A FRAMEWORK FOR INFORMATION QUALITY AND COVERAGE ASSESSMENT FOR TYPE 2 DIABETES WEBSITES

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

DİDEM ÖLÇER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

SEPTEMBER 2020

Approval of the thesis:

A FRAMEWORK FOR INFORMATION QUALITY AND COVERAGE ASSESSMENT FOR TYPE 2 DIABETES WEBSITES

Submitted by DİDEM ÖLÇER in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Information Systems Department, Middle East Technical University** by,

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

Prof. Dr. Sevgi Özkan Yıldırım
Head of Department, **Information Systems** _____

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

Examining Committee Members:

Prof. Dr. Hasan Oğul
Computer Engineering Dept., Çankaya University _____

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

Assoc. Prof. Dr. Aysu Betin Can
Information Systems Dept., METU _____

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

Assoc. Prof. Dr. Burcu Can
Research Institute of Information and Language Processing, Wolverhampton University _____

Date: **22.09.2020**

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: Didem, ÖLÇER

Signature : _____

ABSTRACT

A FRAMEWORK FOR INFORMATION QUALITY AND COVERAGE ASSESSMENT FOR TYPE 2 DIABETES WEBSITES

ÖLÇER, Didem

Ph.D, Department of Information Systems

Supervisor: Assoc. Prof. Dr. Tuğba TAŞKAYA TEMİZEL

September 2020, 102 pages

Health information seekers often use search engines to access high-quality and up-to-date information. However, finding online high-quality health information is increasingly getting difficult due to the high volume of information generated by non-experts in the area. There are manual tools which help end-users in assessing the quality of websites. However, they are labour-intensive. This thesis aims to propose a framework that automatically evaluates the content coverage and quality of health websites according to evidence-based medicine. The thesis has two main contributions. The first one is a method which utilizes quality indicators derived from professional health literacy guidelines to measure information quality. The second contribution includes a method which uses both textual and content-based features with Okapi BM25 and MeSH term expansion to assess information coverage and information quality. Content-based features were acquired using American Diabetes Association's (ADA) guideline, which is an evidence-based practice guideline in diabetes. Specifically, sentences containing auxiliary verbs from ADA guideline were extracted and the weirdness coefficient of terms, 2-grams and 3-grams generated from these sentences were calculated using iWeb corpus. The results showed that the use of both textual and content-based features is effective in classification of high and low-quality websites. In addition, the features derived from professional health guidelines lead to a significant positive impact in classification results.

Keywords: DISCERN, website quality, quality features, information retrieval

ÖZ

TİP 2 DİYABET WEB SİTELERİ İÇİN BİLGİ KALİTESİ VE KAPSAM DEĞERLENDİRMESİ İÇİN BİR ÇERÇEVE

ÖLÇER, Didem

Doktora, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Doç. Dr. Tuğba TAŞKAYA TEMİZEL

Eylül 2020, 102 sayfa

Sağlık bilgisi arayanlar, genellikle yüksek kaliteli ve güncel bilgilere erişmek için arama motorlarını kullanırlar. Bununla birlikte, alanında uzman olmayan kişiler tarafından üretilen yüksek bilgi hacmi nedeniyle çevrimiçi yüksek kaliteli sağlık bilgisi bulmak gittikçe zorlaşmaktadır. Son kullanıcıların web sitelerinin kalitesini değerlendirmelerine yardımcı olan manuel araçlar vardır. Ancak, bu araçlar için zaman ve iş gücüne ihtiyaç vardır. Bu tez, kanıta dayalı tıbba göre sağlık web sitelerinin içerik kapsamını ve kalitesini otomatik olarak değerlendiren bir çerçeve önermeyi amaçlamaktadır. Tezin iki temel katkısı vardır. Birincisi, bilgi kalitesini ölçmek için profesyonel sağlık okuryazarlığı rehberlerinden türetilmiş kalite göstergelerini kullanan bir yöntemdir. İkinci katkı, bilgi kapsamını ve bilgi kalitesini değerlendirmek için Okapi BM25 ve MeSH terim genişletme ile hem metinsel hem de içerik tabanlı özellikleri kullanan bir yöntemi içerir. İçerik tabanlı özellikler diyabette kanıta dayalı bir uygulama rehberi olan Amerikan Diyabet Birliğinin (ADA) rehberi kullanılarak edinildi. Spesifik olarak, ADA kılavuzundan yardımcı fiilleri içeren cümleleri çıkartıldı ve iWeb korpus'u kullanarak bu cümlelerden üretilen terimlerin, 2-gram ve 3-gram terimlerin tuhaflık katsayısıhesaplandı. Sonuçlar, hem metin hem de içerik tabanlı özelliklerin kullanımının yüksek ve düşük kaliteli web sitelerinin sınıflandırılmasında etkili olduğunu göstermiştir. Ayrıca, profesyonel sağlık rehberlerinden türetilen özellikler, sınıflandırma sonuçları üzerinde önemli bir olumlu etkiye yol açmaktadır.

Anahtar Sözcükler: DISCERN, web site kalitesi, kalite özellikleri, bilgi çıkarımı

To my daughter, Beril

To the memories of my grandmother, Tomris

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere appreciation to my supervisor, Assoc. Prof. Dr. Tuğba Taşkaya Temizel for her endless support, valuable guidance, friendly encouragement and deep understanding especially during the hard times throughout this study. Not only being guiding and supportive but also encouraged me on any subjectwhen I fell into desperation throughout the thesis period. I appreciate all her contributions of time and ideas during my Ph.D. experience. It was a great pleasure and honor to have her supervise me.

Besides my supervisor, I would like to thank to the thesis examining committee members Prof. Dr. Hasan Oğul, Assoc. Prof. Dr. Aysu Betin Can, Assoc. Prof. Dr. Burcu Can and Assoc. Prof. Dr. Altan Koçyiğit. I would also like to thank Assist. Prof. Dr. Rahime Belen Sağlam for valuable comments and assistance for helping me accomplish this thesis.

I acknowledge the support of The Scientific and Technical Research Council of Turkey (TÜBİTAK) for providing scholarship during my Ph.D. study.

I would also like to thank all of my colleagues Oğul Göçmen, Buket Ünal, Nihal Uğur, Tülin Erçelebi Ayyıldız, Serian Doma, Duygu Dede Şener, and Çağatay Berke Erdaş for their support and encourage. Also, thanks to Dr. Edip Nehariye for his contribution.

I would like to give my special thanks to my family that always considers my well-being and supports and encourages me at all times. Especially, thanks to my brother Enver Tokmak for his love and support.

Finally, I would like to thank my husband Naim Ölçer for his long-standing support and continuous patience. Last but not the least; my inexpressible debt is to my daughter Beril for her love. It is never enough how much I thank them both.

TABLE OF CONTENTS

ABSTRACT iv

ÖZ v

DEDICATION vi

ACKNOWLEDGEMENTS vii

TABLE OF CONTENTS viii

LIST OF TABLES xi

LIST OF FIGURES xiii

LIST OF ABBREVIATIONS xiv

1. INTRODUCTION 1

1.1. Research Questions 3

1.2. Contribution of the Thesis 5

1.3. Organization of the Thesis 6

2. LITERATURE REVIEW 7

2.1. Quality assessment of health information 8

2.1.1. Manual assessment of online health content 8

2.1.2. Automated assessment of online health content 11

2.2. Quality assesment of Wikipedia articles 12

2.3. Quality term extraction from content 14

2.3.1. Fact-checking 14

2.3.2. Author detection 15

2.4. Professional health literacy guidelines for writing good quality information for patients 15
2.5. Information retrieval and extraction techniques 17
2.5.1. Text-based document ranking 17
2.5.2. Automatic term recognition techniques 18
2.6. Summary 19
3. WEBSITE QUALITY ASSESSMENT 21
3.1. Collection of websites 21
3.2. Evaluation of the websites 23
3.3. Rating of the websites by the experts 23
4. AUTOMATIC QUALITY ASSESSMENT OF WEB-BASED INFORMATION 29
4.1. Introduction 29
4.2. Textual Features 31
4.3. Content-based Features 33
4.3.1. Corpus 33
4.3.2. Term extraction from ADA 33
4.3.3. Query expansion with MeSH terms 34
4.4. Information retrieval techniques 34
4.4.1. Text-based document ranking with Okapi BM25 35
4.4.2. Feature Construction using Vectorization Methods 35
4.5. Experiments and Results 36
4.5.1. Tools 36
4.5.2. Evaluation methodology 37

4.5.3. Results 37
4.6. Discussion 44
4.7. Conclusion 47
5. INFORMATION QUALITY ASSESSMENT USING PROFESSIONAL HEALTH LITERACY GUIDELINE RECOMMENDATIONS 49
5.1 Introduction 49
5.2. Features obtained from professional health literacy guidelines 50
5.3. Experiments and Results 52
5.4. Discussion 58
5.5. Conclusion 59
6. CONCLUSION AND FUTURE WORK 61
REFERENCES 65
APPENDICES 73
APPENDIX A 73
APPENDIX B 82
APPENDIX C 84
APPENDIX D 96
APPENDIX E 97
APPENDIX F 98
CURRICULUM VITAE 101

LIST OF TABLES

Table 1: HONcode certificated websites....23
Table 2: Average score of each DISCERN question....24
Table 3: DISCERN scores....25
Table 4: Descriptive statistics of the websites, which show their DISCERN and coverage scores....25
Table 5: Examples of sentence types....32
Table 6: Analysis of the terms....36
Table 7: Model Naming....36
Table 8: Kendall's Tau Correlation of features and classes. The feature category indicates whether this feature is used in assessing Wikipedia content (W) and/or used in information quality assessment of generic content other than Wikipedia (O). The features (F) and (H) are constructed based on the health websites' writing styles and health literacy guideline recommendations, respectively. 38
Table 9: LOOCV results showing the mean and standard deviation of the results with textual features....41
Table 10: Model Performance Measures, where the top 350 terms are used....43
Table 11: Model Performance Measures, where the top 630 terms are used....43
Table 12: p-values for nxn comparisons of the FSs over VMs for elastic net....44
Table 13: Recommendations for writing easy to read health materials....50
Table 14: Kendall's Tau correlation of features and DISCERN. The feature category indicates whether that feature is used in assessing Wikipedia Content (W) and/or used in information quality assessment of generic content other than Wikipedia (O). The features (H) are constructed based on the health websites' writing styles and health literacy guideline recommendations respectively. 53
Table 15: Cross-validation results....55

Table 16: The ranks of the features for Elastic net 55
Table 17: The ranks of the features for Random Forest.....56
Table 18: Cross-validation results after feature elimination 58
Table 19: The results of the models as alpha changes 58
Table 20: Data Set Information.....82
Table 21: All selected terms and weirdness scores 84
Table 22: Model Performance Measures with textual features.....96
Table 23: Model Performance Measures, where the top 350 terms are used with elastic net.....96
Table 24: The ranks of the textual features for Elastic net.....97

LIST OF FIGURES

Figure 1: Process of selection of the websites 22
Figure 2: Methodology.....30
Figure 3: Existence of candidate terms on websites 40
Figure 4: Accuracy scores of classification algorithms to assess coverage, where different numbers of terms sorted according to their Weirdness coefficients are used (a) The content-based feature set (FS1) with the BM25 scores (VM1) is used. (b) The content-based feature set (FS1) with the percentage of the existence of terms (VM3) is used. 42
Figure 5: Accuracy scores of classification algorithms to assess coverage, where different numbers of terms sorted according to their Weirdness coefficients are used (a) The hybrid feature set (FS2) with the BM25 scores (VM1) is used. (b) The hybrid feature set (FS2) with the percentage of the existence of terms (VM3) is used. 43

LIST OF ABBREVIATIONS

ADA	American Diabetes Association
ATR	Automatic Term Recognition
AQA	Automated Quality Assessment
EBM	Evidence Based Medicine
FKGL	Flesch Kincaid Grade Level
FS	Feature Set
HON	Health on the Net
IDF	Inverse Docement Frequency
JAMA	Journal of the American Medical Association
LOOCV	Leave one out cross validation
MCC	Matthew's Correlation Coefficient
MeSH	Medical Subject Heading
NLP	Natural Language Processing
POS	Part of Speech
SmPC	Summary of Product Characteristics
SMOG	Simple Measure of Gobbledygook
SVM	Support Vector Machine
TF	Term Frequency
VM	Vectorization Method
VSM	Vector Space Model

CHAPTER 1

INTRODUCTION

Many people use search engines to access useful and high-quality health information on the web (Bull, 2010; Fox & Duggan, 2013). According to the Pew Research Center's Internet and American Life Project report, one in three American adults use the internet to retrieve health information (Fox, 2011). Moreover, 72% of Internet users use the web for accessing health information of any kind (Fox & Duggan, 2013). 63% of adults in Great Britain used the Internet to look for health-related information in 2019 and the usage has been rising 9 percentage points from 2018 (Prescott, 2019). Information pollution, however, results from the growth of information generated by non-expert users in the domain. Without access to high-quality information, patients cannot make informed choices about their treatment. Users can either end up leaving their current treatments prescribed by their doctors or spending on ineffective treatments and drugs that are promoted by commercial companies.

The process of interacting with patients, when there are treatment alternatives or uncertainty in outcomes, is called shared decision-making. The context of decision-making especially focuses on life-threatening illnesses and various treatment options with benefits or risks (O'Connor, Llewellyn-Thomas, & Flood, 2004; Mandelblatt, Kreling, Figeuriedo, & Feng, 2006). Many studies show that patients would like to take an active role in treatment decision-making. Marahrens et al. (Marahrens, et al., 2015) investigated shared decision-making preference in treatment choices of patients with type 2 diabetes. Their study found that 74.3% of patients (599 of 806) prefer shared decision-making. Bruce et al. (Bruce, Tucholka, Steffens, & Neuman, 2015) stated that an improvement of patient satisfaction was observed with active participation in treatment decision-making. This participation contributes to the post-treatment quality of life. To participate in treatment decisions, patients should be knowledgeable about the risks and benefits of each treatment. Most patients need guidance on how to consider their values and long-term goals. Therefore, accessing quality health information becomes life-critical.

The importance of information quality has also increased significantly. Misinformation—such as fake news, spam, and hoaxes—is a fundamental concern

on many platforms. The problem has been increasingly manifesting itself on social media, where the spread of misinformation is common. Crowdsourcing platforms such as Wikipedia have devised some solutions to ensure the coverage, consistency, and accuracy of information, yet the problems have not been fully solved. Although information quality on these platforms has been studied extensively, it has still not received enough attention in the health-related domain. Guidelines including DISCERN (Charnock, Shepperd, Needham, & Gann, 1999), HONcode (Health On the Net, 2018), and JAMA (Silberg, Lundberg, & Musacchio, 1997) have been developed to help standardize the reliability of medical and health information and to provide users with a means to assess the quality of written information on treatment choices for a health problem. These guidelines, however, require domain knowledge, and manual assessment of a website takes a significant amount of time. In addition, these guidelines suffer from limitations. For instance, HONcode cannot guarantee the accuracy and completeness of medical information at any given time (Health On the Net, 2018).

There are also professional health literacy guidelines, including instructions and recommendations for writing high-quality healthcare web content (Department of Health, 2003; Medicines and Healthcare products Regulatory Agency, 2014; MedlinePlus, 2018). Professional health literacy guidelines suggest that content should be written to include certain textual features such as using short sentences and paragraphs and having the maximum reading level for the sixth to eighth grade. In addition, all of them have recommendations related to content such as avoiding the use of “jargon” and providing the definition of special terms. However, in the majority of the studies in the literature, only readability has been investigated in only manual assessment for healthcare web content (Fitzsimmons, Michael, Hulley, & Scott, 2010; San Giorgi, de Groot, & Dikkers, 2017).

Health websites are also expected to include content that is in line with the evidence-based information. Evidence-based medicine (EBM) is presented by Evidence-Based Medicine Working Group as a new paradigm for medical practice in 1992. They define EBM as “the conscientious, explicit, and judicious use of the current best evidence in making decisions about individual patients’ care by gathering the best available external clinical evidence from systematic research" (Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996). It aims to ensure that medical decisions are evidence-based that integrate individual clinical expertise and the best external evidence. In order to achieve this, the EBM approach brings together experts who authoritatively review hundreds of studies on topics in their specialty. At the end of these reviews, a list of effective treatments is provided. Different resources for evidence-based decision-making have been reported for different health issues. In the scope of the thesis, “Standards of Medical Care in Diabetes” released by the American Diabetes Association (ADA), which is one of the well-known diabetes-specific gold standards for EBM, is used as a knowledge base for information coverage assessments of health websites (Association, 2016). ADA is intended to provide individuals with components of diabetes care, general treatment goals, and tools to evaluate the quality of care. It performs an extensive clinical literature search

to update the standards annually based on the quality of new evidence by the ADA's Professional Practice Committee.

Assessment of the content quality of health-related websites is an important problem but encounters challenges. (1) It requires an approach that can automatically evaluate the content coverage and information quality of health websites for end-users. How to automatically assess the website quality is an open question. (2) Such an automated approach should provide an interpretation of which textual and content-based features are valuable in assessing the quality of health-related websites.

In this context, the main focus the thesis is to address the above issues by proposing a novel method to assess the quality of health websites. To evaluate the proposed method, extensive experiments are conducted on the dataset of type 2 diabetes websites with multiple feature sets: textual and content-based features set, and comprehensive feature set derived from professional health literacy guidelines. Content-based features were acquired using an evidence-based practice guideline in diabetes and were evaluated based on the presence of terms on websites, but not semantically. A number of textual features inspired by professional health literacy guidelines and commonly used features to assess the quality of the information in other domains have also been used.

1.1. Research Questions

This thesis aims to answer the following research questions:

RQ1: Which textual evidences reveal the quality of health websites? Are linguistic styles of high-quality websites different from the low-quality ones?

RQ1.1: Which of the main textual features derived from generic information quality related studies and professional health literacy guidelines are useful in assessing the quality of websites?

RQ1.2: Is there any relationship between the frequencies of certain sentence types and the information quality of websites? How is the number of auxiliary verbs related to the quality of websites?

RQ2: How content-based evidences are related to the quality and coverage of websites?

RQ3: Does the use of both textual and content-based features perform better in the classification of websites according to their information quality compared to the sole use of textual or content-based features?

RQ4: Is the comprehensive feature set derived from professional health literacy guidelines informative in assessing the information quality of websites? How

successful are these feature sets in the classification of websites according to their information quality?

In this thesis, a framework is proposed to automatically assess the quality and content coverage of health websites related to type 2 diabetes and mainly focuses on written information about treatment choices to assess quality. Two main studies are conducted to answer the above research questions. Each study follows a similar process which includes the shortcomings of existing quality assessment studies, feature extraction, and the evaluation of the proposed model by involving these features.

First, the existing quality assessment of health-related websites and their shortcomings are identified. Accordingly, a novel automatic framework is proposed that investigates the impact of using textual and content-based features in predicting the quality of health-related texts. In order to validate the model and the framework proposed in the thesis, a set of type 2 diabetes-related websites are collected. Two medical domain experts have evaluated these websites in terms of DISCERN and coverage. An assessment manual to be provided to the experts for the scoring process is prepared using ADA which included summaries and key points of selected treatment-related chapters and another one including instructions for rating each question with the DISCERN tool. Medical domain experts assessed the websites in light of these manuals. In order to measure information coverage, ADA is used to extract significant terms as a knowledge base, and professional health literacy guidelines and the features commonly used for assessing information quality in other domains are used to extract textual features. This study investigated whether there is any relation between textual features and information quality, how content-based features are related to the coverage of websites and how content-related medical terms can be obtained that are important for evaluating coverage. Lastly, how textual and content-based features individually and combined are effective in classifying websites according to their information quality is investigated.

In the second study, the relation between comprehensive feature set derived from professional health literacy guidelines in assessing the information quality of websites is investigated. Most of the textual features presented in the first study were derived from the recommendations of professional health literacy guidelines, but numerous visual presentation and writing style related features were omitted from the initial analysis for the simplication purpose. Consequently, in the second part of the study, a comprehensive feature set is constructed and correlation analyses are conducted between these variables. In addition, their effect on website classification according to information quality is studied and the most significant and effective features are identified using the model regression and the mean decrease accuracy coefficients. Lastly, the study also focuses on feature elimination to analyse the importance of features.

1.2. Contribution of the Thesis

The contributions of this thesis are enlisted as follows:

- The features commonly used for assessing information quality in other domains—such as readability, image count, the number of auxiliary verbs, and section count—are used as indicators and this encourages the applicability of the textual features in the health domain. A new set of textual features is proposed inspired mainly from professional health literacy guidelines in addition to the features commonly used for assessing information quality in other domains and their effectiveness is compared with the ones in the literature for type 2 diabetes-related websites. A comprehensive feature set derived from professional health literacy guidelines is proposed. Also based on sentence types used in health websites, a new set of textual features is proposed.

- The framework assesses the degree of information coverage, which is an important indicator of website information quality as an integral part of accuracy. The studies in the related literature have not studied this factor in this domain before.

- The proposed framework requires the use of an evidence-based practice guideline such as ADA guideline and a recently introduced corpus, iWeb to measure content coverage and information quality automatically. The gold standards for evidence-based medicine are also available for other health domains such as the American Heart Association (American Heart Association, 2018) and the National Comprehensive Cancer Network (National Comprehensive Cancer Network, 2018) guidelines. This facilitates the applicability of the proposed method in other health domains without requiring an input from a human expert.

- Both textual and content-based features have been used together for information quality evaluation in the health domain for the first time in the literature. The impact of using both types of features on quality assessment has not been measured specifically for health domain before.

- The framework uses the DISCERN tool that assesses the quality of written information about treatment choices for a health problem rather than using the JAMA or HONcode which mostly look for features such as authorship, references, advertising, or timeliness. DISCERN concerns how well the website supports shared decision-making for patients with treatment choices and properties.

1.3. Organization of the Thesis

The rest of the thesis is divided into five chapters. Chapter 2 reviews the literature surveying the quality criteria for health domain, term extraction, and presents a review of quality instruments for health-related websites.

Chapter 3 introduces the data collection procedure in detail. Collection of websites, evaluation of them, and how websites are ranked are given in this chapter.

Chapter 4 presents a method to automatically evaluate the quality of diabetes websites including textual and content coverage features. The study is three-staged: in the first phase, textual features are analysed, in the second phase, content-based features are analysed and lastly, textual features processed in the first phase and content-based features are used for automated assessment of information quality of health websites.

Chapter 5 presents information quality assessment using professional health literacy guideline recommendations. In this chapter, a comprehensive textual feature set derived from these guidelines for assessing the information quality is used and the results are presented.

Finally, Chapter 6 summarizes the overall conclusions of the thesis, discusses the limitations, and provides planned future directions.

CHAPTER 2

LITERATURE REVIEW

This chapter provides a review of the techniques used in the proposed framework and explains the grounds for the proposed framework. In the first section, the information quality studies in general, and the ones specific to health websites are presented. The information quality assessments in Wikipedia are given in the second section. Then, the related work regarding quality term extraction from content is explained. The chapter ends with information quality instruments for health-related websites, which include recommendations for writing high-quality healthcare web content for end-users.

Information quality can be defined as the accuracy of data for use, and it can deal with several dimensions, which are accuracy, accessibility, completeness, and ease of understanding (Floridi, 2014). Stvilia et al. performed an extensive search on the literature to group the criteria used to evaluate the quality of health information on the web (Stvilia, Mon, & Yi, 2009). According to their study, the most frequently used quality criteria include accuracy, completeness, authority, usefulness, and accessibility provided.

Accuracy criteria states that the content of information should be examined with the evidence available to support the claims. It also seeks the presence of references that can verify information, an examination of whether it contains grammatical errors, and whether the information is based on medical research.

Completeness criteria looks for whether the information is complete and clear, additional information about the domain is provided and whether users can find the complementary information mouse-click away.

Authority criteria requires information about who published the page, the credentials of authors, and whether the writers backed by a known organization or affiliated with a university. It also includes the contact address of the owner, website sponsors and investors, and the conflict of interest declaration.

Usefulness criteria necessitates a well-designed website. It includes the ease of navigation, search capability, and the existence of "help" and "search" tips.

Accessibility criteria concerns with access and security of websites. It requires information about sponsoring group or funding, members of the healthcare community, and policy.

2.1. Quality assessment of health information

Increased usage of health information on websites raises concerns about reliability and quality. Several studies have been conducted to address these concerns. The proposed solutions are mainly carried out using two different approaches: (1) manually checking the contents using guidelines; (2) automatically evaluating the contents using information retrieval and extraction tools.

2.1.1. Manual assessment of online health content

Quality assessment of health information can be manually carried out using guidelines such as HONcode (Health On the Net, 2018), JAMA benchmarks (Silberg, Lundberg, & Musacchio, 1997), and widely-used DISCERN tool (Charnock, Shepperd, Needham, & Gann, 1999; Fahy, Hardikar, Fox, & Mackay, 2014).

DISCERN has been designed to help health consumers and information providers to assess the quality of written information about treatment choices for a health problem (Charnock, Shepperd, Needham, & Gann, 1999). DISCERN has 16 questions, with each rated on a scale from 1 (definite NO) to 5 (definite YES). Any rating in between implies the presence of the element in the question to some degree. These questions are categorized into three sections. Section 1 (questions 1 to 8) assesses the trustworthiness of the information source of the site, Section 2 (questions 9 to 15) concerns how well the site supports shared decision making for patients with respect to treatment options and properties, and Section 3 (question 16) evaluates the overall quality rating of the website, with a rating of 1=Low to 5=High.

DISCERN Questions

SECTION 1: Is the publication reliable?

1) Are the aims clear?
2) Does it achieve its aims?
3) Is it relevant?
4) Is it clear what sources of information were used to compile the publication (other than the author or producer)?
5) Is it clear when the information used or reported in the publication was produced?

6) Is it balanced and unbiased?
7) Does it provide details of additional sources of support and information?
8) Does it refer to areas of uncertainty?

SECTION 2: How good is the quality of information on treatment choices?

9) Does it describe how each treatment works?
10) Does it describe the benefits of each treatment?
11) Does it describe the risks of each treatment?
12) Does it describe what would happen if no treatment is used?
13) Does it describe how the treatment choices affect overall quality of life?
14) Is it clear that there may be more than one possible treatment choice?
15) Does it provide support for shared decision-making?

