12-2018

Exploring Identifiers of Research Articles Related to Food and Disease Using Artificial Intelligence

Marco Ross
Sheridan College

EXPLORING IDENTIFIERS OF RESEARCH ARTICLES RELATED TO FOOD AND DISEASE USING ARTIFICIAL INTELLIGENCE

A Thesis

Presented to

The Faculty of Applied Science and Technology, School of Applied Computing

of

Sheridan College, Institute of Technology and Advanced Learning

by

ROSS, MARCO

In partial fulfillment of requirements

for the degree of

Bachelor of Computer Science (Mobile Computing)

December, 2018

ABSTRACT

EXPLORING IDENTIFIERS OF RESEARCH ARTICLES RELATED TO FOOD AND DISEASE USING ARTIFICIAL INTELLIGENCE

Marco Ross
Sheridan College, 2018

Advisor:
Dr. El Sayed Mahmoud

Currently hundreds of studies in the literature have shown the link between food and reducing the risk of chronic diseases. This study investigates the use of natural language processing and artificial intelligence techniques in developing a classifier that is able to identify, extract and analyze food-health articles automatically. In particular, this research focusses on automatic identification of health articles pertinent to roles of food in lowering the risk of cardiovascular disease, type-2 diabetes and cancer as these three chronic diseases account for 60% of deaths (WHO, 2015). Three hundred food-health articles on that topic were analyzed to help identify a unique key (Identifier) for each set of publications. These keys were employed to construct a classifier that is capable of performing online search for identifying and extracting scientific articles in request. The classifier showed promising results to perform automatic analysis of food-health articles which in turn would help food professionals and researchers to carry out efficient literature search and analysis in a timelier fashion.

ACKNOWLEDGEMENTS

This research was supported by the Centre for Mobile Innovation (CMI) – Sheridan College, ON, Canada. through the work study program. The author gratefully acknowledges Dr. El Sayed Mahmoud for his efforts in guiding the research process and help in putting together this proposal; Dr. El-Sayed M. Abdelaal for supplying the training data for the thesis; We thank the thesis advisory committee members at Sheridan College for their useful feedback on the work; Dr. Edward Sykes, Dr. Aeiman Gadafi, Dr. Richard Pyne, and Dr. Yuchong Rachel Jiang.

TABLE OF CONTENTS

LIST OF TABLES

List of Figures

CHAPTER ONE

1. INTRODUCTION

1.1 The Problem Context

Health professionals in Canada rarely use the results of medical research to promote health and influence policy. This has been shown in a 2007 survey of Canadian health professionals, based on the answers of 928 professionals and managers from Canadian health service organizations. The survey results showed that 57% of the respondents frequently or very frequently received research results (Belkhodia, 2007). These received results never or rarely influenced the health professionals' decisions and choices in fourteen percent of the cases. Additionally, they were also never or rarely transformed into concrete applications in another eleven and half percent of the cases (MEDLINE Fact Sheet, 2018).

The main reason for these low uptake percentages can be attributed to the outright volume of medical research related to food-health being produced on a regular basis. Large numbers of scientific publications make selecting an article about a specific food and disease more difficult. For example, popular biomedical database MEDLINE, produced by the United States National Library of Medicine, contains over 24 million references to biomedical texts alone (MEDLINE Fact Sheet, 2018). Figure 1 demonstrates the rapid growth in biomedical literature from 1986 to 2005 as illustrated in (Reviews on Text Mining in Biomedicine, 2006).

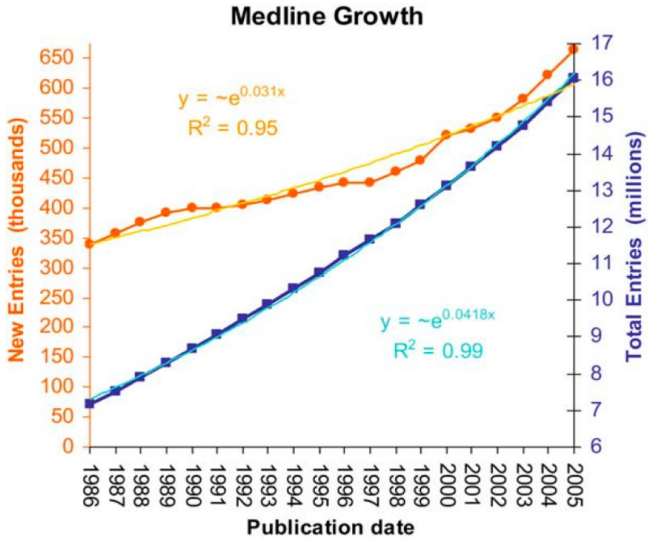


Figure 1. Growth in biomedical literature, 1986-2005

This graph illustrates a growth of approximately 1,800 articles *per day* throughout these nineteen years (Hunter & Cohen, 2006). Since 2005, the literature has seen consistent rise from 16 million articles to 25 million (Hunter & Cohen, 2006), marking a continued increase of approximately 1,800 articles per day from 2005 until present day.

The large number of articles increases the difficulty of finding appropriate papers for a given topic on food and health. It also delays extracting useful information from these articles and perhaps this useful information could be lost in the search process. Such information includes roles of food in health and nutrition, food intake, recommended foods for disease prevention, food protection mechanisms, etc. This information is essential for food or health policies, and food health, nutrient or function claim and food labelling. This research was intended to build an automatic article classifier in the area of roles of food in reducing the risk of three choric diseases: CVD, type-II diabetes and cancer. The study’s goal was to develop a text classification tool that is capable of performing efficient literature search and analysis in a timely fashion.

1.2 Terms and Definitions

Table 1. Terms and Definitions

Article	A peer-reviewed medical article whose subject matter is specifically related to nutrition *and* a particular health outcome relating to one of: CVD, diabetes mellitus, or cancer.
Cardiovascular disease (CVD)	A class of diseases which involve the heart or blood vessels.
Cancer	A disease caused by an uncontrolled division of abnormal cells in a part of the body.
Classifier	An algorithm which is built around a profile created from the use of n-grams. This profile is constructed through the careful analysis and interpretation of hundreds of *n-grams.*
Diabetes mellitus (type-II diabetes)	A disease in which the pancreas does not produce enough insulin, or the body does not properly use the insulin it makes.
Food-health research	Medical research specifically concerned with nutrition and health outcomes.
Health professional	A health practitioner or healthcare provider who provides preventive, curative, promotional, or rehabilitative healthcare services to people in a systematic way.

Health promotion	The process of enabling people to increase control over, and to improve, their health. It moves beyond a focus on individual behaviour towards a wide range of social and environmental interventions.
Health policy	Decisions, plans, and actions undertaken to achieve specific healthcare goals within a society.
Hypotheses	Testable statements that, if true, may explain an observed phenomenon.
Information extraction	The process of automatically assessing documents, data or knowledge bases to extract statements that are *likely* to be true given the available information. IE can be based on defined patterns, machine-learning techniques, statistical analyses or automated reasoning.
Knowledge bases	Databases of statements covering a knowledge domain. Often, statements are represented in a form that permits the automated or manual inference of statements that are not explicitly stated using inference rules.
Natural language processing (NLP)	A branch of computer science and artificial intelligence which is concerned with how computers understand, interpret, and manipulate human language.

Natural Language Toolkit (NLTK)	A leading platform for building Python programs to work with human language data.
N-gram	A contiguous sequence of *n* items from a given sample of text.
Nutrition assessment	An in-depth evaluation of data related to an individual’s food and nutrient intake, lifestyle, and medical history. This data is used to assess the nutritional status of that individual.
Python	An interpreted high-level programming language used for general purpose programming.
Research uptake	All the activities which facilitate and contribute to the use of research evidence by policy-makers, practitioners and other actors.

1.3 Problem Statement

Research that relates nutrition to health risks continues to grow at a consistently linear rate. This constant influx of new research makes it difficult for both health professionals and patients to find papers related to their interest which in turn limits using keep up with the latest research results. This work develops a classifier that identifies food-health articles related to the diseases: CVD, diabetes and cancer. This classifier enables health professionals to search for the articles relevant to their interest. It also could be used as a component in an automatic system for extracting useful information from particular scientific articles.

1.4 Purpose

The purpose of this thesis is to examine the potential for n-grams, a natural language processing technique, to identify food-health articles automatically. It aims to facilitate finding food-health articles related to the diseases: CVD, cancer and type-II diabetes. This helps in creating systems that extract useful information from food-health articles automatically which in turn increases the health professionals' uptake of medical research. The thesis employs n-grams to determine the subject matter of a food-health article, i.e. if an article is related to a particular food or nutrient *and* CVD, type-II diabetes, or cancer.

