and compared in more detail. The reduced set of models is explained in the previous section. The selected models are listed in Table 7. The percentages used to extract candidate terms are treated as different data sets besides repetitions. Moreover, TW and UW cases will not be investigated in detail leaving out only "NO" value for "d" in model naming. Therefore, in this section, naming convention for models become "a.b.c".

Table 7: Higher Performing Models

Model No	**Model**
1	ADA.FS1.VM1
2	ADA.FS1.VM2
3	ADA.FS2.VM1
4	ADA.FS2.VM2
5	ADA.FS3.VM1
6	ADA.FS3.VM2
7	WKP.FS1.VM1
8	WKP.FS1.VM2
9	WKP.FS2.VM1
10	WKP.FS2.VM2
11	WKP.FS3.VM1
12	WKP.FS3.VM2

Table 8: Training Data Set Mean Squared Errors

Model	MSE ($mean \pm std$)				
	%10 Percentage	**%20 Percentage**	**%30 Percentage**	**%40 Percentage**	**%50 Percentage**
1	1.98 ± 0.36	1.48 ± 0.41	1.28 ± 0.41	2.00 ± 0.40	1.83 ± 0.28
2	2.62 ± 0.47	2.44 ± 0.45	2.36 ± 0.53	1.99 ± 1.07	1.19 ± 0.58
3	1.95 ± 0.30	1.03 ± 0.35	0.45 ± 0.16	0.20 ± 0.37	0.55 ± 0.34
4	2.22 ± 0.39	1.78 ± 0.62	0.05 ± 0.25	0.32 ± 0.25	0.32 ± 0.32
5	1.75 ± 0.21	0.98 ± 0.33	0.33 ± 0.14	0.92 ± 0.32	0.22 ± 0.28
6	2.34 ± 0.40	1.37 ± 0.32	0.15 ± 0.14	0.62 ± 0.46	1.76 ± 0.62
7	0.17 ± 0.84	0.01 ± 0.70	0.01 ± 0.80	0.30 ± 0.31	0.32 ± 0.34
8	0.61 ± 0.55	0.31 ± 0.36	0.75 ± 0.51	0.02 ± 0.41	0.02 ± 0.48
9	0.26 ± 0.58	0.28 ± 0.52	0.37 ± 0.38	0.78 ± 0.34	0.57 ± 0.27
10	0.62 ± 0.63	0.36 ± 0.39	0.68 ± 1.31	0.34 ± 0.07	0.04 ± 0.46
11	0.07 ± 0.77	0.09 ± 0.54	0.09 ± 0.38	0.70 ± 0.30	0.65 ± 0.36
12	0.81 ± 0.59	0.70 ± 1.23	0.67 ± 0.60	0.18 ± 0.67	0.09 ± 0.70

A detailed analysis is performed for each repetition and percentage value. Models for each repetition are constructed using λ_{min} values of related repetition. MSE values between the actual and predicted information coverage scores are used as the performance metric. Table 8 presents mean and standard deviation of MSE values obtained from training data set for all models and percentage values among folds and repetitions. Glmnet cross-validation implementation iterates over a range of λ values until one-hundred iterations are performed or percent deviation explained does not change sufficiently. For the training data set results, it is observed that some models have low average MSE values that indicate overfitting. We speculate that a high percent deviation explained may have caused this situation. To prevent this, iterations need to be terminated considering the deviation explained as well as number of iterations and convergence of percent deviation explained. Termination can be at a point where it is low enough to prevent overfitting and still at a sufficient level. Moreover, it is seen that lowest training set errors occur for models that have higher number of features. The effect of feature size can be investigated to understand this behavior. Lastly, distribution of actual scores among train and test sets may be analyzed in detail to comment on models with low training error. It is known that applying cross-validation and regularization in elastic nets would prevent overfitting

even for the case of high number of features (Orr, 1995), (Platt, 1999). The obtained results will be investigated in the future with respect to these effects.

Model performance is measured by using test sets. Mean and standard deviation of MSE values for all models and percentage values are calculated among repetitions and tabulated in Table 9.

Table 9: Model Performance

Model	**MSE ($mean \pm std$)**				
	%10 Percentage	**%20 Percentage**	**%30 Percentage**	**%40 Percentage**	**%50 Percentage**
1	4.98 ± 0.19	5.43 ± 0.17	5.79 ± 0.24	6.02 ± 0.24	5.57 ± 0.32
2	4.85 ± 0.32	5.08 ± 0.42	5.47 ± 0.39	5.77 ± 0.16	5.40 ± 0.37
3	4.56 ± 0.21	4.47 ± 0.39	4.16 ± 0.40	4.33 ± 0.33	4.76 ± 0.37
4	4.88 ± 0.47	4.83 ± 0.55	3.57 ± 0.40	4.30 ± 0.24	5.18 ± 0.47
5	4.35 ± 0.43	4.29 ± 0.53	4.22 ± 0.35	4.59 ± 0.18	4.43 ± 0.32
6	5.08 ± 0.49	4.71 ± 0.42	4.09 ± 0.31	5.50 ± 0.44	5.88 ± 0.63
7	4.53 ± 0.52	3.65 ± 0.61	4.20 ± 0.60	3.63 ± 0.51	3.66 ± 0.47
8	4.23 ± 0.65	4.33 ± 0.62	5.09 ± 0.39	4.83 ± 0.54	5.16 ± 0.52
9	3.70 ± 0.60	4.07 ± 0.44	4.29 ± 0.36	3.82 ± 0.53	3.99 ± 0.42
10	4.88 ± 0.51	4.99 ± 0.83	5.36 ± 0.66	5.14 ± 0.74	5.35 ± 0.79
11	4.09 ± 0.57	4.12 ± 0.47	4.36 ± 0.40	4.38 ± 0.52	4.64 ± 0.45
12	4.85 ± 0.43	5.78 ± 0.67	6.06 ± 0.54	5.76 ± 0.64	5.98 ± 0.63

To investigate the effect of corpus size, the mean MSE values are plotted against percentages for both corpora and the plots can be seen in Figure 7. In this figure, solid lines indicate VM1 and dashed lines indicate VM2. Square markers are for Feature set 1, circle markers are for Feature set 2 and triangle markers are for Feature set 3. Lower MSE mean values imply higher performing models.

When ADA corpus plots are investigated; it is seen that as the Percentage increase, model performance generally decreases for Feature set 1. Other feature sets mostly perform better than Feature set 1 and they have the best performing models at 30 percent. VM1 and VM2 behave in a similar trend for ADA corpus. For better

performing feature sets (FS2 and FS3); VM2 generally performs worse than VM1 except the best performing point at a percentage of 30.

When Wikipedia corpus is considered, there is a clear distinction between VM1 and VM2. VM1 perform much better than VM2. Moreover, the performance for VM1 does not change consistently as the percentage changes, whereas, for VM2, performance tends to decrease as percentage increases. For this corpus, Feature set 1 performs the best, followed by Feature set 2 and lastly Feature set 3. The difference between feature sets is much lower for VM1. These results comply with the findings of the previous section. An additional comment on these plots is that inclusion of verbs performs close to other feature sets for VM1.

It is known that Wikipedia data sets are much larger than ADA data sets. This result may indicate that as the number of candidate terms increase, VM1 becomes more significant in assessing information coverage. This may be due to the fact that disturbance effects of VM2, caused by differing website number of words, become more dominant as the number of candidate terms increase.

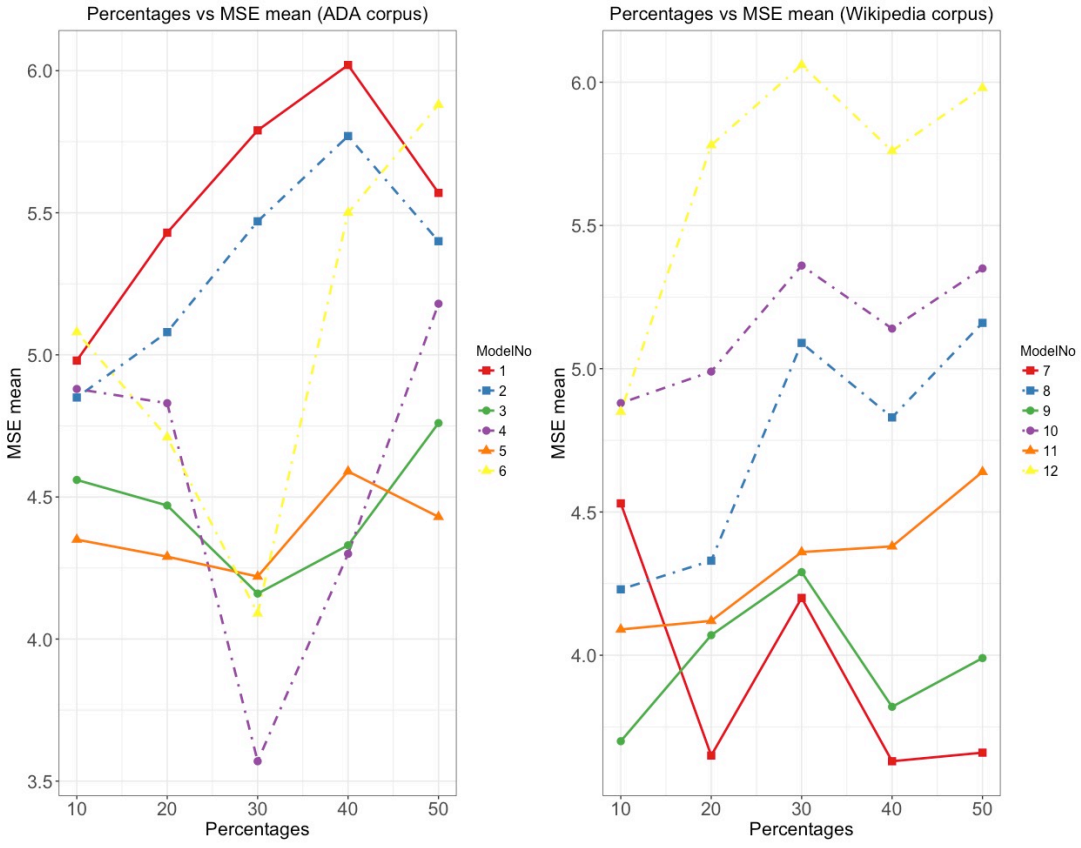


Figure 7 Percentages vs MSE mean values

In Table 9, mean and standard deviation values are calculated among five repetitions for simplicity. Actually, there are 25 data sets consisting of 5 repetitions and 5 percentage values for each model. MSE values are used to compare models statistically. For performing Friedman's test on MSE values, "friedman.test" function of R is used. Output of test is p-value < 2.2e-16. This result implies that the null hypothesis of MSE behavior being the same for all models apart from an effect of data sets can be rejected. In summary, models are statistically significantly different.

In order to compare 12 models pairwise and to decide which ones are performing better, Wilcoxon signed-ranks test is performed by "wilcox.test" function of R. Bonferroni adjusted pairwise p-values are presented in Table 10. Bonferroni correction multiplies the values obtained from Wilcoxon test by the number of all possible pairs and sets an upper limit of one since p-value can be between 0 and 1. In the table, the p-values that do not indicate statistically significant difference for a confidence level of 0.95 (alpha=0.05) are marked red (p>0.05 case). The remaining region, shown in black, indicate statistically significantly different models and these can be used to compare models.

Table 10 Wilcoxon Signed-Ranks Test Results

2	3	4	5	6	7	8	9	10	11	12	Model No
1.000	**0.000**	**0.013**	**0.000**	0.595	**0.000**	1.000	**0.000**	1.000	**0.000**	1.000	**1**
	0.000	**0.001**	**0.000**	**0.040**	**0.000**	0.085	**0.000**	1.000	**0.000**	1.000	**2**
		1.000	1.000	0.054	1.000	**0.009**	1.000	**0.000**	1.000	**0.000**	**3**
			0.280	1.000	1.000	1.000	1.000	**0.004**	1.000	**0.000**	**4**
				0.013	1.000	**0.001**	1.000	**0.000**	1.000	**0.000**	**5**
					1.000	1.000	0.107	0.896	1.000	**0.009**	**6**
						0.466	1.000	**0.000**	1.000	**0.000**	**7**
							0.046	1.000	0.754	**0.017**	**8**
								0.000	1.000	**0.000**	**9**
									0.000	1.000	**10**
										0.000	**11**

Different models are analyzed according to Wilcoxon test results table pairwise and if there is a significant difference, MSE values of 25 data sets are examined to decide which model is performing better.

