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

Marco Ross[1], El Sayed Mahmoud[2]
Faculty of Applied Science & Technology
Sheridan College
Oakville, Canada

El-Sayed M. Abdel-Aal[3]
Guelph Research and Development Centre
Agriculture and Agri-Food Canada
Guelph, Canada

***Abstract*—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. 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 timely fashion.**

***Keywords*—Natural language processing; text classification; ngrams; bioinformatics; knowledge extraction; nutrition assessment; health promotion; research uptake***

## I. Introduction

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 fifty seven percent of the respondents frequently or very frequently received research results. 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 [1].

The main reason for these low uptake percentages could 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 the selection process of an article about a specific food and disease more difficult. For example, the popular biomedical database MEDLINE, produced by the United States National Library of Medicine, contains over 24 million references of biomedical texts alone [2]. Since 2005, the literature has seen consistent from 16 million articles to 24 million [3], 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 goal was to develop a text classification tool that is capable of performing efficient literature search and analysis in a timely fashion.

### A. 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.

### B. Organization of Paper

The remainder of this paper 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 results section will highlight the experimental findings including the analysis of these findings and potential future research.

## II. 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 [4] [5], which target experts in biosciences as their primary audience [6].

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 [6] [7] [8]. 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 ~~chapter~~ 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 [7]. Hirschman and others 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 [7].

A current 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 [6]. 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 s. **Figure 1** 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 [7].