The ultimate goal of this research is to provide a smart search tool that supports automatic analysis of food-health articles.

1.5 Motivation

Improving health promotion and disease prevention through diets with the use of artificial intelligence techniques is the main motivation of this research. The use of artificial intelligence techniques in performing literature search and analysis should improve its efficiency. Currently, the link between diet and health promotion and disease prevention is well established with numerous amounts of publications. This requires techniques and tools to extract and analyze data. The current research should make a difference in the way we manage data, develop strategies and conduct research. The stakeholders of health promotion such as health professionals, dieticians, policy makers, and researchers should benefit from this tool. This research aligns with the Government of Canada's vision of the agriculture-food sector of Canada being among the top five competitors in the agri-food sector, being recognized as the most trusted supplier of safe, sustainable, high-quality agri-food products (Report of Canada's Economic Strategy Tables: Agri-Food, 2018). This is supported by the open data principle adapted by the Government of Canada which is the practice of making machine-readable data freely available, easy to access, and simple to reuse (Open Data 101, 2017). This work makes use of this open data practice by using data partially provided by the Department of Agriculture and Agri-Food of the Government of Canada.

1.6 Proposed Work

This work consists of two main phases. The initial phase involves determining whether or not n-grams can be used to accurately determine the subject matter of a scientific article. This phase specifically focuses on assessing the utility of using n-grams to differentiate among the food-health articles related to cardiovascular disease, type-II diabetes, or cancer. The second phase of this work was determining what *sequence* of n-grams are the most accurate and most effective in determining the subject matter of these food-health articles. There are several different options for the size of the n-grams, and this thesis tested four different n-gram sizes These sizes are unigrams (n=1), bi-grams (n=2), trigrams (n=3), and quadrigrams (n=4). These four n-gram sizes were chosen as they are the de facto standard when using n-grams; while some researchers have used (n=5), we did not see value in include (n=5) given the disappointing results of (n=4).

Consider the following popular Shakespearean phrase from *The Merchant of Venice* and consider its individual words: "All that glitters is not gold". If this phrase was broken down into n-gram sequences, it would result in the n-grams found in Table 2.

Table 2. Example of resultant n-grams using various sequences

1-gram	2-gram	3-gram	4-gram	5-gram
All, that, glitters, is, not, gold	All that, that glitters, glitters is, is not, not gold	All that glitters, that glitters is, glitters is not, is not gold	All that glitters is, that glitters is not, glitters is not gold	All that glitters is not, that glitters is not gold

Table 2 shows that using different size of grams drastically changes the meaning of the phrase. There is a great degree of variance between the resulting values provided by the various sequence sizes and therefore all sequences must be examined in order to gain a clear and accurate understanding of the phrase. Using larger sequence sizes provides greater context to the phrase and allows the reader to better understand the affective meaning of the phrase; however, using a larger size of gram results in having fewer overall grams. This would become a problem when trying to build an accurate classifier for several hundred different medical articles as it was unlikely that the results of higher sequence grams will be commonly found throughout those articles. For example, it is much more likely that medical articles will share common phrases such as 'risk of diabetes' (n=3) than 'sugar and risk of diabetes' (n=5). For this reason, the different sizes of grams must all be considered when building a profile as well as when comparing that trained profile to external articles for testing.

1.7 Thesis Statement

An n-gram algorithm can be developed that is able to automatically identify food-health articles relevant to the three proposed diseases. The most frequent n-gram terms in a food-health article show topics related to the disease targeted by the article. These frequent terms could be used as a unique identifier for the article group related to the targeted disease. This research aims to determine the size of each term and the set of frequent terms that accurately identify the article-targeted disease

1.8 Contributions

This work showed how to use n-grams for building a classifier to recognize food-health articles related to the diseases: CVD, diabetes and cancer. The contributions of this work include:

- Developing a customized converter from different data types to text
- Identifying the n-gram size that enables automatic recognition of a food-health article related to a specific disease.
- Identifying three different lists of n-grams to differentiate among food-health articles of the disease CVD, diabetes and cancer automatically. Each n-grams list is considered a signature for a particular type of food-health articles.
- Developing a measure that quantifies the similarity among food health documents

1.9 Organization of Thesis

The remainder of this thesis consists of a literature review, methodology and results. The literature review focuses on prior research conducted in the fields of text mining, food-healthiness knowledge extraction, and natural language processing (NLP). It will examine the recent literature in these areas as well as full scale surveys and reviews of the general field of food-health and NLP. The methodology section describes the details of methodologies involved in the work. This includes selecting the training data, identifying n-gram sequence sizes, building the article classifier, testing the classifier and performance metrics used. The result section will highlight the experimental findings including the analysis of these findings and potential future research.

Chapter Two

2. Literature Review

The explosive growth of food-health literature has prompted increasing interest in using text mining techniques to address the information overload faced by domain experts. This is reflected by the conception of articles reviewing this work (Reviews on Text Mining in Biomedicine, 2006) (Ananiadou & McNaught, 2006), which target experts in biosciences as their primary audience (Cohen & Hersh, 2005).

The recent proliferation of articles reviewing using text mining for medical applications includes electronic medical records knowledge extraction, epidemic detection through semantic analysis of social media, abbreviations in biomedical text, automatic terminology management in biomedicine, as well as automatic scientific literature analysis as a tool for novel findings and hypotheses from research (Cohen & Hersh, 2005) (Hirschman, et al., 2012) (Rebholz-Schuhmann, Oellrich, & Hoehndorf, 2012). The category of automatic scientific literature analysis as a tool for novel findings from research is of most relevance to this work and will be the focus of this study.

Automated analysis of scientific literature complements the reading of scientific literature by individual researchers because it allows quick access to information contained in large volumes of documents (Hirschman, et al., 2012). Hirschman et al. hypothesize that in the future, it is likely that solutions will be developed that produce and test hypotheses against knowledge bases. This type of solution development in the field of bioinformatics relies heavily on researchers having rapid access to a large corpus of literature readily available which may be automatically analyzed and interpreted (Hirschman, et al., 2012).

The growth of food-health literature has prompted increasing interest in using text mining techniques to address the information overload faced by domain experts. This is reflected by the conception of articles reviewing this work (Reviews on Text Mining in Biomedicine, 2006) (Ananiadou & McNaught, 2006), which target experts in biosciences as their primary audience (Cohen & Hersh, 2005).

A survey of work in biomedical text mining conducted by Cohen and Hersh in 2005 hypothesized that the biggest challenge to biomedical text mining in the coming 5-10 years would be building systems which are useful to researchers (Cohen & Hersh, 2005). A literature review by Rebholz-Schuhmann et al. builds on this hypothesis by suggesting future work in this field should be focused on helping researchers in problem solving of specific real-world scenarios. Figure 2 contains a modified version of a diagram made by Rebholz-Schuhmann et al. which shows the different categories where text mining can help scientific researchers, using food-health relationships as an example (Rebholz-Schuhmann, Oellrich, & Hoehndorf, 2012). Figure 2 distinguishes four primary stages in text-mining solutions: information retrieval, information extraction, building knowledge bases and knowledge discovery (Rebholz-Schuhmann, Oellrich, & Hoehndorf, 2012). Information retrieval could involve a user submitting a query to a search engine and receiving a document fitting to their submitted query in return.