When the difference between VM1 and VM2 are considered, the following models are compared: 1 and 2, 3 and 4, 5 and 6, 7 and 8, 9 and 10, 11 and 12. Among these

comparisons, the significant ones are 5 and 6, 9 and 10, 11 and 12. Models 5, 9 and 11 yield better results when these comparisons are performed according to MSE values. This implies that VM1 performs better than VM2, hence, binary weighting can be selected against term frequencies.

To investigate the effect of linguistic features using PoS tags, ADA corpus models 1, 3, 5 and Wikipedia corpus models 7, 9 and 11 are compared. Among ADA corpus, model 5 performs the best, followed by 3 and lastly 1. We can interpret that Feature set 3 set performs the best followed by Feature set 2. This implies that using diverse linguistic features improves the estimation accuracy. However, Wikipedia corpus models does not cause a significant difference with respect to PoS tags when VM1 is used. Hence, we can conclude that performances of models that use Wikipedia corpus and VM1 are not affected significantly by PoS tags. This may be an indication that keeping corpus size large also improves the estimation accuracy. This is achieved by using diverse linguistic features in ADA and making use of many related webpages in Wikipedia. When VM2 is considered and models 2, 4, 6 and 8, 10, 12 are compared; the results do not change.

Lastly, to examine the effect of corpus and corpus size, the following models are compared: 1 and 7, 3 and 9, 5 and 11. Model 7 performs better than model 1. However, models 3 and 9 are not significantly different just as models 5 and 11. This shows that for Feature set 1 (nouns), Wikipedia is more effective in estimating information coverage. This may be due to larger Wikipedia corpus that enables a higher domain coverage. Addition of adjectives and verbs does not bring a significant change with respect to corpora used.

4.2.3 Analysis of Significant Terms

In the models, the terms with positive coefficients are positively correlated to information coverage scores, meaning as they increase, the score also increases. These are called positive terms. Similarly, negative terms have negative coefficients and they are negatively correlated to information coverage scores. For each model, the terms are ranked using these coefficients as described in Section 3.8. This ranking algorithm enables us to assign higher ranks to the terms that consistently have high coefficients among models.

Model 5 is one of the best performing models and it performs the best at a percentage of 30%. Twenty-five percent of the positive and negative significant terms obtained for this condition are analyzed in this section to be representative in Table 11 and Table 12. The meaning of the terms and/or reason of their significance are interpreted and example phrases indicating their usage are provided. List of all positive and negative significant terms for the aforementioned model are given in Appendix G.

Table 11: Analysis of Representative Positive Terms

Rank	Positive term	Average Coefficient	Evaluation
1	inhibitor	1.259	A commonly used medical term, takes part in names of drugs for diabetes Example phrases: “DPP-4 inhibitors”, “SGLT2 inhibitors”, “alpha-glucosidase inhibitors”
2	acting	1.038	A commonly used medical term, takes part in names of insulin types used for diabetes treatment Example phrases: “rapid-acting insulin”, “short-acting insulin”, “fast-acting insulin”
3	observational	1.742	A supporting term Example phrase: “Observational study”
4	bariatric	0.770	A medical term related to obesity management for treatment of type 2 diabetes and to topic 3: “Bariatric Surgery” Example phrases: “bariatric surgery”
5	gain	0.838	A supporting term related to topic 1: “Diet, Physical Activity, and Behavioral Therapy” Example phrases: “weight gain”
6	physical	0.722	A supporting term related to topic 1: “Diet, Physical Activity, and Behavioral Therapy” Example phrases: “physical exercise”, “physical activity”
7	rapid	0.642	A medical term that takes part in a type of insulin used for diabetes treatment Example phrases: “rapid-acting insulin”
8	therapy	0.450	A commonly used supporting term related to diabetes treatment Example phrases: “insulin therapy”, “combination therapy”, “initial therapy”

Table 11 (continued)

9	provide	0.550	A general term that can be used in various ways Example phrase: “healthcare provider”, “…provide a steady level of insulin…”
10	frequent	0.504	A general term used mostly for side effects of treatments Example phrases: “frequent urination”, “frequent infections”
11	medication	0.464	A commonly used supporting term related to diabetes treatment. Mostly used alone as a noun. Example phrases: “concomitant medication”, “alternative medication”
12	injection	0.430	A medical term that takes part in insulin therapy related text Example phrases: “insulin injection”, “injection pen”
13	alternative	0.474	A term to explain treatment process or treatment options Example phrases: “alternative medication”, “as an alternative to…”, “diabetes medication”
14	food	0.344	A supporting term related to topic 1: “Diet, Physical Activity, and Behavioral Therapy” used to give advice on diet Example phrases: “salty foods”, “fried foods”, “foods that you should limit…” Also used for treatment instructions Example phrases: “after food”, “with food”
15	treatment	0.532	A commonly used supporting term related to diabetes treatment. Example phrases: “insulin treatment”, “your treatment”, “treatment of type 2 diabetes”, “treatment includes…”, “treatment option”, “treatment plan”

Table 11 (continued)

16	schedule	0.403	A supporting term related to treatment process Example phrases: “dosing schedule”, “meal schedule”
17	flexibility	0.410	A supporting term related to treatment process Example phrases: “increase/improve flexibility”, “flexibility around meal timing”
18	mol	0.514	Unit of measures related to diabetes monitoring Example phrase: mmol/mol (for HbA1c)
19	pioglitazone	0.344	A medicine name used for diabetes treatment Example phrases: “Thiazolidinediones Pioglitazone”, “Pioglitazone (Actos)”
20	group	0.336	A general term that can be used in various ways. May not have a semantic indication. Example phrases: “food group”, “age group”, “group of medicines”
21	cap	0.618	A term to explain treatment process or instructions Example phrases: Short for “capsule”
22	consider	0.338	A term to explain treatment process or instructions Example phrases:
23	twice	0.318	A term to explain treatment process or instructions Example phrases: “twice a day”, “twice a year” Also, to give statistics. Example phrase: “…twice as likely…”
24	combination	0.315	A term to explain treatment process or instructions Example phrases: “combination of fruits and vegetables”, “combination therapy”, “combination pills”

Table 11 (continued)

25	approve	0.318	A word to indicate status of medications Example phrases: “FDA approves…”, “…approved by FDA”
26	session	0.295	A supporting term related to topic 1: “Diet, Physical Activity, and Behavioral Therapy” used to give advice on diet Example phrase: “exercise session”
27	intestine	0.208	A term to explain treatment process or instructions Example phrases: “…digestion in the small intestine”, “…breakdown of some sugars in the intestines”
28	target	0.331	A term to explain treatment process or instructions Example phrases: “target blood glucose level”, “target weight”
29	gastric	0.243	A supporting term related to diabetes treatment. Example phrase: “gastric dyspepsia” as side effects of a medicine Also, a medical term related to obesity management for treatment of type 2 diabetes and to topic 3: “Bariatric Surgery” Example phrase: “gastric bypass surgery”
30	effect	0.279	A general term that can be used in various ways. May not have a semantic indication. Example phrases: “long-term effects”, “side effects”
31	version	0.273	A general term that can be used in various ways. May not have a semantic indication. Example phrase: “version of a diet”, “synthetic version of amylin”

Table 12: Analysis of Representative Negative Terms

Rank	Negative term	Average Coefficient	Evaluation
1	present	-0.715	A general term that can be used in various ways. May not have a semantic indication. Example phrases: “if liver disease is present”, “the insulin that is present”
2	thiazolidinedione	-0.426	A medicine name used for diabetes treatment Example phrases: This term is present in sections about diabetes treatment and medicine types The scores of webpages that include this term are 10 for 4 pages, 9 for 5 pages, 7 for 1 page, 6 for 1 page and 5 for 2 pages. Even though this term is not seen consistently on low score pages, it is marked as a negative term.
3	concentrated	-0.723	1. A term to explain medicine properties Example phrases: “concentrated form of insulin glargine” (Webpage id:17, score:4), “lantus in its concentrated form1 (Webpage id: 46, score:10) 2. However, a different occurrence is detected related to dairy consumption: “concentrated animal feeding operations” that do not appear in guidelines. (Webpage id:35, score:3) These are all of the occurrences of this word in data set. The scores of webpages that include this term are 10 for 1 page, 4 for 1 page and 3 for 1 page. Considering the lower frequency of low scores in the whole data set, such an appearance on mostly low score pages may indicate a negative term.

Table 12 (continued)

4	cardiovascular	-0.382	A medical term used generally when explaining side effects or risks Example phrases: “body's cardiovascular and metabolic systems”, “cardiovascular problems”, “cardiovascular disease” Binary weighted vectorization method (VM1) vs Information coverage score for this term is plotted in Figure 8. This figure fails to give a consistent indication about the term’s positive or negative effect on information coverage score.
5	accumulate	-0.418	A supporting term that explains a biological process Example phrases: “glucose accumulates in the blood”, “excess sugar accumulates in the blood and urine” The scores of webpages that include this term are 10 for 1 page, 9 for 1 page and 3 for 2 pages. Considering the lower frequency of low scores in the whole data set, such an appearance on mostly low score pages may indicate a negative term.
6	intolerance	-0.253	A medical term generally used when explaining side effects or risks Example phrases: “glucose intolerance”, “intolerance of dairy or gluten” This phrase is not seen in another form. The scores of webpages that include this term are 10 for 1 page, 9 for 2 pages, 4 for 1 page, 3 for 1 page and 1 for 1 page. Considering the lower frequency of low scores in the whole data set, such an appearance on mostly low score pages may indicate a negative term.

Table 12 (continued)

7	death	-0.281	A supporting term used with explaining side effects or risks The term is also used to present statistical information. Example phrases: “according to death certificate data, diabetes contributed to the deaths of…”, “increased risk of cardiovascular events and death” The scores of webpages that include this term are 12 for 1 page, 10 for 5 pages, 9 for 3 pages, 5 for 3 pages, 4 for 1 page, 3 for 2 pages, 2 for 2 pages, 1 for 1 page and 0 for 1 page Considering the lower frequency of low scores in the whole data set, such an appearance among webpages may indicate a negative term.
8	fig	-1.063	1. A supporting term related to topic 1: “Diet, Physical Activity, and Behavioral Therapy” Example phrase: “…dried apricots, figs or other dried fruit…” (Webpage id: 17, score:4) 2. Abbreviation for “figure” (Webpage id: 42, score:10) These usages may not have a semantic meaning.
9	requirement	-0.285	A term to explain treatment process or instructions Example phrase: “insulin requirements”, “requirements for extra food”, “legal requirements” The scores of webpages that include this term are 12 for 1 page, 10 for 3 pages, 9 for 5 pages, 7 for 2 pages, 5 for 1 page and 3 for 2 pages. Although this term is seen much on high score pages, it is selected as a negative term.

Table 12 (continued)

10	mortality	-0.189	A supporting term used with explaining side effects or risks Example phrases: “risk factor for mortality as smoking”, “…heart disease and higher mortality rates…” The scores of webpages that include this term are 10 for 1 page, 9 for 2 pages, 7 for 1 page, 3 for 1 page and 1 for 1 page. Considering the lower frequency of low scores in the whole data set, such an appearance among webpages may indicate a negative term.
11	min	-0.292	1. A term to explain treatment process or instructions Example phrases: “15 min before meal”, “30 mins” 2. Also, a different occurrence is detected. For Webpage id:24, “met-FOR-min” is written in parenthesis near “metformin”. This results in three separate words as: met, for and min. Therefore, the term “min” is also counted as occurred for this page. The scores of webpages that include this term are 11 for 1 page, 10 for 1 page, 7 for 2 pages and 5 for 1 page. Considering the lower frequency of low scores in the whole data set, such an appearance among webpages may indicate a negative term.