SECTION 3: Overall Rating of the Publication

16) Based on the answers to all of the above questions, rate the overall quality of the publication as a source of information about treatment choices.

The HON (Health on the Internet) Foundation is a non-profit, nongovernmental organization based in Switzerland that promotes reliable online health information (Health On the Net, 2018). According to the foundation's website www.hon.ch, the HONcode of conduct is the oldest and most common quality seal for online health information. HON was created to promote useful and reliable online health information and accredits websites that provide health information. HONcode and its principles aim to improve the quality of health information on the Internet. To be certified, site owners have to apply for a review of their site. The medical professionals of the HON review committee check if the site adheres to all eight of the HONcode ethical principles (According to www.hon.ch: 'Authority' 'Complementarity' 'Confidentiality' 'Attribution' 'Justifiability' 'Transparency' 'Financial disclosure' 'Advertising'). Health-related websites containing all the eight principles of certification are allowed to use the HONcode logo. A certified site receives a biennial review, beginning one year after the initial certification, or following user alert.

JAMA benchmark, established by the Journal of the American Medical Association, also provides a qualitative assessment of websites for a number of criteria: authorship, attribution, currency and disclosure (Silberg, Lundberg, & Musacchio, 1997). Authorship means that the website properly identifies the author's name, links, and credentials. Attribution means an effective reference of content that is served across a website. The currency implies the up-to-dateness of the content, which also necessitates the display of the last published or the last update dates. Finally, the disclosure category indicates that the "owner" site is highlighted with any potential conflict of interest statement.

Currently, with the technological advancements in web technologies, many websites have started to offer information to support patient decision-making for end-users to assess their medical condition and give appropriate recommendations. DISCERN is one of the most frequently used tools which propose manual guidelines for end-users to assess the quality of the treatment choices of websites. According to DISCERN, high-quality information should exhibit the following features: description of treatment options and the benefits and risks of each treatment, description of the effects of the treatment options on day-to-day activity, and suggestions for shared decision making. Learning and applying these guidelines, however, is time-consuming and difficult (Khazaal, Fernandez, Cochand, & Reboh, 2008). (To use this tool; firstly user needs to learn how to rate each question and then calculate the total rating for each website). DISCERN has been used for evaluation of the quality of the information on websites in many domains, such as Mental health-related websites (Khazaal, et al., 2009), Prostate Cancer (Rees, Ford, & Sheard, 2002), Social Phobia (Khazaal, Fernandez, Cochand, & Reboh, 2008), Cocaine Addiction (Khazaal, Chatton, Cochand, & Zullino, 2008), Bipolar Disorder (Morel, Chatton, Cochand, Zullino, & Khazaal, 2008), Type2 diabetes (Weymann, Harter, & Dirmaier, 2014), Childhood Epilepsy (Cerminara, Santarone, Casarelli, Curatolo, & Malhany, 2014), Thumb Sucking (Kiran, et al., 2015), Breast Cancer (Nghiem, Mahmoud, & Som, 2016), Renal Diet (Lambert, Mullan, Mansfield, Koukomous, & Mesiti, 2017), Heart Failure (Cajita, Rodney, Xu, Hladek, & Han, 2017), and Lingual Orthodontics (Olkun, Demirkaya, & Aras, 2019). In these studies, domain experts manually assessed whether search engines are able to retrieve high-quality websites and rank them higher in their search results. At last, these studies reached discrepant conclusions. For instance, the studies (Morel, Chatton, Cochand, Zullino, & Khazaal, 2008; Cerminara, Santarone, Casarelli, Curatolo, & Malhany, 2014; Nghiem, Mahmoud, & Som, 2016; Lambert, Mullan, Mansfield, Koukomous, & Mesiti, 2017) concluded that search engines are limited in this respect; yet the studies (Khazaal, Fernandez, Cochand, & Reboh, 2008; Khazaal, Chatton, Cochand, & Zullino, 2008; Kiran, et al., 2015; Olkun, Demirkaya, & Aras, 2019) claimed the opposite conclusion.

Although these guidelines are good quality indicators, they also have weaknesses. Some studies have found that HONcode and DISCERN are not related to content quality (Morel, Chatton, Cochand, Zullino, & Khazaal, 2008; Khazaal, Chatton, Cochand, & Zullino, 2008). Also, although the HONcode logo exists, many websites do not fully cover all eight criteria (Meric, et al., 2002). JAMA is one of the popular rating tools, but Barker has suggested that websites meeting all its criteria contain more incorrect information than other sites (Barker, Charlton, & Holstege, 2010). Many studies used DISCERN guideline, however DISCERN is criticised to be too time-consuming to manually check, have limited public recognition (Khazaal, Fernandez, Cochand, & Reboh, 2008), have poor inter-rater reliability (Rao, Mohapatra, Mishra, & Joshi, 2012) and be subjective (Nghiem, Mahmoud, & Som, 2016). They suggested that one tool is not enough for quality assessment, but it should be used with other tools.

Trustworthiness

There are also studies which concern with the assessment of the trustworthiness of websites. Trustworthiness is evaluated using the following criteria: currency, references, author information, feedback, and spelling errors. Lee et al. (Lee, Shin, & An, 2018) evaluated the stroke-related websites regarding trustworthiness, readability, and sustainability. In their study, trustworthiness was evaluated by authority, currency, and accuracy criteria. Their results for authority showed that 21% of websites did not present the publisher or contact information. Regarding accuracy, only 33% of websites posted health information based on medical research evidence along with citations. In terms of currency, 79% of the selected websites displayed the published date of the information on their websites.

The trustworthiness of websites focused on nutrition, dieting, and fitness with surface, references, authority, and feedback credibility components were examined (Machackova & Smahel, 2018). Their study found that the importance of these criteria was linked to demographics, the motivation of visiting websites, and online experiences. The surface component was more important for women; for users with higher education, references were important; authority was more important for users who are more active online; and for users with weight loss motivation, the feedback was more important than others.

The comparative effects of spelling errors and inappropriate capitalization on ratings of trustworthiness were also quantified (Witchel, et al., 2020). They tested inappropriate capitalization only, misspelling only, no errors, or the combination of both errors. Their study showed that inappropriate capitalization and misspelling affect trustworthiness. Also, the combination of both errors induced a strong effect on ratings of trustworthiness.

2.1.2. Automated assessment of online health content

Manual assessments of online health content were presented in the previous section. Although these guidelines standardize the assessment of website quality, their results may differ because manual evaluation is subjective. Also, manually assessment takes time to perform quality scores. Instead of manual assessment, automated approaches can be a more rapid and objective evaluation of online information.

Sağlam and Temizel (Sağlam & Temizel, 2014) proposed a tool that evaluates websites' bias and quality based on sentiment analysis, relevance feedback and information retrieval techniques. They extracted the commercial terms from content to detect any potential bias and the weights of terms are computed using Robertson-Sparck-Jones to differentiate the terms that appear mostly in relevant web pages. Next, they used relevance feedback to obtain quality-measuring queries that are related to type 2 diabetes. Finally, they computed a quality score, which was the

average of the weights of all the words in the text. Their proposed framework achieved 77% accuracy in predicting the timeliness of content based on time interval.

Another study made by Griffiths et al. (Griffiths, Tang, Hawking, & Christensen, 2005) developed an automated quality assessment (AQA) procedure, which was designed to rank depression websites automatically. AQA used a domain-specific set of weighted keywords to assess the evidence-based quality of depression websites in accordance with the Oxford University Centre for Evidence-Based Mental Health's guidelines. They evaluated the validity of AQA relative to human-rated evidence-based quality scores and compared it with Google PageRank. Their proposed method, AQA, produced rankings that exhibited a stronger correlation with the human-rated evidence-based scores than Google PageRank's rankings did.

Liu et al. (Liu, Yu, Wu, Qing, & Peng, 2019) proposed a framework that used feature-based and text-based methods for detecting unreliable health-related information on Chinese social media. Their feature-based method was based on specific word frequencies which were extracted with word cloud, the length of title, text length, paragraph length, modal sentences and frequency of punctuation marks. They were found as significant in determining information quality.

In addition, there are studies related to the automated detection of HONcode principles to assess the quality of health websites. Boyer et al. (Boyer, Frossard, Gaudinat, Hanbury, & Falquetd, 2017) proposed an automated system for the detection of eight HONcode principles, which they used as features to test the level of trustworthiness of websites. Their system's evaluations, however, did not match human experts' assessments, because the system could detect structured elements but could not evaluate the meaning of the text. Oroszlányová et al. (Oroszlányová, Teixeira Lopes, Nunes, & Ribeiro, 2018) investigated whether the characteristics of documents such as readability level, last update time, references, privacy policy, and contacts of the author are practical to predict information quality in multivariate analysis using HONcode certification as ground truth of quality. The features of the document, some of them are related to the HONcode principles, were manually assessed by experts except for seven readability level indicators. They built statistical models to find the features that are important to predict quality: they first included every feature, second without the features related to the HONcode principles and third without Utilization Review Accreditation Commission features. They noted that even without including the features related to the HONcode principles in the models, prediction models have had high accuracy.

2.2. Quality assesment of Wikipedia articles

Wikipedia is a large scale collaborative content that communities can produce and the numbers of digital documents being generated are huge and continue to grow. To assess the quality of information in the article, Wikipedia's editorial team tags articles according to Wikipedia's six-level quality taxonomy: *featured article, A-*

class, *good*, *B-class*, *start*, and *stub*. *Stub* implies the lowest quality, and *featured article* indicates the highest quality. For instance, "well written, comprehensive, well researched, neutral, and stable" are the established criteria for pages labelled as the *featured articles* (Wikipedia, 2018). As content generation is rapid and immense, manual assessment of articles is not scalable.

Therefore, numerous solutions have been proposed to assess Wikipedia articles' quality automatically (Rassbach, Pincock, & Mingus, 2007; Blumenstock, 2008; Dalip, Gonçalves, Cristo, & Calado, 2011; Dalip, Lima, Cristo, & Calado, 2014; Shen, Qi, & Baldwin, 2017). Statistical indicators derived from articles—such as word count, readability, editor information, network links between articles, and edit history—are used to measure information quality; of these, the number of editors and the number of edits are specific to Wikipedia, like collaborative articles (Rassbach, Pincock, & Mingus, 2007; Blumenstock, 2008; Morten, Cosley, & Riedl, 2013; Dalip, Lima, Cristo, & Calado, 2014).

Blumenstock (Blumenstock, 2008) used just word count as a measure of quality on Wikipedia articles and achieved 96.31% accuracy in binary classification. Rassbach et al. (Rassbach, Pincock, & Mingus, 2007) trained a maximum entropy classification model on five different tagged articles, using 50 different kinds of quality indicators, including word count, paragraph count, part-of-speech usage, web-specific features, and readability metrics. In addition to textual features, they observed that images on websites are a strong indicator of quality. Dalip et al. (Dalip, Lima, Cristo, & Calado, 2014) proposed a method to assess Wikipedia article quality. They used three different Wikipedia article datasets and 68 features related to history, readability, style, length, network, and structure. They observed that textual features, such as article length and structure features are useful domain-independent attributes for measuring information quality. They concluded that the best set of features varies across datasets (Morten, Cosley, & Riedl, 2013; Dalip, Lima, Cristo, & Calado, 2014).

In addition to these studies, a hybrid approach combining neural network document embedding with structural features such as readability scores, length of the document, number of images, and number of citations was proposed (Shen, Qi, & Baldwin, 2017). They used three different number of articles datasets (20K, 30K and 60K datasets) with six different classes and evaluated features together and separately. Their experimental results indicated that their hybrid model achieved 6.5% higher accuracy on 60K datasets than on the other datasets with respect to structural features. Performances on the 20K and 30K datasets, however, were very close when they used solely structural features with their hybrid method. They concluded that structural features play an important role in the classification of smaller datasets. As mentioned above, the models presented in the studies assessed the quality of articles using different features across different types of articles.

2.3. Quality term extraction from content

Health websites designed for patients comprise definitions, instructions and recommendations. ADA guideline inherently comprises this information for type 2 diabetes. Hence, ADA can be considered as a good candidate for assessing coverage. Coverage refers to how much the topic under examination mentioned at all. It has been manually assessed using the scales proposed in (Nilsson-Ihrfelt, et al., 2004) (Khazaal, Fernandez, Cochand, & Reboh, 2008). Nevertheless, it can be measured, to some degree, by comparing the content of two sources in terms of significant quality terms or phrases.

Mining quality phrases refer to automatically extracting salient phrases from a given corpus with syntactic and semantic word relationship (Liu, Shang, Wang, Ren, & Han, 2015). The selected phrases facilitate query suggestion and information extraction because many concepts and relations are demonstrated in phrases. For text-based information retrieval, phrase mining is a task for text analytics of various domains, such as news, science, documents and social media to find events, organizations, and products. The existing approaches based on natural language processing (NLP), term frequencies, term frequency-inverse document frequency (TF-IDF), C-value, weirdness coefficient, information gain, and chi-squared are also used to extract quality terms and phrases (Al-Thubaity, Alanazi, & Al-Tuwaijri, 2012; Sağlam & Temizel, 2014). However, frequency-based statistics might result in misleading assessment when the frequencies are very close. Al-Thubaity et al. (Al-Thubaity, Alanazi, & Al-Tuwaijri, 2012) proposed a method to classify Arabic special texts using weirdness coefficient as a feature selection method and their study showed promising classification accuracy with respect to chi-squared.

Furthermore, a framework was proposed that extracted quality phrases from text corpora integrated with phrasal segmentation (Liu, Shang, Wang, Ren, & Han, 2015). This approach explores the benefits of phrase extraction and phrasal segmentation. In addition to this approach, they proposed a new automated phrase mining framework, which supports any language as long as a general knowledge base in that language is available, such as Wikipedia (Shang, et al., 2018).

Moreover, Li et al. (Li, Yang, Wang, & Cui, 2017) proposed a method, which considers both intra-cohesion and inter-isolation in mining phrases to guarantee appropriateness and to ensure the completeness of phrases.

2.3.1. Fact-checking

The process of analysing a piece of information, assessing coverage, cross checking it with existing knowledge, and verifying its accuracy is called “fact checking”. Studies refer to background knowledge and open data repositories to extract factuality judgements. To detect fake news, some studies have used NLP to extract

named entities and to detect subjectivity and linguistic approach (Goasdoué, et al., 2013; Conroy, Rubin, & Chen, 2015; Rashkin, Choi, Jang, Volkova, & Choi, 2017).

Rashkin et al. (Rashkin, Choi, Jang, Volkova, & Choi, 2017) analysed various lexical features—such as usage of superlatives, comparatives, and adverbs—to investigate the differences between reliable and less reliable news sources. Their results indicated that first-person and second-person pronouns and subjectivity, superlatives, and modal adverbs appear more frequently in fake news. Conversely, comparatives, money, and numbers are used more in truthful news. The lexical features used in the above studies were not applicable for the research problem in this study, because this thesis focused on the content of health web pages.

2.3.2. Author detection

There are studies for automatic detection of the authorship attribution that has been a subject of increasing research interest in the recent years with many relevant applications such as plagiarism detection (Vysotska, Burov, Lytvyn, & Demchuk, 2018), authorship identification (Gómez-Adorno, Sidorov, Pinto, Vilariño, & Gelbukh, 2016) and authorship profiling (Patra, Banerjee, Das, Saikh, & Bandyopadhyay, 2013). Patra et al. (Patra, Banerjee, Das, Saikh, & Bandyopadhyay, 2013) used lexical, syntactic, and structural features for identifying the gender and age group of the authors. These features are word class frequency (about money, job, sports, etc.), positive and negative word class, stop words frequency, smiley words frequency, foreign language frequency, pronouns frequency, average length of sentences and punctuations. Also, Reddy et al. (Reddy, Vardhan, & Reddy, 2017) proposed a model that calculated document weights with n-grams. Then, they used the most frequent terms and their weights to classify the gender of the author. The lexical features used in the above studies can be applicable to the research problem in study. This thesis, however, focused on the content of health web pages to assess the coverage, not only the prediction of authorship attribution.

2.4. Professional health literacy guidelines for writing good quality information for patients

Professional health literacy guidelines include instructions and recommendations related to writing high-quality healthcare web content and they are prepared to ensure content publishers are producing clear and easy-to-follow health information for end-users regardless of age, background or reading level (Department of Health, 2003; Medicines and Healthcare products Regulatory Agency, 2014; MedlinePlus, 2018). Professional health literacy guidelines suggest that content should be written in plain language. Use of bullets and shortlists, meaningful headings, writing short sentences and paragraphs providing supplementary images, use of a maximum reading level of sixth to eighth grade and avoidance of passive voice are common examples of recommendations offered by these guidelines. In addition, they all have

suggestions related to the content of healthcare web content such as avoiding the use of "jargon", focusing on what patients need to do to manage or treat, and providing only essential information with respect to what patients need to know.

Usability studies state that users are initially interested in learning the basics. Therefore, health website content should directly specify what to do and how to do it. For instance, sentences should be written as imperatives to keep information direct and to the point: "*Check your blood pressure every 2 years, especially if you are aged 40 or older*" (U.S. Department of Health and Human Services, 2010).

Moreover, professional health literacy guidelines emphasize using interactive content, such as questions and answers, to engage users and to divide up the text. In particular, how the number of auxiliary verbs and sentences relates to the assessment of quality was examined because behavioural recommendations principally use declarative and advisable styles of language to emphasize the benefits of taking action. For this purpose, sentences are generally written using auxiliary verbs: "*You may need to get tested before age 50 if colorectal cancer runs in your family. Talk with your doctor and ask about your risk of developing colorectal cancer*" (U.S. Department of Health and Human Services, 2010). Sentences containing auxiliary verbs are often used in the context of the process of making decisions about treatment was also noticed: "*There is strong and consistent evidence that obesity management can delay progression from prediabetes to type 2 diabetes and may be beneficial in the treatment of type 2 diabetes*" (American Diabetes Association, 2016).

In many studies, these guidelines are used to measure patient literacy, to find information from leaflets, or to improve documents. Raynor et al. (Raynor, De Veene, & Bryant, 2014) conducted user testing to understand the effectiveness of the Summary of Product Characteristics (SmPC) document, which is a legal document describing how to use a medicine and targeted for health professionals. The process of user testing examined by general practitioners with iterative processes was tested in 3 rounds and each round the document was revised for improvement. Although their study met regulatory standards, it did not meet user needs. Their study showed that the usage of headings, bullet points, shorter sentences, and the direct style of writing improved the effectiveness of SmPC document in describing essential information.

Dickinson et al. (Dickinson, Raynor, Knapp, & MacDonald, 2016) conducted a user test study that analysed how patients use the headline section in leaflets to find information about medicines. This study is based on the recommendation of Medicines and Healthcare products Regulatory Agency (Medicines and Healthcare products Regulatory Agency, 2014), which stated leaflets are recommended to include a headline section. Their questionnaire study has shown that the participants value the presence of the headline section, which facilitates to find important information about their medicines.

Furthermore, Durrant et al. (Durrant, Taylor, Thompson, Usher, & Jackson, 2019) analysed patient education resources to understand patient literacy in pressure injury. They used the National Health Service Toolkit for Producing Patient Resources (Department of Health, 2003) and compared it to an internationally-advocated leaflet of pressure injury. Their findings showed that leaflets mostly met the required production and content guidelines, but patients did not interact with these materials and demonstrated limited health literacy about pressure injury.

Among the guideline recommendations, although many of them coincide with the features, which are commonly used to measure information quality in the literature, only readability has been widely studied for health documents. Giorgi et al. (San Giorgi, de Groot, & Dikkers, 2017) conducted a study that measures the readability of online English health information about recurrent respiratory papillomatosis and they concluded that none of the websites met the recommended values of readability. Moreover, a study assessing the readability of online Parkinson's disease information found that only 1% of the top 100 Parkinson disease information websites were fully comprehensible to the average adult (Fitzsimmons, Michael, Hulley, & Scott, 2010).

2.5. Information retrieval and extraction techniques

In this section, an introduction of information retrieval and extraction techniques is presented. An overview of text-based ranking techniques and automatic term recognition techniques is given.

2.5.1. Text-based document ranking

Text-based document ranking is the primary way to manage high-volume documents that provide a search system. Usually, it assigns scores to documents based on the distribution of query terms both in the document and in the corpus. A query term can be a single word, tuples of words (n-), or synonyms. There are three main types of ranking models: Boolean matching, vector space model (VSM), and probability model.

In Boolean matching models, it identifies a document relevant to a query; however, it disregards the degree of matching. In VSM, a document is represented in a multidimensional field with a perpendicular dimensional vector for each term in the corpus (Salton, Wong, & Yang, 1975). A distance function is used that calculates the degree of overlap between the query and the document. The vector can consist of ones and zeros, which indicate the presence or absence of a term in the document respectively. The term and document statistics can also be used.

Okapi BM25, an example of probabilistic retrieval models, considers document length for normalization since term frequency values tend to be larger in longer documents and it causes a potential bias where longer document will be assigned to a higher score (Robertson, Walker, Beaulieu, & Willett, 1999). In order to recover

from that effect, they derived the document length normalization. Document length normalization is important especially for the corpora in which the length of documents highly differs. Okapi BM25 (Eq.1) computes the score for each word in the query as follows;

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i).\frac{f(q_i,D).(k_1+1)}{f(q_i,D)+k_1.(1-b+b.\frac{|D|}{avgdl})} \quad \text{(Eq.1)}$$

where $f(q_i, D)$ is q_i's term frequency in the document D, $|D|$ is the length of the document D in words, and *avgdl* is the average document length in the text collection from which documents are drawn. k_1 and b are free parameters and set to values of 1.2 for k_1, 0.75 for b, which were found as the most reasonable parameters in several experiments in the literature (Manning, Raghavan, & Schütze, 2008). $IDF(q_i)$ is the inverse document frequency (IDF) that measures how many documents a term appears in and then measures how special term is for the query term q_i.

2.5.2. Automatic term recognition techniques

Automatic term recognition (ATR) has various application areas that have been studied for a long time. It is widely used for information retrieval applications. Depending on the nature of the text data and its implementation, ATR may concentrate on different features of the text, such as specific speech part labels for purposes. It can also be classified according to the use of background information, such as the general corpus area. ATR has two types of properties: unithood and termhood. The unithood measures indicate the colocation strength of units that comprise a term. Mutual information, log-likelihood, and t-test are examples of unithood measures. The termhood measures indicate association strength of a term to domain concepts (Kageura & Umino, 1996). TF-IDF and weirdness are examples of termhood measure. Weirdness coefficient is a measure, which is used to identify candidate terms based on the assumption that distribution of terms in a specialized corpus and in a general corpus differs significantly from each other (Ahmad, Gillam, & Tostevin, 1999). It is expressed by Eq.2:

$$Weirdness = \frac{f_s/\sum t_s}{f_g/\sum t_g} \quad \text{(Eq.2)}$$

where f_s is the frequency of word in the specialized corpus and f_g is the frequency of word in the general corpus; t_s is the total number of words in specialized corpus and t_g is the total number of words in the general one.

2.6. Summary

The first section has reviewed the recent approaches for quality assessment of health domain. These approaches were divided into two main categories, manual assessments and automated assessments. Based on the findings from the literature, these approaches suffer from two main drawbacks. First, most studies assess the quality of health information manually; however, firstly a user needs to learn how to rate each question and then calculate the total rating for each website. Also, to evaluate quality of websites users need domain knowledge about disease. Second, automated assessment studies are limited to evaluating the quality of web pages specifically in terms of content and coverage, as they consider features such as last update time, the number of references, objectivity or subjectivity, presence of advertisements, authorship, text length, and term frequencies.

In the second section, background information about assessing quality of different types of collaborative content articles has been given. Those studies have showed that word count was one of the significant baselines of assessing website quality. This thesis study is motivated to investigate text structure features, which can be better than using word count in assessing the quality and coverage of the information on health websites.

The third section has presented the studies related to extract special terms related to content of documents. This thesis study speculates that extracting health-related phrases while evaluating the quality and coverage of websites gives insight about assessment.

The fourth section has mentioned the studies related to professional health literacy guidelines for writing good quality information. Those studies have showed that investigating whether features that adhere to these guidelines are indicative of high-quality health-related content. In addition to these features derived from these guidelines, this study is motivated to investigate the use of various sentence types on high-quality health websites which have not been studied in the literature before.

Finally, in the last section, background information about information retrieval techniques that are employed in the proposed methods has been provided within the scope of the thesis.

CHAPTER 3

WEBSITE QUALITY ASSESSMENT

This chapter explains in detail the construction of the corpus of related websites and the rating methodology. As previously mentioned, this study focused on the type 2 diabetes domain. In type 2 diabetes, a chronic condition gradually leads to severe microvascular, macrovascular, neurological, and ophthalmological complications. The presence of interventions aimed at reducing the risk of these complications is particularly important. In addition, this domain was chosen as a case study because the treatment options and knowledge in the domain do not rapidly change as like other diseases such as cancer.

3.1. Collection of websites

For the collection of websites, keyword searches were performed on Google and Yahoo search engines in June 2016 and August 2016. Six keywords were selected based on type 2 diabetes and DISCERN questions. The following keywords used: *"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"*. A wide range of search keywords is used to ensure reaching websites that can be used to assess the coverage of websites related to DISCERN questions.

The ***first 30*** websites in English returned from each keyword query were retrieved. Many studies in the literature selected the first 20 pages as most people only refer to the first search page for their inquiries (Eysenbach, Powell, Kuss, & Sa, 2002); however, websites that were irrelevant, not accessible, requiring access fee, not in English, journal articles, news, video, personal experiment sites, sponsored links, advertisements, academic press, abstracts, and forum items, were excluded as the main scope of the proposed methodology was not to develop an information retrieval method. 180 websites from each search engine were visited, making a total of 360 websites; of these, 230 were eligible under the exclusion criteria. Removal of overlaps and duplicates of websites left 48 unique websites. In addition, 12 additional websites (not appeared in the search results) about type 2 diabetes studied

in the previous studies were included (Seidman, Steinwachs, & Rubin, 2003; Sağlam & Temizel, 2014). Only the HTML files of the selected websites were downloaded and stored. The search history and cookie information of the computer used for the study was erased before the search, which is in line with the study (Nghiem, Mahmoud, & Som, 2016). Figure 1 shows the process of selection of websites.