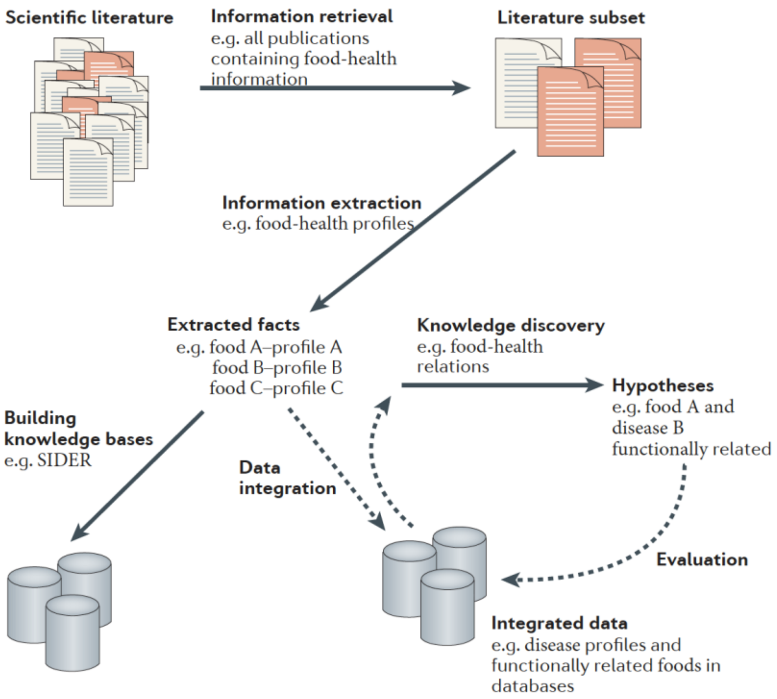


Figure 2. Categories of text mining solutions

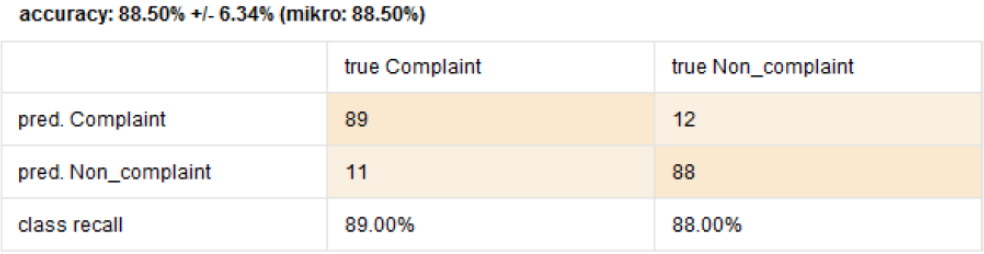
Information extraction involves the identification of entities, such as diseases or foods, as well as the identification of complex relationships between these entities (Rebholz-Schuhmann, Oellrich, & Hoehndorf, 2012). Scientific facts extracted from literature may be used for the purposes of populating databases or data curation. From these extractions, knowledge bases can be built that contain the collected statements together with collected evidence in the form of references to the literature (Hirschman, et al., 2012) (Rebholz-Schuhmann, Oellrich, & Hoehndorf, 2012). Knowledge discovery involves identifying undiscovered or hidden knowledge by applying data-mining algorithms to the collection of facts gathered from the literature. From here, text-mining results may be used to suggest new hypotheses automatically which can be used to either validate or disprove existing hypotheses or to help direct future research (Rebholz-Schuhmann, Oellrich, & Hoehndorf, 2012).

This work automatically identifies whether a given medical article is related to food and CVD, type-II diabetes, or cancer, and therefore the category of text mining is most similar to information extraction. This research assumes that databases already exist from which users can query. The ultimate goal of this work was to develop a tool that is able to automatically identify food-health articles relevant to the three proposed diseases which facilitates extracting useful information from medical literature and building knowledge bases.

2.1 N-Grams in Analyzing Text

N-grams have been used for classifying text in a variety of different domains and industries for the past decades including tourism, malicious code detection, speech corpus generation, automatic evaluation of text summaries, and automatic bug reporting classification (Beebe, Maddox, & Liu, 2013). They have proven to be a simple and accurate method for classifying various types of texts and they are known for being language independent (Chumwatana & Chuaychoo, 2016).

Chumwatana and Chuaychoo (Chumwatana & Chuaychoo, 2016) studied the potential for n-grams to automatically classify emails sent to businesses belonging to the tourism industry in Thailand in order to classify them into two categories of: complaint and non-complaint emails. They used 200 emails for training and employing the 3-gram method, they were able to accurately detect whether an email was a complaint or not in 88.50% of instances. These results can be seen in Figure 3 (Chumwatana & Chuaychoo, 2016).

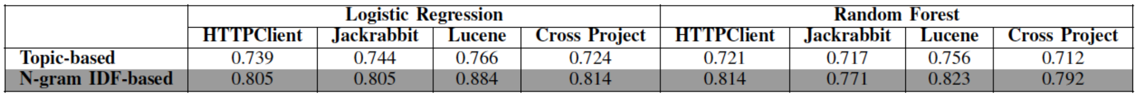


accuracy: 88.50% +/- 6.34% (mikro: 88.50%)

	true Complaint	true Non_complaint
pred. Complaint	89	12
pred. Non_complaint	11	88
class recall	89.00%	88.00%

Figure 3. Accuracy performance of prediction in training process

Further research by Terdchanakulet al. in 2017 studied the accuracy of n-grams in classifying bug reports due to the time-consuming nature of software companies manually classifying bug reports (Terdchanakul, Hata, Phannachitta, & Matsumoto, 2017). According to their research data, they were able to achieve better results using n-grams over a topic modelling based-approach to text classification for over 11,000 reports (Terdchanakul, Hata, Phannachitta, & Matsumoto, 2017). Figure 4 shows the F-measure values comparing the n-gram results to the topic-based results, illustrating that the F-measures for the n-grams were more accurate in every single category of bug reporting studied when using n > 1 for their n-grams (Terdchanakul, Hata, Phannachitta, & Matsumoto, 2017).



	Logistic Regression				Random Forest			
	HTTPClient	**Jackrabbit**	**Lucene**	**Cross Project**	**HTTPClient**	**Jackrabbit**	**Lucene**	**Cross Project**
Topic-based	0.739	0.744	0.766	0.724	0.721	0.717	0.756	0.712
N-gram IDF-based	0.805	0.805	0.884	0.814	0.814	0.771	0.823	0.792

Figure 4. F-scores of n-gram based vs topic-based training classifications

Abou-Assaleh et al. used single character n-grams to classify benign and malicious code in order to detect viruses on a computer before they are able to cause damage to the machine. They were able to achieve an incredible 100% accuracy on their training data of sixty-five distinct Windows executable files (25 with malignant code and 40 with benign code). Using 3-fold cross validation, which is the process of dividing the data into three distinct sets and using two sets for training and one for testing repeated three times, resulted in 98% accuracy of malignant code detection (Abou-Assaleh, Cercone, Keselj, & Sweidan, 2004). N-grams have had successful widespread use in a variety of domains for the purposes of text classification and therefore were chosen as the approach for this study.

2.2 Approaches

Research in using natural language processing for healthcare has used various approaches including bag-of-words, rule-based classification, text corpuses created from domain-specific sources, binary classification, named entity classification, and sometimes a combination of two or more of these methods.

Wiegand et al. in 2013 set out to detect reliable statements about food-health relationships from natural language text based on a specially annotated web corpus from forum entries discussing the healthiness of certain food items on food-related websites (Wiegand & Klakow, Towards Contextual Healthiness Classification of Food Items - A Linguistic Approach, 2013)**.** They categorized food items from the forum posts of 418,558 web pages as either *suitable, beneficial, preventative, unsuitable, causation* (negative impact) and also categorized the posts as reliable or unreliable based on whether a citation was provided and based on the authority (e.g. a physician) from which the information came. Results of their study showed that a mixture of bag-of-words and high-level classifiers provided the best F-score results when measuring accuracy of the trained data set, as illustrated in Figure 5 (Wiegand & Klakow, Towards Contextual Healthiness Classification of Food Items - A Linguistic Approach, 2013).

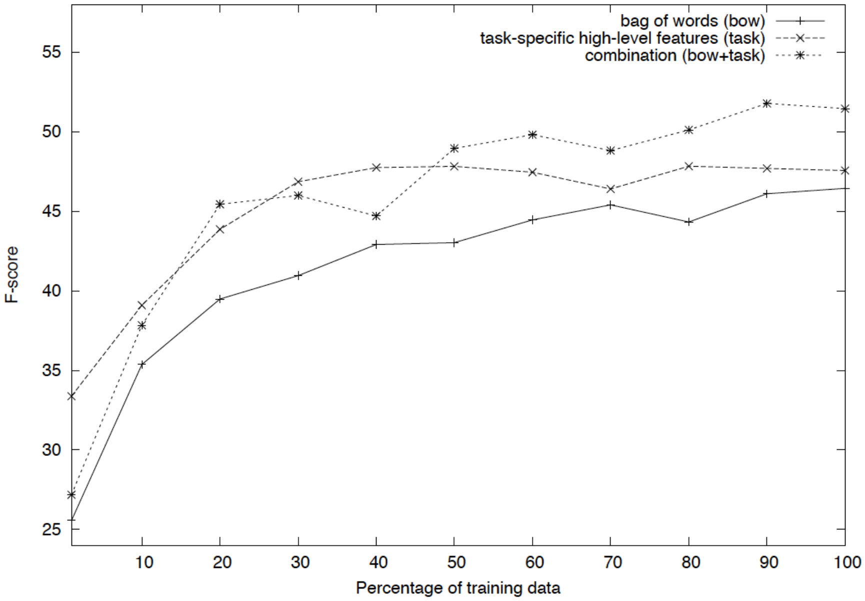


Figure 5. F-score comparison of bag-of-words and task-specific features

Wiegand et al. found that strong polar expressions and intensifiers are fairly predictive and complement bag-of-words (Wiegand & Klakow, 2013).