For some negative terms, terms with the same meaning may exist in the candidate term set. For example, “minimum” exists which may have similar meaning as “min”. Also, “tzd” used as an abbreviation of “thiazolidinediones” exists in the candidate term set as well as “thiazolidinediones”. The semantic effects of these terms are not handled together which is a limitation of this study. If these terms could be joined, the results may differ and be more accurate.

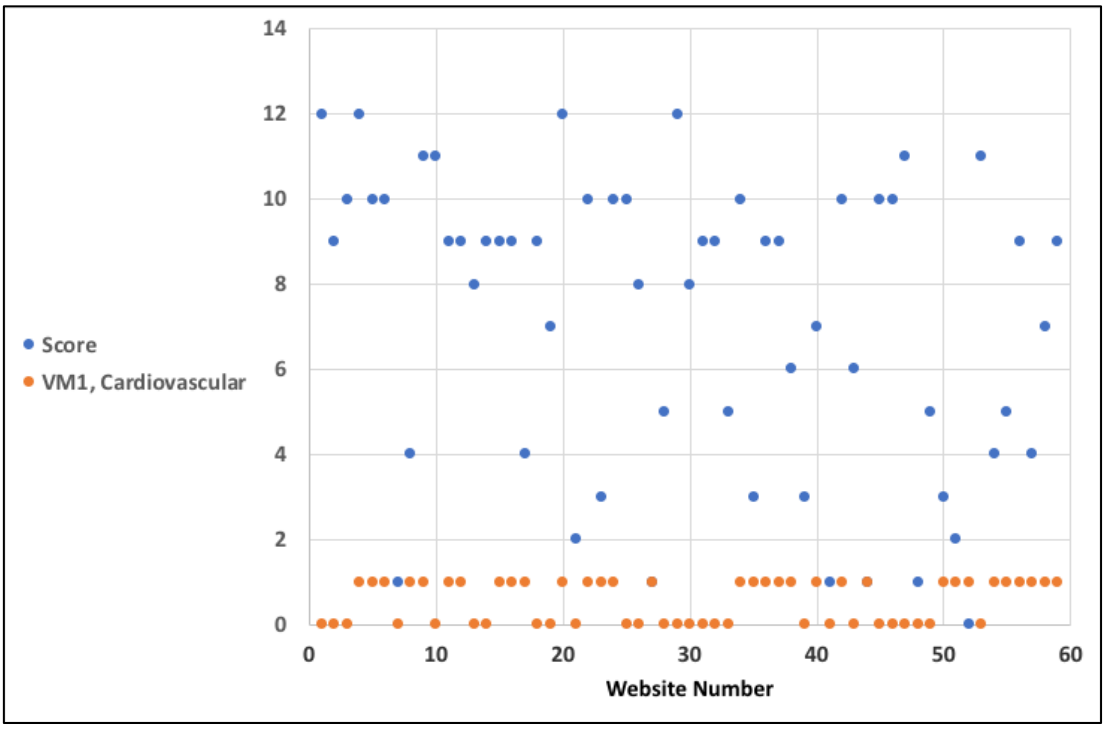


Figure 8: Information Coverage Score and VM1 values for the term, "Cardiovascular"

4.3.Summary

This chapter presents the results of the study and explores the findings in three subsections: Firstly, model construction using different features is studies. Insights are gained and higher performing models are detected for detailed analysis. Secondly, detailed evaluation of higher performing models as well as their significance assessment is performed. This way, models and features can be compared and deductions can be made. Finally, significant term analysis is performed by exploring extracted salient words. The following chapter, Chapter 5, concludes the report by summarizing the content and listing answers to research questions and contributions of the thesis. Finally, recommendations for future work are elaborated.

CHAPTER 5

CONCLUSION AND FUTURE WORK

An automated approach to estimate information coverage of type 2 diabetes websites is presented in this thesis. Information coverage is a measure of the extent of addressing necessary issues in a context. It is a metric used for information quality of documents because in order for a content to be useful, it must be complete. Many studies treat coverage as an integral part of accuracy (Eysenbach, Powell, Kuss, & Sa, 2002).

In an era of digital information, both the content and usage of Internet is increasing rapidly. As a consequence, people frequently search for health information online. However, because all websites are not prepared by medical experts, some pages may be misleading for online health seekers. They can even be harmful for users such that patients may discontinue their current treatments or start following unsupported treatments.

Mining quality terms is a fundamental task in natural language processing. In order to deduce knowledge about information quality of a website, identifying terms that are representative of a domain can provide insight about the domain as well as the data set. Moreover, they can be quantified to estimate quality measures. This study is based on mining quality terms to assess information coverage of websites.

For the purpose of automatically assessing information coverage of websites, a complete data mining approach is adopted to reach information starting from raw data. The applied methodology covers the following main data mining steps:

- Data collection
- Preprocessing of collected data
- Feature construction
- Feature selection
- Model construction

- Model evaluation
- Information retrieval on domain of interest

The presented study contributes in the following ways:

1. An automated approach to estimate information coverage of type 2 diabetes websites is provided. A comprehensive data mining approach is implemented to achieve this task. The ability to perform this methodology automatically, indicates that it can be easily and directly adapted for use in other domains.

2. Utilization of knowledge bases in the process of information coverage assessment is investigated. It is observed that knowledge bases provide factual information that can be used as a baseline to evaluate content in its domain.

3. Two knowledge bases are studied: ADA, which is a structured, formal and detailed document on diabetes domain, and Wikipedia, which is not formal but is detailed like ADA. The main difference between these two corpora when realizing the study was that ADA chapters are directly used to assign topics to be evaluated for coverage. This might have caused a dependency in the results to favor ADA. Nonetheless, no significant difference between knowledge bases is detected especially when the features are kept diverse. Constructing a large Wikipedia corpus may have been useful to eliminate this drawback.

4. Significant terms are explored to gain insight on domain-specific information. These terms are considered representative terms of the domain. In the scope of automatic website quality assessment, further studies can be carried out to generate phrases using these terms or to investigate the structure of phrases around these terms.

5. To comment on the usability and scalability of knowledge base usage, we state that a formal guideline such as ADA may not be present for other domains. This makes Wikipedia advantageous since it is a readily available and extensive open source. One drawback of Wikipedia over ADA is that it is computationally less effective because of the large corpus size.

6. Linguistic features are investigated in the scope of this thesis. For the larger corpus, Wikipedia, linguistic features do not have a significant effect on model performance. For ADA corpus, it is observed that using a diverse set of linguistic features improves the results. These findings are interpreted such that keeping a large corpus size can improve estimation accuracy.

7. Vectorization methods for candidate terms are studied. It is inferred that binary weighting performs better than term frequencies.

The most important limitation of this study was extracting only unigrams. Extending this study to include n-grams and identifying important phrases in addition to words may be more explanatory in understanding the domain and improve the results of models to perform information coverage evaluation.

Ambiguation also creates a limitation. For example; “min” is a term that may have a meaning of “minimum” or “minute” both of which may be important terms for describing treatment process. If these terms could be differentiated by performing contextual analysis, the results may differ and be more accurate.

There are also limitations due to lemmatization and PoS tagging errors. Inconsistent lemmatization or assigning wrong PoS for words results in errors in evaluation of linguistic features.

Although knowledge bases provide many advantages such as automating the process by enabling extraction of factual information, the need for a trustworthy knowledge base can be a limitation. For this reason, Wikipedia content is also studied as an easily accessible open source and found to generate comparable results.

Another limitation was discarding numerical terms. Some numerical terms such as limits of medical measurements to decide a treatment method may be distinctive for this domain. Incorporating numerical terms in feature construction and investigating their role on information coverage may prove beneficial for future studies.

One other limitation of the study was usage of medical terms in different forms on websites. For example; “DPP4”, “DPP-4”, “DPP-IV”, “dipeptidyl peptidase-4 inhibitor” and “dipeptidyl peptidase-4” are all different expressions of the same term. Similarly, “sulfonylureas” and “sulphonylureas” are both used to refer to the same term. As a future study, a method that detects these different forms and treats them the same can improve the results.

As a future work, the findings of this study can be utilized to further develop automated methodologies for checking correctness of online health information on treatment of

type 2 diabetes. This research uses a bag-of-words approach and does not take the position of words into account. This approach removes the necessity of structured content. However, both the independent and relative positions of words in text is also distinctive since it introduces a semantic meaning. This approach would be especially useful for deduction of information correctness, that is, fact-checking, which may be the next research topic after identifying significate terms. The structure of text around the extracted salient terms can act as a baseline for such studies.

Lastly, future work can be carried out to study overlapping content of websites. Similarities of website data can be analyzed to detect whether a website has original and unique content or it is copied from another website.

REFERENCES

Ölçer, D. (2018). *A framework for automatic assessment of web-based information quality to support patient decision making on health.* Middle East Technical University, Information Systems, Ankara.

American Diabetes Association. (2016). Standards of medical care in diabetes - 2016. *Diabetes Care, 39*(Suppl. 1).

Association, A. P. (2001). 200031: Criteria for assessing the quality of health information on the Internet. *American journal of public health, 91*(3), 513.

Astrakhantsev, N. A., Fedorenko, D. G., & Turdakov, D. Y. (2015). Methods for automatic term recognition in domain-specific text collections: A survey. *Programming and Computer Software, 41*(6), 336-349.

Banasiak, N. C., & Meadows-Oliver, M. (2017). Evaluating asthma websites using the Brief DISCERN instrument. *Journal of asthma and allergy, 10*, 191-196.

Baur, C., & Prue, C. (2014, September). The CDC Clear Communication Index is a new evidence-based tool to prepare and review health information. *Health Promotion Practice, 15*(5), 629-637.

Belen Saglam, R., & Taskaya Temizel, T. (2015). A framework for automatic information quality ranking of diabetes websites. *Informatics for Health and Social Care, 40*(1), 45-66.

Boyer, C., & Dolamic, L. (2014). Feasibility of automated detection of HONcode conformity for health-related websites. *IJACSA, 5*(3), 69-74.

Cerminara, C., Santarone, M. E., Casarelli, L., Curatolo, P., & El Malhany, N. (2014). Use of the DISCERN tool for evaluating web searches in childhood epilepsy. *Epilepsy & behaviour, 41*, 119-121.

Charnock, D. (1998). *The DISCERN Handbook. Quality criteria for consumer health information on treatment choices.* Oxford, UK: Radcliffe Medical Press Ltd.

Charnock, D., Shepperd, S., Needham, G., & Gann, R. (1999). DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of Epidemiology & Community Health, 53*(2), 105-111.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. *PLoS one, 10*(6), e0128193.

Conrado, M. d., Pardo, T. A., & Rezende, S. O. (2013). A machine learning approach to automatic term extraction using a rich feature set. *Proceedings of the 2013 NAACL HLT Student Research Workshop*, (pp. 16-23). Atlanta, Georgia.

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*(7), 1-30.

Devine, T., Broderick, J., Harris, L. M., Wu, H., & Hilfiker, S. W. (2016). Making quality health websites a national public health priority: toward quality standards. *Journal of Medical Internet Research, 18*(8).

Dueñas-Garcia, O. F., Kandadai, P., Flynn, M. K., Patterson, D., Saini, J., & O'Dell, K. (2015). Patient-focused websites related to stress urinary incontinence and pelvic organ prolapse: a DISCERN quality analysis. *International urogynecological association, 26*(6), 875-880.

Dutta-Bergman, M. J. (2004). The impact of completeness and web use motivation on the credibility of e-health information. *Journal of Communication, 54*(2), 253-269.

Elwyn, G., O'Connor, A., Stacey, D., Volk, R., Edwards, A., Coulter, A., . . . Holmes-Rovner, M. (2006). Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. *Bmj, 333*(7565), 417.

Evidence-Based Medicine Working Group. (1992). Evidence-based medicine. A new approach to teaching the practice of medicine. *Jama, 268*(17), 2420.

Eysenbach, G., Powell, J., Kuss, O., & Sa, E.-R. (2002, May 22/29). Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. *Jama, 287*(20), 2691-2700.