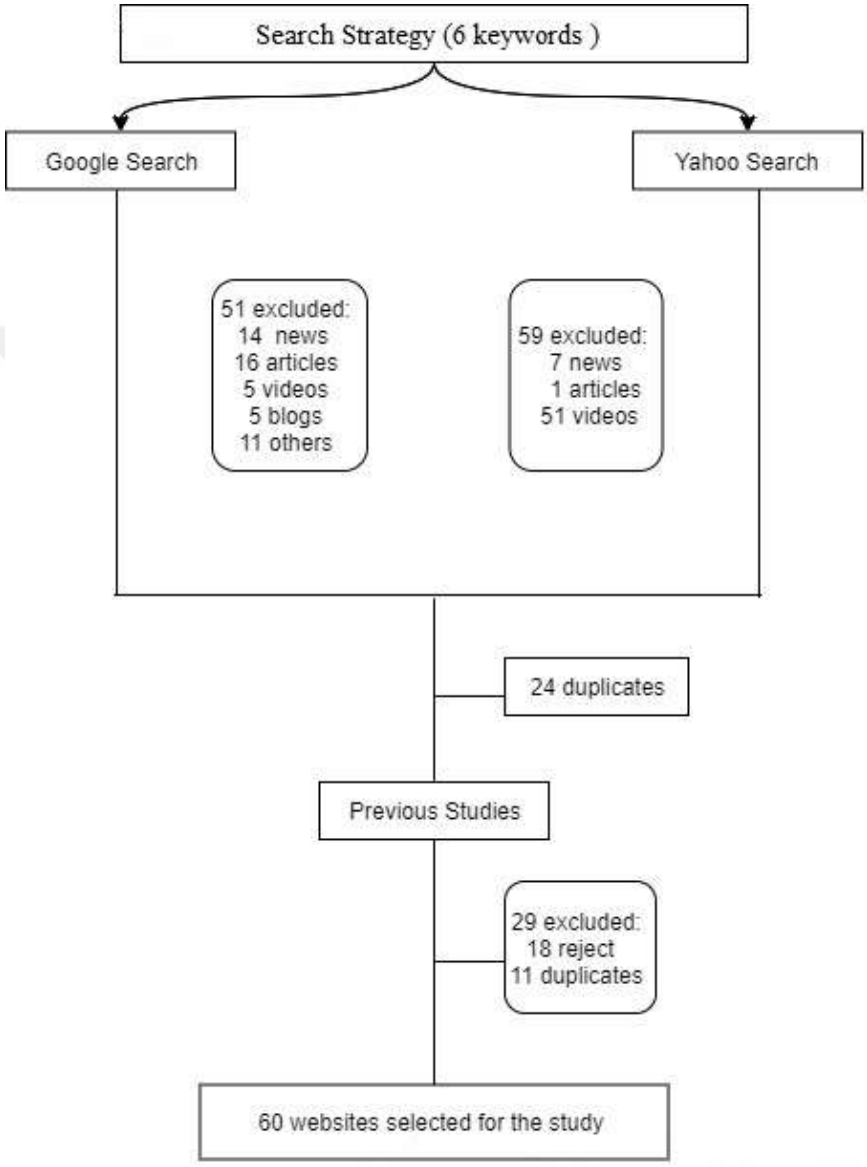


Figure 1: Process of selection of the websites

Web pages were downloaded and stored with HTTrack Website Copier with an HTML form. HTTrack is a free and easy-to-use offline browser utility. It allows downloading a World Wide Web site from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to computer.

Each web page was manually checked whether it was related to type 2 diabetes or not. The irrelevant pages were excluded due to the comprehensiveness criteria. 60 websites including a total of 433 web pages were obtained in the end. The HTML pages were also pre-processed before using them. After manually inspecting the content of websites, references and links using “ref” and “links” division tagging were separated from each other. Also, advertisement related fields, sidebars, navigation sections, and HTML fields that include unrelated information such as meta, script, and style, were removed so that they did not affect the analyses.

3.2. Evaluation of the websites

The origin of the sites is as follows (government: 6, university: 2, commercial: 29, nonprofit organization: 15, others: 8).

Presence or absence of Health on the Net (HON) code was also considered (Health On the Net, 2018). In search results, only 16 of 60 websites had HONcode certification. Table 1 shows that HONcode certificated websites.

Table 1: HONcode certificated websites

HONcode Certificated Websites
http://www.webmd.com
http://www.healthline.com/
http://www.everydayhealth.com/
http://patient.info/
http://www.webmd.boots.com/
https://www.drugwatch.com/
http://www.medicalnewstoday.com/
http://www.news-medical.net/
http://www.dlife.com/
https://diatribe.org/
http://www.bupa.co.uk/
https://www.drugs.com/
http://www.diabetesselfmanagement.com/
https://www.verywell.com/
http://www.emedicinehealth.com/

3.3. Rating of the websites by the experts

In this study, two different evaluations were conducted. The first evaluation concerned the assessment of content using DISCERN. The second evaluation concerned the assessment of coverage using evidence-based health information.

Although the main focus in this study was on coverage, data regarding the correctness of the content were also collected.

Two general practitioners, each with six years' experience, rated each downloaded website using a manual including instructions for rating each question with the DISCERN tool and another one including instructions for coverage assessment. The coverage assessment manual was prepared by using ADA's gold standard. The DISCERN and coverage assessment manuals for the experts are given in Appendix A. The experts assessed the websites in light of these manuals, detailed chapters of the ADA, and their expertise. The reason why experts rated instead of end-users is that experts could better determine whether the treatment options on the web pages are reported correctly and completely.

In the first approach, the DISCERN tool has been used for the assessment of content. DISCERN has 16 questions with three sections which were explained in detail in Section 2.1.1. For this study, Section 2 and Section 3 were considered, as they are related to the focus of this study. Therefore, the rating results were out of a possible 40 points, with a minimum score of 8 points, allocated according to the answers to 8 questions.

Table 2 presents how the websites performed for each question; it shows the average scores out of five. The question about "Is it clear that there may be more than one possible treatment choice?" scored highest; followed by the question about "Does it describe how each treatment works?" The lowest scoring questions were "Does it describe what would happen if no treatment is used?" and "Does it describe how the treatment choices affect the overall quality of life?"

Table 2: Average score of each DISCERN question

Discern Questions	**Mean Score(1-5)**
9. Does it describe how each treatment works?	3.75
10. Does it describe the benefits of each treatment?	3.64
11. Does it describe the risks of each treatment?	3.61
12. Does it describe what would happen if no treatment is used?	2.93
13. Does it describe how the treatment choices affect overall quality of life?	3.11
14. Is it clear that there may be more than one possible treatment choice?	3.95
15. Does it provide support for shared decision making?	3.38
16. Based on the answers to all of the above questions, rate the overall quality of the publication as a source of information about treatment choices.	3.59

Table 3: DISCERN scores

	Total(n=60)
8-13 (very poor)	11
14-19 (poor)	7
20-25 (fair)	8
26-31 (good)	5
32-40 (excellent)	29
Average DISCERN score	26.77

In the previous studies (Nghiem, Mahmoud, & Som, 2016) (Grant, Rodger, & Hoffmann, 2015), the researchers rated the websites as "excellent" (63-80), "good" (51-62), "fair" (39-50), "poor" (27-38), or "very poor" (15-26), based on three sections where the maximum score was 80; however, in this study only Section 2 and Section 3 were considered, where the maximum score was 40. Therefore, to be in line with previous studies, the scoring was readjusted as "very poor" (8-13), "poor" (14-19), "fair" (20-25), "good" (26-31), and "excellent" (32-40). The reason for halving these thresholds was that there are 16 questions in DISCERN and eight of them were used for evaluation in this thesis. The good and excellent levels were merged into one category labelled "high-quality," and the other three levels were merged into one category labelled "low-quality". The merging operation was conducted so as to avoid the sparsity in certain categories. The number of websites that were categorized as very poor, poor, fair, good, and excellent is shown in Table 3. The descriptive statistics of the websites are presented in Table 4. It demonstrates that the information quality of the websites returned by search engines varies. For instance, 18 out of 48 of the websites were labelled as "low-quality."

Table 4: Descriptive statistics of the websites, which show their DISCERN and coverage scores

		Mean	**Median**	**Standard Deviation**	**Min**	**Max**
All websites	DISCERN	26.77	30.00	11.60	8.00	40.00
	Coverage	14.93	17.50	7.14	0.00	24.00
Websites returned from search engines	DISCERN	28.52	32.50	10.46	8.00	40.00
	Coverage	16.04	18	6.13	0.00	24.00
Websites obtained from (Sağlam and Temizel, 2014) (Seidman, et al., 2003)	DISCERN	19.75	12.5	13.69	8.00	40.00
	Coverage	10.50	9	9.27	0.00	20.00

The second approach was to evaluate content coverage based on evidence-based-health information (Khazaal, Chatton, Cochand, & Zullino, 2008). The ADA guideline was used for evidence-based health information, which is one of the standards of medical care provided for diabetes (Association, 2016). The ADA's clinical practice recommendations are published each year and based on a complete review of the relevant literature on diabetes. They are divided up into several sections to highlight important topic areas, such as glycaemia targets, lifestyle

management, or classification and diagnosis of diabetes. Revisions are made each year, and a brief summary is given for each revision, in which new sections or sections that have been revisited are enumerated. Two main topics from the ADA guideline for the treatment of type 2 diabetes were selected: *"Obesity Management for the Treatment of Type 2 Diabetes"* (Section 6) and *"Pharmacological Therapy for Type 2 Diabetes"* (Section 7 subsection). Content assessment was examined around the following six points:

For Obesity Management:

- Diet, Physical Activity, and Behavioural Therapy;
- Pharmacotherapy;
- Bariatric Surgery;

For Pharmacological Therapy:

- Initial Therapy;
- Combination Therapy and
- Insulin Therapy.

The scale proposed for coverage assessment in the literature (Nilsson-Ihrfelt, et al., 2004; Khazaal, Fernandez, Cochand, & Reboh, 2008) was applied. Taking into account these two main topics, scale requires domain experts to rate each website content coverage as ''none'', ''minimal'' and ''sufficient'' (0–2 points). A categorization of "sufficient" requires a detailed discussion of the topic. When coverage is categorized as "minimal", there is an indiscriminate discussion about the topic. When the topic is not mentioned at all, it is categorized as "none". Correctness refers to the presence of information needed to completely correspond to evidence-based knowledge and it has been assessed using a scale ''mostly not'', ''mostly'' and ''completely right'' (0–2 points) (Nilsson-Ihrfelt, et al., 2004). A categorization of "completely right" means all information is completely in line with the evidence-based knowledge and current best medical practice. When correctness is categorized as "mostly", there is only negligible incorrectness accepted. When most of the information is regarded as incorrect, it is categorized as "mostly not". The content quality scores of the collected websites corresponded to the sum of the six studied aspects, with a possible total of 24 points.

For coverage assessment, a similar procedure to DISCERN labelling was carried out. As a result, 41 of 60 sites having a score higher than the mean score were labelled as "high-quality". 12 out of 48 of the websites which returned from search engines were labelled as "low-quality" as shown in Table 4. The selected websites, their DISCERN, and coverage assessment scores assigned to each website by the experts are given in Appendix B.

In order to examine the degree of agreement of different ratings kappa was used, which is a measure of the degree of consistency (Sim & Wright, 2005). After the initial scores of the websites were received from the experts, the degree of agreement among their ratings was measured using the kappa statistics. The results showed that

the agreement levels of 11 websites were lower than the moderate level. Hence, these websites were re-evaluated by the experts, and an agreed score was reached for each website.

CHAPTER 4

AUTOMATIC QUALITY ASSESSMENT OF WEB-BASED INFORMATION

The previous sections concluded that the studies in the literature are limited as they do not provide any tools that automatically evaluate the quality of the content aspect. This chapter introduces the proposed methodology for quality assessment. First, a brief introduction of the methodology is explained. In the second section, the construction of the textual features under the six categories are presented. The construction of content-based features is explained in the third section. Then, the text-based document ranking with Okapi BM25 and vectorization methods are explained. After that, the experimental results are provided. Finally, the chapter ends with the discussion and conclusion.

4.1. Introduction

The methodology included pre-processing, the construction of the textual and content-based features, the vectorization method, and the model fitting steps. The methodology employed in this study is summarized in Figure 2.

The crawled websites were initially pre-processed before conducting the analyses. If a webpage of the website was not related to type 2 diabetes, it was excluded due to the comprehensiveness criteria and then remaining webpages were gathered together in a document. Some textual features such as paragraph count and image count were obtained using the HTML tags. To extract content-based features, the content was first tokenized, and then all upper-case letters were transformed into lower-case letters. Only one sentence was kept if multiple occurrences of the same sentence were found in the text. In addition, stop words and punctuations were removed from each document, and lemmatization was applied. Finally, the candidate terms were generated from ADA, and their presence was investigated on the websites.

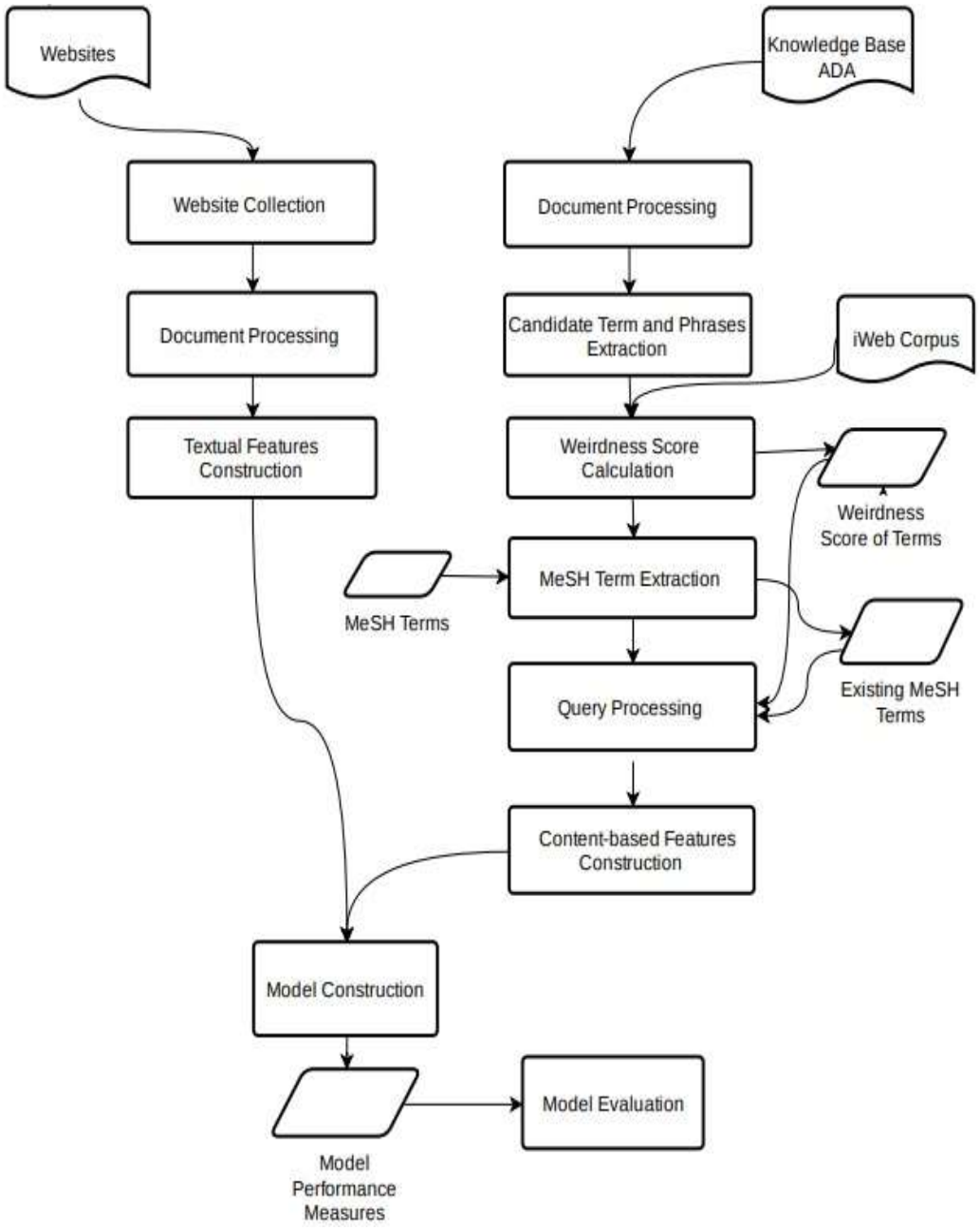


Figure 2: Methodology

The study in this chapter investigates whether the features that adhere to the professional health literacy guidelines and to the content-based by using EBM are indicative of high-quality health-related content. Mainly, the aim of the study is to offer a solution for the first three research questions: *"Which textual evidences reveal the quality of health websites? Are linguistic styles of high-quality websites different from the low-quality ones?", "How content-based evidences are related to the quality and coverage of websites?" and "Does the use of both textual and content-*

based features perform better in the classification of websites according to their information quality compared to the sole use of textual or content-based features?"

4.2. Textual Features

The textual features consist of the distinctive formal (particularly stylistic and structural) aspects of an utterance, text, or artwork in any medium (Oxford Dictionary, 2019). They were investigated under six categories, which four of six were adapted from (Dalip, Gonçalves, Cristo, & Calado, 2011; Dalip, Lima, Cristo, & Calado, 2014) and two were introduced for the first time in this study: Text Structure, Text Style, Text Length, Text Sentence Type, Level of Readability and Lexical Variety. All categories are described as follows:

- ***Text Structure*** features are to describe the structure of how well a text is organized such as the length of paragraphs, the use of images, the use of bullets and the number of sections. The HTML tags were counted to obtain the number of structural features; for paragraphs <p>, for sections <h1> to <h6>, for bullets <li> and for images <img>.

- ***Text Style*** features are used to capture the way authors write their articles through their word usage; they try to capture how the text is written, how author used the language and whether the text has some specific characteristics. For instance, the usage of short sentences indicates clarity in writing. Also, professional health literacy guidelines encourage the use of active voice instead of passive voice (Department of Health, 2003; Medicines and Healthcare products Regulatory Agency, 2014; MedlinePlus, 2018). (Best Practice Guidance on Patient Information Leaflets, 2014; Toolkit for producing patient information, 2003; U.S. Department of Health and Human Services, 2015; MedlinePlus, 2018). Text style features aim to capture such specific usages.

- ***Text Length*** features used by (Rassbach, Pincock, & Mingus, 2007; Blumenstock, 2008; Dalip, Lima, Cristo, & Calado, 2014) such as word count, sentence count, and character count are related to the size of the text in different aspects. These features are useful in assessing whether the information is complete and comprehensive.

- ***The Level of Readability*** features are used as a way to verify if the text is well written, understandable, and free of unnecessary complexity. Flesch Kincaid Grade Level (FKGL) and Simple Measure of Gobbledygook (SMOG) indices are frequently used. The formula for FKGL is given at Equation 3. In the calculation of SMOG, at least 30 sentences in a row near the beginning, in the middle and in the end are selected from text and then used the formula at Equation 4. The motivation behind this is health websites are expected to provide understandable health information regardless of age,

background or reading level (Toolkit for producing patient information, 2003; Best Practice Guidance on Patient Information Leaflets, 2014; MedlinePlus Trusted Health Information for You, 2018).

$$FKGL = 0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\ \left(\frac{total\ syllables}{total\ words}\right) - 15.59 \qquad \text{(Eq. 3)}$$

$$SMOG = 1.0430\sqrt{number\ of\ polysyllables * \frac{30}{number\ of\ sentences}} + 3.1291 \quad \text{(Eq. 4)}$$

Apart from adopted features, the following feature set was developed:

- ***Text Sentence Type*** features are constructed to analyse the types of sentences. Unlike other studies, sentence types were used as a feature in this study. The motivation behind this is health websites generally use imperative and declarative sentences to guide patients. Five different types of sentences in English were considered: Imperative, Interrogative, Exclamatory, Existential and Declarative. If the first word of the sentence is a verb, then the sentence is labelled as imperative. The sentence is marked as an interrogative sentence when it has a question mark. If it has an exclamation mark, it is regarded as an exclamatory sentence. An existential sentence is a sentence that asserts the existence or nonexistence of something; if a sentence starts with existential "there", it is considered as an existential sentence. The rest of the sentences apart from these in the text are labelled as declarative. Table 5 shows an example for each sentence type.

Table 5: Examples of sentence types

Sentence types	**Examples**
Imperative sentences	"*Monitor* your blood glucose every three to four hours."
Interrogative sentences	"How do insulin pumps work*?*"
Exclamatory sentences	"Initially take it slow you don't want to start off too hard, if you are not used to the exercise you will be sore the next day and this will not make exercising a fun experience*!*"
Existential sentences	"*There* are different types of insulin depending on how quickly they work, when they peak, and how long they last."
Declarative sentences	"Inside the pancreas, beta cells make the hormone insulin."

- ***Lexical variety*** is the normalized measure of the unique words used in a text with all words. The motivation behind this is if a website is comprehensive, the lexical variety is expected to be high.

$$Lexical\ variety = \frac{number of\ unique\ words}{number\ of\ words} \quad \text{(Eq. 5)}$$

The complete list of features is presented in Table 8.

4.3. Content-based Features

The previous sections concluded that the studies in the literature are limited as they do not provide any tools that automatically evaluate the quality of the content aspect. This chapter introduces the proposed methodology for quality assessment.

4.3.1. Corpus

ADA guideline is one of the standards of medical care provided for diabetes (Association, 2016). ADA Clinical Practice Recommendations are published each year and based on a complete review of the relevant literature about diabetes. Revisions are made each year and a brief summary is given for each revision in which new sections or sections that are revisited are stated. In this study, the guideline published in 2016 was used and related sections about treatments are covered. Two main topics were chosen from the ADA guideline for the treatment of type 2 diabetes: *Obesity Management for the Treatment of Type 2 Diabetes* (Section 6) and *Pharmacological Therapy for Type 2 Diabetes* (Section 7 subsection).

The iWeb (The iWeb corpus, 2018), which is one of only three large web-based corpora, contain more than 12-13 billion words in 22 million web pages. It has many features that make it unique, even compared to the other Brigham Young University corpora such as COCA and BYU-BNC. It provides three main ways of searching: browsing a frequency list of the top 60,000 words in the corpus, including searches by word form, part of speech, ranges in the 60,000 word list, and even by pronunciation; searching by individual word, and providing collocates, topics, clusters (n-grams), websites, concordance lines, and related words for each of these words; searching for phrases, substrings, lemmas, part of speech, synonyms, and customized wordlists. For the general corpus, iWeb is used in this study. In addition, 100 websites have been selected randomly by creating a special health corpus with the keyword "*health*". The top-level domains of the websites were as follows: .gov (8), .edu (2), .com (44), .org (37), others (8) (i.e. .au, .net).

4.3.2. Term extraction from ADA

Sentences containing auxiliary verbs from ADA's related sections about the treatment of type 2 diabetes were extracted and each was employed to form the

queries. ADA was used as a specialized corpus and iWeb was used as a general corpus. Part-of-speech tagger (specifically Stanford Part-Of-Speech (POS) Tagger (Kristina & Christopher D., 2000)) was applied. The candidate feature terms were restricted as following: NN, JJ; for 2-grams: NN NN, JJ NN; for 3-grams: NN NN NN, JJ NN NN, JJ JJ NN, where NN and JJ are the POS tags for nouns and adjectives respectively. Finally, the weirdness coefficient of each candidate term was calculated.

Weirdness coefficient (Equation 2) is used to identify candidate terms based on the assumption that distribution of terms in a specialized corpus and in a general corpus differs significantly from each other (Ahmad, Gillam, & Tostevin, 1999). Recall that in this formula, f_s is the frequency of the word in the specialized corpus and f_g is the frequency of the word in the general corpus; t_s is the total number of words in specialized corpus and t_g is the total number of words in the general one. To evaluate the effect of corpus size on estimating information coverage of websites, candidate terms were used incrementally. They were sorted in descending order according to their weirdness coefficients. The first 70 terms have zero frequency in the iWeb corpus, but a finite frequency in the ADA corpus; hence weirdness coefficients for these terms were resulted in infinity. These 70 terms were given a surplus of seventy-first term's score instead of infinity. As a result, the feature sets included the top 70 to 630 terms with an increment of 70. All selected terms and weirdness scores are presented in Appendix C in detail.

4.3.3. Query expansion with MeSH terms

Query expansion refers to the process of reformulating a query to improve retrieval performance, typically by including additional terminology that is synonymous. In the context of query expansion, Medical Subject Headings (MeSH) was used, which is designed by the National Library of Medicine (Medical Subject Headings, 2018) for indexing and searching of the MEDLINE database of journal citations. The MeSH terms were extracted from each sentence of the ADA. The number of terms MeSH derived from sentences containing auxiliary verbs is 69. For instance, if a sentence has a "*Metformin*" term, the query will be automatically mapped to the MeSH terms "*Metformin hydrochloride*" and "*Glucophage*". Thus, not only sentences having the word "*Metformin*" will be retrieved, any other sentences indexed with the MeSH terms "*Glucophage*" and "*Metformin hydrochloride*" will also be obtained.

4.4. Information retrieval techniques

In this section, the text retrieval technique assigns scores to documents based on the distribution of query terms within both the document and the corpus. The following steps are applied to construct these features: text-based document ranking with Okapi BM25 and vectorization methods.

4.4.1. Text-based document ranking with Okapi BM25

Okapi BM25 was utilized for the retrieval task which is introduced in Section 2.5.1. It considers document length for normalization since term frequency values tend to be larger in longer documents and it causes a potential bias that longer documents will be assigned to a higher score (Robertson, Walker, Beaulieu, & Willett, 1999). In order to compensate that effect, they derived the document length normalization. Document length normalization is important especially for the corpora in which the length of documents highly differs. Therefore, in this study, Okapi was used to treat each site as a document.