A more recent study by Wiegand et al. in 2015 expanded on their previous research, this time exploring the feasibility of extracting suitable and unsuitable food items for particular health conditions from natural language text (Wiegand, 2015). This was done by classifying certain foods into the categories of "suitable" and "non-suitable" based on online forum entries on food-related websites (Wiegand & Klakow, 2015). Their approach included using a bag-of-words feature set which they found to be more effective at producing a classifier than simply using a healthiness lexicon (Wiegand & Klakow, 2015).

BioCaster is a text mining system for detecting and tracking the distribution of infectious disease outbreaks from linguistic signals on the web (Collier, Doan, Kawazoe, & Matsuda Goodwin, 2008). BioCaster uses topic classification as well as named entity recognition (NER) on data analyzed from over 1700 RSS feeds, which it then plots on a Google map using geocoded information from the RSS feeds. BioCaster was able to achieve an 94.8% F-score accuracy based on naive Bayes raw text, n-grams, semantic tag-based features, and feature selection, outperforming the NER system using SVM which achieved an F-score of 76.6% (Collier, Doan, Kawazoe, & Matsuda Goodwin, 2008).

2.3 Metrics

The most commonly used metrics in determining the effectiveness of a natural language processing method include *precision, recall, F-score, Receiver operating characteristic (ROC) curve, and multiclass classification*.

Recall is often used in conjunction with precision and F-score to measure relevance. With precision, recall, and F-score, four main classifications of retrieval are used to determine relevance:

1. True positive (TP): correctly retrieving a relevant instance
2. False positive (FP): incorrectly retrieving a relevant instance
3. True negative (TN): correctly *not* retrieving an instance when a condition is not met
4. False negative (FN): incorrectly *not* retrieving a relevant instance when a condition is met

Recall measures the fraction of relevant instances which have been retrieved over the total amount *of all relevant instances.* For example, if someone were to ask you to recall how many times you have gone for dinner at a particular restaurant which you have visited *ten* times, and you recall having gone there *seven* times, then your recall rate can be said to be 0.7 (70%). This can be seen in the following formula:

> *how many you recall*: (7) ÷ *how many you recall additional to how many you missed*: (7 + 3) = 0.70. Recall can therefore be measured as *TP ÷ (TP + FN).*

Precision measures the fraction of relevant instances among the *retrieved instances*. Continuing on the same example from above, if you were to recall having visited that same restaurant *fifteen* times, your precision can be said to be 0.6667 (66.67%). This can be represented as:

> *how many relevant instances exist* (10) ÷ *number of relevant instances additional to the number of incorrectly recalled instances* (10 + 5).

Recognizing this, the formula for precision may be represented as *TP ÷ (TP + FP)* (Powers, 2011).

Sometimes, an algorithm may have high recall but low precision, or even high precision but low recall. For this reason, the F-score exists which is a mean of the two metrics. F-score may be represented in a formula as (TP + TN) ÷ (TP + TN + FP + FN). Receiver operating characteristic (ROC) is another metric used in data mining which plots the true positives against the false positives at different thresholds (Hanley & McNeil, 1982). ROC determines the probability that a randomly selected instances will be correctly identified with a higher degree of certainty over a randomly selected *non-instance* (Hanley & McNeil, 1982).

We have examined the role of n-grams in natural language processing and text mining, we have explored several different approaches to text mining used in the field of healthcare and have studied some of the most commonly used metrics to measure the performance of these approaches.

CHAPTER THREE

3. METHODOLOGY

This chapter introduces the proposed research methods in details. This includes the steps for building the article classifier, metrics, time and space complexity of the algorithm, and the testing strategy.

3.1 Proposed Classifier

The steps for building the proposed classifier include: (1) creating n-gram lists with various n-gram sizes for each disease category of CVD, type-II diabetes, and cancer (2) determining the effective list of most frequent n-grams in each category (3) identifying the effective n-gram size for detecting the subject of an article. This process is illustrated in Figure 6.

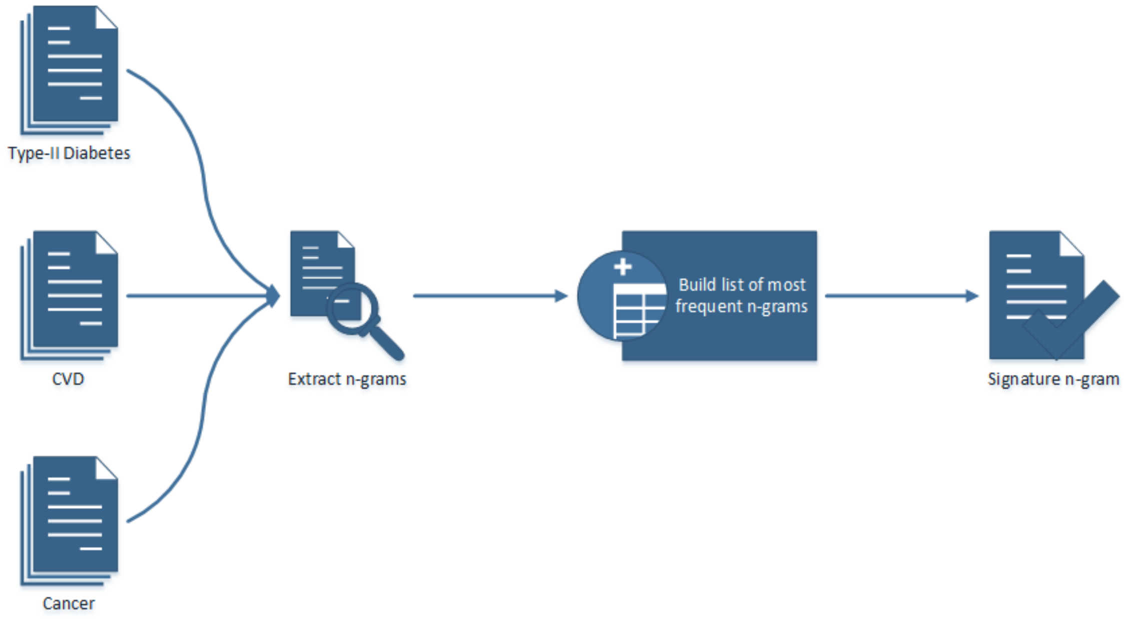


Figure 6. N-gram Classifier Process

For the sake of simplicity, the diagram in Figure 6 shows the process as if *one n-gram classifier* was created from all three disease categories, when in fact three *separate* classifiers were created using this same process.

3.1.1 Creating a Classifier

In order to create an accurate food-health article classifier for each disease category, one-hundred peer-reviewed articles which relate to a certain food and that specific disease were be manually selected for each of CVD, type-II diabetes, and cancer, resulting in a total of 300 unique articles. One-hundred articles has been determined to be an appropriate sample size according to (Beleites, Neugebauer, Bocklitz, Krafft, & Popp, 2013) as illustrated in Figure 7.