Fast, A. M., Deibert, C. M., Hruby, G. W., & Glassberg, K. I. (2013). Evaluating the quality of Internet health resources in pediatric urology. *Journal of pediatric urology, 9*(2), 151-156.

Fox, S., & Duggan, M. (2013). *Health Online 2013.* Washington, DC: Pew Research Center's Internet & American Life Project.

Graves, L. (2018). *FACTSHEET: Understanding the promise and limits of automated fact-checking.* Reuters Institute for the Study of Journalism.

Griffiths, K. M., & Christensen, H. (2005). Website Quality Indicators for Consumers. *Journal of Medical Internet Research, 7*(5).

Griffiths, K. M., Tang, T. T., Hawking, D., & Christensen, H. (2005). Automated assessment of the quality of depression websites. *Journal of Medical Internet Research, 7*(5).

Guan, J. (2016). *A study of the use of keyword and keyphrase extraction techniques for answering biomedical questions.* Sydney: Macquire University.

Hastie, T., & Junyang, Q. (2016, September 13). *Glmnet vignette.* Retrieved 2018, from http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf

Jing, L. P., Huang, H. K., & Shi, H. B. (2002). Improved feature selection approach TFIDF in text mining. *Proceedings of the First International Conference on Machine Learning and Cybernetics* (pp. 944-946). Beijing: IEEE.

Kaicker, J., Debono, V. B., Dang, W., Buckley, N., & Thabane, L. (2010). Assessment of the quality and variability of health information on chronic pain websites using the DISCERN instrument. *BMC Medicine, 8*(1), 59.

Khazaal, Y., Chatton, A., Cochand, S., & Zullino, D. (2008). Quality of web-based information on cocaine addiction. *Patient education and counseling, 72*(2), 336-341.

Khazaal, Y., Chatton, A., Cochand, S., Coquard, O., Fernandez, S., Khan, R., . . . Zullino, D. (2009). Brief DISCERN, six questions for the evaluation of evidence-based content of health-related websites. *Patient education and counselling, 77*(1), 33-37.

Khazaal, Y., Chatton, A., Cochand, S., Jermann, F., Osiek, C., Bondolfi, G., & Zullino, D. (2008). Quality of web-based information on pathological gambling. *Journal of gambling studies, 24*(3), 357-366.

Khazaal, Y., Chatton, A., Zullino, D., & Khan, R. (2012). HON label and DISCERN as content quality indicators of health-related websites. *Psychiatric Quarterly, 83*(1), 15-27.

Khazaal, Y., Fernandez, S., Cochand, S., Reboh, I., & Zullino, D. (2008). Quality of web-based information on social phobia: a cross-sectional study. *Depression and anxiety, 25*(5), 461-465.

Kiran, D. P., Bargale, S., Pandya, P., Bhatt, K., Barad, N., Shah, N., . . . Ramesh, K. (2015). Evaluation of Health on the Net seal label and DISCERN as content quality indicators for patients seeking information about thumb sucking habit. *Journal of pharmacy & bioallied sciences, 7*(Suppl 2), S481.

L Sackett, D., Rosenberg, W., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ, 312*, 71-72.

Lee, S., & Kim, H.-j. (2008). News keyword extraction for topic tracking. *Networked Computing and Advanced Information Management* (pp. 5544-559). Fourth International Conference on .

Li, W., & Chen, H. (n.d.). Retrieved from Artificial Intelligence Laboratory, The University of Arizona: https://ai.arizona.edu/sites/ai/files/resources/logistic_regression_and_elastic_net.pptx

Marafino, B. J., Boscardin, W. J., & Dudley, R. A. (2015). Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *Journal of Biomedical Informatics, 54*, 114-120.

Marshall, E., & Marquier, B. (n.d.). *Friedman test in R.* Statstutor Community Project, Sheffield University.

McCool, M. E., Wahl, J., Schlecht, I., & Apfelbacher, C. (2015). Evaluating Written Patient Information for Eczema in German: Comparing the Reliability of Two Instruments, DISCERN and EQIP. *PloS one, 10*(10), e0139895.

Morel, V., Chatton, A., Cochand, S., Zullino, D., & Khazaal, Y. (2008). Quality of web-based information on bipolar disorder. *Journal of affective disorders, 110*(3), 265-269.

Moult, B., Franck, L. S., & Brady, H. (2004). Ensuring quality Information for patients: development and preliminary validation of a newinstrument to improve the quality of written healthcare information. *Health Expectations, 7*(2), 165-175.

National Health Service England. (n.d.). Retrieved August 2018, from The Information Standard About: https://www.england.nhs.uk/tis/about/

Nilsson-Ihrfelt, E., Fjallskog, M.-L., Blomqvist, C., Ahlgren, J., Edlund, P., Hansen, J., . . . Andersson, G. (2004). Breast cancer on the Internet: the quality of Swedish breast cancer websites. *The Breast, 13*(5), 376-382.

Orr, M. J. (1995). Regularization in the selection of radial basis function centers. *Neural computation, 7*(3), 606-623.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers, 10*(3), 61-74.

Rees, C. E., Ford, J. E., & Sheard, C. E. (2002). Evaluating the reliability of DISCERN: a tool for assessing the quality of written patient information on treatment choices. *Patient education and counselling, 47*(3), 273-275.

Saeed, F., & Anderson, I. (2017). Evaluating the quality and readability of Internet information on meningiomas. *World Neurosurgery, 97*, 312-316.

Sandoval, A. M., Diaz, J., Llanos, L. C., & Redondo, T. (2018). Biomedical term extraction: NLP techniques in computational medicine. *International Journal of Interactive Multimedia and Artificial Intelligence*.

Sarkar, K., Nasipuri, M., & Ghose, S. (2010). A New Approach to Keyphrase Extraction Using Neural Networks. *International Journal of Computer Science, 7*(2), 16-25.

Sondhi, P., Vydiswaran, V. G., & Zhai, C. (2012). Reliability prediction of webpages in the medical domain. *European conference on information retrieval* (pp. 219-231). Berlin, Heidelberg: Springer.

Teisseyre, P. (2017). CCnet: Joint multi-label classification and feature selection using classifier chains and elastic net regularization. *Neurocomputing, 235*, 98-111.

Thakurdesai, P. A., Kole, P. L., & Pareek, R. P. (2004). Evaluation of the quality and contents of diabetes mellitus patient education on Internet. *Patient Education and Counseling, 53*(3), 309-313.

Wang, F., Lan, M., & Wu, Y. (2017). ECNU at SemEval-2017 Task 8: Rumour evaluation using effective features and supervised ensemble models. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)* (pp. 491-496). Vancouver, Canada: Association for Computational Linguistics.

Yuan, Y., Gao, J., & Zhang, Y. (2017). Supervised learning for robust term extraction. *Asian Language Processing (IALP), 2017 International Conference*, 302-305.

Zhang, Z., Gao, J., & Ciravegna, F. (2018). SemRe-Rank: Improving Automatic Term Extraction by Incorporating Semantic Relatedness with Personalised PageRank. *ACM Transactions on Knowledge Discovery from Data (TKDD), 12*(5), 57.

Zhang, Z., Iria, J., Brewster, C., & Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms.

Zhang, Z., Zhang, Z., & Li, H. (2015). Predictors of the authenticity of Internet health rumours. *Health Information and Libraries Journal*(32), 195-205.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR), 51*(2), 32.

APPENDICES

APPENDIX A

MANUAL ASSESSMENT OF INFORMATION COVERAGE

Coverage scores are addressed under six topics:

1. Diet, Physical Activity, and Behavioral Therapy
2. Pharmacotherapy
3. Bariatric Surgery
4. Initial Therapy
5. Combination Therapy
6. Insulin Therapy

These topics are generated using subsections of ADA's gold standard (American Diabetes Association, 2016) chapters as explained in detail in Chapter 3.

A guideline, prepared by using ADA's gold standard, is provided to domain experts to inform them about the issues that need to be considered for assigning a score. This guideline is elaborated in Figure 9 - Figure 11. The topics used are highlighted in yellow in the figures. Experts are expected to use this guideline as well as ADA content and their expertise to make an assessment.

As an example, the main text content of a low-quality website with information coverage score of 1 is presented in Figure 12 for page 1 and Figure 13 for page 2 of the website (webpage id=7). The figures depict main text content of the website. The content that is not shown in the figures include subsections that consist of only links. They are under headings such as, "latest news", "symptoms", "diagnosis and tests", "prevention and risk factors", "treatments and therapies", "living with", "related issues", "specifics", "genetics", "statistics and research", "clinical trials". Some of these links are internal and some of them are external. The website is scored according to its main text content.

The website receives a total score of one (1) which is for the "Diet, Physical Activity, and Behavioral Therapy" section. The discussion of the expectations with respect to topics are given below:

1. Diet, Physical Activity, and Behavioral Therapy: The website mentions that type 2 diabetes risk increases for people who are old or obese, has a family history of diabetes or do not exercise. It also states that meal planning and physical activity improves the condition, however; it fails to address specific instructions or guidelines that are present in the guideline and EBM.
2. Pharmacotherapy: The medication options for overweight or obese patients are expected to be covered for this section.
3. Bariatric Surgery: Surgery options for severe obesity and conditions to undergo operation are anticipated to be mentioned for this topic.
4. Initial Therapy: Information about lifestyle changes as well as initial medicine use as monotherapy are expected to address this topic.
5. Combination Therapy: For this section, the following information is expected: combination therapy such as dual or triple medication therapy, the conditions to decide which one to follow and advantages as well as disadvantages of these therapies
6. Insulin Therapy: This topic is expected to cover types of insulin therapy and their usage criteria, usage methodology as well as their advantages and disadvantages.

The website fails to cover topics 2-6 explained above. They touch upon some issues; however, it is further away from being informative. The website uses short and non-specific sentences overall. It fails to provide evidences and present facts. We should note that this is the case for evaluation of main text content on websites. A study to include evaluation of the content at links the websites provide would return different results.

OBESITY MANAGEMENT FOR THE TREATMENT OF TYPE 2 DIABETES (section 6)

Treatment for overweight and obesity in type 2 diabetes

Treatment	BMI category (kg/m²)				
	23.0* or 25.0–26.9	27.0–29.9	30.0–34.9	35.0–39.9	≥40
Diet, physical activity, and behavioral therapy	‡	‡	‡	‡	‡
Pharmacotherapy		‡	‡	‡	‡
Bariatric surgery				‡	‡

- **Diet, Physical Activity, and Behavioral Therapy**

Among overweight or obese patients with type 2 diabetes and inadequate glycemic, blood pressure and lipid control, and/or other obesity-related medical conditions, lifestyle changes that result in modest and sustained weight loss produce clinically meaningful reductions in blood glucose, A1C, and triglyceride.

  - **Lifestyle Interventions**

Weight loss can be attained with lifestyle programs that achieve a 500–750 kcal/day energy deficit or provide approximately 1,200–1,500 kcal/day for women and 1,500–1,800 kcal/day for men, adjusted for the individual's baseline body weight. Although benefits may be seen with as little as 5% weight loss, sustained weight loss of ≥7% is optimal.

- **Pharmacotherapy**

When considering pharmacological treatments for overweight or obese patients with type 2 diabetes, providers should first consider their choice of glucose-lowering medications.

  - **Concomitant Medications**

Providers should carefully review the patient's concomitant medications and, whenever possible, minimize or provide alternatives for medications that promote weight gain.

  - **Approved Medications**

The U.S. Food and Drug Administration (FDA) has approved five weight loss medications (or combination medications) for long-term use by patients with BMI ≥27 kg/m² with one or more obesity-associated comorbid conditions and by patients with BMI ≥30 kg/m² who are motivated to lose weight.

  - **Assessing Efficacy and Safety**

Efficacy and safety should be assessed at least monthly for the first 3 months of treatment.