Okapi BM25 computes the score for each word in the query as given in Equation 1. However, since high and low-quality websites coexist in the dataset, calculating the IDF might create a bias. While the IDF gives importance to the rareness of a term, it ignores the usefulness of the term. Moreover, as health-related websites related to type 2 diabetes were used, the common medical terms which exist in each document were ignored. For instance, when the collected diabetes-2 dataset is solely used, the "*Glucose*" term gets a negative BM25 score because it is often found in the dataset. The importance of that term cannot be ignored just because the score is a negative value. For the reasons above, a special corpus as described in Section 4.3.1 was used to calculate the $IDF(q_i)$ in Equation 1 and the final score for a document is set to the sum of all term frequency weights.

In this study, Okapi BM25 was modified for the retrieval task. For instance, if the query is "*physical activity*", Okapi has divided the terms as "*physical*" and "*activity*" and calculated each score differently and then summed up the scores. However, we calculated the score if the "*physical activity*" was in the document.

4.4.2. Feature Construction using Vectorization Methods

In this study, the following three quantization techniques were used:

- *The BM25 score:* The BM25 scores calculated as described above are used.
- *The percentage of the terms:* The number of each query term in a document is divided into the total number of words in the query.
- *The logarithm of* weirdness coefficient*:* The range of weirdness coefficients varies exponentially. Hence, when the raw scores are used, a significantly increased importance is given to the terms with the highest coefficients. For instance, when the weirdness coefficient of the "*Blood Glucose Level*" and "*Bariatric Surgery*" terms in Table 6 were compared, although there is a big difference between them, both terms are equally important in the domain; hence the term with a low score is not a term that can be ignored compared to

the one with a very high score. Therefore, the scores were normalized by taking the logarithm of each.

As can be seen from Table 6, the inclusion of non-Mesh terms is important since it is expected that the content should also include information related to lifestyle or blood glucose level about type 2 diabetes.

Table 6: Analysis of the terms

Term	BM25 score	Weirdness coefficient (WC)	Log(WC)	Mesh Term
Risk	0	38	1.58	X
Patient	0	79	1.90	X
Provider	0	96	1.98	
Hyperglycemia	2.66	1378	3.14	X
Blood Glucose Level	5.38	47416	4.68	
Bariatric Surgery	1.37	6059	3.78	X

Different feature sets and different vectorization methods for model construction were used. The models are named in the format "a.b.c" according to Table 7.

Table 7: Model Naming

Numbering element	a. Features Set	b. Vectorization Method	c. Weirdness coefficients
Possible Values	**FS1:** Content Features	**VM1:** BM25 scores	**70..630**
	FS2: Textual + Content Features (Hybrid)	**VM2:** Logarithm of WS	**(with an increment of 70)**
		VM3: Percentage Matrix	

4.5. Experiments and Results

Using the features described in above, a set of experiments using four vectorization methods with two different feature sets were performed. The experimental tools, the evaluation methodology, and the results obtained were described in this section.

4.5.1. Tools

In this study, Python and R platforms were utilized. The websites required a pre-processing before they could be analysed by algorithms. Separate entries were created in the database for HTML and text versions of the websites. In order to remove HTML tags and scripts tags from HTML pages, the "BeautifulSoup" Parser was used. After the main text was obtained, sentence segmentation and word

tokenization were applied using the "NLTK" toolkit. The Stanford Part-Of-Speech Tagger toolkit (Kristina & Christopher D., 2000) which assigns POS to each word was also used. In extracting dependency parser of sentences, Stanford dependency parser for English syntax (Marie-Catherine & Christopher D., 2008) was used. The assessment of websites' readability scores was calculated by Python's "textstat" package. R for model fitting was used. Leave one out cross validation (LOOCV) was employed due to the limited number of websites and to overcome overfitting. For testing, one fold is reserved as an unseen dataset and the remaining was reserved for training and validation purposes.

4.5.2. Evaluation methodology

An automatic classification system for type 2 diabetes websites based on textual features and content-based features was developed using several classification algorithms. In this thesis, binary classification was considered for high- and low-quality websites as described in Section 3.3. The parameters were set as follows for the classifiers: For random forest classifier, the number of trees (*ntree*) is set to 200 because the performance has not dropped significantly while adding trees from 200 to 1000. *Mtry*, a number of variables available for splitting at each tree node, is set to the square root of the number of input variables. For elastic net classifier, *alpha* is set to 0.5 and *lambda* is set to the average of all the folds' largest values of lambdas such that error is within one standard error of the minimum using cross-validation to control the strength of the penalty. Support Vector Machines (*SVM*) classification is carried out with C-classification using radial kernel that is effective when the data is not linearly separable. All the features were normalized with the number of words or sentences before modelling.

The performance of the proposed classification methods was evaluated using Matthews correlation coefficient (MCC) and accuracy. In addition, recall, precison and F1-measure are given in Appendix D.

4.5.3. Results

In this section, the performances of the proposed framework were presented in relation to the research questions that were posed.

RQ1. Relation between textual features and website quality

With respect to the first research question, whether there is any relation between the textual evidence and the quality of the website was investigated. To achieve it, Kendall's Tau correlation was computed. As shown in Table 8, many features are significantly correlated. In line with the study (Blumenstock, 2008), word count is also found to be statistically significant. As some features are expected to be correlated with the number of words, this analysis aims to pay more attention either to the features, which have higher correlation coefficients than the number of words or to the features, which have significant correlations but not highly correlated with

the number of words while presenting the results. In this way, the most differentiating and significant features were intended to highlight.

Table 8: Kendall's Tau Correlation of features and classes. The feature category indicates whether this feature is used in assessing Wikipedia content (W) and/or used in information quality assessment of generic content other than Wikipedia (O). The features (F) and (H) are constructed based on the health websites' writing styles and health literacy guideline recommendations, respectively.

			DISCERN CLASS		**COVERAGE CLASS**	
	Kendall's tau	**Feature Category**	**Correlation Coefficient**	**p-value**	**Correlation Coefficient**	**p-value**
Text Structure Features	Section count	W	.458**	.000	.436**	.000
	Paragraph count	H	.455**	.000	.383**	.000
	Bullet count	W,H	.338**	.002	.341**	.001
	Length of shortest section	W	-.290**	.008	-.331**	.002
	Number of images	W,H	.274*	.017	.260*	.023
	Average section length	W	-.053	.622	-.001	.994
	Average paragraph length	W,H	-.126	.271	-.004	.975
Text Style Features	Number of uses of verb *to be*	W	.468**	.000	.464**	.000
	Number of passive voice sentences	W,H	.473**	.000	.434**	.000
	Number of nouns	W	.432**	.000	.464**	.000
	Number of auxiliary verbs	W	.432**	.000	.488**	.000
	Number of verbs	W	.416**	.000	.451**	.000
	Number of adjectives	F	.418**	.000	.421**	.000
	Number of sentences starting with a pronoun	W	.392**	.000	.402**	.000
	Short sentence rate	W,H	.088	.424	.065	.556
Sentence Type Features	Number of declarative sentences	F	.434**	.000	.456**	.000
	Number of existential sentences	F	.343**	.002	.404**	.000
	Number of interrogative sentences	W,H	.348**	.001	.423**	.000
	Number of imperative sentences	H	.319**	.003	.329**	.002
	Number of exclamatory sentences	F	-.135	.265	-.133	.275
Text Length Features	Character count	W	.433**	.000	.471**	.000
	Word count	W,H	.425**	.000	.462**	.000
	Sentence count	W	.404**	.000	.417**	.000
Readability Features	Flesch-Kincaid Grade Level	W,H,O	.069	.526	.059	.583
	SMOG	W,H,O	.154	.156	.154	.155
Variety	Lexical variety	H	-.408**	.000	-.439**	.000

**. Correlation is significant at the 0.01 level.

*. Correlation is significant at the 0.05 level.

The most significant category of the features related to quality is found as text style features where the number of uses of verb "to be", the number of nouns and the number of auxiliary verbs have higher correlation value than the word count even though these features are highly correlated with the word count ($r>0.8$, $p<0.01$).

Regarding the RQ1.1, within the textual features derived from professional health literacy guidelines, it is found that the bullet count, the section count and the number of images are positively correlated; contrary to expectation, short sentence rate, average paragraph length and readability features are not found statistically significant with the information quality of websites. Even though the use of passive voice sentences is discouraged in professional health literacy guidelines, it is observed that the higher the information quality, the higher the use of passive voice sentences.

The results show that there is a statistically significant relation between information quality and the sentence types (RQ1.2). All sentence types excluding exclamatory sentences are found statistically significant. The correlation coefficients of all sentence types except the declarative ones are lower than the coefficients of the number of words and the number of sentences. However, there is only moderate correlation between the number of imperative ($r=0.563$, $p<0.01$) and the number of interrogative ($r=0.434$, $p<0.01$) features with the number of words. Moreover, the results showed that lexical variety is negatively correlated. The higher the information quality, the lower the lexical variety.

The highest correlation with the coverage scores is observed with the number of auxiliary verbs. Although text length features do still have higher correlation coefficients compared to the other features, the sentence types are observed as more determinant features for assessing coverage compared to DISCERN as they have higher scores.

RQ2. Relation between content-based features and website quality

The second research question was investigated on how content-based evidences are related to the coverage of websites and how content related medical terms that are important for evaluating coverage can be acquired. For this purpose, terms from ADA were extracted and their weirdness coefficients were calculated using iWeb corpus. These coefficients show how the medical content-based terms differ from common terms. For instance, the term "*alpha-glucosidase*" has one of the highest scores that have zero frequency in the iWeb corpus and a finite frequency in the ADA corpus. On the other hand, the term "*economic*" has one of the lowest coefficients, as it appears on any content type such as finance, economy, health so on.

After the terms were sorted in descending order based on their weirdness coefficients, the total number of unique terms within each term list (70 to 630 terms with an increment of 70) for each website was calculated i.e. out of the 70 terms how

many terms appear in a specific website. The websites are ordered from lowest to highest according to their information coverage score in Figure 3 (first 19 websites are low-quality; the others high-quality). As seen in Figure 3, the terms with high weirdness coefficients exist more often on high-quality websites. It is observed that the first 140 terms are rare terms for both high-quality and low-quality. While investigating whether there is any relation between the term count and the quality of the website, the first 70 and 140 terms are uncorrelated with the quality, others are highly correlated ($r>0.361$, $p<0.01$). On the other hand, as the number of terms increases, the frequency of terms is observed as increasing from low-quality websites to high-quality websites. For high-quality websites, up to 9 of the first 210 terms are seen but 53 terms are seen when the first 350 terms are studied. Unlike high-quality websites, 3 terms out of 210 terms and 15 out of 350 terms are observed.

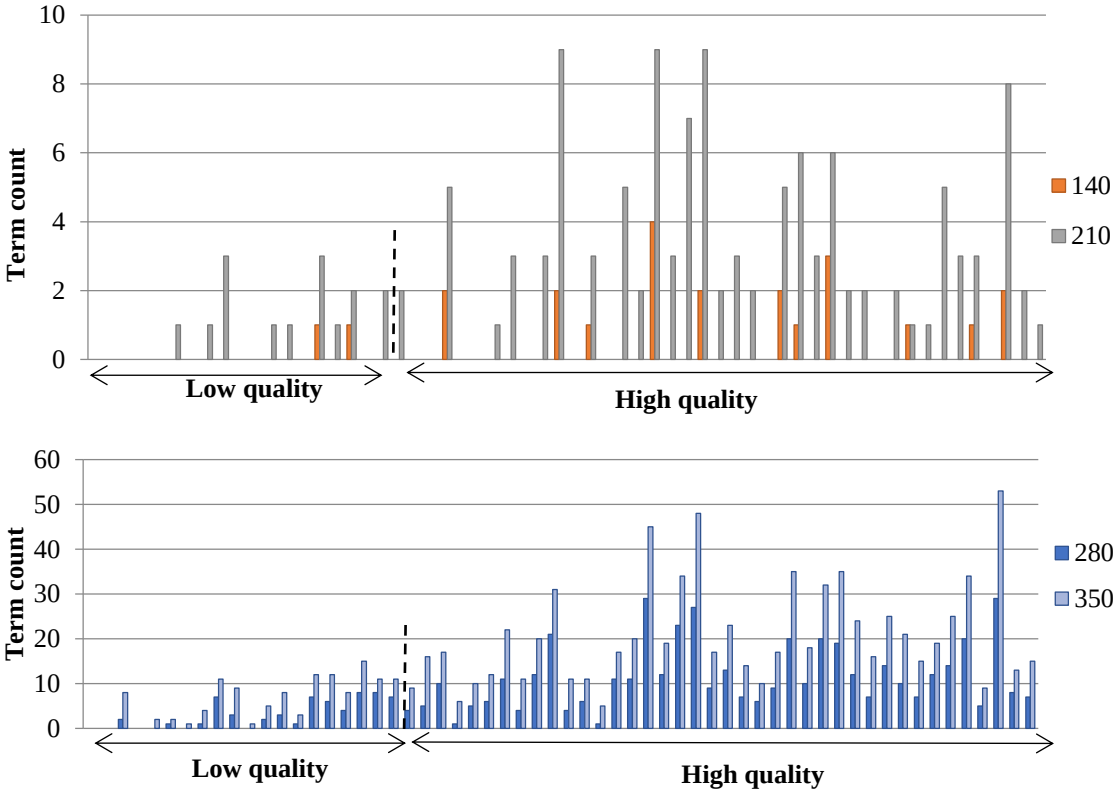


Figure 3: Existence of candidate terms on websites

RQ3: Automatic classification of website quality with textual and content-based features

In the last research question, how textual and content-based features separately and together are effective in classifying websites according to their information quality was investigated. First of all, the effect of the textual features was investigated. The baseline models using only word count feature were constructed and the comparative results were presented in Table 9. The effects of the textual features on the performance of the classification results were also examined. All the features in Table 8 were normalized with the number of words or sentences before modelling.

Table 9: LOOCV results showing the mean and standard deviation of the results with textual features

	DISCERN		**Coverage Assessment**	
	MCC	**Accuracy**	**MCC**	**Accuracy**
Elastic net	0.50±0.87	0.75±0.44	0.77±0.65	0.88±0.32
Random forest	0.47±0.89	0.73±0.45	0.60±0.81	0.80±0.40
SVM	0.43±0.91	0.72±0.45	0.60±0.81	0.80±0.40
Baseline (Word count)	0.23±0.98	0.62±0.49	0.37±0.94	0.68±0.47

More generally, the results of all the models are statistically above the baseline accuracy of 0.57 (the performance obtained when all the websites in the dataset were classified as high-quality). Also, the results show that the models using textual features significantly outperform the models using solely word count feature. The highest accuracy was obtained 75% for DISCERN and 88% for coverage, using the elastic net algorithm. SVM performed the worst in DISCERN. The high performance values were obtained with elastic net due to its ability to address collinearity and to work successfully with high-dimensional but small datasets. Furthermore, elastic net performs feature selection and deals with grouping effects, thereby rendering classifiers more easily interpretable (Liu, Sun, Wang, Liu, & Tao, 2016). It makes feature selection inherently where the features that have non-significant coefficients are shrinking toward zero and are not present in the final model. In each fold, the rank of each selected feature is determined based on the magnitude of its coefficient. The feature with the highest value received the first rank indicating its significance. All the features selected by the models were ranked in each fold according to their magnitudes. If a feature was not selected, its rank was assigned to the one point higher than the number of selected features. Then, the average and standard deviation of their ranks were calculated within all the folds. The most important five features were found as follows: number of adjectives, number of exclamatory sentences, number of auxiliary verbs, number of existential sentences, and lexical variety. The ranks of the features obtained with elastic net model are presented in Appendix E.

Second, the performance of the content-based features was evaluated. As depicted in Table 7, there are several parameters that might affect the performance.

Consequently, 54 models were constructed to classify a website according to its coverage. The weirdness coefficients of all the terms from the relevant ADA guideline sections were calculated, and the terms were sorted in descending order. The MCC scores for elastic net, random forest, and SVM using different term lists are presented in Figure 4. The top 70 and 140 terms were ignored because the results of all the models were zero with these features. This is attributable to the fact that few of these words appear in the texts, as can be seen in Figure 3. The performance of 18 possible models, where the top 350 and 630 terms were used, is tabulated in Table 10 and Table 11. All the algorithms yielded better results for content-based features than the textual features. The results indicated that the top 350 terms appeared to perform the best. The highest accuracy was obtained as 92% for coverage, using the content-based features. The ranking results showed that the first three queries of the related sections on the treatment that were described in Section 3 are "Insulin Therapy", "Bariatric Surgery", and "Pharmacotherapy".

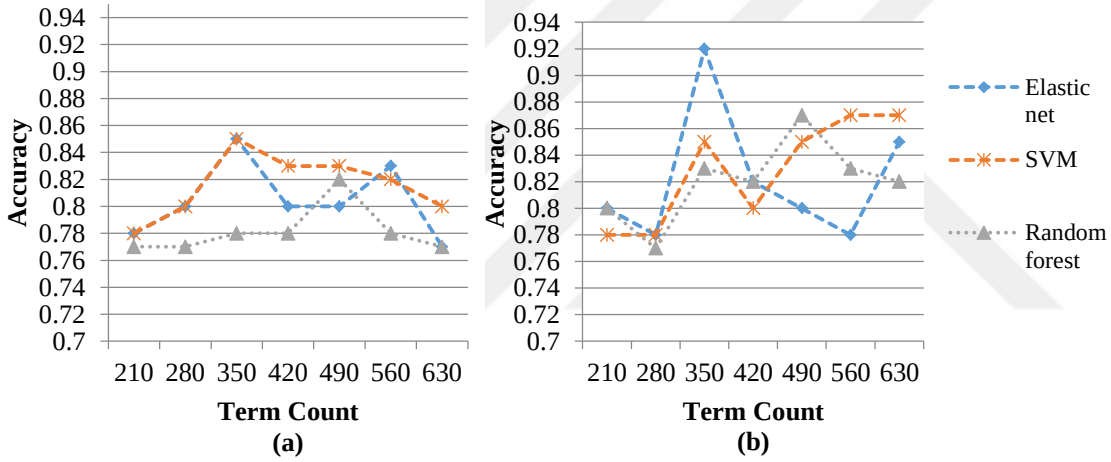


Figure 4: Accuracy scores of classification algorithms to assess coverage, where different numbers of terms sorted according to their Weirdness coefficients are used (a) The content-based feature set (FS1) with the BM25 scores (VM1) is used. (b) The content-based feature set (FS1) with the percentage of the existence of terms (VM3) is used.

Thirdly, the performance of the models where the hybrid feature set (FS2) was used was investigated. Figure 5 depicts the hybrid feature set (FS2) with the BM25 scores (VM1) and the percentage of the existence of terms (VM3). The results revealed that the models constructed with both content-based and textual features (FS2) performed better than the models using textual features alone. In addition, the elastic net and random forest model achieved a higher performance than the SVM model. SVM resulted 32% accuracy for 210 and 280 terms; since it could not find the optimal hyperplane in the given feature set, it classified all the classes as low quality.

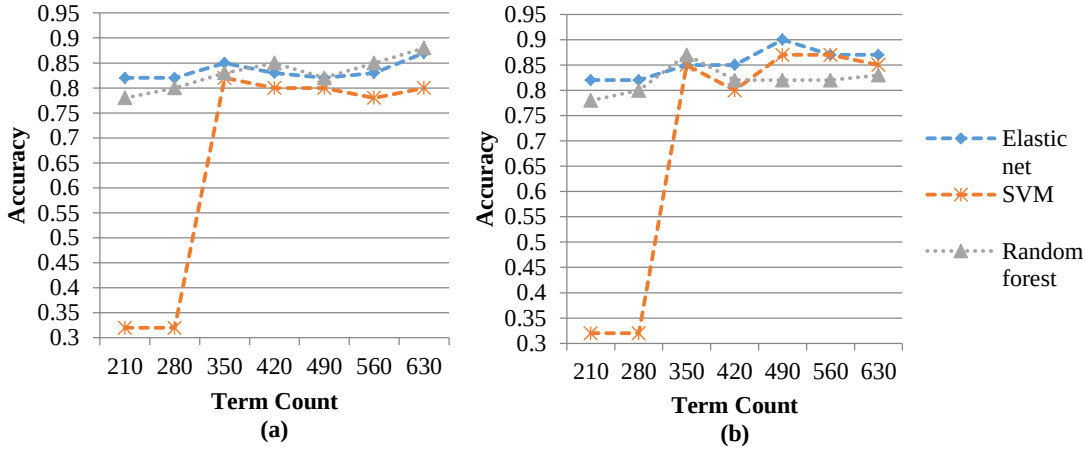


Figure 5: Accuracy scores of classification algorithms to assess coverage, where different numbers of terms sorted according to their Weirdness coefficients are used (a) The hybrid feature set (FS2) with the BM25 scores (VM1) is used. (b) The hybrid feature set (FS2) with the percentage of the existence of terms (VM3) is used.

Table 10: Model Performance Measures, where the top 350 terms are used

			Elastic		Random forest		SVM	
			MCC	**Accuracy**	**MCC**	**Accuracy**	**MCC**	**Accuracy**
350	**FS1**	**VM1**	0.70 ±0.72	0.85 ±0.36	0.57 ±0.83	0.78 ±0.42	0.70 ±0.72	0.85 ±0.36
		VM2	0.63 ±0.78	0.82 ±0.39	0.63 ±0.78	0.82 ±0.39	0.73 ±0.69	0.87 ±0.34
		VM3	0.83 ±0.56	0.92 ±0.28	0.67 ±0.75	0.83 ±0.38	0.70 ±0.72	0.85 ±0.36
	FS2	**VM1**	0.70 ±0.72	0.85 ±0.36	0.67 ±0.75	0.83 ±0.38	0.63 ±0.78	0.82 ±0.39
		VM2	0.70 ±0.72	0.85 ±0.36	0.67 ±0.75	0.83 ±0.38	0.70 ±0.72	0.85 ±0.36
		VM3	0.70 ±0.72	0.85 ±0.36	0.73 ±0.69	0.87 ±0.34	0.70 ±0.72	0.85 ±0.36

Table 11: Model Performance Measures, where the top 630 terms are used

			Elastic		Random forest		SVM	
			MCC	**Accuracy**	**MCC**	**Accuracy**	**MCC**	**Accuracy**
630	**FS1**	**VM1**	0.53 ±0.85	0.77 ±0.43	0.52 ±0.85	0.77 ±0.43	0.60 ±0.81	0.80 ±0.40
		VM2	0.60 ±0.80	0.80 ±0.40	0.63 ±0.78	0.82 ±0.39	0.67 ±0.75	0.83 ±0.38
		VM3	0.70 ±0.72	0.85 ±0.36	0.63 ±0.78	0.82 ±0.39	0.73 ±0.69	0.87 ±0.34
	FS2	**VM1**	0.67 ±0.75	0.83 ±0.38	0.77 ±0.65	0.88 ±0.32	0.60 ±0.81	0.80 ±0.40
		VM2	0.70 ±0.72	0.85 ±0.36	0.70 ±0.72	0.85 ±0.36	0.67 ±0.75	0.83 ±0.37
		VM3	0.73 ±0.69	0.87 ±0.34	0.67 ±0.75	0.83 ±0.38	0.70 ±0.72	0.85 ±0.36

In order to assess the effectiveness of the models, vectorization methods, and feature sets, Friedman's test was performed, which is a non-parametric equivalent to repeated measures ANOVA to control multiple statistical significance (Demšar, 2006). To examine the significance of differences, Wilcoxon's signed rank test, which is a non-parametric alternative to the paired t-test, was used on MCC scores.

The test ranks the differences in performances of two classifiers for each dataset, 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. The results of this test are adjusted by Bonferroni correction. The analyses both for the top 350 and 630 terms were repeated to prevent any bias resulting from the choice of term selection. For each model, vectorization method and feature set, pairwise comparisons were carried out based on loocv. For instance, elastic net's performance with FS1 and VM1 was compared with elastic net's performance with other feature set (FS) and VM combinations.

Friedman test is significant for all comparisons between the models (p-value<0.05). However, no statistically significant differences are observed within vectorization methods and FS for all models, except 630 terms of elastic net. Initially, the impact of using only content (FS1) and hybrid features (FS2) was compared. According to elastic net results, the p-values of all pairwise comparisons of models are placed in Table 12. While comparing FS1 and FS2, however, any of the p-values were below 0.05 indicating that respective models did not differ significantly. The accuracy of each fold was used for comparison based on each model.

Table 12: p-values for nxn comparisons of the FSs over VMs for elastic net

630 terms			
FS1		**FS2**	**p-value**
VM1	*vs*	VM1	1
VM1	*vs*	VM2	1
VM1	*vs*	VM3	1
VM2	*vs*	VM1	1
VM2	*vs*	VM2	1
VM2	*vs*	VM3	1
VM3	vs	VM1	1
VM3	vs	VM2	1
VM3	vs	VM3	1

When the impact of vectorization method is assessed for models, the results also show that there are no statistically significant performance differences between the models built with different vectorization methods. However, in overall when all FS and VM combinations were considered, no VM appeared to perform consistently superior at all times.

4.6. Discussion

This study aimed to propose an automated framework to identify and evaluate the quality of the content of websites that provide information about type 2 diabetes according to DISCERN and EBM. Specifically, the websites' textual features —such as linguistic styles, types of sentences, and word count—and content-based features in the health domain were investigated. The findings can be summarized as follows:

RQ1. Relation between textual features and website quality

One of the important indicators of information quality on health websites is the length of the text specifically word count in general domain. The accuracy was acquired as 88% using solely textual features although in the literature they obtained up to 96% such as in (Blumenstock, 2008). However, word count is not an entirely reliable measure because the misclassified websites were observed as low-quality give sufficient information in a short and concise way whereas the websites misclassified as high-quality included information redundancy, advertisements, information on other topics but not the necessary relevant information. Therefore, it is important that information coverage should be assessed taking into consideration the content-based features.