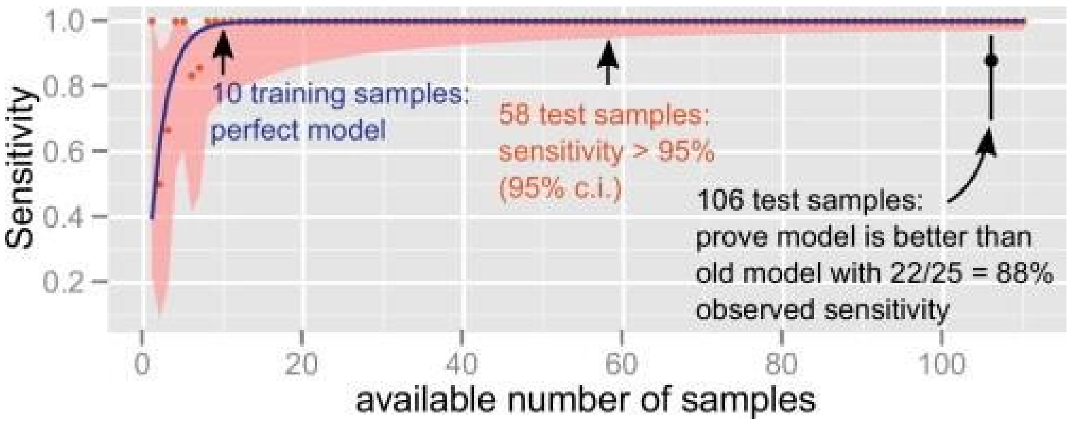
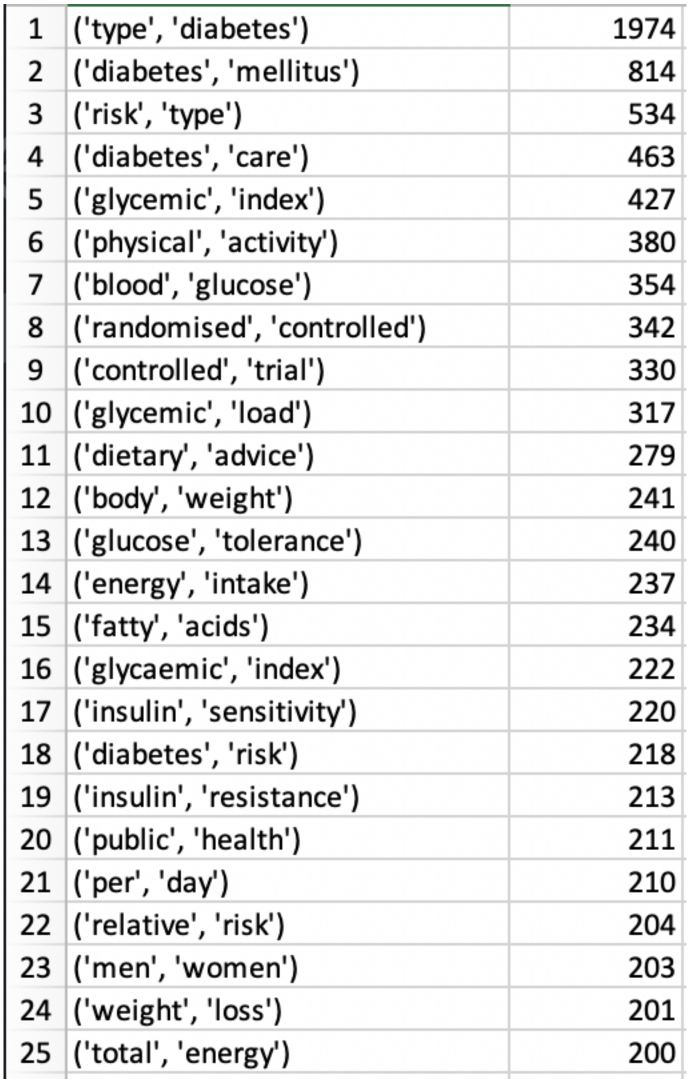


Figure 7. Classification sample size sensitivity

After one-hundred articles are gathered from each of the respective diseases, an n-gram algorithm was applied to the articles in order to extract the n-gram lists from them, thus providing the building blocks of a signature to be refined in the next steps.

3.1.2 Determining Most Frequent n-grams

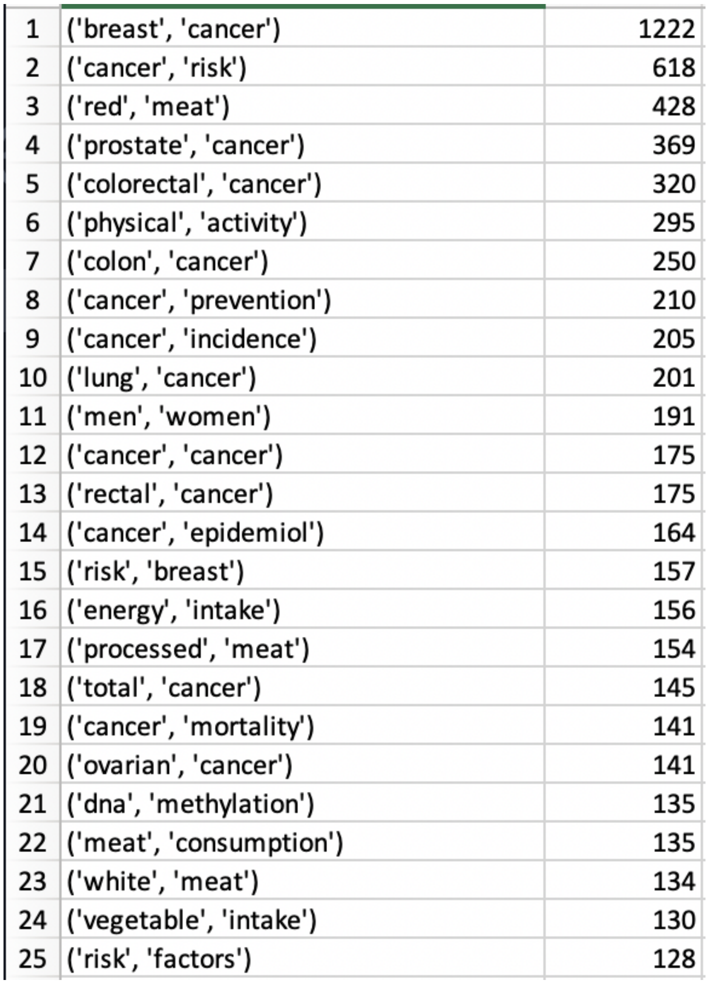
After n-grams have been gathered from each of the 4 chosen sequences sizes of n-grams (n=1, n=2, n=3, n=4), the most frequently n-grams in each of the respective articles are used as an identifier for the food-health articles related to that disease. The amount of the commonly found n-grams were determined experimentally as it was not immediately obvious how many unique n-grams will be found, nor is it obvious how many of the top most commonly found ones will be enough to accurately build the classifier. There were thousands of unique n-grams and thus the most appropriate allocation of the most commonly found n-grams was determined once the n-grams have been generated and analyzed experimentally. In Figure 8, one can see the top 25 most frequently found n-grams with their frequencies in the diabetes training data which were extracted from 70 diabetes articles.



1	('type', 'diabetes')	1974
2	('diabetes', 'mellitus')	814
3	('risk', 'type')	534
4	('diabetes', 'care')	463
5	('glycemic', 'index')	427
6	('physical', 'activity')	380
7	('blood', 'glucose')	354
8	('randomised', 'controlled')	342
9	('controlled', 'trial')	330
10	('glycemic', 'load')	317
11	('dietary', 'advice')	279
12	('body', 'weight')	241
13	('glucose', 'tolerance')	240
14	('energy', 'intake')	237
15	('fatty', 'acids')	234
16	('glycaemic', 'index')	222
17	('insulin', 'sensitivity')	220
18	('diabetes', 'risk')	218
19	('insulin', 'resistance')	213
20	('public', 'health')	211
21	('per', 'day')	210
22	('relative', 'risk')	204
23	('men', 'women')	203
24	('weight', 'loss')	201
25	('total', 'energy')	200

Figure 8. Bigrams of Diabetes (n=2)

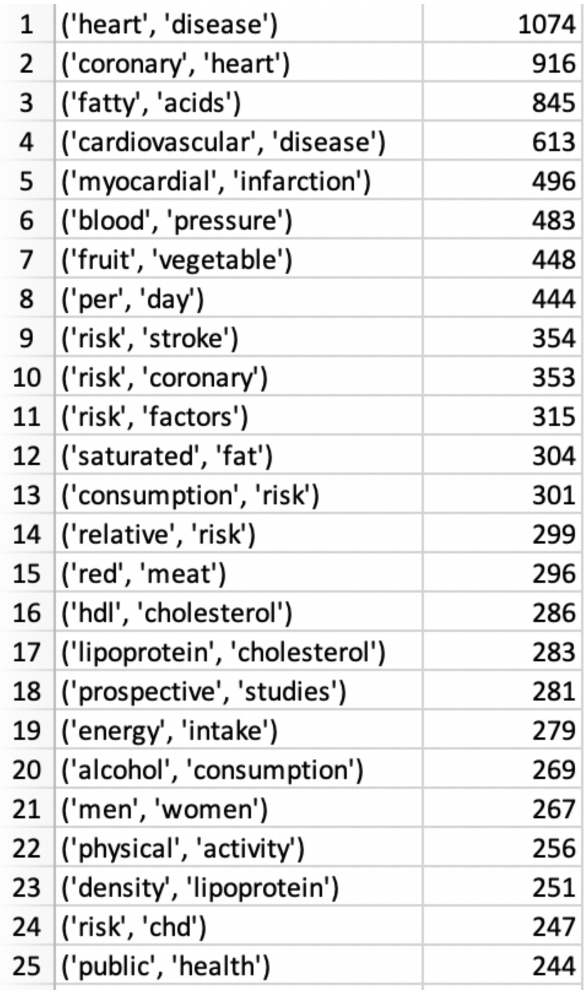
An example of the 25 most commonly found bigrams with their frequencies in cancer articles can be found in Figure 9. One can see the differences between the most commonly found cancer grams compared to the most commonly found diabetes grams.