Figure 9: Guideline, Page 1

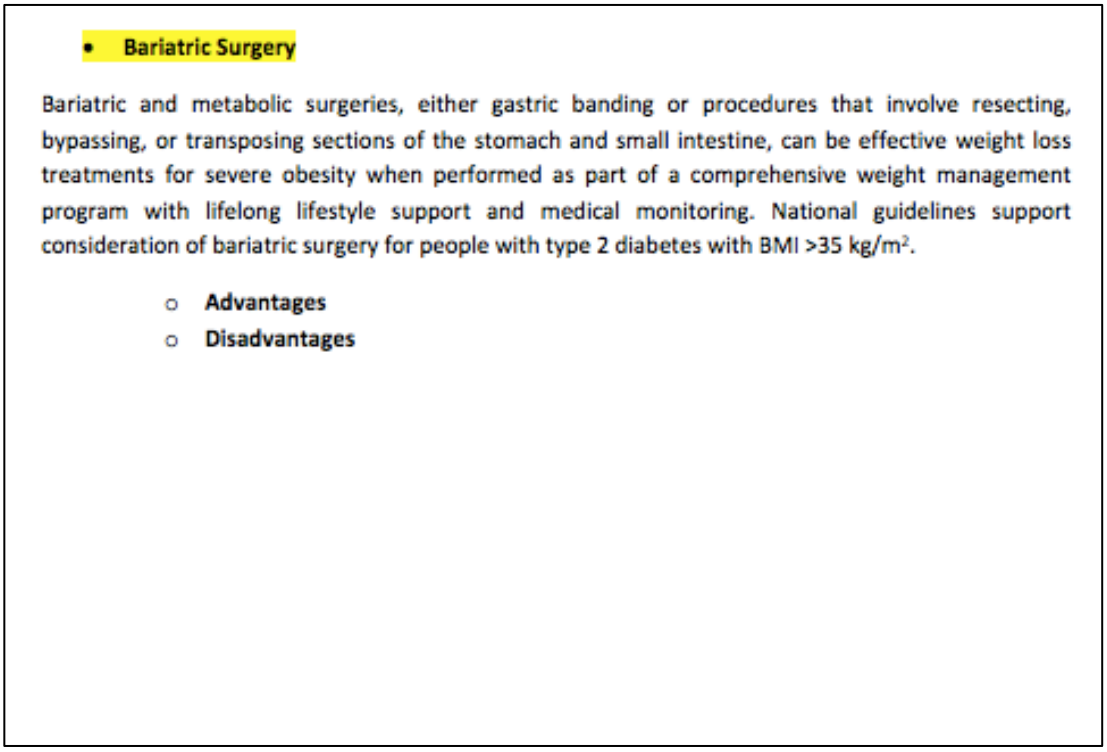


Figure 10: Guideline, Page 2

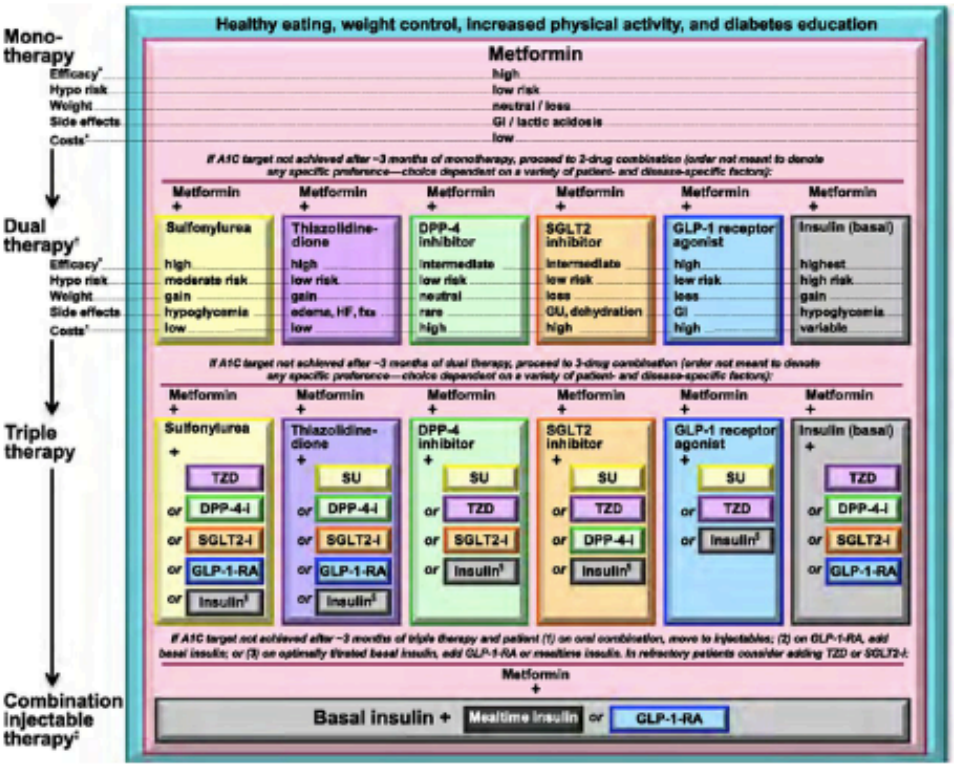
PHARMACOLOGICAL THERAPY FOR TYPE 2 DIABETES (section 7)

- **Initial Therapy**

Most patients should begin with **lifestyle changes.**

When lifestyle efforts alone do not achieve or maintain glycemic goals, **metformin monotherapy.**

In patients with metformin intolerance or contraindications, consider an initial drug from other classes depicted in Figure under **"Dual therapy"** and proceed accordingly.



- **Combination Therapy**
- **Insulin Therapy**

For patients with type 2 diabetes who are not achieving glycemic goals, providers should promptly initiate insulin therapy.

  - **Basal Insulin**
  - **Bolus Insulin**
  - **Continuous Subcutaneous Insulin Infusion ("basal–bolus" therapy)**
  - **Inhaled Insulin**

- **Bariatric Surgery**

Figure 11: Guideline, Page 3

Summary

Diabetes means your blood glucose, or blood sugar, levels are too high. With type 2 diabetes, the more common type, your body does not make or use insulin well. Insulin is a hormone that helps glucose get into your cells to give them energy. Without insulin, too much glucose stays in your blood. Over time, high blood glucose can lead to serious problems with your heart, eyes, kidneys, nerves, and gums and teeth.

You have a higher risk of type 2 diabetes if you are older, obese, have a family history of diabetes, or do not exercise. Having prediabetes also increases your risk. Prediabetes means that your blood sugar is higher than normal but not high enough to be called diabetes.

The symptoms of type 2 diabetes appear slowly. Some people do not notice symptoms at all. The symptoms can include

- Being very thirsty
- Urinating often
- Feeling very hungry or tired
- Losing weight without trying
- Having sores that heal slowly
- Having blurry eyesight

Blood tests can show if you have diabetes. One type of test, the A1C, can also check on how you are managing your diabetes. Many people can manage their diabetes through healthy eating, physical activity, and blood glucose testing. Some people also need to take diabetes medicines.

NIH: National Institute of Diabetes and Digestive and Kidney Diseases

Start Here

- Diabetes Mellitus Type 2: Overview (Beyond the Basics) (UpToDate)
- Facts about Type 2 Video (American Diabetes Association) Also in Spanish
- Type 2 Diabetes (Mayo Foundation for Medical Education and Research)
- Type 2 Diabetes: What You Need to Know From the National Institutes of Health (National Institute of Diabetes and Digestive and Kidney Diseases) - In English and Spanish
- Types of Diabetes From the National Institutes of Health Easy-to-Read (National Institute of Diabetes and Digestive and Kidney Diseases) Also in Spanish

Figure 12: Example Website, Page 1

Diabetes Medicines

Summary

Diabetes means your blood glucose, or blood sugar, levels are too high. If you can't control your diabetes with wise food choices and physical activity, you may need diabetes medicines. The kind of medicine you take depends on your type of diabetes, your schedule, and your other health conditions.

With type 1 diabetes, your pancreas does not make insulin. Insulin is a hormone that helps glucose get into your cells to give them energy. Without insulin, too much glucose stays in your blood. If you have type 1 diabetes, you will need to take insulin.

Type 2 diabetes, the most common type, can start when the body doesn't use insulin as it should. If your body can't keep up with the need for insulin, you may need to take pills. Along with meal planning and physical activity, diabetes pills help people with type 2 diabetes or gestational diabetes keep their blood glucose levels on target. Several kinds of pills are available. Each works in a different way. Many people take two or three kinds of pills. Some people take combination pills. Combination pills contain two kinds of diabetes medicine in one tablet. Some people take pills and insulin.

NIH: National Institute of Diabetes and Digestive and Kidney Diseases

Start Here

- Diabetes Medicines From the National Institutes of Health Easy-to-Read (National Institute of Diabetes and Digestive and Kidney Diseases) Also in Spanish
- Diabetes: Insulin Therapy (American Academy of Family Physicians) Also in Spanish
- Insulin Basics (American Diabetes Association) Also in Spanish
- Taking Medication (American Association of Diabetes Educators) - PDF Also in Spanish
- Women and Diabetes -- Diabetes Medicines Easy-to-Read (Food and Drug Administration)

Latest News

- FDA Approves 1st 'Artificial Pancreas' for Type 1 Diabetes (09/28/2016, HealthDay)
- Intensive Type 2 Diabetes Treatment Can Extend Survival (09/08/2016, HealthDay)

Figure 13: Example Website, Page 2

APPENDIX B

DATA SET INFORMATION

Table 13: Data Set Information

Website id	Website	Number of Web Pages	Total Number of Words	Information Coverage Score
1	http://www.diabetes.org/diabetes-basics/type-2/?loc=db-slabnav	20	7489	12
2	http://www.webmd.com/diabetes/type-2-diabetes-guide/type-2-diabetes	12	7957	9
3	https://www.diabetesaustralia.com.au/type-2-diabetes	7	5208	10
4	http://www.healthline.com/health/diabetes	7	10414	12
5	http://www.nhs.uk/Conditions/Diabetes/Pages/Diabetes.aspx	7	7817	10
6	http://www.mayoclinic.org/diseases-conditions/type-2-diabetes/home/ovc-20169860	5	5396	10
7	https://www.nlm.nih.gov/medlineplus/	2	435	1
8	http://www.diabetes.org.nz/about_diabetes/type_2_diabetes	4	3733	4
9	http://www.joslin.org/info/managing-diabetes.html	30	14596	11
10	http://www.everydayhealth.com/type-2-diabetes/guide/	5	3743	11
11	http://kidshealth.org/en/parents/type2.html?WT.ac=ctg#catendocrine	12	11074	9

Table 13 (continued)

12	http://patient.info/health/type-2-diabetes	4	10388	9
13	https://www.diabetes.ie/living-with-diabetes/living-with-type-2/	5	4013	8
14	http://www.webmd.boots.com/diabetes/type-2-diabetes-guide/default.htm	9	4105	9
15	http://www.endocrineweb.com/conditicon/type-2-diabetes/type-2-diabetes-overview	10	5770	9
16	https://www.drugwatch.com/actos/type-2-diabetes/	3	3752	9
17	https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes	4	8486	4
18	http://www.medicalnewstoday.com/info/diabetes/type2diabetes.php	5	3955	9
19	http://www.diabetes.ca/diabetes-and-you/living-with-type-2-diabetes	5	3579	7
20	https://www.diabetes.org.uk/Guide-to-diabetes/What-is-diabetes/	14	6151	12
21	https://healthfinder.gov/HealthTopics/Category/health-conditions-and-diseases/diabetes/take-steps-to-prevent-type-2-diabetes	5	5241	2
22	http://www.diabetes.co.uk/type2-diabetes.html	8	7122	10
23	http://www.news-medical.net/health/What-is-Type-2-Diabetes.aspx	7	2865	3
24	http://www.dlife.com/diabetes/type-2	34	11742	10
25	https://diatribe.org/type-2-diabetes	3	3103	10

Table 13 (continued)