Regarding RQ1.1, since professional health literacy guidelines recommend avoiding usage of passive voice sentences, a negative correlation between passive sentences and quality and coverage scores was expected. However, in dataset, positive moderate correlations were obtained. In addition, while professional health literacy guidelines suggest the use of short sentences, a significant correlation between the percentage of short sentences and the quality related scores could not be found. These websites are mostly written by academic writers who do not want to focus on who acts but on who receives or experiences the action instead. Thus, the passive voice usage is more appropriate than the active voice for academic index/health sciences library and dispersion of using active and passive voice identified through descriptive content analysis is 51% and 49% respectively (Bada & Ulum, 2018). In the collected dataset out of 60 websites, 57 numbers of websites use passive voice and three sites that do not use passive are low-quality websites. As a consequence, the majority of websites do not follow professional health literacy guidelines.

The expected readability scores of websites are recommended in the range of 6 to 8 according to the professional health literacy guidelines. Nevertheless, only two low-quality sites and two high-quality sites had the desired readability score in collected dataset. FKGL readability score for the remaining 56 websites indicated difficult readability; a mean grade level 10.5 for high-quality websites and for SMOG index, the mean grade level was 9.2. There is no correlation between readability and the information quality and coverage scores, which are in line with the study of Giorgi et al (San Giorgi, de Groot, & Dikkers, 2017). In other words, a high-quality website is not necessarily well readable and vice versa. Also, in Wikipedia studies (Dalip, Gonçalves, Cristo, & Calado, 2011; Dalip, Lima, Cristo, & Calado, 2014) readability features were found as insignificant for assessing quality and the reasons were attributed to their incapability of measuring how well a given text covers the domain information. Moreover, in (Dalip, Lima, Cristo, & Calado, 2014) study, they found that although none text style features were selected as common features on the three datasets they used, the number of auxiliary verbs, the number of passive voice sentences, and the number of interrogative sentences were found significant on two different datasets. In addition, negative moderate correlations on lexical variety were obtained. As lexical variety is calculated, unique words are normalized by the total

number of words; so lower lexical variety is correlated with higher information quality. Hence, if a website is comprehensive, lexical diversity is expected to be high.

Text structure features derived from generic information related studies and professional health literacy guidelines were also found as important indicators for assessing quality of text for both health and general domain. Section count (Dalip, Gonçalves, Cristo, & Calado, 2011; Dalip, Lima, Cristo, & Calado, 2014), and image count (Dalip, Gonçalves, Cristo, & Calado, 2011; Dalip, Lima, Cristo, & Calado, 2014; Rassbach, Pincock, & Mingus, 2007) were indicated as important features implying the organization of text and they are useful for assessing text quality.

Regarding RQ1.2, the correlation results show that except exclamatory sentences all sentence types have significant positive correlations with the information quality and coverage scores. Despite the moderate correlation between the sentence types and the word count, they can still be considered as valuable features for measuring information quality of health websites.

RQ2. Relation between content-based features and website quality

The content-based analysis showed that the terms used in EBM guidelines and websites differ. Almost all the top 140 terms with the highest weirdness scores could not be found on any website indicating that very specific medical terms in ADA guideline do not appear in health websites designed for end users. On the other hand, when a higher number of terms were selected from ADA guideline, they successfully generated the relevant queries for assessing information coverage with the help of MESH term expansion. It was observed that high-quality websites include more domain specific terms compared to low-quality websites.

RQ3: Automatic classification of website quality with textual and content-based features

When classifying websites with textual features only, content-based features only and both of them, using both textual and content-based features gives better results. However, no statistical difference was observed between the hybrid model and content only model. On the other hand, based on the mean of accuracy values, a higher score for the hybrid model to assess coverage was achieved.

Furthermore, coverage assessment class achieved better results than DISCERN in the classification of websites with textual features only. As a reason for this, websites are observed to have different quality scores for DISCERN and information coverage. In DISCERN scoring, websites obtain a higher score if all the treatment options are provided; whereas explanations of only a small number of them will suffice. For example, while there are many treatment options related to the disease, it is enough

to explain only two previously mentioned ones in detailed. Conversely, for information coverage, the ADA guideline includes several essential statements under "Obesity Management and Pharmacological Therapy," which might be missing on their web pages. Consequently, the comprehensiveness of the topic might be impeded. Therefore, the use of an evidence-based guideline can be valuable in assessing information coverage. The use of textual features derived from Wikipedia and/or professional health literacy guidelines has a positive effect on the classification results.

4.7. Conclusion

This chapter presents an automatic information quality and coverage evaluation framework based on textual and content-based features, which were used for the first time in the health domain in this research. Although the highest classification accuracy is obtained as 91%, the main focus was to investigate the features which indicate information quality.

Like Wikipedia studies, this study revealed that textual features are valuable as quality indicators. The features constructed using professional health literacy guidelines were significantly correlated with information quality. The direction of the correlation of some of the features, however, ran counter to initial expectations. For example, passive voice is discouraged in writing, but it was frequently used in the analysed documents; hence, we uncovered a significant positive relation between passive voice usage and website quality, even though the expected relation was negative. Evidently, health websites do not strictly follow the recommendations of the professional health literacy guidelines. The frequency of certain sentence types as features complement the widely used textual features, such as word and sentence counts, in information quality assessment.

Three different vectorization methods with different feature sets and models were studied. The vectorization method where the percentage of the existence of terms is used appears to perform marginally better. The impact of the vectorization method might be better observed with a larger dataset.

CHAPTER 5

INFORMATION QUALITY ASSESSMENT USING PROFESSIONAL HEALTH LITERACY GUIDELINE RECOMMENDATIONS

The textual features introduced in Chapter 4 were derived from the recommendations of the professional health literacy guidelines, however, the majority of the visual presentation and writing style features were excluded from the initial analyses. In this chapter, a comprehensive feature set derived from professional health literacy guidelines is used.

5.1 Introduction

Professional health literacy guidelines include instructions and recommendations about writing high-quality healthcare content for content publishers. These guidelines facilitate producing clear and easy-to-follow health information for end-users regardless of their age, background, or reading level (Department of Health, 2003; Medicines and Healthcare products Regulatory Agency, 2014; MedlinePlus, 2018).

All these studies in the literature, however, are limited in evaluating the quality of such leaflets comprehensively, as they partially studied the recommendations of the guidelines such as the presence and use of headings, bullet points, and short sentences (Raynor, De Veene, & Bryant, 2014; Durrant, Taylor, Thompson, Usher, & Jackson, 2019). Also, these studies were mostly applied as user-test study and not used for automatic website quality assessment. There are limited features that are commonly used to measure information quality in the literature, and the widely studied feature is that of readability.

The study in this chapter investigates whether features that adhere to these professional health literacy guidelines are indicative of high-quality health-related content. Mainly, the aim of the study is to offer a solution for the fourth research question: *"Is the comprehensive feature set derived from professional health literacy guidelines informative in assessing the information quality of websites?"* and *"How successful are these feature sets in the classification of websites according to their information quality?"*

5.2. Features obtained from professional health literacy guidelines

The features were constructed using the recommendations of the professional health literacy guidelines specifically to measure the quality of the websites. In this study, "MedlinePlus-How to write easy-to-read health materials" (MedlinePlus, 2018) and "NHS-Toolkit for producing patient information" (Department of Health, 2003), which were explained detail in Section 2.4, were used to construct the features.

Table 13 shows a summary of the guidelines' recommendations on how health materials should be written. The recommendations were divided into three sections as follows: "General points", "Language and writing style" and "Visual presentation style". All recommendations of the referred guidelines are presented in Appendix F in detail.

Table 13: Recommendations for writing easy to read health materials

General points	
G1	No longer than six grade reading level (seven or eight may be)
G2	Remove words, sentences or paragraphs; to make text more understandable
G3	Incorporate web accessibility
G4	Be up to date
G5	Give facts about risks, side effects and benefits
G6	Tell other information and sources
Language and writing style	
L1	Keep active voice; avoid passive
L2	Use simple words; avoid jargon (medical terms, abbreviations, acronym)
L3	Use short sentences (10-15 words)
L4	Say in a positive way (such as instead of writing "Don't eat lots of red meat", prefer using "Eat less red meat")
L5	Be consistent in using terms; avoid using synonyms
L6	Use bullets instead of blocks of text
L7	Use personal pronouns "you" instead of "the patient"
L8	Divide up text; a question and answer format
L9	Do not use long paragraphs; divide them up using headings and new paragraphs
L10	Avoid instructions; give examples and cases for clarification
Visual presentation style	
V1	Use bolded headings and subheadings
V2	Use colours; appeal to target audience
V3	Use images in line with communication
V4	Avoid upper case letters, italics and underlining
V5	Balance the use of text; try for about 40-50% white space
V6	Use sans serif fonts and fix font size
V7	Justify left margin
V8	Avoid text on top of shaded background, photos or patterns; can be difficult to read

Most of the recommendations aim to assess whether the text is well written, understandable, and does not exhibit unnecessary complexity. For this purpose, professional health literacy guidelines suggest the use of a maximum of six to eight reading levels (G1). Too many words, sentences, and paragraphs can also make websites difficult to understand; therefore, avoiding wordiness is suggested (G2). To obtain the features, FKGL and SMOG indices were calculated for readability. Also,

words, sentences, and paragraphs on websites were counted. The items (G4, G5) concern with giving up-to-date information regarding the risks, side effects and benefits of treatments. Therefore, the following terms in each sentence were counted:

- *"risk, side effect, disadvantage, warning, precaution, implication, contraindication, adverse effect, adverse reaction, adverse event, drug interaction, caution, overdose, susceptibility, hypersensitivity, and allergic reaction "* for risk-related sentences;

- *"advantage, and benefit"* for benefit related sentences;

- *"last updated, last reviewed, last modified, and last edited"* for the existence of date information, and whether the published date is within the last year.

These terms were selected after manually inspecting the high-quality websites' term usage. Also, WordNet and MeSH were used to find alternative terms. WordNet, however, returns different results that are not used in medical domains, such as *endangerment* and *jeopardy* terms returned for risk synonyms. On the other hand, the majority of these selected terms are also present in MeSH. Similarly, the terms used to specify the update time on websites were selected and the last update dates, one year before the download date, were considered as up-to-dateness. In addition, the guidelines recommend that an accessible website helps people with reading and learning disabilities (G3). There are many different criteria for measuring web accessibility such as providing alternative texts of images and titles, providing user controls for moving content or providing captions for videos with audio, etc. The comprehensiveness criteria are out of scope in this study as its coverage is far-reaching. Available resources and other information are important to engage the reader (G6). To be specific, references and links are provided using "ref" and "links" division tagging respectively (detailed in Chapter 3). External links provided using this tagging are counted if the URL contains *"HTTP"*. Also, if the URL contains *".pdf"*, it is counted towards the calculation of the number of "different types of sources (pdf)" variable. The video was also counted for alternative sources, but videos found on websites were tagged differently for each website. For this reason, if the URL contains *"video"*, it is counted towards the calculation of the number of "different types of sources (video)" variable.

For language and writing style, guidelines suggest using personal pronouns instead of using *patient* word. Saying things in a positive way can help capture readers' attention and make text patient-friendly (L4, L7). Therefore, the term *"patient"*, and sentences start with "PRP" and "PRP$" for pronoun PoS tags were counted separately. Nevertheless, sentiment analysis is used to determine whether the language used on the websites is positive. In order to determine affective content of each sentence, a lexicon and a rule-based sentiment analysis tool, the Valence Aware Dictionary for sEntiment Reasoner (VADER) was used (Hutto & Gilbert, 2014). For sentiment analysis, VADER sets the valence score to each word in the lexicon that could be positive, negative, or neutral words. The valence score of each word determines the sentiment score more accurately. It also takes into consideration word order and degree modifiers. The sentiment score of a sentence, the compound score, is obtained summing the valence scores of each word in the lexicon, adjusted

according to the rules. The compound score is in the range between -1(negative) and +1 (positive). If the score is between -0.05 and 0.05, the sentence is considered as neutral. According to the compound score obtained as a result of this analysis, each sentence is counted as having positive or negative or neutral sentiment. Additionally, it is important to use few and short sentences to increase comprehension (L3). For instance, dividing up text into question-answer format or using more headings for logically structures websites (L8, L9). Bullet lists also divide text to make it more readable (L6). The short sentence rate was calculated as the percentage of sentences whose length is less than 15 words. The features correspond to the number of <h1-h6> tags for section heading, the number of <li> for bullet list, and the number of sentences with a question mark were counted as interrogative sentences. Instead of passive voice, using active voice helps people associate concepts with actions (L1). Avoiding medical jargon, acronym, and synonyms of terms, and providing only essential information concerning what patients need to know provide consistency and make reading easier (L2, L5). For this purpose, MeSH was used to identify medical terms and their synonyms. If a medical term needs to be used, it is recommended to explain the term first (L10). Therefore, the presence of special terms about description, specifically the terms *"example, instance, tip, referred to as and mean"* were counted. A description could be written using "is-a" clause such as *"Insulin is a hormone that is usually made in the pancreas."* However, we did not study them in the context of this thesis as they are hard to identify in the text and such clauses are used infrequently. A medical term can be also explained within parenthesis; therefore, the number of the parenthesis was counted. In addition, a regular expression was constructed to count the number of acronym words. The expression matches all words based on whether they have two to four uppercase letters, and if there is a dot between them. As mentioned in the previous chapter, recommendations are often given as imperative sentences and professional health literacy guidelines suggest that things to be said positively (L4). Therefore, in this study, imperative sentences which started with "VB" and "VBP" for the verb PoS tags, and also negative imperative sentences were counted.

For visual presentation and representation, the recommended items (V1-V8) were measured using the count of the respective HTML tags: b, strong, colour, img, i, u, style and sub-attributes of style tag text-align, font-family, font-size and background.

Totally, 49 different features were used to assess the quality of a website. The complete list of features is presented in Table 14.

5.3. Experiments and Results

Using the features described at Section 5.2, a set of experiments was performed using this feature set. In addition to the tools described in Section 4.5.1, to classify sentences' sentiment scores on websites, VADER was used, which is open-sourced lexicon and rule-based sentiment analysis tool (Hutto & Gilbert, 2014). The evaluation strategy given in Section 4.5.2 was used.

The experimental results were presented in relation to the fourth research question that was posed.

RQ4. Relation between the features and the website quality

Kendall's Tau correlation was computed to investigate whether there is any relation between the features and DISCERN. As shown in Table 14, many features are significantly correlated.

It was observed that the benefit related sentences count exhibits the highest correlation with quality. Although the text length features also display higher correlation coefficients than the other general point features, the benefit and risk related sentences count appear to be the determinant features for assessing quality. Contrary to the expectations, the feature regarding information currency, the external and different type of sources, and readability features were not found to be statistically significantly correlated with the information quality of websites.

The results demonstrated that information quality is statistically related to language and writing style. All features except short sentence rate and average paragraph length were found to be correlated with the website quality. Even though the use of the medical terms, acronyms and synonyms were discouraged in professional health literacy guidelines, it was observed that the higher the information quality, the higher the usage of them. The bullet count, section count, and the number of interrogative sentences recommended for the division of information are positively correlated as expected. In addition, the usage of parenthesis and the number of description sentences appear to be important features for assessing quality.

Bolded word count, uppercase word count, the whitespace count and the number of images were found positively correlated. On the other hand, the presences of underline and font size were found negatively correlated as expected.

Table 14: Kendall's Tau correlation of features and DISCERN. The feature category indicates whether that feature is used in assessing Wikipedia Content (W) and/or used in information quality assessment of generic content other than Wikipedia (O). The features (H) are constructed based on the health websites' writing styles and health literacy guideline recommendations respectively.

Kendall's tau		**Feature Category**	**Correlation Coefficient**	**p-value**
General points				
G1.1	Flesch Kincaid Grade Level	W,H,O	.069	.526
G1.2	SMOG	W,H,O	.154	.156
G2.1	Word count	W,H	.425**	.000
G2.2	Sentence count	W	.404**	.000
G2.3	Paragraph count	H	.455**	.000
G4	Information currency (up-to-dateness)	W,H,O	.114	.369
G5.1	Risk related sentence count	H	.464**	.000
G5.2	Benefit related sentence count	H	.515**	.000
G6.1	Different type of sources (pdf)	H	.165	.169
G6.2	Different type of sources (video)	H	-.005	.969
G6.3	Internal sources count	W	.326**	.003
G6.4	External sources count	W	.149	.165

Language and writing style				
L1	Number of passive voice sentences	W,H	.473**	.000
L2.1	Medical term count	H	.452**	.000
L2.2	Medical term variety	H	-.245*	.022
L2.3	Acronym word count	H	.362**	.001
L3	Short sentence rate	W,H	-.085	.433
L4.1	Positive sentence count	H	.406**	.000
L4.2	Negative sentence count	H	.438**	.000
L4.3	Neutral sentence count	H	.431**	.000
L4.1	Normalized positive sentence count	H	.066	.541
L4.2	Normalized negative sentence count	H	.034	.754
L4.3	Normalized neutral sentence count	H	-.073	.493
L4.4	Negative imperative sentence count	H	.285*	.014
L5	Usage of synonym	H	.267*	.025
L6	Bullet count	W,H	.338**	.002
L7.1	Number of sentences starting with a pronoun	W,H	.392**	.000
L7.2	Number of uses of term “patient”	H	0.28*	0.01
L8	Number of interrogative sentences	W,H	.348**	.001
L9.1	Average paragraph length	W,H	-.126	.271
L8,L9.2	Section count	W,H	.458**	.000
L10.1	Number of imperative sentences	H	.319**	.003
L10.2	Number of description sentences	H	.382**	.000
L10.3	Usage of parenthesis	H	.429**	.000
Visual presentation style				
V1.1	Bold count	H	.298**	.006
V1.2	Bold word count	H	.260*	.017
V2	Existence of colour	H	-.194	.130
V3	Number of images	W,H	.274*	.017
V4.1	Italic count	H	.117	.338
V4.2	Italic word count	H	.121	.320
V4.3	Underline count	H	-.260*	.044
V4.4	Underline word count	H	-.260*	.044
V4.5	Uppercase word count	H	.474**	.000
V5.1	Whitespace count	H	.422**	.000
V5.2	Percentage of whitespace	H	-.138	.200
V6.1	Font face	H	-.207	.105
V6.2	Font size	H	-.319*	.012
V7	Left margin alignment	H	-.098	.446
V8	Background	H	.038	.771

**. Correlation is significant at the 0.01 level.
*. Correlation is significant at the 0.05 level.

RQ4.1: Classification of website quality with features

This study also investigated how these features are effective in classifying websites according to their information quality. The means and standard deviations of the performance values were calculated based on the folds. More generally, the results of the two algorithms are statistically above the baseline accuracy of 0.57 as shown in Table 15 (the performance obtained when all the websites in the dataset were classified as high-quality).

Table 15: Cross-validation results

Elastic net		Random Forest	
MCC	**Accuracy**	**MCC**	**Accuracy**
0.40±0.92	0.70±0.46	0.53±0.85	0.77±0.43

To identify significant features for predicting website quality, the model regression coefficients were also investigated for elastic net and the mean decrease accuracy coefficients for random forest models. Elastic net makes feature selection inherently where insignificant features end up having near zero coefficients. The coefficients of insignificant features shrink toward zero and do not present in the final model. The rank of each selected feature is determined based on the magnitude of its coefficient in each fold. If a feature was not present in the final model, its rank was assigned to the one point higher than the number of selected features. Then, the average and standard deviation of their ranks were calculated. For random forest, a similar procedure to the elastic net ranking was carried out. The more important feature has a large mean decrease in accuracy. If accuracy decreases significantly due to the exclusion of a single feature, this excluded feature is considered as important. All features were ranked where the highest value received the first rank. The ranks of the features obtained with elastic net and random forest models were shown in Table 16 and Table 17 respectively.

Table 16: The ranks of the features for Elastic net

	Mean	**Min**	**Max**	**Median**
Number of interrogative sentences	2.33±1.63	1	9	2
Percentage of whitespace	3.53±4.42	1	16	1
Number of passive voice sentences	3.72±1.68	1	12	4
Font Size	4.23±1.14	1	7	5
Medical term count	5.38±1.34	2	11	6
Medical term variety	6.98±5.77	1	16	3
Font face	8.15±3.25	4	16	7
Up to date	8.82±2.53	6	16	8
Underline count	9.27±1.90	7	15	9
Section count	9.40±2.54	3	18	10
Underline word count	9.52±1.64	7	17	9
Risk related sentence count	11.15±3.51	4	31	12
Average paragraph length	11.28±2.23	6	19	12
Paragraph count	12.67±3.03	5	27	13
Usage of parenthesis	13.95±2.84	7	28	14
Normalized neutral sentence count	14.60±3.03	7	30	15
Neutral sentence count	15.03±2.71	5	18	16
Normalized positive sentence count	15.05±2.65	6	18	16
Number of sentences starting with a pronoun	15.08±2.55	8	18	16
Left margin alignment	15.10±2.50	8	18	16
Background	15.12±2.46	8	18	16
SMOG	15.17±2.39	8	18	16
Negative imperative sentence count	15.22±2.89	8	26	16
Different type of sources (pdf)	15.22±2.37	8	18	16
Number of imperative sentences	15.23±3.44	4	31	16
Number of description sentences	15.23±2.59	8	23	16

Uppercase word count	15.27±2.46	8	21	16
Number of images	15.28±2.45	8	20	16
Positive sentence count	15.32±2.53	8	22	16
Italic word count	15.35±2.63	8	24	16
Italic count	15.37±2.69	8	25	16
Bullet count	15.42±3.11	8	31	16
External sources count	15.43±2.97	8	29	16
Benefit related sentence count	15.47±3.13	8	31	16
Negative sentence count	15.47±3.13	8	31	16
Bold count	15.47±3.13	8	31	16
Bold word count	15.47±3.13	8	31	16
Word count	15.47±3.13	8	31	16
Sentence count	15.47±3.13	8	31	16
Existence of colour	15.47±3.13	8	31	16
Internal sources count	15.47±3.13	8	31	16
Whitespace count	15.47±3.13	8	31	16
Acronym word count	15.47±3.13	8	31	16
Short sentence rate	15.47±3.13	8	31	16
Normalized negative sentence count	15.47±3.13	8	31	16
Usage of synonym	15.47±3.13	8	31	16
Number of uses of term "patient"	15.47±3.13	8	31	16
Flesch Kincaid Grade Level	15.47±3.13	8	31	16
Different type of sources (video)	15.47±3.13	8	31	16

Table 17: The ranks of the features for Random Forest

	Mean	**Min**	**Max**	**Median**
Benefit related sentence count	1.57±1.46	1	10	1
Risk related sentence count	2.90±2.11	1	15	2
Section count	4.35±3.30	1	15	3
Uppercase word count	8.35±5.08	2	33	7.5
Word count	9.00±6.53	2	38	7
Positive sentence count	9.65±6.98	2	45	8
Whitespace count	10.57±7.32	1	30	8
Paragraph count	12.48±8.33	3	48	11
Sentence count	13.95±6.64	3	36	13.5
Internal sources count	13.98±7.65	3	29	12.5
Neutral sentence count	15.57±8.44	3	38	13
Usage of parenthesis	15.58±8.71	4	40	13
Number of description sentences	16.20±10.18	3	46	13
Medical term count	17.17±10.54	3	46	14
Negative sentence count	18.48±9.47	3	45	17
Bullet count	18.62±10.73	3	49	16
Negative imperative sentence count	20.37±10.54	4	48	18.5
Acronym word count	20.85±10.99	5	47	19
Normalized positive sentence count	22.82±10.74	4	47	23
Number of images	22.83±10.38	5	47	22.5
Percentage of whitespace	26.40±10.66	3	46	26
Bold word count	26.55±10.09	6	47	26.5
Bold count	27.72±12.03	4	49	27.5
Up to date	28.08±9.62	10	47	30
Number of passive voice sentences	28.35±11.40	9	49	27
Average paragraph length	28.40±10.55	7	45	29.5
Underline word count	29.02±9.62	10	46	29.5

Font size	29.33±11.11	4	48	31
Underline count	29.85±6.77	13	44	30
Existence of colour	29.92±10.15	10	48	30
External sources count	30.27±10.45	11	49	29.5
Background	31.77±9.02	13	46	32
Other sources (pdf)	32.30±10.56	11	49	33
Left margin alignment	32.72±7.26	13	45	33
Font face	32.78±9.30	14	47	34
Number of uses of term "patient"	33.07±10.93	8	49	35
Italic count	33.25±10.86	9	49	33.5
Other sources (video)	33.60±11.37	12	49	36
Normalized neutral sentence count	33.95±11.41	8	49	36.5
Medical term variety	34.03±10.05	11	49	35.5
Italic word count	34.70±9.99	13	49	37
Number of sentences starting with a pronoun	35.05±11.67	6	49	39
Usage of synonym	35.10±9.83	10	49	37.5
Flesch Kincaid Grade Level	37.77±8.95	12	49	40
SMOG	37.93±9.48	12	49	40
Number of imperative sentences	38.22±9.10	11	49	40
Number of interrogative sentences	38.28±9.56	15	49	42.5
Normalized negative sentence count	40.28±8.92	11	49	42.5
Short sentence rate	41.03±7.93	14	49	43

It was observed that *section count, usage of parenthesis, risk related sentence count, medical term count and paragraph count* were significant for both classification models. As can be seen in Table 17, even though all random forest ranking features were positively significant with quality; *font size, medical term variety, underline count and underline word count* had negative moderate correlation and *percentage of white space, font face, up to date and average paragraph length* had no correlation with quality using elastic net. As the reason for the results, the processes of the two models are different, elastic net is linear and random forest is nonlinear, and hence the feature selection is carried out differently. Elastic net performs feature selection based on the magnitude of its coefficient and deals with grouping effects, thereby rendering classifiers more easily interpretable (Liu, Sun, Wang, Liu, & Tao, 2016). Unlike elastic net, random forest computes weighted impurity for each feature and ranks features according to the measure. For both classification methods, many of the selected features in the top 15 features are related to the language and writing style. The result demonstrated that random forest found more features significant in "General points" category compared to elastic net. Only information currency, risk related sentence count and paragraph count were found as significant using elastic net, which ranked 8^{th}, 12^{th} and 14^{th} respectively. On the other hand, elastic net found more features significant in "Visual presentation style than random forest.