1	('breast', 'cancer')	1222
2	('cancer', 'risk')	618
3	('red', 'meat')	428
4	('prostate', 'cancer')	369
5	('colorectal', 'cancer')	320
6	('physical', 'activity')	295
7	('colon', 'cancer')	250
8	('cancer', 'prevention')	210
9	('cancer', 'incidence')	205
10	('lung', 'cancer')	201
11	('men', 'women')	191
12	('cancer', 'cancer')	175
13	('rectal', 'cancer')	175
14	('cancer', 'epidemiol')	164
15	('risk', 'breast')	157
16	('energy', 'intake')	156
17	('processed', 'meat')	154
18	('total', 'cancer')	145
19	('cancer', 'mortality')	141
20	('ovarian', 'cancer')	141
21	('dna', 'methylation')	135
22	('meat', 'consumption')	135
23	('white', 'meat')	134
24	('vegetable', 'intake')	130
25	('risk', 'factors')	128

Figure 9. Bigrams of Cancer (n=2)

The 25 most commonly found CVD bigrams with their frequencies can be seen in Figure 10, once again highlighting the difference between the three resulting bigram sets from the three disease categories.



1	('heart', 'disease')	1074
2	('coronary', 'heart')	916
3	('fatty', 'acids')	845
4	('cardiovascular', 'disease')	613
5	('myocardial', 'infarction')	496
6	('blood', 'pressure')	483
7	('fruit', 'vegetable')	448
8	('per', 'day')	444
9	('risk', 'stroke')	354
10	('risk', 'coronary')	353
11	('risk', 'factors')	315
12	('saturated', 'fat')	304
13	('consumption', 'risk')	301
14	('relative', 'risk')	299
15	('red', 'meat')	296
16	('hdl', 'cholesterol')	286
17	('lipoprotein', 'cholesterol')	283
18	('prospective', 'studies')	281
19	('energy', 'intake')	279
20	('alcohol', 'consumption')	269
21	('men', 'women')	267
22	('physical', 'activity')	256
23	('density', 'lipoprotein')	251
24	('risk', 'chd')	247
25	('public', 'health')	244

Figure 10. Bigrams of CVD (n=2)

3.1.3 Determining Best n-gram Size

The most effective sequence size of the n-grams is determined experimentally. Once again, it was not immediately apparent which size will be the most accurate in classifying a food-health article. N-gram list of larger sequence sizes (e.g. n=5) provide more coherent phrases in natural language, yet they are very specific and unlikely to be commonly found throughout the sample of articles we will use. Likewise, n-grams of much smaller sizes (e.g. n=1) may not be specific enough to differentiate between a food article related to CVD and a food article related to type-II diabetes.

3.2 Testing Strategy

Each classifier for each disease was tested by using manually selected food-health articles which have not been presented to the algorithm. We used the 70/30 split which is the de facto standard for training and testing machine learning algorithms as seen in (K. Weinberger, 2009). This means that 70% of the data are used to train the algorithm, while 30% are used towards testing it. The 30% that have been used to test are articles which are hidden from the algorithm. If the algorithm is able to correctly classify the articles after training, then it will be considered a success.

3.2.1 Performance Metrics

The performance metrics which are used to determine the relevance and accuracy of the algorithm are precision and recall. Using true positive, true negative, false positive, and false negative, determines the accuracy of the classifier. When it receives a medical article as input, does it correctly classify the article or not? That is the only performance metric which was required in order to determine its accuracy.

3.3 Data

The data used for this study was manually gathered, peer-reviewed medical articles which specifically discuss health outcomes related to certain types of foods as their subject matter. The articles have been gathered manually because the nature of the data required is very specific and therefore not readily available in large quantities of word corpora such as social media, for example.

The sources of the data are popular medical databases such as PubMed/MEDLINE and Cochrane Library, as well as multidisciplinary scholarly databases such as ScienceDirect, Web of Science, JSTOR, and Google Scholar. These databases contained articles from popular medical journals including the New England Journal of Medicine (NEJM), British Medical Journal (BMJ), JAMA Network, American Diabetes Association, the American Journal of Clinical Nutrition, American Medical Association (AMA), Ovid Lippincott Williams and Wilkins (OLWW), and more. The portals and databases which we accessed these articles through can be found in Figure 8 below, showing a graphical distribution of the online sources used for gathering the training and testing data.

The data used for both training and testing are pre-processed before being used in the final implementation of the algorithm.

The first stage of preprocessing is converting the medical articles from PDF format to plaintext format, using UTF-8 encoding. The articles are normally retrieved in PDF format, so in order to facilitate extracting n-grams from them, we converted the articles to plain text. An existing Python package ‘pdf2txt’ which uses another Python package ‘pdfminer’ is used to batch convert the PDFs to plain text.

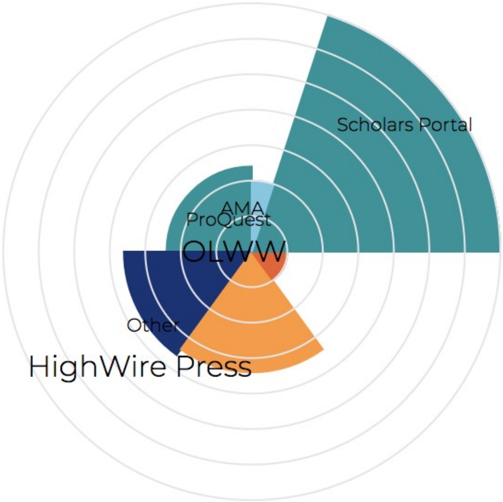


Figure 11. Data Source Distribution

Next, the second stage of preprocessing is normalization which involves tokenizing the text, converting the entire body of text into lowercase, removing non-alphabet characters. in addition, we removed 266 stop words which include unremarkable words such as *aren’t*, *the*, *a*, *as*, and *because*, as well as words which were found repeatedly in the articles such as *journal*, *clinical*, *research*, and *published*. Mean Magnitude of Relative Error (MMRE) is one of the most widely used evaluation criterion for assessing the performance of software prediction models (Foss, Stensrud, Kitchenham, & Myrtveit, 2003). This method involves using the estimated effort required to develop a software less the actual effort to create the software, divided by the actual effort. It is quite similar to the formula for calculating precision in the field of information retrieval. MMRE differs from the method of accuracy determination used in this research primarily due to its application. MMRE is more applicable to determining the accuracy of software estimation when considering man hours and money required to build a software system, whereas precision is more applicable to accuracy of retrieving documents based on a condition.

3.4 Complexity Analysis

Time and space complexity analysis was conducted on the algorithm to determine its growth requirements for both time and space with greater sized inputs.

3.4.1 Space Complexity

Space complexity analysis was conducted on this algorithm by simplifying the problem to being a mathematical one. If we assume that ***y*** is the number of generated n-grams, which are an indicator for the space required, and ***n*** is the number of words being input to the algorithm, and ***m*** is the size of the n-gram (n=1, 2, 3, or 4) then we can express the space complexity as the following formula:

$$\boldsymbol{y = n - m + 1}$$

The number of generated n-grams (tokens) can be expressed as having ***n*** minus ***m*** plus one. Therefore, when adding more words, more memory is required linearly. ***m*** is a constant, and so by following the rules of Big-O notation, we can say that the space complexity of this algorithm is O(n).

3.4.1 Time Complexity

The time complexity equation is a derivation of the space complexity algorithm. If we assume that ***y*** is the number of operations required to generate n-gram terms of size ***m*** from an article containing ***n*** number of words, then our time complexity formula can be expressed as follows:

$$y = (n - m - 1) * m$$

The number of operations, ***y***, consists of two operations, extracting the words from the article and concatenating the words into tokens. Once again, following what is known about the rules of Big-O notation, in our formula when ***n*** grows, we can ignore ***m*** because it becomes very small relative to ***n***, and ***m*** always remains constant. Therefore, our algorithm has a time complexity of O(n).