26	https://www.niddk.nih.gov/health-information/diabetes	6	11064	8
27	http://www.idf.org/about-diabetes	2	821	1
28	http://www.bupa.co.uk/health-information/directory/t/type-2-diabetes	1	3194	5
29	https://www.drugs.com/enc/type-2-diabetes.html	4	6334	12
30	http://www.medicinenet.com/type_2_diabetes/article.htm	6	2553	8
31	https://www.lvhn.org/conditions_treattreat/diabetes/type_2_diabetes/overviov	5	2200	9
32	https://dtc.ucsf.edu/types-of-diabetes/type2/	32	24003	9
33	http://www.health.com/type-2-diabetes	6	1761	5
34	http://www.uptodate.com/contents/diadiabe-mellitus-type-2-treatment-beyond-the-basics	4	10854	10
35	http://diabetes.mercola.com/	5	11402	3
36	http://www.diabeticlivingonline.com/type-2-diabetes	4	7407	9
37	http://www.diabetesselfmanagement.com/about-diabetes/types-of-diabetes/type-2-diabetes-3/	7	15727	9
38	http://www.wikihow.com/Manage-Type-2-Diabetes	2	5461	6
39	http://www.uofmhealth.org/conditions-treatments/type-2-diabetes	1	661	3
40	http://www.diabeticlifestyle.com/type-2-diabetes/type-2-diabetes-causes-symptoms-diagnosis	7	4698	7
41	http://www.defeatdiabetes.org/signs-and-symptoms-of-diabetes/	1	320	1

Table 13 (continued)

42	http://www.diabetesnet.com/about-diabetes/types-diabetes/type-2	41	34130	10
43	https://www.msdiabetes.org/types-diabetes-0	6	2373	6
44	http://www.onlinemedinfo.com/diabetdi.html	7	3280	1
45	http://www.merckmanuals.com/home/hormonal-and-metabolic-disorders/diabetes-mellitus-dm-and-disorders-of-blood-sugar-metabolism/diabetes-mellitus-dm	2	6425	10
46	https://www.verywell.com/type-2-diabetes-4014632	9	8126	10
47	http://www.netdoctor.co.uk/conditions/diabetes/a829/type-2-diabetes/	4	6759	11
48	http://www.lillydiabetes.com/about-diabetes.aspx	1	221	1
49	http://www.healingwell.com/library/diabetes/	3	5814	5
50	http://www.drmirkin.com/diabetes	8	3792	3
51	http://choosehealth.utah.gov/your-health/health-conditions/diabetes.php	2	1240	2
52	http://www.health.govt.nz/your-health/conditions-and-treatments/diseases-and-illnesses/diabetes?mega=Your%20health&title=Diabetes	1	433	0
53	http://muschealth.staywellsolutionsonline.com/Conditions/Diabetes/	9	9834	11
54	http://www.drwhitaker.com/4-natural-type-2-diabetes-treatments/	5	1945	4

Table 13 (continued)

55	http://www.consumerreports.org/cro/2201/12/treating-type-2-diabetes/index.htm	1	6410	5
56	http://umm.edu/health/medical/reports/articles/diabetes-type-2	2	19285	9
57	http://dm2.newlifeoutlook.com/risks-of-anti-diabetic-medications/	4	2154	4
58	http://www.emedicinehealth.com/diabetes_mellitus_type_1_and_type_2/page2_em.htm	6	1544	7
59	http://asweetlife.org/diabetes/type-2-diabetes/	5	3822	9

APPENDIX C

LIST OF TITLES FOR RETRIEVED WIKIPEDIA PAGES

Table 14: Titles of Retrieved Wikipedia Pages

No	Webpage Title	No	Webpage Title
1	Metabolic disorder	79	Cushing's syndrome
2	Hyperglycemia	80	Hyperthyroidism
3	Insulin resistance	81	Pheochromocytoma
4	Insulin	82	Cancer
5	Polydipsia	83	Glucagonoma
6	Polyuria	84	Testosterone
7	Weight loss	85	Cell biology
8	Polyphagia	86	Liver
9	Cardiovascular disease	87	Glucose
10	Stroke	88	Lipid
11	Diabetic retinopathy	89	Adipocyte
12	Visual impairment	90	Incretin
13	Kidney failure	91	Glucagon
14	Amputation	92	Central nervous system
15	Hyperosmolar hyperglycemic state	93	Impaired fasting glucose
16	Diabetic ketoacidosis	94	Impaired glucose tolerance
17	Obesity	95	World Health Organization
18	Heredity	96	American Diabetes Association
19	Diabetes mellitus	97	Retinopathy
20	Diabetes mellitus type 1	98	Pancreatic islets
21	Gestational diabetes	99	Antibody
22	Autoimmunity	100	C-peptide
23	Beta cell	101	Screening (medicine)

Table 14 (continued)

24	Pancreas	102	United States Preventive Services Task Force
25	Glucose test	103	Blood pressure
26	Glucose tolerance test	104	First-degree relatives
27	Glycated hemoglobin	105	Polycystic ovary syndrome
28	Exercise	106	Metabolic syndrome
29	Diabetic diet	107	Acarbose
30	Metformin	108	Vitamin D
31	Bariatric surgery	109	Hypertension
32	Blurred vision	110	Hypercholesterolemia
33	Itch	111	Microalbuminuria
34	Peripheral neuropathy	112	Hypoglycemia
35	Vaginitis	113	Unnecessary health care
36	Fatigue	114	Eye examination
37	Altered level of consciousness	115	Periodontal disease
38	Hypotension	116	Scaling and root planing
39	Coronary artery disease	117	Aerobic exercise
40	Hospital	118	Strength training
41	Chronic kidney disease	119	Low-carbohydrate diet
42	Postoperative cognitive dysfunction	120	Very-low-calorie diet
43	Dementia	121	Vegetarianism
44	Alzheimer's disease	122	Anti-diabetic medication
45	Vascular dementia	123	Sulfonylurea
46	Acanthosis nigricans	124	Thiazolidinedione
47	Sexual dysfunction	125	Dipeptidyl peptidase-4 inhibitor
48	Diet (nutrition)	126	Gliflozin
49	Metabolism	127	Glucagon-like peptide-1 receptor agonist
50	Nutrition	128	Rosiglitazone
51	DNA methylation	129	ACE inhibitor
52	Prevotella	130	Kidney disease

Table 14 (continued)

53	Bacteroides	131	Angiotensin II receptor blocker
54	Overweight	132	Insulin glargine
55	Body mass index	133	Insulin detemir
56	Psychological stress	134	NPH insulin
57	Urbanization	135	Pregnancy
58	Waist–hip ratio	136	Developed country
59	Saturated fat	137	Developing country
60	Trans fat	138	Least Developed Countries
61	Polyunsaturated fat	139	South Asian ethnic groups
62	Monounsaturated fat	140	Pacific Islander
63	White rice	141	Latino
64	Persistent organic pollutant	142	Indigenous peoples of the Americas
65	Gene	143	Western lifestyle
66	Twin	144	Childhood obesity
67	TCF7L2	145	Epidemic
68	Allele	146	Circa
69	Genetic disorder	147	Common Era
70	Maturity onset diabetes of the young	148	List of physicians named Apollonius
71	Donohue syndrome	149	Roman Empire
72	Rabson–Mendenhall syndrome	150	Galen
73	Glucocorticoid	151	Sushruta
74	Thiazide	152	Charaka
75	Beta blocker	153	Diabetes insipidus
76	Atypical antipsychotic	154	Frederick Banting
77	Statin	155	Charles Best (medical scientist)
78	Acromegaly		

APPENDIX D

LIST OF STOPWORDS

Table 15: List of Stopwords

i	did	each
me	doing	few
my	a	more
myself	an	most
we	the	other
our	and	some
ours	but	such
ourselves	if	no
you	or	nor
your	because	not
yours	as	only
yourself	until	own
yourselves	while	same
he	of	so
him	at	than
his	by	too
himself	for	very
she	with	s
her	about	t
hers	against	can
herself	between	will
it	into	just
its	through	don
itself	during	should

Table 15 (continued)

they	before	now
them	after	d
their	above	ll
theirs	below	m
themselves	to	o
what	from	re
which	up	ve
who	down	y
whom	in	ain
this	out	aren
that	on	couldn
these	off	didn
those	over	doesn
am	under	hadn
is	again	hasn
are	further	haven
was	then	isn
were	once	ma
be	here	mightn
been	there	mustn
being	when	needn
have	where	shan
has	why	shouldn
had	how	wasn
having	all	weren
do	any	won
does	both	wouldn

APPENDIX E

LIST OF PART-OF-SPEECH (PoS) TAGS EXTRACTED

Table 16: Extracted PoS Tags

Linguistic Features	Starts with	PoS tags
Nouns	"NN"	NN: Noun, singular or mass NNS: Noun, plural NNP: Proper noun, singular NNPS: Proper noun, plural
Adjectives	"JJ"	JJ: Adjective JJR: Adjective, comparative JJS: Adjective, superlative
Verbs	"VB"	VB: Verb, base form VBD: Verb, past tense VBG: Verb, gerund or present participle VBN: Verb, past participle VBP: Verb, non-3rd person singular present VBZ: Verb, 3rd person singular present

APPENDIX F

MODEL PERFORMANCE MEASURES

In this appendix; performance measures of all models for ADA and Wikipedia corpus are presented. In the tables, the best performing model is marked red, followed by green and then blue followed by orange.

Table 17 Performance Measures of All Models

Model	Correlation	MSE
ADA.FS1.VM1.NO.10	0.763	4.982
ADA.FS1.VM1.TW.10	0.763	4.997
ADA.FS1.VM1.UW.10	0.763	4.991
ADA.FS1.VM2.NO.10	0.772	4.901
ADA.FS1.VM2.TW.10	0.772	4.901
ADA.FS1.VM2.UW.10	0.772	4.901
ADA.FS1.VM3.NO.10	0.377	10.309
ADA.FS1.VM3.TW.10	0.404	10.078
ADA.FS1.VM3.UW.10	0.462	9.441
ADA.FS1.VM4.NO.10	0.506	8.929
ADA.FS1.VM4.TW.10	0.504	8.960
ADA.FS1.VM4.UW.10	0.542	8.483
ADA.FS2.VM1.NO.10	0.785	4.592
ADA.FS2.VM1.TW.10	0.785	4.591
ADA.FS2.VM1.UW.10	0.785	4.595
ADA.FS2.VM2.NO.10	0.768	4.947
ADA.FS2.VM2.TW.10	0.768	4.947

Table 17 (continued)

ADA.FS2.VM2.UW.10	0.768	4.947
ADA.FS2.VM3.NO.10	0.377	10.278
ADA.FS2.VM3.TW.10	0.390	10.158
ADA.FS2.VM3.UW.10	0.448	9.591
ADA.FS2.VM4.NO.10	0.478	9.253
ADA.FS2.VM4.TW.10	0.483	9.203
ADA.FS2.VM4.UW.10	0.525	8.727
ADA.FS3.VM1.NO.10	0.800	4.356
ADA.FS3.VM1.TW.10	0.800	4.356
ADA.FS3.VM1.UW.10	0.800	4.358
ADA.FS3.VM2.NO.10	0.760	5.125
ADA.FS3.VM2.TW.10	0.760	5.125
ADA.FS3.VM2.UW.10	0.760	5.125
ADA.FS3.VM3.NO.10	0.351	10.561
ADA.FS3.VM3.TW.10	0.357	10.507
ADA.FS3.VM3.UW.10	0.404	10.027
ADA.FS3.VM4.NO.10	0.436	9.695
ADA.FS3.VM4.TW.10	0.439	9.663
ADA.FS3.VM4.UW.10	0.485	9.149
ADA.FS1.VM1.NO.20	0.737	5.478
ADA.FS1.VM1.TW.20	0.737	5.478
ADA.FS1.VM1.UW.20	0.737	5.473
ADA.FS1.VM2.NO.20	0.757	5.239
ADA.FS1.VM2.TW.20	0.757	5.239
ADA.FS1.VM2.UW.20	0.757	5.239
ADA.FS1.VM3.NO.20	0.431	9.749
ADA.FS1.VM3.TW.20	0.432	9.751
ADA.FS1.VM3.UW.20	0.488	9.217

Table 17 (continued)