In addition, common high rank characteristics were observed for both models. These six features are marked as italics in Table 16 and 17. Except for *the usage of synonym* and *the number of uses of the term "patient",* there is no correlation between the DISCERN class of these features. These two features have a positive moderate correlation with the website quality. The classification model results after removing these high rank features are presented in Table 18.

Table 18: Cross-validation results after feature elimination

Elastic net		Random Forest	
MCC	**Accuracy**	**MCC**	**Accuracy**
0.43±0.91	0.72±0.45	0.57±0.83	0.78±0.42

As can be seen in Table 18, there is a difference in the results obtained after feature elimination. For this reason, ridge and lasso regression were considered to analyse the importance of features. Ridge regression always keeps all features in the model (alpha=0). On the other hand, Lasso performs feature selection; selects a feature from a correlated feature group, and ignores the others (alpha=1); unlike elastic net. Table 19 presents the results of the models as alpha changes. The best results were obtained when alpha was always the lowest.

Table 19: The results of the models as alpha changes

	All features		**After feature elimination**	
Alpha	**Accuracy**	**MCC**	**Accuracy**	**MCC**
0	**0.77±0.43**	**0.53±0.85**	**0.75±0.44**	**0.50±0.87**
0.1	0.75±0.44	0.50±0.87	0.77±0.43	0.53±0.85
0.2	0.73±0.45	0.47±0.89	0.73±0.44	0.46±0.89
0.3	0.72±0.45	0.43±0.91	0.73±0.45	0.47±0.90
0.4	0.72±0.45	0.43±0.91	0.70±0.46	0.40±0.92
0.5	**0.70±0.46**	**0.40±0.92**	**0.72±0.45**	**0.43±0.91**
0.6	0.72±0.45	0.43±0.91	0.72±0.45	0.43±0.91
0.7	0.70±0.46	0.40±0.92	0.68±0.47	0.36±0.94
0.8	0.73±0.45	0.47±0.89	0.70±0.46	0.40±0.92
0.9	0.70±0.46	0.40±0.92	0.68±0.47	0.37±0.94
1	**0.70±0.46**	**0.40±0.92**	**0.70±0.46**	**0.40±0.93**

5.4. Discussion

This study presented whether health websites are written according to professional health literacy guidelines and evaluated the quality of the websites that provide information about type 2 diabetes using features from guidelines according to DISCERN.

Although guidelines are opposed to the use of excessive number of words, in the literature such studies have achieved higher accuracy in classification (Blumenstock, 2008) (Liu, Yu, Wu, Qing, & Peng, 2019). Less word usage may indicate that the issue is not fully complete and comprehensive. In addition, moderate negative correlations between medical term variety and website quality were obtained; hence, the lower the medical term variety, the higher the information quality. Websites may include information redundancy, advertisements, and information on other topics but

not the necessary relevant information; therefore, high-quality websites would display higher medical term diversity. It can conclude that the higher both word count and medical word count, the higher the information quality.

The correlation results showed that benefit-related sentences exhibited significant positive correlations with the quality in line with the expectation of guidelines. The ranking results also showed that the benefit-related sentences can be considered as a valuable indicator for measuring the quality of information on health websites. Also, in the literature, the studies reportedly relying on the heading feature showed that it is a valuable feature to find relevant information (Raynor, De Veene, & Bryant, 2014; Dickinson, Raynor, Knapp, & MacDonald, 2016). As same as in this study, section count that was the count of each heading and subheading is one of the important features. As a consequence, it can be concluded that the division of information can improve the quality of information.

The expected readability scores of websites should fall in the range of 6 to 8. Nevertheless, only two low-quality sites and two high-quality sites had the desired readability scores in the dataset. The FKGL readability scores for the remaining 56 websites indicated difficult readability, with high-quality websites exhibiting suitability for a mean grade level of 10.5. The SMOG index also indicated difficult readability, with high-quality websites exhibiting suitability for a mean grade level of 9.2. There was no correlation between the readability scores and the information quality; this finding is in line with the study of Giorgi et al. (San Giorgi, de Groot, & Dikkers, 2017). In other words, a high-quality website is not necessarily easy to read, and vice versa.

The result of the classification showed that the majority of the language and writing style features were significant indicators for the quality assessment of websites. When the sentiment characteristics of the sentences and the number of description sentences were analysed differently from the previous study, it was observed that the use of neutral and positive sentences had an impact on the website quality. At the same time, usage of parenthesis and the number of description sentences were observed as useful features in the ranking. Even though the acronym word count and synonym usage have positive correlation with website quality, it was observed that they did not have low rank as expected compared to professional health literacy guidelines.

In addition, uppercase word count, whitespace count and font size of the visual presentation style features are also significant indicators.

Regularized linear models are effective for feature interpretation and selection. According to the results of the study, all the features are important, regardless of how they are ranked, and they affect the prediction results positively.

5.5. Conclusion

This study presents whether the comprehensive feature set derived from professional health literacy guidelines is informative in assessing the quality of information on

websites. This study revealed that these features are valuable as quality indicators. The majority of features constructed using professional health literacy guidelines are observed to be significantly correlated with information quality. The direction of the correlation of some of the features, however, ran counter to initial expectations such as the use of medical terms, synonyms, and acronyms. In future work, the system could be validated in different health domains.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this thesis, a novel automatic information quality and coverage evaluation framework was proposed based on textual and content-based features. The proposed models were used for the first time in the health domain. In this thesis, the type 2 diabetes domain was chosen for evaluation. In-depth analyses and comparisons were presented.

The results of the studies included in the thesis have emphasized the importance of textual and content-based features in assessing information quality. The contributions of the thesis can be summarized as follows:

- A novel framework is proposed for the evaluation of information quality and coverage.

- A new set of textual features is proposed inspired mainly from professional health literacy guidelines in addition to the features commonly used for assessing information quality in other domains. In addition, textual features are proposed based on sentence types used in health websites. These features were significantly correlated with information quality.

- The framework assessing the degree of information coverage requires the use of an evidence-based practice guideline such as ADA guideline, which has not been used before in assessing quality. The results of the study showed that high-quality websites include more domain-specific terms compared to low-quality websites.

- Both textual and content-based features have not been measured together on information quality evaluation for the health domain before.A comprehensive feature set is proposed derived from professional health literacy guidelines. It is observed that this feature set is valuable as quality indicators.

Researchers who would like to study this topic in the future could benefit from the results of the study in several ways. For instance, the effective features on assessing quality of information of health websites obtained in this thesis might be used for developing more effective search engines for health content. Given the generality of the model, it can be applied to variety of healthcare domains. This thesis demonstrated that evidence-based guidelines can be helpful in constructing queries to measure coverage.

Today, people are in search of accurate information more than ever due to the current fight with epidemic. The new coronavirus has reportedly caused the spread of false or misleading health information on the Internet (Cuan-Baltazar, Muñoz-Perez, Robledo-Vega, Pérez-Zepeda, & Soto-Vega, 2020). They manually analysed 110

websites based on HONcode, JAMA, DISCERN and readability scores and they concluded that there was no quality information about coronavirus. Kouzy et al. (Kouzy, et al., 2020) conducted a search on Twitter to analyse the magnitude of misinformation about coronavirus. They concluded that the medical misinformation was being propagated at an alarming rate.

These recent studies also show once more that it is important to access health information of high quality on the internet. HONcode or readability values are not sufficient alone and the health information needs to be evaluated based on evidence. The features used in this study can be used to differentiate websites of high quality that contain accurate information from the ones that include false medical claims, counterfeit drugs, and tests.

Finally, in future studies, the system could be validated in different health areas. The method can be also extended to work on social media texts such as Twitter posts and Facebook messages. Moreover, a further effort is required to be user-friendly in the future.

Limitations

The content evaluation depended on the websites that the search engines returned. One limitation of this study is that the collected dataset included returned websites from the search engines that employ methods of ranking. Second, it was challenging to review the websites comprehensively. Some of the sites contained many pages, often with links to external sites or pages. Some of these websites also contained links to PDF documents that might contain relevant information but were not necessarily authored by the website owner. For this reason, external PDF documents were not included in the evaluation, in line with previous studies (Grant, Rodger, & Hoffmann, 2015).

The misspellings in the documents were not corrected, even though they existed. Furthermore, the tagging structures of websites vary. For instance, some of them used <p> instead of <h1> or <li>. This was handled during pre-processing; however, there might be other tagging structures, which we might have missed during processing.

Another limitation is the use of binary classification due to the imbalanced distribution of the quality scores in the collected dataset. The regression methods could have been applied if the dataset has comprised more balanced quality scores.

The analysis of low-quality websites could have been better validated if more low-quality web sites were collected. Although textual features are useful for the classification of the quality of websites, content based features are more effective than textual features for identifying low-quality websites.

Although both textual and content features are found as important for identification of the quality of health websites, they are not sufficient to measure the truthfulness of certain facts. For instance, a certain medical fact could be written incorrectly and still the website comprising such misinformation could be classified as high-quality with

the proposed features. Hence, a further study could be employed for developing fact-checking methods for assessing health related content.

Finally, term sequences might be effective in the content-based analysis. Although the term sequences were considered as 2-grams and 3-grams, the fact that one sentence comes before another, in general, may disrupt the integrity of meaning. A more comprehensive study might be needed to validate the sequences of terms.

REFERENCES

Ahmad, K., Gillam, L., & Tostevin, L. (1999). University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). *In TREC*, (pp. 1-8).

Al-Thubaity, A., Alanazi, A. H., & Al-Tuwaijri, H. (2012). Weirdness coefficient as a feature selection method for arabic special domain text classification. *International Conference on Asian Language Processing*, (pp. 69-72).

American Heart Association. (2018). *American Heart Association*. Retrieved September 2019, from https://professional.heart.org/professional/GuidelinesStatements/UCM_316885_Guidelines-Statements.jsp

Association, A. D. (2016). Standards of Medical Care in Diabetes. *Diabetes Care, 39*.

Bada, E., & Ulum, Ö. G. (2018). Utilization of active and passive constructions in English academic writing. *Journal of Human Sciences,* 15(1), 413-421.

Barker, S., Charlton, N. P., & Holstege, C. P. (2010). Accuracy of internet recommendations for prehospital care of venomous snake bites. *Wilderness and Environmental Medicine*, 21(4), 298-302, DOI: https://doi.org/10.1016/j.wem.2010.08.016.

Blumenstock, J. E. (2008). Size matters: word count as a measure of quality on wikipedia. In *Proceedings of the 17th International Conference on World Wide Web*, (pp. 1095-1096), ACM.

Boyer, C., Frossard, C., Gaudinat, A., Hanbury, A., & Falquetd, G. (2017). How to sort trustworthy health online information? Improvements of the automated detection of HONcode criteria. *Procedia Computer Science,* 121, 940-949, DOI: https://doi.org/10.1016/j.procs.2017.11.122.

Bruce, J. G., Tucholka, J. L., Steffens, N. M., & Neuman, H. B. (2015). Quality of online information to support patient decision-making in breast cancer surgery. *Journal of Surgical Oncology*, 112(6), 575-580.

Bull, S. (2010). *Technology-based health promotion.* Sage.

Cajita, M. I., Rodney, T., Xu, J., Hladek, M., & Han, H. R. (2017). Quality and health literacy demand of online heart failure information. *The Journal of Cardiovascular Nursing*,32(2), 156.

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

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.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the Association for Information Science and Technology*, 52(1), 1-4.

Cuan-Baltazar, J. Y., Muñoz-Perez, M. J., Robledo-Vega, C., Pérez-Zepeda, M. F., & Soto-Vega, E. (2020). Misinformation of COVID-19 on the internet: infodemiology study. *JMIR Public Health and Surveillance*, 6(2), *DOI: 10.2196/18444*.

Dalip, D. H., Gonçalves, M. A., Cristo, M., & Calado, P. (2011). Automatic assessment of document quality in web collaborative digital libraries. *Journal of Data and Information Quality (JDIQ)*, 2(3), 1-30, DOI: https://doi.org/10.1145/2063504.2063507

Dalip, D. H., Lima, H., Cristo, M., & Calado, P. (2014). Quality assessment of collaborative content with minimal information. *IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, (pp. 201-210), IEEE.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1-30.

Department of Health. (2003). *Toolkit for producing patient information.* Retrieved March 2018, from https://www.uea.ac.uk/documents/246046/0/Toolkit+for+producing+patient+information.pdf

Dickinson, R., Raynor, D. K., Knapp, P., & MacDonald, J. (2016). Do patients use a headline section in a leaflet to find key information about their medicines? Findings from a user-test study. *Therapeutic Innovation & Regulatory Science*, 50(5), 581-591.

Durrant, L. A., Taylor, J., Thompson, H., Usher, K., & Jackson, D. (2019). Health literacy in pressure injury: Findings from a mixed-methods study of community-based patients and carers. *Nursing & Health Sciences*, 21(1), 37-43.

Eysenbach, G., Powell, J., Kuss, O., & Sa, E. R. (2002). Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review.*Jama,*287(20),2691-2700,DOI: 10.1001/jama.287.20.2691.

Fahy, E., Hardikar, R., Fox, A., & Mackay, S. (2014). Quality of patient health information on the internet: reviewing a complex and evolving landscape. *Australasian Medical Journal*, 7(1), 24-28.

Fitzsimmons, P. R., Michael, B. D., Hulley, J. L., & Scott, G. O. (2010). A readability assessment of online Parkinson's disease information. *The Journal of the Royal College of Physicians of Edinburgh*, 40(4), 292-296.

Floridi, L. I. (2014). *The philosophy of information quality* . Cham: Springer.

Fox, S. (2011). *Health topics: 80% of internet users look for health information online.* Pew Internet & American Life Project.

Fox, S., & Duggan, M. (2013). Health online 2013. *Health*, 1-55.

Goasdoué, F., Karanasos, K., Katsis, Y., Leblay, J., Manolescu, I., & Zampetakis, S. (2013). Fact checking and analyzing the web. *In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, (pp. 997-1000).

Gómez-Adorno, H., Sidorov, G., Pinto, D., Vilariño, D., & Gelbukh, A. (2016). Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. *Sensors*, 16(9), 1374.

Grant, N., Rodger, S., & Hoffmann, T. (2015). Evaluation of Autism-Related Health Information on the Web. *Journal of Applied Research in Intellectual Disabilities*, 28(4), 276-282, DOI: https://doi.org/10.1111/jar.12127.

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), DOI: 10.2196/jmir.7.5.e59.

Health On the Net. (2018). *Trustworthy health and medical information: the Health On the Net*. Retrieved March 2018, from https://www.hon.ch/HONcode/

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *In Eighth International Conference on Weblogs and Social Media (ICWSM-14).*

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, 3(2), 259-289.

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, DOI: https://doi.org/10.1016/j.pec.2008.03.002.

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 Counseling*, 77(1), 33-37, DOI: https://doi.org/10.1016/j.pec.2009.02.016.

Khazaal, Y., Fernandez, S., Cochand, S., & Reboh, I. (2008). Quality of web-based information On social phobia: a cross sectional study. *Depression and Anxiety*, 25(5), 461-465, DOI: https://doi.org/10.1002/da.20381.

Kiran, D. S., 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(2), 481.

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M. B., Karam, B., Adib, E., & Baddour, K. (2020). Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus,* 12(3), DOI: 10.7759/cureus.7255.

Kristina, T., & Christopher D., M. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, (pp. 55-60).

Lambert, K., Mullan, J., Mansfield, K., Koukomous, A., & Mesiti, L. (2017). Evaluation of the quality and health literacy demand of online renal diet information. *Journal of Human Nutrition and Dietetics,* 30(5), 634-645.

Lee, M. A., Shin, C. N., & An, K. (2018). Trustworthiness, Readability, and Suitability of Web-Based Information for Stroke Prevention and Self-Management for Korean Americans: Critical Evaluation. *Interactive Journal of Medical Research,* 7(2), DOI: 10.2196/10440 .

Li, B., Yang, X., Wang, B., & Cui, W. (2017). Efficiently mining high quality phrases from texts. *In Thirty-First AAAI Conference on Artificial Intelligence,* (pp.3474-3481).

Liu, J., Shang, J., Wang, C., Ren, X., & Han, J. (2015). Mining quality phrases from massive text corpora. *In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data,* (pp.1729-1744).

Liu, Q., Sun, Y., Wang, C., Liu, T., & Tao, D. (2016). Elastic net hypergraph learning for image clustering and semi-supervised classification. *IEEE Transactions on Image Processing,* 26(1), 452-463.

Liu, Y., Yu, K., Wu, X., Qing, L., & Peng, Y. (2019). Analysis and detection of health-related misinformation on Chinese social media. *IEEE Access,* 7, 154480-154489.

Machackova, H., & Smahel, D. (2018). The perceived importance of credibility cues for the assessment of the trustworthiness of online information by visitors of health-related websites: The role of individual factors. *Telematics and Informatics, 35*(5), 1534-1541.

Mandelblatt, J., Kreling, B., Figeuriedo, M., & Feng, S. (2006). What is the impact of shared decision making on treatment and outcomes for older women with breast cancer? *Journal of Clinical Oncology*, 24(30), 4908–4913.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Probabilistic information retrieval. In *Introduction to Information Retrieval*, (pp. 220-235).

Marahrens, L., Roeck, D., Kern, R., Ziemssen, T., Fritsche, A., & Ziemssen, F. (2015). Do persons with diabetes want to decide on their own when treating diabetic retinopathy? The intended role of the ophthalmologist - results of the DiabCheckOCT+ Study. *Investigative Ophthalmology & Visual Science,* 56(7), 1372.

Marie-Catherine, d. M., & Christopher D., M. (2008). The Stanford typed dependencies representation. *In COLING 2008 Workshop on Cross-framework and Cross-domain Parser Evaluation,* (pp. 1-8).

Medical Subject Headings. (2018). *Medical Subject Headings*. Retrieved June 2018, from https://www.nlm.nih.gov/mesh

Medicines and Healthcare products Regulatory Agency. (29 December 2014). *Best Practice Guidance on Patient Information Leaflets.* Medicines and Healthcare products Regulatory Agency.

MedlinePlus. (2018). *MedlinePlus Trusted Health Information for You*. Retrieved June 2018, from https://medlineplus.gov/etr.html

Meric, F., Bernstam, E. V., Mirza, N. Q., Hunt, K. K., Ames, F. C., Ross, M. I.,& Singletary, S. E. (2002). Breast cancer on the world wide web: cross sectional survey of quality of information and popularity of websites. *BMJ*, 324(7337), 577-581.

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, DOI: https://doi.org/10.1016/j.jad.2008.01.007.

Morten, W.-W., Cosley, D., & Riedl, J. (2013). Tell me more: an actionable quality model for Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*, (pp. 1-10).

National Comprehensive Cancer Network. (2018). *NCCN Guidelines*. Retrieved September 2019, from https://www.nccn.org/professionals/physician_gls/default.aspx

Nghiem, A. Z., Mahmoud, Y., & Som, R. (2016). Evaluating the quality of internet information for breast cancer. *The Breast* , 25, 34-37, DOI: https://doi.org/10.1016/j.breast.2015.10.001.

Nilsson-Ihrfelt, E., Fjällskog, 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.

O'Connor, A., Llewellyn-Thomas, H., & Flood, A. (2004). Modifying unwarranted variations in health care: shared decision making using patient decision aids. *Health Affairs*, 23(2), 63-72.

Olkun, H. K., Demirkaya, A. A., & Aras, B. (2019). The quality of Internet information on lingual orthodontics in the English language, with DISCERN and JAMA, *Journal of Orthodontics*, 46(1), 20-26.

Oroszlányová, M., Teixeira Lopes, C., Nunes, S., & Ribeiro, C. (2018). Predicting the quality of health web documents using their characteristics. *Online Information Review,* 42(7), 1024-1047.

Oxford Dictionary. (2019). *Textual features-Oxford Reference*. Retrieved March 2019,from http://www.oxfordreference.com/view/10.1093/oi/authority.20110803103357643

Patra, B. G., Banerjee, S., Das, D., Saikh, T., & Bandyopadhyay, S. (2013). Automatic author profiling based on linguistic and stylistic features. *Notebook for PAN at CLEF*.

Prescott, C. (2019). Internet access—households and individuals: 2019. *London: Office for National Statistics*.

Rao, N. R., Mohapatra, M., Mishra, S., & Joshi, A. (2012). Evaluation of dengue-related health information on the internet. *Perspectives in Health Information Management*, 9(Summer):1c.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* (pp.2931-2937).

Rassbach, L., Pincock, T., & Mingus, B. (2007). Exploring the feasibility of automatically rating online article quality. In *Proceedings of the 2007 International Wikimedia Conference (WikiMania).* Taipei, Taiwan.

Raynor, D. K., De Veene, P., & Bryant, D. (2014). The effectiveness of the Summary of Product Characteristics (SmPC) and recommendations for improvement. *Therapeutic Innovation & Regulatory Science*, 48(2), 255-265.

Reddy, T. R., Vardhan, B. V., & Reddy, P. V. (2017). N-gram approach for gender prediction. *In 2017 IEEE 7th International Advance Computing Conference (IACC)* (pp. 860-865). IEEE.

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 Counseling,* 47(3), 273-275, DOI: https://doi.org/10.1016/S0738-3991(01)00225-7.

Robertson, S. E., Walker, S., Beaulieu, M., & Willett, P. (1999). Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. *Nist Special Publication SP*, 500, 253-264.

Sackett, D. L., Rosenberg, W. M., 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, DOI: https://doi.org/10.1136/bmj.312.7023.71.

Sağlam, R. B., & Temizel, T. T. (2014). Automatic information timeliness assessment of diabetes web sites by evidence based medicine. *Computer Methods and Programs in Biomedicine,* 117(2), 104-113.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM,* 18(11), 613-620.

San Giorgi, M. R., de Groot, O. S., & Dikkers, F. G. (2017). Quality and readability assessment of websites related to recurrent respiratory papillomatosis. *The Laryngoscope,* 127(10), 2293-2297.

Seidman, J. J., Steinwachs, D., & Rubin, H. R. (2003). Design and testing of a tool for evaluating the quality of diabetes consumer-information Web sites. *Journal of Medical Internet Research*, 5(4), DOI: 10.2196/jmir.5.4.e30.

Shang, J., Liu, J., Jiang, M., Ren, X., Voss, C. R., & Han, J. (2018). Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10), 1825-1837.

Shen, A., Qi, J., & Baldwin, T. (2017). A hybrid model for quality assessment of Wikipedia articles. *In Proceedings of the Australasian Language Technology Association Workshop* , (pp. 43-52).

Silberg, W. M., Lundberg, G. D., & Musacchio, R. A. (1997). Assessing, controlling, and assuring the quality of medical information on the Internet: Caveat lector et viewor—Let the reader and viewer beware. *Jama,* 277(15), 1244-1245.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, 85(3), 257-268.

Stvilia, B., Mon, L., & Yi, Y. J. (2009). A model for online consumer health information quality. *Journal of the American Society for Information Science and Technology*, 60(9), 1781-1791.

The iWeb corpus. (2018). *iWeb: The 14 Billion World Web Corpus*. Retrieved December 2018, from https://www.english-corpora.org/iweb/

U.S. Department of Health and Human Services, O. o. (2015). *Health Literacy Online: A Guide to Writing and Designing Easy-to-Use Health Web Sites.* Retrieved June 2018, from https://health.gov/healthliteracyonline/2010/Web_Guide_Health_Lit_Online.pdf

Vysotska, V., Burov, Y., Lytvyn, V., & Demchuk, A. (2018). Defining author's style for plagiarism detection in Academic Environment. *In 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, (pp. 128-133).

Wang, P., & Li, X. (2019). Assessing the quality of information on wikipedia: A deep-learning approach. *Journal of the Association for Information Science and Technology*, 71(1), 16-28.

Weymann, N., Harter, M., & Dirmaier, J. (2014). Quality of online information on type 2 diabetes: a cross-sectional study. *Health Promotion International*, 30(4), 821-831.

Wikipedia. (2018). *Wikipedia: Featured article criteria*. Retrieved February 2017, from https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

Witchel, H. J., Thompson, G. A., Jones, C. I., Westling, C. E., Romero, J., Nicotra, A., & Critchley, H. D. (2020). Spelling Errors and Shouting Capitalization Lead to Additive Penalties to Trustworthiness of Online Health Information: Randomized Experiment With Laypersons. *Journal of Medical Internet Research*, 22(6), DOI: 10.2196/15171.

APPENDICES

APPENDIX A

DISCERN AND CONTENT ASSESSMENT MANUALS

1) DISCERN Manual

Bölüm 2: Tedavi seçenekleri hakkında bilgi kalitesi ne kadar iyi?

9) Her tedavinin nasıl çalıştığı açıklanıyor mu?

Bu soruyu cevaplarken dikkat edilecek hususlar ve sorunun önemi:

- İyi kaliteli bir yayın vermiş olduğu tedavi seçeneği ile ilgili vücudunuz da oluşacak değişikleri, tedavinin; hastalığın ve semptomlarının nasıl etkilediğini, nasıl tedavi ettiğini açıklamalıdır. Tedavi seçeneğinin nasıl çalıştığının anlatılması, kişinin tedaviyi anlaması ve hangi seçeneğin kendine uygun olduğu kararına varabilmesi önemlidir.

Puanlama

- **5:** Her bir tedavi seçeneğinin tanımında tedavinin nasıl bir süreç olduğunu detaylı olarak içeriyorsa
- **2-4:** Tedavi seçeneklerinin bir kısmının detaylı verilmesi veya detaylarda anlaşılmayan eksik kalan durumlar olduğunda
- **1:** Tedaviler hakkındaki tanımların hiçbirinde detaylı olarak nasıl bir süreç olduğunu açıklamıyorsa

İpucu: Bu soru cevaplanırken tedavinin tanımında; tedavi süresince kişinin vücudunda ne gibi etkileri olup olmadığının anlatılması önemlidir.

Örnek:
Bir tedavi için 5 puanlık örnek
“ X tedavisi, sizin durumunuzu stabilize edecektir. Tedavi, iki ila üç hafta arasında bir süre boyunca her gün yüksek bir doz enjeksiyon olarak verilir. Tedavi kanda dolaşarak vücudun her yerindeki hücrelere ulaşır ve virüsün yeniden üremesini ve yayılmasını önler. “

*******Bir yayın birden fazla tedavi seçeneği içerebilir, her bir tedavi seçeneği için bu şekilde açıklama varsa 5 puan verilmelidir.