CHAPTER FOUR

4. RESULTS AND ANALYSIS

The most frequent Bigrams extracted from food-health articles are three unique identifiers that can be used effectively to enable the automatic identification and classification of the food-health articles related to the three diseases. The *n-gram size* (*n*=2) and *the* length of the n-grams list (*l*=800) have been found to be more effective in identifying food-health articles related to any of the three diseases compared to unigrams, trigrams and quadgrams for various n-gram-list lengths.

The effectiveness of the most frequent 800 bigrams have been tested by using them as an identifier for developing a food-health article classifier. The highest accuracy of the classifier is 90.00%. The overall average accuracy of all the various combinations for the value of n and number of n's are shown in Figure 12. This graph depicts the varying degrees of accuracy resulting from different combinations of these two variables. Figure 12 shows that most combinations of bigrams with 200 or more n-grams retrieved results with the highest average classification accuracy. Trigrams (n=3) are a close second in accuracy but only with using a set of 200 or more n-grams. Quadgrams (n=4) did much more poorly by contrast overall, particularly with a lower value of the n-gram list length. In contrast to quadgrams, unigrams (n=1) seem to be accurate *only* with a lower n-gram list length, with an accuracy drop off once the length is increased.

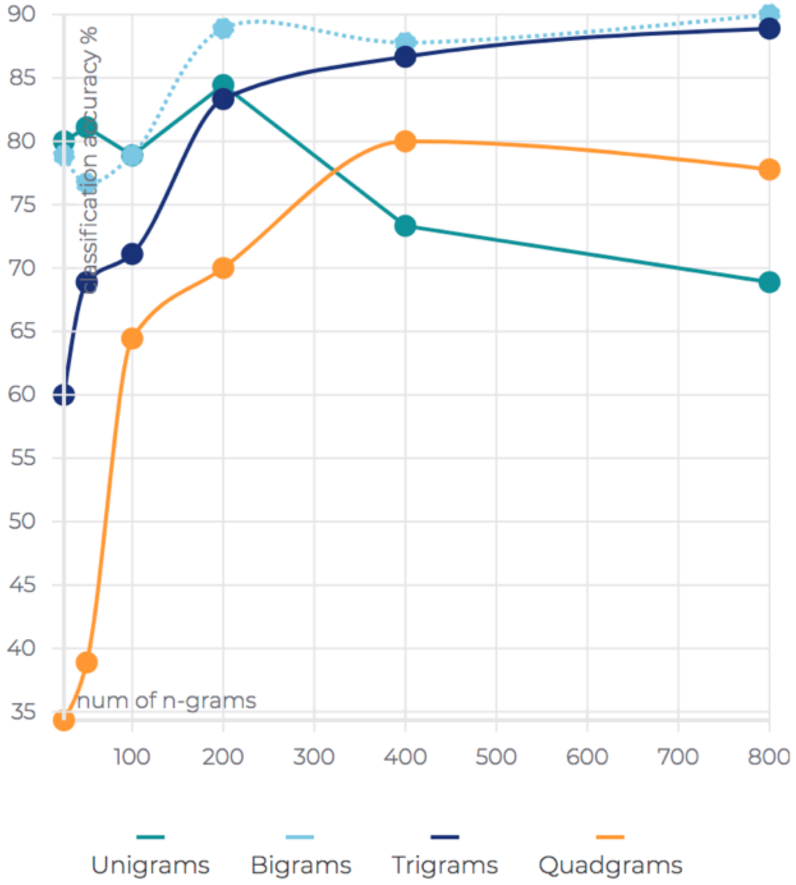


Figure 12. Average accuracy of all combinations

When manipulating n-gram size only and not accounting for n-gram list length, bigrams win out with an average accuracy of 83.5%, while unigrams and trigrams have close similar accuracies of 77.8% and 76.5%, respectively. This is shown in Figure 9 that compares the classifier average accuracy when isolating for n-gram size, without controlling for n-gram list length. Quadgrams showed the biggest reduction in accuracy with an average accuracy of 60.9%. This significant drop in accuracy could be attributed to the decrease in the n-gram list length due to the increase in the n-gram size. This affects the ability of the classifier to find notable differences within the articles. The writing styles of the authors could be another reason behind this drop in the accuracy when finding four words in a row being similar are hard to come by. Additionally, the subject matter of the articles all vary quite a bit even when studying the same diseases. For example, there were a few articles retrieved for the CVD portion of the data which talked about different cardiovascular diseases and their relationship to fish consumption in particular. One article talked about fish consumption and its relationship to risk of myocardial infraction (heart attacks), another spoke about fish and its relation to reduced progression of coronary artery atherosclerosis, and another talked about fish and omega-3 consumption in relation to risk of cerebrovascular disease. This simple example shows that even though an article may study CVD while also talking about fish consumption, it can take many different approaches to doing so. For this reason, quadgrams may be too generic and not as commonly found in order to be an effective method of classifying articles.

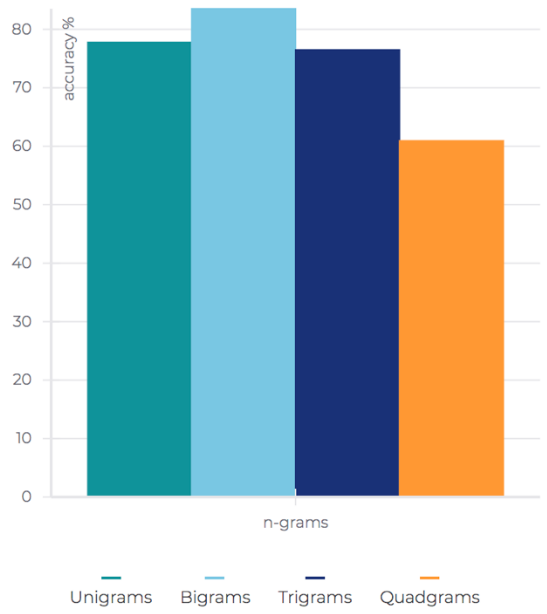


Figure 13. Average accuracy for various n-gram sizes

Alternatively, when isolating for n-gram list length while not controlling for n-gram size, the overall accuracy of varying degrees of the n-gram list length provides interesting results, shown in Figure 10. This figure shows the average classification accuracy for n-gram list length values of 25 through 800. It is interesting to note that while the accuracy gradually increases as we use a higher length value, classification accuracy plateaus after a certain point of 200 n-gram, and only varies by a tenth of a percent between length values of 200, 400 and 800. The respective accuracies of these values are 81.7%, 81.9%, and 81.4%. It is certainly a notable difference from the resulting accuracy of the length values 25 (63.3%) and 50 (66.4%). Perhaps the more notable implication from these values is that increasing the length does not result in higher classification accuracy beyond a certain point. This could be explained by examining how many n-grams are repeatedly found at the bottom of the list when looking at high lengths of n-gram list. Using the cancer training data as an example, we see that using bigrams with a length value of 800, ('breast', 'cancer') is the most frequently found n-gram with a frequency of 1222, with ('cancer', 'risk') coming second with 618 matches. By contrast, the 799th and 800th most commonly found bigrams are ('low', 'folate') and ('lipid, 'metabolism') with a frequency of 15 each. Additionally, the 200th ('cancer', 'patients') and 400th ('dietary', 'indexes') most commonly found bigrams only appear 38 and 23 times, respectively in the entire corpus of training data. This could explain why increasing the length beyond 200 does not drastically increase the classification accuracy, because the data becomes more diluted at this point and contains many more unique n-grams that are very specific to that single test article and may not necessarily be found within the training data.

Another interesting observation from the results of the test data is which individual disease topics had the highest average and highest achievable accuracies. CVD had the highest *achievable* accuracy (HAA) of 100% classification accuracy using n-gram size of 3 and a n-gram list length of 400, which can be noted as (3, 400), while cancer's HAA was 90.0% with a 4-way tie between (2, 100), (2, 400), (2, 800), (3, 800), and diabetes' HAA was 86.7% using (1, 200). This is certainly remarkable because it appears that certain combinations of n-gram size and n-gram list length result in different accuracies for each disease.