ADA.FS1.VM4.NO.20	0.447	9.680
ADA.FS1.VM4.TW.20	0.451	9.630
ADA.FS1.VM4.UW.20	0.507	9.088
ADA.FS2.VM1.NO.20	0.790	4.512
ADA.FS2.VM1.TW.20	0.790	4.512
ADA.FS2.VM1.UW.20	0.790	4.512
ADA.FS2.VM2.NO.20	0.776	4.843
ADA.FS2.VM2.TW.20	0.776	4.843
ADA.FS2.VM2.UW.20	0.776	4.843
ADA.FS2.VM3.NO.20	0.449	9.639
ADA.FS2.VM3.TW.20	0.454	9.613
ADA.FS2.VM3.UW.20	0.501	9.167
ADA.FS2.VM4.NO.20	0.405	10.046
ADA.FS2.VM4.TW.20	0.413	9.978
ADA.FS2.VM4.UW.20	0.484	9.359
ADA.FS3.VM1.NO.20	0.801	4.335
ADA.FS3.VM1.TW.20	0.801	4.335
ADA.FS3.VM1.UW.20	0.801	4.335
ADA.FS3.VM2.NO.20	0.783	4.716
ADA.FS3.VM2.TW.20	0.783	4.716
ADA.FS3.VM2.UW.20	0.783	4.716
ADA.FS3.VM3.NO.20	0.421	9.880
ADA.FS3.VM3.TW.20	0.428	9.804
ADA.FS3.VM3.UW.20	0.471	9.414
ADA.FS3.VM4.NO.20	0.494	9.052
ADA.FS3.VM4.TW.20	0.494	9.052
ADA.FS3.VM4.UW.20	0.495	9.077
ADA.FS1.VM1.NO.30	0.716	5.838

Table 17 (continued)

ADA.FS1.VM1.TW.30	0.716	5.838
ADA.FS1.VM1.UW.30	0.716	5.838
ADA.FS1.VM2.NO.30	0.737	5.582
ADA.FS1.VM2.TW.30	0.737	5.582
ADA.FS1.VM2.UW.30	0.737	5.582
ADA.FS1.VM3.NO.30	0.429	9.838
ADA.FS1.VM3.TW.30	0.430	9.832
ADA.FS1.VM3.UW.30	0.485	9.267
ADA.FS1.VM4.NO.30	0.420	9.928
ADA.FS1.VM4.TW.30	0.427	9.854
ADA.FS1.VM4.UW.30	0.483	9.324
ADA.FS2.VM1.NO.30	**0.803**	**4.257**
ADA.FS2.VM1.TW.30	**0.803**	**4.257**
ADA.FS2.VM1.UW.30	**0.803**	**4.257**
ADA.FS2.VM2.NO.30	**0.818**	**3.962**
ADA.FS2.VM2.TW.30	**0.818**	**3.962**
ADA.FS2.VM2.UW.30	**0.818**	**3.962**
ADA.FS2.VM3.NO.30	0.524	8.855
ADA.FS2.VM3.TW.30	0.535	8.726
ADA.FS2.VM3.UW.30	0.573	8.335
ADA.FS2.VM4.NO.30	0.486	9.220
ADA.FS2.VM4.TW.30	0.487	9.202
ADA.FS2.VM4.UW.30	0.526	8.809
ADA.FS3.VM1.NO.30	**0.804**	**4.233**
ADA.FS3.VM1.TW.30	**0.804**	**4.233**
ADA.FS3.VM1.UW.30	**0.804**	**4.230**
ADA.FS3.VM2.NO.30	**0.810**	**4.116**
ADA.FS3.VM2.TW.30	**0.810**	**4.116**

Table 17 (continued)

ADA.FS3.VM2.UW.30	**0.810**	**4.116**
ADA.FS3.VM3.NO.30	0.460	9.447
ADA.FS3.VM3.TW.30	0.467	9.366
ADA.FS3.VM3.UW.30	0.528	8.764
ADA.FS3.VM4.NO.30	0.481	9.248
ADA.FS3.VM4.TW.30	0.481	9.248
ADA.FS3.VM4.UW.30	0.498	9.056
ADA.FS1.VM1.NO.40	0.705	6.029
ADA.FS1.VM1.TW.40	0.705	6.029
ADA.FS1.VM1.UW.40	0.705	6.034
ADA.FS1.VM2.NO.40	0.721	5.828
ADA.FS1.VM2.TW.40	0.721	5.828
ADA.FS1.VM2.UW.40	0.721	5.828
ADA.FS1.VM3.NO.40	0.402	10.171
ADA.FS1.VM3.TW.40	0.410	10.105
ADA.FS1.VM3.UW.40	0.475	9.526
ADA.FS1.VM4.NO.40	0.483	9.233
ADA.FS1.VM4.TW.40	0.484	9.228
ADA.FS1.VM4.UW.40	0.484	9.223
ADA.FS2.VM1.NO.40	0.797	4.368
ADA.FS2.VM1.TW.40	0.797	4.368
ADA.FS2.VM1.UW.40	0.797	4.368
ADA.FS2.VM2.NO.40	0.800	4.316
ADA.FS2.VM2.TW.40	0.800	4.316
ADA.FS2.VM2.UW.40	0.800	4.316
ADA.FS2.VM3.NO.40	0.494	9.213
ADA.FS2.VM3.TW.40	0.503	9.166
ADA.FS2.VM3.UW.40	0.554	8.637

Table 17 (continued)

ADA.FS2.VM4.NO.40	0.489	9.156
ADA.FS2.VM4.TW.40	0.490	9.144
ADA.FS2.VM4.UW.40	0.506	8.993
ADA.FS3.VM1.NO.40	0.784	4.634
ADA.FS3.VM1.TW.40	0.784	4.634
ADA.FS3.VM1.UW.40	0.784	4.634
ADA.FS3.VM2.NO.40	0.735	5.498
ADA.FS3.VM2.TW.40	0.735	5.498
ADA.FS3.VM2.UW.40	0.735	5.498
ADA.FS3.VM3.NO.40	0.400	10.087
ADA.FS3.VM3.TW.40	0.411	9.979
ADA.FS3.VM3.UW.40	0.489	9.281
ADA.FS3.VM4.NO.40	0.488	9.192
ADA.FS3.VM4.TW.40	0.488	9.192
ADA.FS3.VM4.UW.40	0.504	8.995
ADA.FS1.VM1.NO.50	0.734	5.566
ADA.FS1.VM1.TW.50	0.734	5.566
ADA.FS1.VM1.UW.50	0.734	5.566
ADA.FS1.VM2.NO.50	0.737	5.459
ADA.FS1.VM2.TW.50	0.737	5.459
ADA.FS1.VM2.UW.50	0.737	5.459
ADA.FS1.VM3.NO.50	0.370	10.406
ADA.FS1.VM3.TW.50	0.381	10.319
ADA.FS1.VM3.UW.50	0.453	9.749
ADA.FS1.VM4.NO.50	0.486	9.197
ADA.FS1.VM4.TW.50	0.486	9.193
ADA.FS1.VM4.UW.50	0.484	9.206
ADA.FS2.VM1.NO.50	0.776	4.762

Table 17 (continued)

ADA.FS2.VM1.TW.50	0.776	4.762
ADA.FS2.VM1.UW.50	0.776	4.762
ADA.FS2.VM2.NO.50	0.749	5.259
ADA.FS2.VM2.TW.50	0.749	5.259
ADA.FS2.VM2.UW.50	0.749	5.259
ADA.FS2.VM3.NO.50	0.562	8.542
ADA.FS2.VM3.TW.50	0.571	8.435
ADA.FS2.VM3.UW.50	0.619	7.854
ADA.FS2.VM4.NO.50	0.548	8.516
ADA.FS2.VM4.TW.50	0.548	8.516
ADA.FS2.VM4.UW.50	0.551	8.454
ADA.FS3.VM1.NO.50	0.794	4.410
ADA.FS3.VM1.TW.50	0.794	4.410
ADA.FS3.VM1.UW.50	0.794	4.410
ADA.FS3.VM2.NO.50	0.699	6.152
ADA.FS3.VM2.TW.50	0.699	6.152
ADA.FS3.VM2.UW.50	0.699	6.152
ADA.FS3.VM3.NO.50	0.477	9.365
ADA.FS3.VM3.TW.50	0.486	9.289
ADA.FS3.VM3.UW.50	0.553	8.659
ADA.FS3.VM4.NO.50	0.514	8.927
ADA.FS3.VM4.TW.50	0.514	8.927
ADA.FS3.VM4.UW.50	0.524	8.796
WKP.FS1.VM1.NO.10	0.792	4.617
WKP.FS1.VM1.TW.10	0.792	4.617
WKP.FS1.VM1.UW.10	0.792	4.617
WKP.FS1.VM2.NO.10	0.796	4.400

Table 17 (continued)

WKP.FS1.VM2.TW.10	0.796	4.400
WKP.FS1.VM2.UW.10	0.796	4.400
WKP.FS1.VM3.NO.10	0.489	9.327
WKP.FS1.VM3.TW.10	0.493	9.323
WKP.FS1.VM3.UW.10	0.548	8.685
WKP.FS1.VM4.NO.10	0.530	8.639
WKP.FS1.VM4.TW.10	0.529	8.645
WKP.FS1.VM4.UW.10	0.549	8.394
WKP.FS2.VM1.NO.10	**0.830**	**3.924**
WKP.FS2.VM1.TW.10	**0.830**	**3.924**
WKP.FS2.VM1.UW.10	**0.830**	**3.924**
WKP.FS2.VM2.NO.10	0.762	5.027
WKP.FS2.VM2.TW.10	0.762	5.027
WKP.FS2.VM2.UW.10	0.762	5.027
WKP.FS2.VM3.NO.10	0.435	9.955
WKP.FS2.VM3.TW.10	0.441	9.905
WKP.FS2.VM3.UW.10	0.494	9.403
WKP.FS2.VM4.NO.10	0.519	8.784
WKP.FS2.VM4.TW.10	0.519	8.784
WKP.FS2.VM4.UW.10	0.542	8.505
WKP.FS3.VM1.NO.10	0.808	4.277
WKP.FS3.VM1.TW.10	0.808	4.277
WKP.FS3.VM1.UW.10	0.808	4.277
WKP.FS3.VM2.NO.10	0.765	4.992
WKP.FS3.VM2.TW.10	0.765	4.992
WKP.FS3.VM2.UW.10	0.765	4.992
WKP.FS3.VM3.NO.10	0.434	9.950
WKP.FS3.VM3.TW.10	0.444	9.864

Table 17 (continued)

WKP.FS3.VM3.UW.10	0.490	9.376
WKP.FS3.VM4.NO.10	0.554	8.359
WKP.FS3.VM4.TW.10	0.554	8.359
WKP.FS3.VM4.UW.10	0.554	8.344
WKP.FS1.VM1.NO.20	0.825	3.936
WKP.FS1.VM1.TW.20	0.825	3.936
WKP.FS1.VM1.UW.20	0.825	3.936
WKP.FS1.VM2.NO.20	0.798	4.348
WKP.FS1.VM2.TW.20	0.798	4.348
WKP.FS1.VM2.UW.20	0.798	4.348
WKP.FS1.VM3.NO.20	0.396	10.241
WKP.FS1.VM3.TW.20	0.410	10.142
WKP.FS1.VM3.UW.20	0.473	9.502
WKP.FS1.VM4.NO.20	0.521	8.756
WKP.FS1.VM4.TW.20	0.521	8.756
WKP.FS1.VM4.UW.20	0.542	8.503
WKP.FS2.VM1.NO.20	0.819	4.072
WKP.FS2.VM1.TW.20	0.819	4.072
WKP.FS2.VM1.UW.20	0.819	4.072
WKP.FS2.VM2.NO.20	0.757	5.123
WKP.FS2.VM2.TW.20	0.757	5.123
WKP.FS2.VM2.UW.20	0.757	5.123
WKP.FS2.VM3.NO.20	0.500	9.219
WKP.FS2.VM3.TW.20	0.504	9.163
WKP.FS2.VM3.UW.20	0.545	8.725
WKP.FS2.VM4.NO.20	0.533	8.664
WKP.FS2.VM4.TW.20	0.533	8.664
WKP.FS2.VM4.UW.20	0.550	8.427

Table 17 (continued)