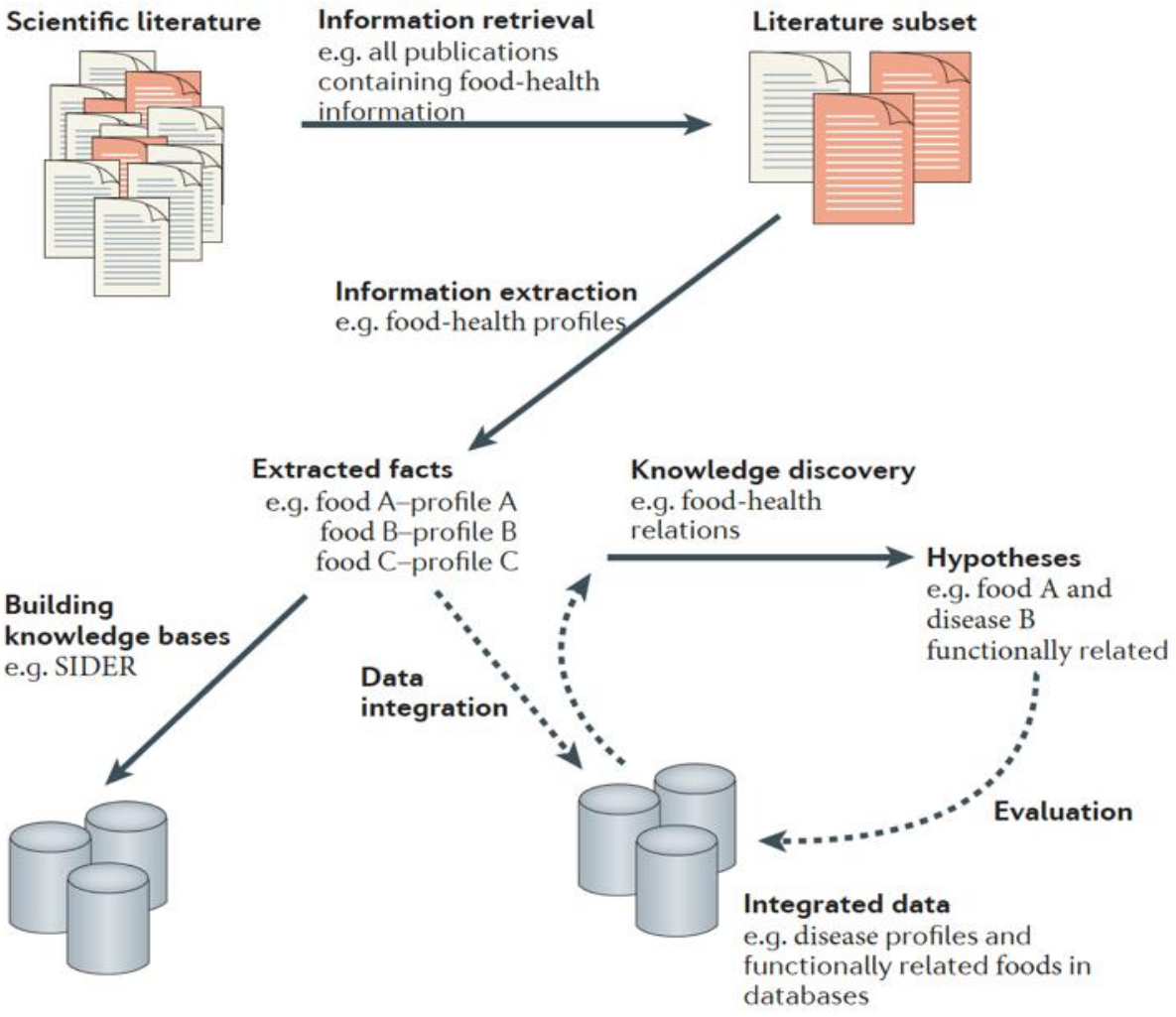


Fig. 1. Categories of Text Mining Solutions.

**Figure 1** distinguishes four primary stages in text-mining solutions: information retrieval, information extraction, building knowledge bases and knowledge discovery [8].

Information retrieval could involve a user submitting a query to a search engine and receiving a document fitting to their submitted query in return. Information extraction involves the identification of entities, such as diseases or foods, as well as the identification of complex relationships between these entities [8]. 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 [7] [8]. 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 [8].

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.

N-grams has been used for finding identifiers of disease outbreaks in reasons of entering the emergency department room [9] and fining identifiers for customer intent [10]. This work investigates how to use N-grams for finding identifiers for food-health articles related to the proposed diseases.

## III. Methodology

### A. 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 2.**

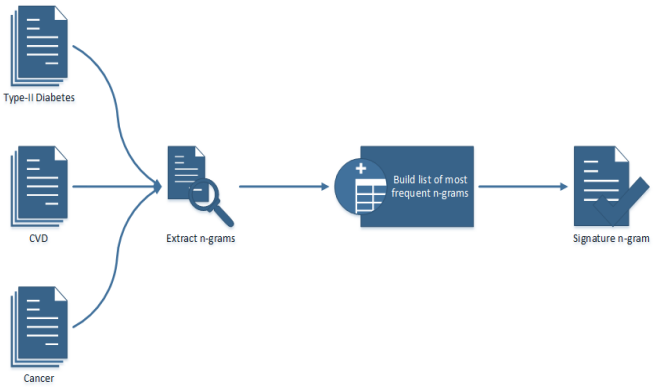


Fig. 2. Steps to Identify A Unique Key For Food-Health Articles Related to the Diseases: CVD, Cancer and Diabetes.

For the sake of simplicity, the diagram in **Figure 2** shows the process as if one n-gram classifier will be created from all three disease categories, when in fact three *separate* classifiers will be created using this same process.

### B. 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 will 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 [11].

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

### C. 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 are determined experimentally as it is 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 will likely be hundreds, possibly thousands of unique n-grams and thus the most appropriate allocation of the most commonly found n-grams will be determined once the n-grams have been generated and analyzed experimentally.

### D. Determining Best N-Gram Size

The most effective sequence size of the n-grams is determined experimentally. Once again, it is 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.

### E. Testing Strategy

Each classifier for each disease is 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 [12]. 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.

### F. 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 will be required in order to determine its accuracy.

### G. Data

The data used for this study are 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 3** 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.