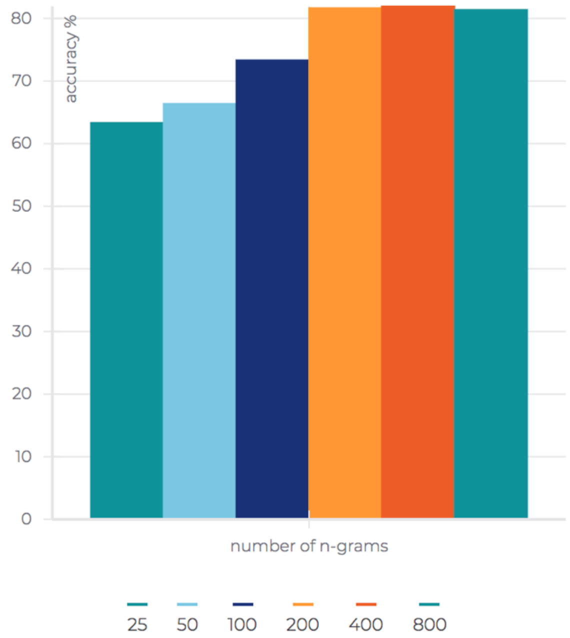


Figure 14. Average Accuracy for various n-gram list lengths

Across the 26 different combinations of n-gram size (1-4) and n-gram list length (25-800), CVD alone had the highest average accuracy of 87.4% while cancer and diabetes lagged behind with 69.3% and 69.2%, respectively. Thus, we could conclude that the CVD training data was either more unique or that the diabetes and cancer data was not unique enough. The latter seems to be the more likely, as the classifier was not able to distinguish between the test data belonging to diabetes more often than it incorrectly classified it. That is to say, when a diabetes article was not correctly recognized by the classifier as a diabetes article, it was because the test article had *an equal number of matches* from both the diabetes and CVD training data, not because it flat out incorrectly guessed the subject matter of the article. This could be because diabetes and CVD tend to have many overlapping terms and risk factors in medicine.

For the cancer testing data, the low overall accuracy could be explained through the fact that there is not very much research available linking food to cancer, and when there is, there are so many different types of cancers that these articles study including breast cancer, lung cancer, colorectal cancer, kidney cancer, and prostate cancer. This may have led to a failure to classify the articles correctly on a consistent basis due to the training data being so diverse.

CHAPTER FIVE

5. CONCLUSION

5.1 Conclusion

This research is the first report to describe the use of natural language processing and artificial intelligence techniques to extract and analyze data from literature via an automatic classifier. The developed classifier was able to achieve an average accuracy of 90.00% using n=2 and l=800. The developed classifier could change the way we manage data, develop strategies and conduct research. The classifier tool would be useful for a broad range of stakeholders including health professionals, dieticians, policy makers and researchers. More research is underway to further develop this classifier into one that is able to find trends in food and health, in order to develop novel hypotheses and support existing ones. Additionally, some features will be built in to filter articles on the basis of inclusion/exclusion criteria provided by authorities.

5.2 Future Work

The established identifiers are the fundamental step of the automatic extraction of useful information from the food-health articles related to specific diseases. The next steps will focus on analysis and mining the contents of the identified articles for specific disease. Data warehousing, big data techniques will be investigated to store and organize the extracted data in multidimensional databases. These databases could be used by food or nutrition researchers and other stakeholders to identify research gaps and to guide future strategies in food and health for both private and public sector.

5.3 Limitations

Some potential improvements to the algorithm which can be the focus of future research include weighting, synonyms, word stemming, and sentiment analysis.

The weighting of the tokens based on their frequency has the potential to increase the algorithm's accuracy by providing more weight to the higher frequency tokens found in Figures 8, 9, and 10. The more frequent ones being assigned more weight has the potential to increase the accuracy because it can provide a narrower approach to classification, by eliminating low frequency tokens.

Synonyms of words all counting as the same word also has the potential to increase the accuracy of the algorithm. Currently for example, the words *cancer* and *carcinoma* are read by the algorithm as two different words. If we expanded synonyms to even include related words such as *cancer, carcinoma, carcinogenic, oncology,* then that has the potential to increase the classification accuracy because it will once again narrow the use of the words more.

Word stemming is also a technique which has the potential to improve accuracy by reducing words to just their stems without their suffixes, e.g. cancer cancer**s**, cancer**ous**, cancer**ously**, cancer**ed** could all mean the same thing, *cancer.* Once again this has the potential to improve classification accuracy by making fewer words mean the same thing, thus trimming the word corpus and providing more context to a greater range of words.

Finally, sentiment analysis has the potential to take the research in a different direction by finding the relationship between foods and their diseases. For example, this algorithm is able to identify that white button mushrooms are related in some way to cancer incidence, but it is unable to distinguish whether this relationship is a positive or negative one. Am I supposed to avoid white button mushrooms if I want to reduce my risk of cancer or should I be eating more of them?

REFERENCES

Abou-Assaleh, T., Cercone, N., Keselj, V., & Sweidan, R. (2004). N-gram-based detection of new malicious code. *Proceedings of the 28th Annual International Computer Software and Applications Conference, 2004. COMPSAC 2004.*, 41-42.

Ananiadou, S., & McNaught, J. (2006). Text mining for biology and biomedicine. *Scitech Book News*, 286.

Beebe, N. L., Maddox, L. A., & Liu, L. (2013). Sceadan: Using Concatenated N-Gram Vectors for Improved File and Data Type Classification. *IEEE Transactions on Information Forensics and Security, 8*(9), 1519-1530.

Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C., & Popp, J. (2013). Sample size planning for classification models. *Analytica Chimica Acta, 760*, 25-33.

Belkhodia, O., Amara, N., Landry, R., & Ouimet, M. (n.d.). The Extent and Organizational Determinants of Research Utilization in Canadian Health Services Organizations. *Science Communication, 28*(3), 377-417.

Chumwatana, T., & Chuaychoo, I. (2016). Automatic filtering non-English complaint emails in tourism industry using N-gram extraction and classification techniques. *2016 4th International Symposium on Computational and Business Intelligence (ISCBI)*, 216-220.

Cohen, A., & Hersh, W. (2005). A survey of current work in biomedical text mining. *Briefings in bioinformatics, 6*(1), 57-71.

Collier, N., Doan, S., Kawazoe, A., & Matsuda Goodwin, R. (2008). BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics, 24*(24), 2940–2941.

Foss, T., Stensrud, E., Kitchenham, B., & Myrtveit, I. (2003). A simulatio study of the model evaluation criterion MMRE. *IEEE transactions on software engineering, 29*(11), 985-995.

Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1).

Hirschman, L., C Burns, G. A., Krallinger, M., Arighi, C., Bretonnel Cohen, K., Valencia, A., . . . Wiegers, T. (2012). Text mining for the biocuration workflow. *Database: The Journal of Biological Databases and Curation, 1*.

Hunter, L., & Cohen, K. (2006). Biomedical Language Processing: What's Beyond PubMed? *Molecular Cell, 21*(5), 589-594.

K. Weinberger, J. B. (2009). Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal Of Machine Learning Research, 17*(1), 207-244.

MEDLINE Fact Sheet. (2018). (U.S. National Library of Medicine) Retrieved April 5, 2018, from https://www.nlm.nih.gov/pubs/factsheets/medline.html

Open Data 101. (2017, December 19). Retrieved from https://open.canada.ca/en/open-data-principles

Powers, D. M. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies, 2*(1), 37-63.

Rebholz-Schuhmann, D., Oellrich, A., & Hoehndorf, R. (2012). Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews.Genetics, 13*(12), 829-839.

Report of Canada's Economic Strategy Tables: Agri-food - Economic Strategy Tables. (2018, September, 28) Retrieved December 10, 2018, from https://www.ic.gc.ca/eic/site/098.nsf/eng/00022.html

Reviews on Text Mining in Biomedicine. (2006). (Biomedical Literature and Text Mining Publications) Retrieved April 11, 2018, from http://blimp.cs.queensu.ca/cateR_1.html

Terdchanakul, P., Hata, H., Phannachitta, P., & Matsumoto, K. (2017). Bug or Not? Bug Report Classification Using N-Gram IDF. *2017 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 534-538.

WHO | Overview - Preventing chronic diseases: a vital investment. (2015). *WHO*. Retrieved from https://www.who.int/chp/chronic_disease_report/part1/en/index1.html

Wiegand, M., & Klakow, D. (2013). Towards Contextual Healthiness Classification of Food Items - A Linguistic Approach. *International Joint Conference on Natural Language Processing*, 19-27.

Wiegand, M., & Klakow, D. (2015). Detecting conditional healthiness of food items from natural language text. *Language Resources and Evaluation, 49*(4), 777-830.