10) Her tedavinin yararları açıklanıyor mu?
Bu soruyu cevaplarken dikkat edilecek hususlar ve sorunun önemi:

- İyi kaliteli bir yayın vermiş olduğu her bir tedavi seçeneğinin yararlarını açıklamalıdır. Birçok tedavi seçeneğinin bazı yararları ya da avantajları bulunmaktadır. Bir tedavinin sonucunda ne olacağını bilmek ve sizin için önemli olan bir sonuçla ilişkili olup olmadığı kararına varmak için tedavi seçeneklerinin faydalarının farkında olmak önemlidir.

Puanlama
Soru; hastalığın tedavisi ya da yayında açıklanan tedaviler ile ilgili herhangi bir yararın açıklamasının söz konusu olup olmadığını değerlendirmek içindir. Soru; yarar boyutu veya kimlerin en büyük olasılıkla yarar sağlayacağı ile ilgili değildir. Soru; tedavinin bütün yararlarını içerip içermediği gibi bir kıyas yaparak puanlanmamalıdır.

- **5:** Her bir tedavi seçeneğinin yararlarını içeriyorsa
- **2-4:** Tedavi seçeneklerinin bir kısmının yararlarını içeriyorsa
- **1:** Tedaviler hakkında hiçbir yarar olduğu açıklanmıyorsa

İpucu: Bu soru cevaplanırken tedavinin yararları; kısa vadeli ve uzun vadeli, kontrol etmek veya semptomlarından kurtulmak, durumun tekrarını önlemek ve durumu ortadan kaldıracak durumlar ile ilgili bilgileri içermelidir.

Örnek:
Bir tedavi için 5 puanlık örnek
"Çoğu kişide bu tedavi birkaç hafta içinde hastalığı tamamıyla iyileştirmektedir."
"Kısa vadede, X tedavisi hastalıktan kaynaklanan kilo kaybını durdurabilir. Aynı zamanda ağrı ve nefes darlığı semptomlarını azaltabilir. Bu durum tam olarak tedavi edilemez iken, X tedavisi uzun vade de istikrar sağlayabilir ve kötüleşen durumu önleyecektir."
******* Bir yayın birden fazla tedavi seçeneği içerebilir, her bir tedavi seçeneği için bu şekilde açıklama varsa 5 puan verilmelidir.

11) Her tedavinin riskleri açıklanıyor mu?
Bu soruyu cevaplarken dikkat edilecek hususlar ve sorunun önemi:

- İyi kaliteli bir yayın her tedavi için tedavinin risklerini anlatmalıdır. Birçok tedavi seçeneğinin bazı riskleri ya da dezavantajları bulunmaktadır. Bir tedavinin sonucunda ne olacağını bilmek ve sizin için önemli olan bir sonuçla ilişkili olup olmadığı kararına varmak için tedavi seçeneklerinin risklerinin farkında olmak önemlidir.

Puanlama
Soru; hastalığın tedavisi ya da yayında açıklanan tedaviler ile ilgili herhangi bir riskinin açıklamasının söz konusu olup olmadığını değerlendirmek içindir. Soru; risk boyutu veya kimlerin en büyük olasılıkla risk sağlayacağı ile ilgili değildir. Soru; tedavinin bütün risklerini içerip içermediği gibi bir kıyas yaparak puanlanmamalıdır.

- **5:** Her bir tedavi seçeneğinin risklerini içeriyorsa
- **2-4:** Tedavi seçeneklerinin bir kısmının risklerini içeriyorsa
- **1:** Tedaviler hakkında hiçbir riskin olduğu açıklanmıyorsa

İpucu: Bu soru cevaplanırken tedavinin riskleri; kısa vadeli ve uzun vadeli, yan etkileri, komplikasyonları ve tedavinin olumsuz tepkileri gibi bilgileri içermelidir.

Örnek:
"Tedavinin yan etkileri arasında baş dönmesi, kas ağrısı ve uykusuzluk yer almaktadır."

"Araştırmalara göre tedavinin yan etkileri genellikle hafif ve geri dönüşümlü olmuştur; ağız kuruluğu ve mide bulantısı dâhil. Tedavi; kısa bir süre içinde organlar üzerinde herhangi büyük zararlı etkiye sahip değildir. Ancak, bazı hastalarda uzun süreli kullanımdan sonra böbrek problemleri gelişebilir; tedavi dikkatle izlenmelidir."
******* Bir yayın birden fazla tedavi seçeneği içerebilir, her bir tedavi seçeneği için bu şekilde açıklama varsa 5 puan verilmelidir.

12) Herhangi bir tedavi uygulanmazsa ne olacağı açıklanıyor mu?
Bu soruyu cevaplarken dikkat edilecek hususlar ve sorunun önemi:
- İyi kaliteli bir yayın tedavi uygulanmaması durumda neler olacağının açıklamasını içermelidir. Her hangi bir tedavi uygulanmaması (tedavi sonrası süreci izleme, tedavinin ertelenmesi, tedaviden tam olarak vazgeçilmesi vs.) sonucunda ne olacağını bilmek ve sizin için önemli olan bir sonuçla ilişkili olup olmadığı kararına varmak için tedavi uygulanmadığında ki sonuçların farkında olmak önemlidir.

Puanlama
Soru; hastalık için hiçbir tedavi uygulanmaması durumunda neler olacağı ile ilgili herhangi bir açıklamanın söz konusu olup olmadığını değerlendirmek içindir. Soru; yarar ve risk boyutu veya kimlerin en büyük olasılıkla etkileneceği ilgili değildir. Soru; tedavi uygulanmadığında oluşacak bütün yararların veya risklerin içerip içermediği gibi bir kıyas yaparak puanlanmamalıdır.
- **5:** Herhangi bir tedavi seçeneğinin kullanılmaması durumda ki yararları ve riskleri detaylı olarak içeriyorsa
- **2-4:** Herhangi bir tedavi seçeneğinin kullanılmaması durumda ki yararlar ve risklerin bulunmasına rağmen anlaşılmaz veya tam bilgi içermiyorsa
- **1:** Tedavi seçeneğinin kullanılmaması hakkında hiçbir risk veya yarar olduğunu açıklanmıyorsa

İpucu: Bu soru cevaplanırken tedavi uygulanmadığı sürece oluşan riskler ve yararlar ya da kesin olarak tedaviden feragat edildiğinde ki riskler ve yararlar ile ilgili bilgileri içermelidir.

Örnek:
"Bu hastalık hızla ilerleyen, yaşamı tehdit eden bir durumdur ve erken tedavi tavsiye edilir. Ne yazık ki, tedavi geciktirilmesi kalbinizde uzun vadeli hasara yol açabilir. "

******* Bir yayın birden fazla tedavi seçeneği içerebilir, her bir tedavi seçeneği için bu şekilde açıklama varsa 5 puan verilmelidir.

13) Tedavi seçeneklerinin genel yaşam kalitesini nasıl etkilediği açıklanıyor mu?
Bu soruyu cevaplarken dikkat edilecek hususlar ve sorunun önemi:

- İyi kaliteli bir yayın her hangi bir tedavi seçeneğinin sadece risk veya yararları belirtilmemeli; tedavi sürecinin genel olarak günlük hayatı nasıl etkilediği anlatılmalıdır. Bu soru, bir tedavi seçeneğine karar vermeden önce, yaşam tarzınız ya da koşullarınız üzerinde önemli etkilere sahip olup olmadığı aynı zamanda size yakın olanları nasıl etkilediği hakkında bilgi sahibi olunması açısından önemlidir.

Puanlama
Soru verilmiş olan tedavi seçenekleri ile ilgilidir. Eğer soru 12'ye 1 puandan fazla verdiyseniz, tedavi uygulanmama durumu da bir tedavi seçeneği olarak kabul edilir.

- **5:** Tedavi seçeneklerinin genel olarak yaşam kalitesini nasıl etkilediği ile ilgili açık bir referans içeriyorsa
- **2-4:** Tedavi seçeneklerinin genel olarak yaşam kalitesini nasıl etkilediği ile ilgili referans içermesine rağmen anlaşılmaz veya tam bilgi içermiyorsa
- **1:** Tedavi seçeneği ile yaşam kalitesini nasıl etkilediği hakkında herhangi bir ilişki vermemişse

İpucu: Bu soru cevaplanırken tedavi seçeneğinin günlük hayata olan etkileri ve aileniz, arkadaşlarınız veya bakımınızı sağlayan kişi ile olan ilişkinize olan etkilerini içermelidir.

Örnek:
"Tedavinin gerçekleştirilmesinden önce, hastalar her hafta hastanede izlenir ve iyileşme süresi birkaç gün ila birkaç hafta arasında değişebilir. Eğer bu tedavi şeklini seçiyor iseniz işten uzun süre izin almanız ve eve geldiğinizde sizin için bakım sağlayacak birine sahip olmanız gerekebilir."
*******Tedavinin genel yaşam kalitesine olan etkileri ile ilgili bu şekilde açıklama varsa 5 puan verilmelidir.

14) Birden fazla tedavi seçeneği olduğu açık mı?
Bu soruyu cevaplarken dikkat edilecek hususlar ve sorunun önemi:

- Bazı tedavi seçenekleri herkese uygun olmayabilir, bazı kişilere daha uygun olabilir, ancak yazı da belirtilmese bile her zaman farklı tedavi seçeneklerinin var olabileceğinden bahsedilmelidir.

Puanlama
Bu soru cevaplandırılırken sizin kararlarınız önemli olacaktır. Yazıyı bir bütün olarak düşünün ve bunu her tüketici veya hasta için tedavi seçiminde karar vermesine yardımcı olacak bilgiyi içerip içermediğine bakarak karar vermeniz size yardımcı olacaktır.

- **5:** Yazı açıkça birden fazla tedavi seçeneği olduğunu belirtiyorsa
- **2-4:** Yazı tedavi seçenekleri olduğunu belirtmesine rağmen anlaşılmaz veya tam bilgi içermiyorsa
- **1:** Tedavi seçenekleri olduğu ile ilgili hiçbir gösterge yoksa

İpucu: Bu soru cevaplanırken anlatılan tedavi seçeneğinin en çok kimler için uygun olduğu; bir tedavi seçeneğini seçerken diğer seçenekler ile ilgili yeterli bilgiye sahip olunup ona göre karar verilmesi gerekliliği ile ilgili bilgiler içermelidir.

Örnek:
"Bu yazı yeni bir tedavi seçeneği ile ilgili bilgiler içermektedir. Bu tedavinin kimlere uygun olduğu, nasıl yan etkileri olduğu, tedavi sonucunda neler olacağı açıkça anlatılmıştır. Diğer tedavi seçenekleri ise açıkça belirtilmekte ve okuyucular için detaylı bilgi için farklı referanslar verilmiştir."
*******Tedavi seçenekleri ile ilgili bu şekilde açıklamalar varsa 5 puan verilmelidir.

"Bu yazı da var olan hastalık için bir tedavi anlatacaktır. Diğer tedavi seçenekleri için hepsinin yan etkisi olduğu veya tedavi uygulanmazsa büyük kayıpların olacağı belirtilmektedir. Bu tedavinin herkese uygun olduğu belirtilmiştir."
*******Tedavi seçenekleri ile ilgili bu şekilde açıklamalar varsa 1 puan verilmelidir.

15) Karar verme (shared decision making) için destek sağlıyor mu?
Bu soruyu cevaplarken dikkat edilecek hususlar ve sorunun önemi:

- Yayın sizin için en iyi tedavi seçeneğini seçmeniz açısından bakım dâhil her durumu açıklamalıdır. Sağlanan bir tedavi seçeneği ya da seçenekleri hakkındaki bilgileri bir sağlık profesyoneli ile istişare ederek hazırlanması ile ilgili ya da tedavi seçenekleri ile ilgili aileniz, arkadaşlarınız veya bakıcılarınızı etkileyebilecek konularda bilgi sağlamalıdır.

Puanlama
Yayını bir bütün olarak düşünün ve seçmiş olduğunuz tedavi seçenekleri ile ilgili görüşmek üzere belirli konularda bir liste hazırlamanızda etkin olup olmadığına karar verin. Bu konular daha ziyade eksiklikler ve boşluklardan kaynaklanan durumlar olmaktansa, yayında açıkça belirtilen durumlar olmalıdır. Ne kadar yüksek puan vereceğiniz yayının tedavi seçenekleri hakkında vermiş olduğu bilgilerin sizin paylaşım kararlarınıza ne kadar destek olduğuna bağlıdır. Eğer kendi tedavi kararınız hakkında puanlama yapmak istemiyorsanız, tedavi seçenekleri hakkında daha fazla bilgi isteyen bir bakıcı veya akraba bakış açısıyla yayını puanlayın.

- **5:** Yazı tedaviye karar verme açısından çok iyi destek sağlıyorsa
- **2-4:** Yazı tedaviye karar verme açısından destek sağlıyorsa
- **1:** Yazı tedaviye karar verme açısından hiçbir destek sağlamıyorsa

İpucu: Bu soru cevaplanırken anlatılan tedavi seçeneği için aileniz, arkadaşlarınız ya ada diğer sağlık görevlileri ile bazı durumlar hakkında öneriler almanızdan bahsetmelidir.

Örnek:
"Tedavi birkaç yıl sürebilir ve bu süre içinde hamile kalmanız tavsiye edilmez. Yakında bir bebek sahibi olmayı planlıyor ya da gebe olduğunuzu düşünüyorsanız tedaviye başlamadan önce, doktorunuzla konuşmalısınız. Eğer bir eşiniz varsa, doktorunuzla bu konuları birlikte tartışmayı deneyin."
*******Tedavi seçenekleri ile ilgili bu şekilde açıklamalar varsa 5 puan verilmelidir.

Bölüm 3: Yazıyı genel olarak puanlama

16) Cevaplamış olduğunuz bütün soruları temel alarak, yazının tedavi seçenekleri ile ilgili vermiş olduğu bilginin genel kalitesini oylayın.
Bu soruyu cevaplarken dikkat edilecek hususlar ve sorunun önemi:

- Her yayının bazı eksiklikleri olabilir. Ancak, soruları yanıtladıktan sonra genel olarak yayına baktığınız zaman; yeterli bilgi olduğunu, tedavi seçeneklerini uygun bir şekilde anlattığı hissini sizde bırakmalıdır.

Puanlama
Bu soru cevaplandırılırken sizin kararlarınız önemli olacaktır.

- **5:** Genelde bütün sorulara 4 ve üzeri puanladıysanız genel olarak yayın iyi bir kalite de denebilir.
- **3:** Yazı için eşit sayıda yüksek ve düşük puanlamalar var ise yazı için orta kalitede denebilir.
- **1:** Genelde sorular 2 ve altında puanlandıysa genel olarak yayın kötü kalite de denebilir.

2) Content Coverage Manual

A guideline is prepared by using ADA's gold standard (Association, 2016) to domain experts to inform them about the issues that need to be considered for assigning a score. The following text is retrieved partially from the guideline as it is.

Obesity Management For The Treatment Of Type 2 Diabetes (Section 6)

Treatment for overweight and obesity in type 2 diabetes

Treatment	BMI category (kg/m^2)				
	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/m^2 with one or more obesity-associated comorbid conditions and by patients with BMI ≥30 kg/m^2 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.

- **Bariatric Surgery**

Bariatric and metabolic surgeries, either gastric banding or procedures that involve resecting, bypassing, or transposing sections of the stomach and small intestine, can be effective weight loss treatments for severe obesity when performed as part of a comprehensive weight management program with lifelong lifestyle support and medical monitoring. National guidelines support consideration of bariatric surgery for people with type 2 diabetes with BMI >35 kg/m^2.

  - **Advantages**
  - **Disadvantages**

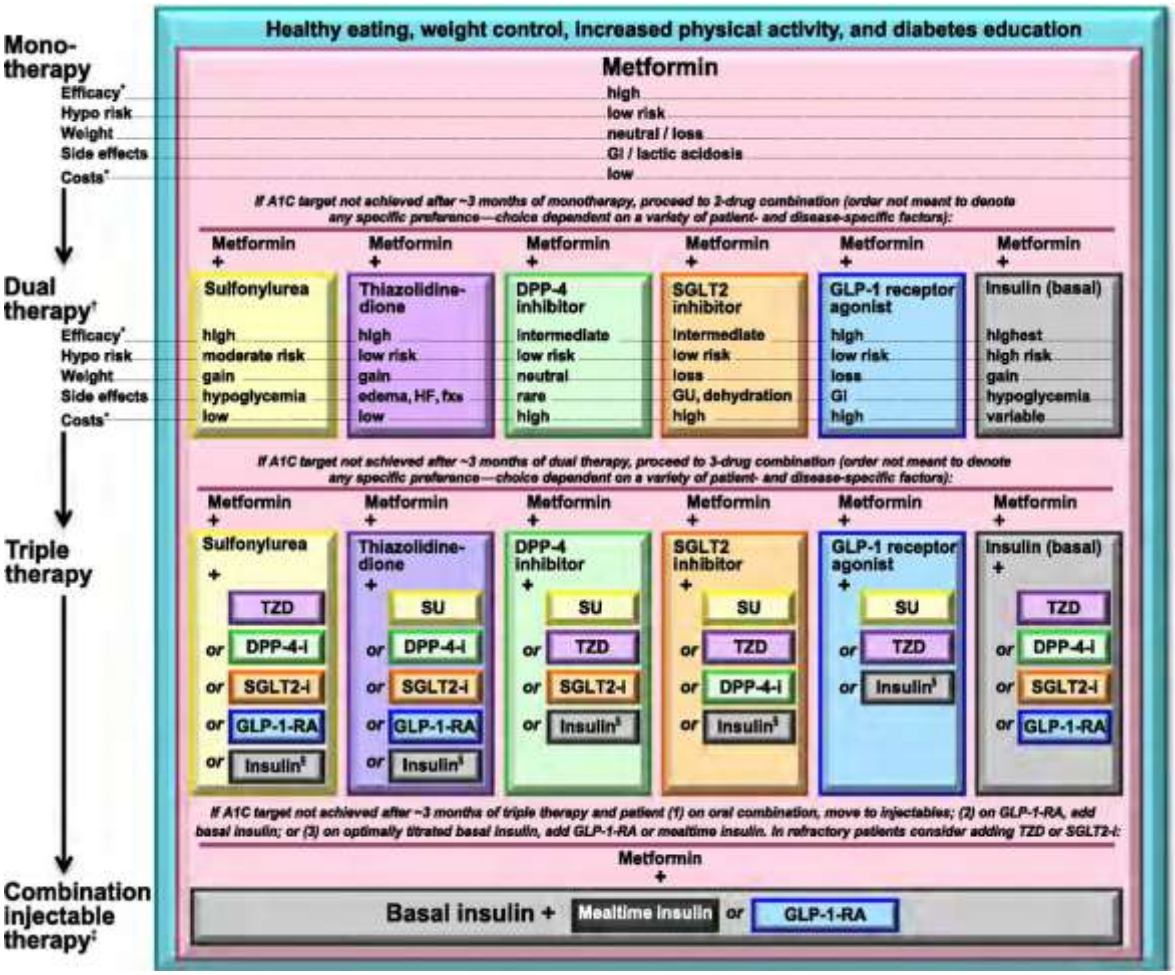
Pharmacological Therapy for Type 2 Diabetes (Section 7)

- **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**

Scoring in section 6

APPENDIX B

DATASET INFORMATION

Table 20: Data Set Information

Website	DISCERN	Coverage Assessment
http://www.diabetes.org	40	24
http://www.webmd.com	39	18
https://www.diabetesaustralia.com.au	39	20
http://www.healthline.com	39	24
https://en.wikipedia.org/wiki/Diabetes_mellitus_type_2	38	24
http://www.nhs.uk	38	20
http://www.mayoclinic.org	38	22
http://www.diabetes.org.nz	9	2
http://www.diabetes.org.nz	16	8
http://www.joslin.org	39	21
http://www.everydayhealth.com	38	22
http://kidshealth.org	25	18
http://patient.info	39	17
https://www.diabetes.ie	33	16
http://www.webmd.boots.com	23	18
http://www.endocrineweb.com	34	18
https://www.drugwatch.com	33	19
https://www.betterhealth.vic.gov.au	23	8
http://www.medicalnewstoday.com	33	17
http://www.diabetes.ca	27	12
https://www.diabetes.org.uk	40	24
https://healthfinder.gov	11	4
http://www.diabetes.co.uk	40	20
http://www.news-medical.net	14	5
http://www.dlife.com	38	20
https://diatribe.org	36	20
https://www.niddk.nih.gov	32	16
http://www.idf.org	8	2
http://www.bupa.co.uk	21	15
https://www.drugs.com	40	24
http://www.medicinenet.com	37	17
https://www.lvhn.org	22	18
https://dtc.ucsf.edu	32	19
http://www.health.com	19	15
http://www.uptodate.com	36	20
http://diabetes.mercola.com/	8	6
http://www.diabeticlivingonline.com/	20	17
http://www.diabetesselfmanagement.com	26	18
http://www.wikihow.com/Manage-Type-2-Diabetes	16	12
http://www.uofmhealth.org	11	6
http://www.diabeticlifestyle.com	30	17
http://www.defeatdiabetes.org	8	1
http://www.diabetesnet.com	40	20

https://www.msdiabetes.org	10	12
http://www.onlinemedinfo.com	10	1
http://www.merckmanuals.com	39	20
https://www.verywell.com	40	20
http://www.netdoctor.co.uk	40	22
http://www.lillydiabetes.com	8	1
http://www.healingwell.com	21	16
http://www.drmirkin.com	15	6
http://choosehealth.utah.gov	8	4
http://www.health.govt.nz	8	0
http://muschealth.staywellsolutionsonline.com	30	23
http://www.drwhitaker.com	18	8
http://www.consumerreports.org	26	15
http://umm.edu/health/medical	33	19
http://dm2.newlifeoutlook.com	14	7
http://www.emedicinehealth.com	21	19
http://asweetlife.org	37	19

APPENDIX C

SELECTED TERMS AND WEIRDNESS SCORES

Table 21: All selected terms and weirdness scores

Term	ADA frequency	iWeb frequency	Score
min week	3	0	403846154
encounter bmi	2	0	403846154
aglucosidase	1	0	403846154
aglucosidase inhibitors	1	0	403846154
analogs people	1	0	403846154
bromocriptine pramlintide	1	0	403846154
catabolic features	1	0	403846154
classes sulfonylurea	1	0	403846154
colesevelam bromocriptine	1	0	403846154
glp1 analog	1	0	403846154
goals insulin	1	0	403846154
goals metformin	1	0	403846154
goals providers	1	0	403846154
inhibitor glp1	1	0	403846154
inhibitors colesevelam	1	0	403846154
mmol mol	1	0	403846154
noninsulin agents	1	0	403846154
nonsurgical subjects	1	0	403846154
ongoing glycemia	1	0	403846154
prandial use	1	0	403846154
regimen flexibility	1	0	403846154
secretagogues meglitinides	1	0	403846154
sulfonylurea dpp4	1	0	403846154
sulfonylureas dpp4	1	0	403846154
aglucosidase inhibitors colesevelam	1	0	403846154
behavioral lifestyle intervention	1	0	403846154
behavioral lifestyle interventions	2	0	403846154
biphasic insulin analogs	1	0	403846154
catabolic features weight	1	0	403846154
classes sulfonylurea dpp4	1	0	403846154
colesevelam bromocriptine pramlintide	1	0	403846154
combination injectable therapy	1	0	403846154
comprehensive weight maintenance	2	0	403846154
diet physical activity	5	0	403846154
doses adjunctive use	1	0	403846154
dpp4 inhibitor glp1	1	0	403846154
drug classes sulfonylurea	1	0	403846154
effectiveness study grade	1	0	403846154
extreme dietary energy	1	0	403846154
glycemic goals insulin	1	0	403846154
glycemic goals metformin	1	0	403846154
glycemic goals providers	1	0	403846154
goals insulin therapy	1	0	403846154

goals metformin monotherapy	1	0	403846154
high intensity lifestyle	2	0	403846154
inhibitor glp1 analog	1	0	403846154
inhibitors colesevelam bromocriptine	1	0	403846154
insulin analogs people	1	0	403846154
intensity lifestyle interventions	2	0	403846154
intensive behavioral lifestyle	3	0	403846154
irregular meal schedules	1	0	403846154
kcal day energy	3	0	403846154
lifelong lifestyle support	1	0	403846154
limited dosing range	1	0	403846154
longstanding evidence base	1	0	403846154
loss such programs	1	0	403846154
loss sustained weight	1	0	403846154
min week minimum	1	0	403846154
month high intensity	2	0	403846154
necessary energy deficit	1	0	403846154
ongoing glycemia reduction	1	0	403846154
other medical psychosocial	1	0	403846154
patient encounter bmi	2	0	403846154
rapid acting secretagogues	1	0	403846154
routine patient encounter	1	0	403846154
serial lung function	1	0	403846154
sulfonylurea dpp4 inhibitor	1	0	403846154
term comprehensive weight	1	0	403846154
year comprehensive weight	2	0	403846154
day energy deficit	3	1	403846154
least monthly contact	2	1	269230769
acceptable fasting blood	1	1	134615385
loss maintenance programs	1	1	134615385
lower gfr range	1	1	134615385
weight loss ketosis	1	1	134615385
weight maintenance counseling	1	1	134615385
bolus insulin dosing	1	2	67307692
cost effectiveness models	1	2	67307692
late postprandial hypoglycemia	1	2	67307692
moderate glycemic effect	1	2	67307692
suboptimal blood glucose	1	2	67307692
total meal replacements	2	5	53846154
features weight loss	1	3	44871795
longer term mortality	1	3	44871795
complex insulin regimens	1	4	33653846
glp1 receptor agonists	1	4	33653846
body weight weekly	2	8	33653846
glycemia reduction approaches	1	5	26923077
weight maintenance programs	1	6	22435897
insufficient weight loss	1	7	19230769
kcal day	7	3	17543860
basal insulin analogs	1	8	16826923
postprandial glucose excursions	2	16	16826923
close medical monitoring	2	17	15837104
ml min	2	1	15037594
dietary energy restriction	1	9	14957265
greater short term	1	12	11217949
initiate insulin therapy	1	12	11217949