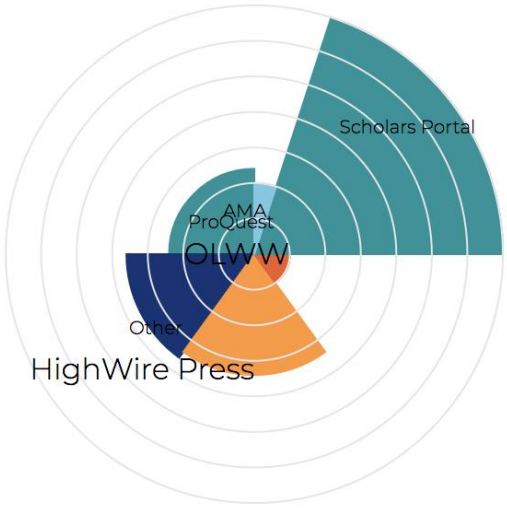


Fig. 3. Data Source Distribution.

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 [13]. 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.

## IV. 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.

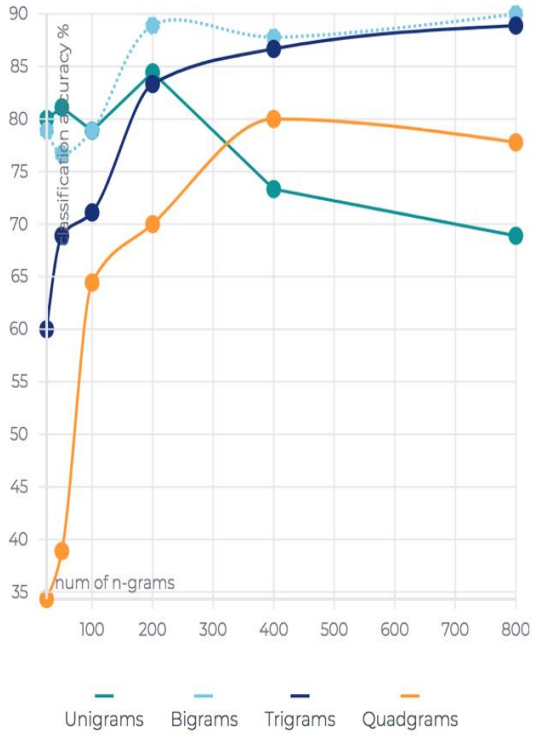


Fig. 4. Overall Classification Accuracy.

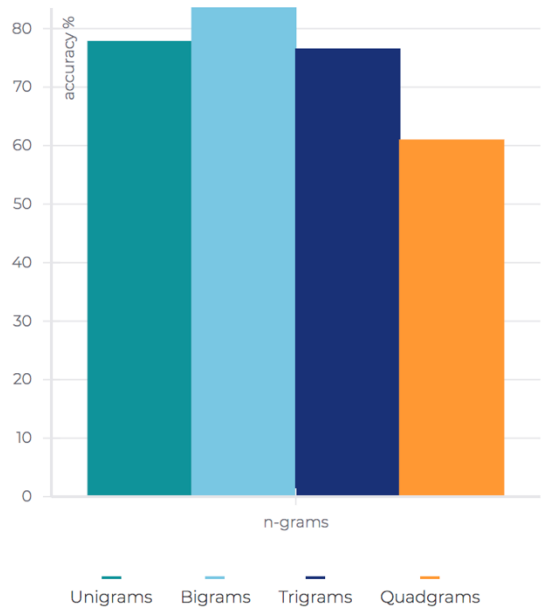


Fig. 5. Average Accuracy for Various N-Gram Sizes.

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 4**. This graph depicts the varying degrees of accuracy resulting from different combinations of these two variables. **Figure 4** 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 poorly 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.

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 by **Figure 5** 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.

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 6**. 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 799$^{th}$ and 800$^{th}$ most commonly found bigrams are ('low', 'folate') and ('lipid, 'metabolism') with a frequency of 15 each. Additionally, the 200$^{th}$ ('cancer', 'patients') and 400$^{th}$ ('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.

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.