WKP.FS3.VM1.NO.20	0.813	4.229
WKP.FS3.VM1.TW.20	0.813	4.229
WKP.FS3.VM1.UW.20	0.813	4.229
WKP.FS3.VM2.NO.20	0.709	5.952
WKP.FS3.VM2.TW.20	0.709	5.952
WKP.FS3.VM2.UW.20	0.709	5.952
WKP.FS3.VM3.NO.20	0.485	9.372
WKP.FS3.VM3.TW.20	0.488	9.346
WKP.FS3.VM3.UW.20	0.534	8.823
WKP.FS3.VM4.NO.20	0.564	8.226
WKP.FS3.VM4.TW.20	0.564	8.226
WKP.FS3.VM4.UW.20	0.555	8.326
WKP.FS1.VM1.NO.30	0.791	4.614
WKP.FS1.VM1.TW.30	0.791	4.614
WKP.FS1.VM1.UW.30	0.791	4.614
WKP.FS1.VM2.NO.30	0.757	5.097
WKP.FS1.VM2.TW.30	0.757	5.097
WKP.FS1.VM2.UW.30	0.757	5.097
WKP.FS1.VM3.NO.30	0.419	10.006
WKP.FS1.VM3.TW.30	0.431	9.885
WKP.FS1.VM3.UW.30	0.511	9.040
WKP.FS1.VM4.NO.30	0.514	8.844
WKP.FS1.VM4.TW.30	0.514	8.844
WKP.FS1.VM4.UW.30	0.519	8.795
WKP.FS2.VM1.NO.30	0.811	4.274
WKP.FS2.VM1.TW.30	0.811	4.274
WKP.FS2.VM1.UW.30	0.811	4.274
WKP.FS2.VM2.NO.30	0.728	5.638

Table 17 (continued)

WKP.FS2.VM2.TW.30	0.728	5.638
WKP.FS2.VM2.UW.30	0.728	5.638
WKP.FS2.VM3.NO.30	0.517	9.054
WKP.FS2.VM3.TW.30	0.523	8.985
WKP.FS2.VM3.UW.30	0.575	8.273
WKP.FS2.VM4.NO.30	0.545	8.512
WKP.FS2.VM4.TW.30	0.545	8.512
WKP.FS2.VM4.UW.30	0.543	8.501
WKP.FS3.VM1.NO.30	0.795	4.526
WKP.FS3.VM1.TW.30	0.795	4.526
WKP.FS3.VM1.UW.30	0.795	4.526
WKP.FS3.VM2.NO.30	0.691	6.262
WKP.FS3.VM2.TW.30	0.691	6.262
WKP.FS3.VM2.UW.30	0.691	6.262
WKP.FS3.VM3.NO.30	0.406	10.246
WKP.FS3.VM3.TW.30	0.405	10.249
WKP.FS3.VM3.UW.30	0.492	9.349
WKP.FS3.VM4.NO.30	0.515	8.885
WKP.FS3.VM4.TW.30	0.515	8.885
WKP.FS3.VM4.UW.30	0.523	8.782
WKP.FS1.VM1.NO.40	**0.849**	**3.639**
WKP.FS1.VM1.TW.40	**0.849**	**3.639**
WKP.FS1.VM1.UW.40	**0.849**	**3.639**
WKP.FS1.VM2.NO.40	0.766	4.966
WKP.FS1.VM2.TW.40	0.766	4.966
WKP.FS1.VM2.UW.40	0.766	4.966
WKP.FS1.VM3.NO.40	0.425	9.938
WKP.FS1.VM3.TW.40	0.417	10.026

Table 17 (continued)

WKP.FS1.VM3.UW.40	0.502	9.197
WKP.FS1.VM4.NO.40	0.485	9.192
WKP.FS1.VM4.TW.40	0.485	9.192
WKP.FS1.VM4.UW.40	0.499	9.021
WKP.FS2.VM1.NO.40	**0.835**	**3.844**
WKP.FS2.VM1.TW.40	**0.835**	**3.844**
WKP.FS2.VM1.UW.40	**0.835**	**3.844**
WKP.FS2.VM2.NO.40	0.755	5.150
WKP.FS2.VM2.TW.40	0.755	5.150
WKP.FS2.VM2.UW.40	0.755	5.150
WKP.FS2.VM3.NO.40	0.468	9.687
WKP.FS2.VM3.TW.40	0.475	9.589
WKP.FS2.VM3.UW.40	0.545	8.688
WKP.FS2.VM4.NO.40	0.541	8.496
WKP.FS2.VM4.TW.40	0.541	8.496
WKP.FS2.VM4.UW.40	0.545	8.446
WKP.FS3.VM1.NO.40	0.802	4.392
WKP.FS3.VM1.TW.40	0.802	4.392
WKP.FS3.VM1.UW.40	0.802	4.392
WKP.FS3.VM2.NO.40	0.717	5.837
WKP.FS3.VM2.TW.40	0.717	5.837
WKP.FS3.VM2.UW.40	0.717	5.837
WKP.FS3.VM3.NO.40	0.429	10.462
WKP.FS3.VM3.TW.40	0.430	10.320
WKP.FS3.VM3.UW.40	0.495	9.314
WKP.FS3.VM4.NO.40	0.503	8.955
WKP.FS3.VM4.TW.40	0.503	8.955
WKP.FS3.VM4.UW.40	0.511	8.880

Table 17 (continued)

WKP.FS1.VM1.NO.50	**0.844**	**3.689**
WKP.FS1.VM1.TW.50	**0.844**	**3.689**
WKP.FS1.VM1.UW.50	**0.844**	**3.689**
WKP.FS1.VM2.NO.50	0.764	5.009
WKP.FS1.VM2.TW.50	0.764	5.009
WKP.FS1.VM2.UW.50	0.764	5.009
WKP.FS1.VM3.NO.50	0.410	10.091
WKP.FS1.VM3.TW.50	0.415	10.042
WKP.FS1.VM3.UW.50	0.500	9.224
WKP.FS1.VM4.NO.50	0.469	9.382
WKP.FS1.VM4.TW.50	0.469	9.382
WKP.FS1.VM4.UW.50	0.493	9.094
WKP.FS2.VM1.NO.50	0.825	3.982
WKP.FS2.VM1.TW.50	0.825	3.982
WKP.FS2.VM1.UW.50	0.825	3.982
WKP.FS2.VM2.NO.50	0.749	5.266
WKP.FS2.VM2.TW.50	0.749	5.266
WKP.FS2.VM2.UW.50	0.749	5.266
WKP.FS2.VM3.NO.50	0.417	10.110
WKP.FS2.VM3.TW.50	0.422	10.058
WKP.FS2.VM3.UW.50	0.497	9.260
WKP.FS2.VM4.NO.50	0.485	9.183
WKP.FS2.VM4.TW.50	0.485	9.183
WKP.FS2.VM4.UW.50	0.498	9.030
WKP.FS3.VM1.NO.50	0.784	4.655
WKP.FS3.VM1.TW.50	0.784	4.655
WKP.FS3.VM1.UW.50	0.784	4.655
WKP.FS3.VM2.NO.50	0.702	6.109

Table 17 (continued)

WKP.FS3.VM2.TW.50	0.702	6.109
WKP.FS3.VM2.UW.50	0.702	6.109
WKP.FS3.VM3.NO.50	0.420	10.429
WKP.FS3.VM3.TW.50	0.415	10.338
WKP.FS3.VM3.UW.50	0.497	9.209
WKP.FS3.VM4.NO.50	0.502	8.962
WKP.FS3.VM4.TW.50	0.502	8.962
WKP.FS3.VM4.UW.50	0.512	8.870

APPENDIX G

SIGNIFICANT TERMS AND CORRESPONDING COEFFICIENTS

In this appendix; ranked significant terms separated according to their coefficients as positive and negative are presented. This sample is for model 5 and 30% percentage.

Table 18: Positive Significant Terms

Rank	Positive term	Coefficient	Rank	Positive term	Coefficient
1	inhibitor	1.259	63	placebo	0.304
2	acting	1.038	64	laparoscopic	0.237
3	observational	1.742	65	longer	0.248
4	bariatric	0.770	66	surgery	0.273
5	gain	0.838	67	set	0.368
6	physical	0.722	68	production	0.188
7	rapid	0.642	69	diet	0.183
8	therapy	0.450	70	maintenance	0.770
9	provide	0.550	71	insulin	0.156
10	frequent	0.504	72	activate	0.172
11	medication	0.464	73	result	0.185
12	injection	0.430	74	advise	0.229
13	alternative	0.474	75	mmol	0.148
14	food	0.344	76	reach	0.080
15	treatment	0.532	77	assess	0.214
16	schedule	0.403	78	secretion	0.142
17	flexibility	0.410	79	mealtime	0.545
18	mol	0.514	80	mean	0.272
19	pioglitazone	0.344	81	maximum	0.173

Table 18 (continued)

20	group	0.336	82	support	0.126
21	cap	0.618	83	range	0.073
22	consider	0.338	84	surgical	0.395
23	twice	0.318	85	headache	0.103
24	combination	0.315	86	vomiting	0.082
25	approve	0.318	87	hypoglycemia	0.173
26	session	0.295	88	receptor	0.302
27	intestine	0.208	89	rosiglitazone	0.110
28	target	0.331	90	sulfonylurea	0.088
29	gastric	0.243	91	management	0.347
30	effect	0.279	92	section	0.135
31	version	0.273	93	premixed	0.249
32	adjust	0.238	94	diagnosis	0.222
33	base	0.243	95	cover	0.280
34	tab	0.375	96	individual	0.084
35	smbg	0.298	97	compare	0.076
36	short	0.203	98	hypo	0.042
37	year	0.239	99	kg	0.046
38	calorie	0.244	100	dos	0.144
39	day	0.322	101	stomach	0.041
40	bypass	0.293	102	specific	0.056
41	class	0.279	103	focus	0.044
42	lose	0.231	104	slow	0.159
43	choice	0.201	105	require	0.052
44	antidepressant	0.982	106	counseling	0.095
45	electronic	0.328	107	like	0.040
46	medical	0.194	108	obesity	0.046
47	benefit	0.266	109	effectiveness	0.035
48	su	0.267	110	provider	0.043
49	begin	0.235	111	daily	0.042

Table 18 (continued)

50	action	0.195	112	drug	0.028
51	add	0.207	113	volume	0.039
52	injectable	0.356	114	contraindicate	0.014
53	effort	0.274	115	energy	0.072
54	achieve	0.155	116	thyroid	0.057
55	metformin	0.236	117	intensity	0.004
56	month	0.276	118	acid	0.011
57	bolus	0.180	119	en	0.015
58	follow	0.365	120	monitoring	0.003
59	patient	0.198	121	intermediate	0.000
60	blind	0.255	122	pressure	0.054
61	goal	0.130	123	pharmacological	0.001
62	minimum	0.176			

Table 19: Negative Significant Terms

Rank	Negative term	Coefficient	Rank	Negative term	Coefficient
1	present	-0.715	23	agent	-0.394
2	thiazolidinedione	-0.426	24	safety	-0.165
3	concentrated	-0.723	25	postprandial	-0.334
4	cardiovascular	-0.382	26	contraindicate	-0.154
5	accumulate	-0.418	27	europe	-0.061
6	intolerance	-0.253	28	deficit	-0.130
7	death	-0.281	29	acid	-0.099
8	fig	-1.063	30	average	-0.271
9	requirement	-0.285	31	woman	-0.112
10	mortality	-0.189	32	controlled	-0.264
11	min	-0.292	33	clinical	-0.224
12	maintenance	-0.415	34	noninsulin	-0.038
13	initial	-0.158	35	outcome	-0.067
14	inexpensive	-0.521	36	weight	-0.140
15	intestinal	-0.303	37	order	-0.074
16	ldl	-0.345	38	caution	-0.056
17	event	-0.179	39	serotonin	-0.160
18	preference	-0.168	40	trial	-0.118
19	drive	-0.258	41	equal	-0.064
20	data	-0.301	42	dependent	-0.076
21	profile	-0.385	43	gastrointestinal	-0.061
22	initiate	-0.239	44	comprehensive	-0.043