medical care settings	2	25	10769231
health economic outcomes	1	13	10355030
filtration rate gfr	1	15	8974359
goals long term	1	16	8413462
loss ketosis	1	1	7518797
medical psychosocial	1	1	7518797
metabolic surgeries	1	1	7518797
serial lung	1	1	7518797
u200 degludec	1	1	7518797
u300 glargine	1	1	7518797
glucose lowering medications	1	24	5608974
comparative effectiveness study	1	26	5177515
intensity lifestyle	2	3	5012531
physical activity goal	1	27	4985755
patient centered approach	1	34	3959276
mg dl	1	2	3759398
mmol l	1	2	3759398
major drug classes	1	37	3638254
short term weight	2	87	3094607
baseline body weight	1	46	2926421
weight loss counseling	1	47	2864157
comprehensive weight management	1	53	2539913
acceptable fasting	1	3	2506266
limited dosing	1	3	2506266
maintenance counseling	1	3	2506266
weight loss medications	2	118	2281617
behavioral lifestyle	3	11	2050581
diabetes providers	1	4	1879699
weight loss treatments	1	74	1819127
comprehensive weight loss	1	91	1479290
glycemia reduction	1	6	1253133
lispro mix	1	6	1253133
lung function testing	1	109	1235004
long term effectiveness	1	111	1212751
optimal medical therapy	1	111	1212751
low calorie diets	3	437	924133
complex insulin	1	9	835422
study grade	1	9	835422
weight loss maintenance	1	165	815851
gfr range	1	10	751880
concentrated preparations	1	11	683527
high carbohydrate foods	1	199	676459
term weight loss	2	424	634978
effectiveness models	1	12	626566
glp1 receptor	1	12	626566
longstanding evidence	1	12	626566
trained interventionists	1	14	537057
rate gfr	1	15	501253
suboptimal blood	1	15	501253
trained interventionist	1	15	501253
glycemic goals	3	49	460335
dpp4 inhibitor	1	18	417711
pharmacodynamic profiles	1	18	417711
sustained weight loss	1	353	381347
lifestyle efforts	1	20	375940

lower gfr	1	20	375940
lifelong lifestyle	1	21	358038
initiate insulin	1	22	341763
ongoing procedures	1	22	341763
tolerability issues	2	49	306890
glucose excursions	2	52	289184
weight weekly	2	52	289184
term mortality	1	27	278474
biphasic insulin	1	30	250627
u200	1	31	242542
features weight	1	35	214823
injectable therapy	1	36	208855
weight management program	1	658	204583
modest efficacy	1	37	203211
irregular meal	1	40	187970
glycemic effect	1	42	179019
extreme dietary	1	43	174856
moderate glycemic	1	44	170882
postprandial hypoglycemia	1	44	170882
noninsulin	1	45	167084
comprehensive weight	4	184	163452
u300	1	47	159974
metformin monotherapy	1	48	156642
dpp4 inhibitors	1	49	153445
successful weight loss	1	908	148255
effective weight loss	1	942	142904
glp1	2	112	134264
total meal	2	124	121271
meal schedules	1	68	110571
part mix	1	73	102997
glucose toxicity	1	77	97647
loss counseling	1	77	97647
glomerular filtration rate	1	1420	94800
blood glucose control	1	1448	92966
day energy	3	255	88456
pharmacological therapy	3	258	87428
adjunctive use	1	88	85441
possible medications	1	91	82624
insulin regimens	1	93	80847
long term benefits	1	1673	80463
weight loss goals	2	3412	78907
diet physical	5	478	78649
counseling setting	1	97	77513
reduction approaches	1	102	73714
insulin analogs	2	208	72296
baseline body	1	104	72296
monthly contact	2	220	68353
diet choice	1	111	67737
dosing range	1	111	67737
loss medications	2	228	65954
longer studies	1	116	64817
trained practitioners	2	245	61378
postprandial glucose	2	255	58971
term effectiveness	1	128	58741
higher bmis	1	130	57837

effectiveness study	1	131	57395
metabolic profiles	1	135	55695
retrospective analyses	1	137	54882
insufficient weight	1	139	54092
energy deficit	4	598	50293
delay progression	1	158	47587
blood glucose level	1	2839	47416
alternative medications	2	325	46270
lowering medications	1	163	46128
lifestyle support	1	164	45846
energy restriction	1	169	44490
loss maintenance	1	175	42965
meglitinides	1	177	42479
routine patient	1	178	42240
colesevelam	1	185	40642
lifestyle interventions	4	779	38607
pramlintide	1	200	37594
calorie diets	3	606	37222
degludec	1	202	37222
insulin dosing	1	204	36857
basal insulin	5	1042	36079
bolus insulin	1	215	34971
rapid acting	1	229	32833
behavioral strategies	2	461	32620
lifestyle counseling	1	232	32409
prandial	1	241	31198
accurate dosing	1	248	30318
activity goal	1	251	29955
glucose lowering	1	257	29256
obesity management	1	262	28698
term weight	2	554	27144
combination regimen	1	278	27046
secretagogues	1	278	27046
csii	1	280	26853
insulin administration	1	286	26289
lifestyle programs	1	295	25487
newer agents	1	305	24652
dietary energy	1	309	24333
cause mortality	1	318	23644
modeling studies	1	341	22049
patient encounter	2	694	21668
sglt2 inhibitors	1	349	21544
medical monitoring	3	1063	21220
centered approach	1	369	20376
obese patients	6	2329	19370
pharmacological treatments	1	400	18797
sustained weight	1	403	18657
concomitant medications	1	405	18565
glycemia	1	433	17364
dpp4	2	922	16310
insulin therapy	3	1538	14666
pharmacological agents	1	527	14267
smaller volumes	1	534	14080
lifestyle intervention	1	541	13898
lispro	1	546	13771

thiazolidinediones	1	553	13596
sulfonylureas	2	1115	13487
behavioral counseling	1	561	13402
necessary energy	1	567	13261
meal replacements	2	1149	13088
sulfonylurea	2	1185	12690
postprandial	3	1822	12380
insulin doses	1	608	12366
progressive nature	1	610	12326
receptor agonists	1	632	11897
consistent evidence	1	716	10501
severe obesity	1	720	10443
severe hypoglycemia	1	760	9893
weight maintenance	2	1543	9746
gastric banding	1	809	9294
week minimum	1	817	9203
drug classes	1	846	8887
loss treatments	1	846	8887
reproductive years	1	863	8712
kcal	7	6104	8622
sglt2	1	893	8420
such interventions	1	914	8226
maintenance programs	2	1910	7873
carbohydrate foods	1	1040	7230
bromocriptine	1	1076	6988
successful weight	1	1155	6510
better absorption	1	1156	6504
comparative effectiveness	1	1209	6219
bmis	1	1225	6138
interventionists	1	1228	6123
bariatric surgery	6	7445	6059
ongoing monitoring	2	2501	6013
treatment approaches	2	2588	5811
small studies	1	1305	5762
glargine	1	1330	5653
mmol	2	2689	5592
effective weight	1	1354	5553
lifestyle modifications	1	1389	5413
pharmacodynamic	1	1462	5143
economic outcomes	1	1518	4953
biphasic	1	1525	4930
major drug	1	1552	4845
pioglitazone	1	1555	4835
filtration rate	1	1609	4673
calorie diet	2	3302	4554
observational data	1	1788	4205
high carbohydrate	1	1852	4060
loss goals	2	3753	4007
adjuncts	2	4021	3740
glomerular filtration	1	2020	3722
patient record	1	2024	3715
tolerability	2	4207	3574
term benefits	1	2180	3449
bariatric	7	15399	3418
cohort studies	1	2454	3064

gfr	2	4976	3022
analogs	2	5049	2978
pharmacological	5	12664	2969
glycemic control	1	2602	2890
low calorie	3	8557	2636
nph	1	2982	2521
intervention strategies	1	3006	2501
reliable method	1	3028	2483
suboptimal	2	6060	2481
glucose control	1	3036	2477
cardiovascular events	1	3063	2455
medical therapy	1	3184	2361
adjunctive	1	3434	2190
glycemic	5	17880	2103
behavioral therapy	2	7488	2008
such patients	1	3766	1996
nonsurgical	1	3890	1933
product label	1	3946	1905
hypoglycemia	4	15885	1893
such studies	1	3981	1889
prediabetes	1	4039	1862
longer duration	1	4109	1830
interventionist	1	4123	1824
care settings	2	8378	1795
glomerular	1	4430	1697
metformin	4	17735	1696
glucose level	1	4484	1677
group sessions	1	4502	1670
specific situations	1	4591	1638
a1c	2	9184	1637
such programs	2	9245	1627
comorbidities	1	4873	1543
basal	6	31488	1433
catabolic	1	5353	1405
hyperglycemia	1	5458	1378
evidence base	1	5492	1369
cost saving	1	5668	1327
monotherapy	1	5905	1273
potential risks	2	12140	1239
cigarette smoking	1	6150	1223
function testing	1	6155	1222
lung function	1	6155	1222
low value	1	6190	1215
cost effectiveness	1	6200	1213
agonists	1	6321	1189
other agents	1	6937	1084
potential benefits	2	14199	1059
bolus	1	7133	1054
high intensity	3	22135	1019
higher doses	1	7944	946
obese	7	60200	874
high fat	1	8968	838
comparator	1	8989	836
weight loss	21	190209	830
bmi	5	45314	830

contraindications	1	9881	761
ketosis	1	10067	747
physical activity	8	80562	747
insulin	17	182560	700
weight management	1	11547	651
mol	1	11975	628
medical record	1	12232	615
alcohol use	1	12308	611
injectable	1	12481	602
dosing	3	38975	579
health status	1	13622	552
m2	3	41688	541
body weight	4	58040	518
concomitant	1	16181	465
regimens	1	16355	460
inhibitors	3	49284	458
substance use	1	16556	454
psychosocial	1	16814	447
calorie	5	85711	439
most patients	1	17226	436
overweight	5	88120	427
banding	1	17848	421
lifestyle changes	1	19978	376
blood glucose	2	40306	373
interventions	6	124061	364
behavioral	8	168226	358
diabetes	19	400701	357
regimen	3	64331	351
recent studies	1	21493	350
small intestine	1	21516	349
glucose	7	153896	342
other drugs	1	23121	325
management program	1	23756	317
observational	1	24707	304
excursions	2	53337	282
diets	4	109002	276
short term	4	113260	266
high cost	1	28646	262
medications	10	292698	257
lower cost	1	30428	247
replacements	2	63214	238
efficacy	3	94830	238
contraception	1	31770	237
longer term	1	32455	232
longstanding	1	33503	224
deficit	4	135940	221
cardiovascular disease	1	34663	217
medical care	2	74865	201
retrospective	1	37865	199
min	5	190316	198
dl	1	38168	197
gastric	1	40374	186
inhibitor	1	40997	183
lifestyle	12	503829	179
intolerance	1	44597	169

obesity	3	135890	166
high levels	2	94525	159
metabolic	2	96771	155
carbohydrate	1	51289	147
intestine	1	51339	146
dehydration	1	53739	140
weight gain	1	57548	131
mortality	2	115870	130
counseling	4	235230	128
surgeries	1	59623	126
filtration	1	60786	124
initiation	1	61638	122
intensive	3	185725	121
weight	32	2037850	118
long term	5	318592	118
ml	2	128789	117
cohort	1	65368	115
therapy	11	720373	115
cardiovascular	2	133227	113
cost effective	1	68953	109
toxicity	1	69661	108
clinical trials	1	72286	104
comparative	1	72697	103
medication	4	292698	103
receptor	1	75898	99
readiness	1	77548	97
preparations	1	77684	97
providers	8	627000	96
doses	2	157910	95
feasible	1	80148	94
irregular	1	81058	93
intensity	3	246890	91
loss	21	1746951	90
practitioners	2	178655	84
reproductive	1	91678	82
patients	20	1905804	79
effectiveness	3	286320	79
nausea	1	97642	77
insufficient	1	98638	76
cigarette	1	98997	76
patient	13	1292474	76
analyses	1	103996	72
therapies	1	104753	72
monitoring	5	545905	69
initiate	1	109262	69
diet	8	875909	69
profiles	2	225599	67
lifelong	1	113442	66
restriction	1	113808	66
potent	1	115829	65
optimal	2	244962	61
baseline	1	122993	61
approaches	3	369105	61
absorption	1	124606	60
surgery	6	755201	60

ongoing	4	522216	58
newer	2	261645	57
intervention	2	267348	56
more complex	1	134191	56
agents	4	575483	52
analog	1	151405	50
modifications	1	154523	49
dietary	1	154916	49
modest	1	155435	48
modeling	1	155799	48
consistent	1	157184	48
encounter	2	321447	47
assumptions	1	161108	47
comprehensive	4	645034	47
term	10	1656359	45
knowledgeable	1	172545	44
inexpensive	1	173743	43
strategies	4	699627	43
schedules	1	177182	42
lung	1	179453	42
progression	1	182553	41
punishment	1	190763	39
sessions	3	574808	39
severe	3	580251	39
participation	2	388285	39
activity	8	1554302	39
treatments	2	389065	39
volumes	1	198522	38
risks	3	598176	38
surgical	1	203289	37
consumption	2	410299	37
physical	8	1710958	35
type	17	3643365	35
meal	3	658511	34
procedures	3	674507	33
serial	1	227037	33
progressive	1	234636	32
alternatives	1	235296	32
mg	1	238039	32
studies	6	1471404	31
preferences	1	252458	30
side effects	1	253406	30
moderate	1	262907	29
maintenance	3	835502	27
goals	5	1400725	27
neutral	1	281328	27
adjustment	1	282447	27
trials	1	300495	25
alternative	3	905787	25
medical	8	2421913	25
absence	1	309638	24
dependent	1	318027	24
beneficial	1	327145	23
duration	1	330074	23
acceptable	1	331463	23

smoking	1	345035	22
diagnosis	1	349572	22
stomach	1	349639	22
substance	1	350918	21
rapid	1	368527	20
effective	5	1851834	20
outcomes	1	383540	20
pregnant	1	385240	20
safety	5	1956886	19
dose	1	391490	19
delay	1	392132	19
latter	1	416594	18
foods	2	841972	18
flexibility	1	433377	17
subjects	1	435944	17
individuals	3	1338105	17
programs	5	2347209	16
reduction	1	469461	16
routine	1	471742	16
combination	2	963749	16
frequency	1	494056	15
benefits	4	2032659	15
administration	2	1016793	15
evidence	3	1542837	15
treatment	4	2064973	15
sections	1	517378	15
potential	4	2133678	14
drug	2	1080568	14
threat	1	542692	14
minimum	2	1085677	14
reliable	1	554935	14
label	1	556767	14
extreme	1	572968	13
mix	2	1169326	13
energy	5	2923774	13
settings	2	1232270	12
procedure	1	620639	12
risk	4	2484305	12
situations	1	634335	12
cost	5	3193709	12
alcohol	1	657937	11
choice	3	1992588	11
short	4	2743822	11
height	1	700543	11
low	4	2894692	10
accurate	1	733103	10
adults	1	743158	10
drugs	1	761131	10
focus	3	2285318	10
presence	1	810046	9
response	2	1628532	9
clinical	1	855608	9
failure	1	857306	9
fat	1	858221	9
months	4	3621205	8

high	8	7331043	8
month	3	2845117	8
grade	1	954612	8
symptoms	1	958828	8
blood	2	1935293	8
effect	2	1979464	8
volume	1	989962	8
efforts	1	998364	8
remains	1	1005575	7
present	2	2013482	7
helpful	1	1008347	7
control	4	4050121	7
cover	2	2031830	7
body	4	4212929	7
levels	2	2116194	7
expensive	1	1065605	7
classes	1	1073586	7
record	2	2150221	7
lower	2	2201112	7
individual	2	2256002	7
critical	1	1137706	7
gain	1	1152575	7
comfortable	1	1160116	6
advantage	1	1170485	6
models	1	1176331	6
women	3	3592830	6
smaller	1	1202254	6
total	2	2432167	6
complex	1	1238994	6
testing	1	1245387	6
status	1	1256658	6
target	1	1279199	6
higher	2	2571877	6
economic	1	1295544	6
balance	1	1297577	6
least	3	4039409	6

APPENDIX D

RESULTS OF THE FEATURES

Table 22: Model Performance Measures with textual features

	DISCERN			Coverage Assessment		
	Recall	**Precision**	**F1-measure**	**Recall**	**Precision**	**F1-measure**
Elastic net	0.62	0.76	0.68	0.63	1.00	0.77
Random forest	0.58	0.71	0.64	0.53	0.77	0.63
SVM	0.54	0.74	0.62	0.37	1.00	0.54

Table 23: Model Performance Measures, where the top 350 terms are used with elastic net

	Coverage Assessment			Hybrid Assessment		
	Recall	**Precision**	**F1-measure**	**Recall**	**Precision**	**F1-measure**
VM1	0.68	0.81	0.74	0.63	0.86	0.73
VM2	0.58	0.79	0.67	0.53	1.00	0.69
VM3	0.84	0.89	0.86	0.58	0.92	0.71

The results showed that high-quality websites were correctly classified using these features compared to the low-quality websites.

APPENDIX E

RANKS OF THE FEATURES

Table 24: The ranks of the textual features for Elastic net

	Mean	**Min**	**Max**	**Median**
Number of adjectives	*1.05±0.22*	*1*	*2*	*1*
Number of exclamatory sentences	*1.95±0.22*	*1*	*2*	*2*
Number of auxiliary verbs	*3.28±0.61*	*3*	*6*	*3*
Number of existential sentences	*3.85±0.57*	*3*	*6*	*4*
Lexical variety	*4.93±0.31*	*4*	*6*	*5*
Number of interrogative sentences	5.93±0.25	5	6	6
Number of imperative sentences	7.28±1.35	7	15	7
SMOG	7.93±0.25	7	8	8
Length of shortest section	9.18±0.47	8	11	9
Paragraph count	11.00±0.98	10	16	11
Flesch-Kincaid Grade Level	11.90±2.68	8	16	10
Section count	12.10±1.36	9	17	12
Sentence count	12.50±0.81	11	14	13
Word count	13.80±0.73	12	15	14
Character count	14.80±0.73	13	16	15
Short sentence rate	15.10±1.83	9	17	16
Number of images	15.20±1.57	10	17	16
Average paragraph length	15.70±0.80	13	17	16
Bullet count	15.70±0.81	13	17	16
Average section length	15.80±0.73	14	17	16
Number of uses of verb to be	15.80±0.73	14	17	16
Number of passive voice sentences	15.80±0.73	14	17	16
Number of nouns	15.80±0.73	14	17	16
Number of verbs	15.80±0.73	14	17	16
Number of sentences starting with a pronoun	15.80±0.73	14	17	16
Number of declarative sentences	15.80±0.73	14	17	16

APPENDIX F

PROFESSIONAL HEALTH GUIDELINES

MedlinePlus -- How to Write Easy-to-Read Health Materials
General Points
Now that you have learned about your audience, consider them carefully when writing. Cultural, age, and gender differences may have an impact on your content. For example, the writing style and graphics may be different for an HIV/AIDS brochure for teens than for adults over age 50.
Keep within a range of about a 7th or 8th grade reading level.
Grab your readers' attention from the very beginning. Make the first few sentences stand out by using compelling, clear language.
Structure the material logically. Start by stating the topic. Include the most important points about what you want people to learn. Then write the body of the material. Finish with a summary of the material and restate the key concepts.
Include specific actions users can take. Emphasize the benefits of taking action. For example, "Proper use of asthma inhalers helps you breathe better."
Examples and stories help engage readers.
Avoid making assumptions about people who read at a limited level. Don't talk down to the reader. Maintain an adult perspective while using easy-to-read language.
Write your draft, then review and edit several times. Remove words, sentences or paragraphs that do not help the reader understand the concepts. Materials that are too wordy will make them more difficult to understand.
Consider incorporating Web accessibility best practices. An accessible Web site helps people with reading and learning disabilities. For more information on Web accessibility, see the WebAIM (Web Accessibility in Mind) site from the Center for Persons with Disabilities at Utah State University.
Language and writing style
Find alternatives for complex words, medical jargon, abbreviations, and acronyms. At times there may not be an alternative or you may want to teach patients the terms because their healthcare providers will be using them. In those cases, teach the terms by explaining the concept first in plain language. Then introduce the new word(s). It is also helpful to provide a simple pronunciation guide. For example, "A normal heart beat starts in the upper right chamber of the heart, or atrium (ay-tree-yim)."
Avoid abstract language in giving instructions for actions. For example, instead of "Don't lift anything heavy," use "Don't lift anything heavier than a gallon of milk (about 10 pounds)."
Use varied sentence length to make the material interesting, but keep sentences simple. A general guideline is to limit most sentences to 10-15 words.
Use words like "you" instead of "the patient". A personal voice can help engage readers.
Where appropriate, use bulleted lists instead of blocks of text to make information more readable. Dense blocks of text can be difficult to read.
Use the active voice. Direct language helps people associate concepts with actions. Example: Active: Amanda used her inhaler today. Passive: The inhaler was used by Amanda today.
Be consistent with terms. For example, using "drugs" and "medications" interchangeably in

the same document can cause confusion.
When possible, say things in a positive way. For example, use "Eat less red meat" instead of "Don't eat lots of red meat."
Visual Presentation and Representation
Use colors that are appealing to your target audience. Be aware, however, that some people may have vision disabilities. For example, some people with color blindness cannot distinguish red from green.
Use illustrations and photos with concise captions. Keep captions close to images.
Avoid graphs and charts unless they are essential to convey the concepts. If you do use them, make sure they are simple and clear.
Balance the use of text, graphics, and clear or "white space". Try for about 40-50% white space.
Avoid italics of more than a few words at a time.
Use sans serif fonts such as Arial, Verdana, Frutiger, Tahoma, or Helvetica. San serif fonts are considered easy to read even at smaller sizes. Also see the font accessibility page from WebAIM.
Use bolded headings and subheadings to separate and highlight document sections.
Justify the left margin only. This means the left margin should be straight and the right margin should be "ragged."
Use column widths of about 30-50 characters long (including spaces).
Text on top of shaded backgrounds, photos, or patterns can be difficult to read and may also be inaccessible to many users with disabilities.

NHS – Toolkit for producing patient information
When writing information for patients, you should remember the following points.
Use everyday language. Avoid jargon and acronyms, and use plain language to make it easier to read. As many as seven million people (roughly one in five adults) in England have difficulties with basic literacy and numeracy, but that does not mean you have to be patronising or use childish language (source: Lord Moser's 1999 report 'A Fresh Start').
Use patient-friendly text. Use personal pronouns such as 'we' and 'you'. Do not use frightening language, for example, 'electrodes will be put on your chest'. If it is difficult to avoid using some medical terminology, such as 'nuclear medicine', give an explanation.
Be relevant to individual patients.
Information should be in context with other information given to patients, for example, letters, leaflets and appointments.
Reinforce the information that patients have been told at the clinic.
Avoid instructions. For example, do not just say 'do not eat anything for six hours before an operation' - explain why.
Be helpful. Help people make decisions by giving them facts about the risks, side effects and benefits.
Do not confuse people by covering several treatments and conditions in the same leaflet.
Tell people what other information, resources and supports are available.
Be up to date. Give the most recent practice and latest phone numbers.
Let people know if the information is available in other formats, for example, on audiotape.
To make text more inviting to read, use the following.
Short sentences – in general no more than 15 to 20 words long.
Lower-case letters, where possible, as they are easier to read. Exceptions to this are proper names and the first letter in a sentence.
Present and active tenses, where possible, for example, 'your appointment is on…' not 'your appointment has been made for…'

A question and answer format is helpful to divide up text.
Bulleted or numbered points to divide up complicated information.
Small blocks of text. Do not use long paragraphs - divide them up using headings and new paragraphs.
White space makes the information easier to read.
Large bold font emphasises text. Avoid UPPER CASE letters, italics and underlining as they make the text more difficult to read.
Numbers from one to nine are easier to read if they are written in words, and numbers from 10 can be represented as numbers.
A font size of no less than 12 point (see the print guidelines on page 9).
Diagrams and pictures are very effective and should be in line with our communication principles. Where appropriate, use them to illustrate the text, remember to label them and do not print over them. You should not use clip art as it does not add to the reputation of a professional organisation. Visit www.nhs.uk/syndication/photo.asp to use the NHS photo library.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Ölçer, Didem

Nationality: Turkish (TC)

Date and Place of Birth: 25 June 1987, Ankara

Marital Status: Married

EDUCATION

Degree	Institution	Year of Graduation
Ph.D.	Information Sytems, METU	2020
MS	Computer Engineering, Başkent University	2011
BS	Computer Engineering, Çankaya University	2009
High School	Çankaya High School	2004

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2010-Present	Başkent University, Computer Eng.	Research Assistant

PUBLICATIONS

1. **Ölçer D,** Taşkaya-Temizel T., Quality Assessment of Web-Based Information on Type 2 Diabetes, *Online Information Review*, 2020. (Submitted)
2. Erdaş Ç. B, **Ölçer D,** Evaluation of Feature Reduction and Classification Methods for Activity Recognition. *11th International Statistics Congress (ISC'2019)*: Muğla; 4/10/2019 - 8/10/2019

3. Erdaş Ç. B, **Ölçer D,** Human Identification with Deep LSTM by Using Biometric Walking Patterns. *11th International Statistics Congress (ISC'2019)*: Muğla; 4/10/2019 - 8/10/2019

4. **Ölçer D,** Açarı S, Sümer E. Human Motion Recognition from Video Images. *4th International Conference on Advanced Technology & Sciences (ICAT'Rome)*: İtaly-Roma; 23/11/2016 - 25/11/2016

5. **Ölçer D**, Oğul H. Combined analysis of sequence and time-series expression data can reveal cancer-related microRNA. *International Symposium on Health Informatics and Bioinformatics, HIBIT'13*: Ankara; 25/09/2013 - 27/09/2013

6. **Ölçer D,** Oğul H. Clustering microRNAs from sequence and time-series expression. *Fifth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies(BIOTECHNO'13)*: Lisbon; 24/03/2013 - 27/03/2013