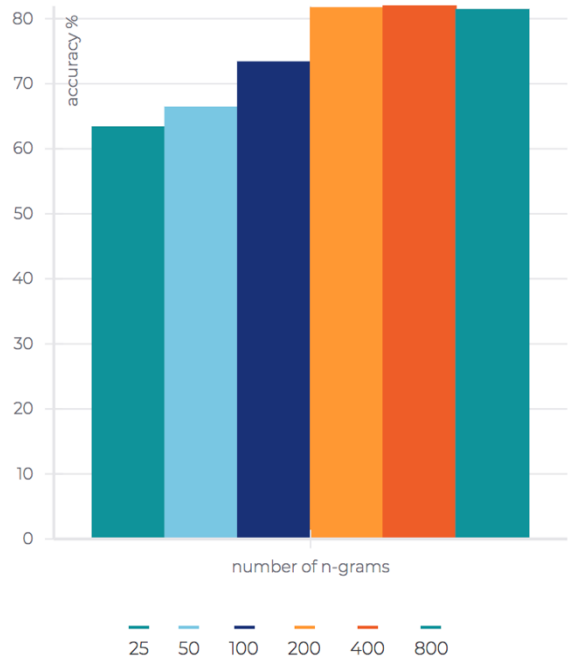


Fig. 6. Average Accuracy for Various n-Gram List Lengths.

## V. 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 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.

## VI. Next Steps

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.

## Acknowledgment

This work is resulted from Marco Ross's undergraduate thesis of the Honours Bachelor of Computer Science (Mobile Computing). This research was supported by the Centre for Mobile Innovation (CMI) – Sheridan College, ON, Canada. through the work study program. We thank the thesis advisory committee members at Sheridan college for their useful feedback on the work.

## References

[1] O. Belkhodia, N. Amara, R. Landry and M. Ouimet, "The Extent and Organizational Determinants of Research Utilization in Canadian Health Services Organizations," *Science Communication*, vol. 28, no. 3, pp. 377-417.

[2] "MEDLINE Fact Sheet," U.S. National Library of Medicine, 2018. [Online]. Available: https://www.nlm.nih.gov/pubs/factsheets/medline.html. [Accessed 5 April 2018].

[3] L. Hunter and K. Cohen, "Biomedical Language Processing: What's Beyond PubMed?," *Molecular Cell*, vol. 21, no. 5, pp. 589-594, 2006.

[4] "Reviews on Text Mining in Biomedicine," Biomedical Literature and Text Mining Publications, 2006. [Online]. Available: http://blimp.cs.queensu.ca/cateR_1.html. [Accessed 11 April 2018].

[5] S. Ananiadou and J. McNaught, "Text mining for biology and biomedicine," *Scitech Book News*, p. 286, 2006.

[6] A. Cohen and W. Hersh, "A survey of current work in biomedical text mining," *Briefings in bioinformatics*, vol. 6, no. 1, pp. 57-71, 2005.

[7] L. Hirschman, G. A. P. C Burns, M. Krallinger, C. Arighi, K. Bretonnel Cohen, A. Valencia, C. H. Wu, A. Chatr-Aryamontri, K. G. Dowell, E. Huala, A. Lourenço, R. Nash, A.-L. Veuthey and T. Wiegers, "Text mining for the biocuration workflow," *Database: The Journal of Biological Databases and Curation*, vol. 1, 2012.

[8] D. Rebholz-Schuhmann, A. Oellrich and R. Hoehndorf, "Text-mining solutions for biomedical research: enabling integrative biology," *Nature Reviews.Genetics*, vol. 13, no. 12, pp. 829-839, 2012.

[9] E. S. Mahmoud and S. Deborah, "Identifying Syndromic Fingerprints in Reason Fields in Emergency Department or Telehealth Records using N-grams for Similarity Analysis," *Advances in Disease Surveillance*, vol. 4, p. 55, 2007.

[10] S. Akulick and E. S. Mahmoud, "Intent detection through text mining and analysis," in *Future Technologies Conference (FTC)*, Vancouver, 2017.

[11] C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft and J. Popp, "Sample size planning for classification models," *Analytica Chimica Acta*, vol. 760, pp. 25-33, 2013.

[12] J. B. L. S. K. Weinberger, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *Journal Of Machine Learning Research*, vol. 17, no. 1, pp. 207-244, 2009.

[13] T. Foss, E. Stensrud, B. Kitchenham and I. Myrtveit, "A simulation study of the model evaluation criterion MMRE," *IEEE transactions on software engineering*, vol. 29, no. 11, pp. 985-995, 2003.