Automating Feature Extraction of Food Health relationships

Alexis Mayet
Department of Advanced Computing Sciences
Faculty of Science and Engineering
Maastricht University
Maastricht, The Netherlands

***Abstract*—This project tackles the challenge of assisting human expert in substantiation processes using NLP, namely in this case the substantiation of food-health claims submitted to the European Food Safety Authority (EFSA). To achieve this, it is required to automate Knowledge Graph creation from biomedical literature. To this end, a Linear Support Vector Machine classifier was trained such that food-health relations are correctly identified and classified. This model achieves an accuracy of 82.91%. However, this model has several limitations such as a label imbalance in the data set leading to bias in the training, as well as a number of guiding assumptions that reduce the scope of application of the model. Such a tool can improve the efficiency of human experts in the process of substantiating food-health claims, which in term can lead to improve regulatory market conditions for manufacturers and consumers.**

I. Introduction

Diet plays a fundamental role in the development of diseases and health conditions in human health. It constitutes the individual's needs for energy and nutrients to develop and maintain its body. Hippocrates already had intuitions about prescribing a nutritious diet to improve health conditions. This constitutes general knowledge, therefore consumers have an increasing desire to be aware of the nutrients they are consuming and to understand their impact on health.

However, in recent years, medicine has put an emphasis on research and prevention of health conditions in the context of genomics. This is due to the fact that technologies and methods for this approach have greatly improved in accessibility and efficiency. This emphasis has been the cause of a lack of research in the context of the exposome [1]. This can also be attributed to the difficulty of the process of establishing and communicating knowledge about the relationship between food and health. There has been a rise in studies establishing knowledge in this domain, namely with the apparition of the term "functional foods" which refers to foods with specific beneficial functions [2]. The large number of claims and the challenge of communicating these to consumers have resulted in the necessity of legal frameworks to validate the scientific accuracy of those claims and to ensure that consumers have the opportunity to make rational diet choices towards healthy foods, which would in turn incentivize industries to produce healthier products [3]. Along with this legal framework, there is a need for regulations regarding the communication of the components of a product, and the impact of those components on human health [4], as there is a large risk of consumer misunderstanding. Towards this goal, different regulatory institutions across the world have introduced policies concerning food-health claims : the US have introduced the Food and Drug Administration (FDA) , Japan has introduced the Nutrition Improvement Laws, and the European Union (EU) introduced the regulations 1924/2006 [5], the latter of which is the main focus of this study.

This EU regulation lays out conditions for the usage of health claims in the food industry, more specifically advertising, and establishes a system of scientific evaluation across all member states. This system of scientific evaluation is embodied by the European Food Safety Authority (EFSA), which is responsible for reviewing any food or health claim made by a company, and establishing a verdict concerning the validity of the claim. This review is refered to as a substantion process, as it aims to establish the veracity of the claim. All such claims are freely available online in the EFSA database [12]. However, the application of the regulation requires an important amount of resources. Each claim submitted must be evaluated by domain experts, who assess the evidence provided and determine whether it is sufficiently substantiated, and therefore whether it will be authorized. The impact on the market and the companies involved in case of authorization is also assessed. This process is estimated to cost the EU between 4.51 and 7.56 million euros, without taking the cost additional clinical trials into account [4] [6]. Reference [11] have found that employing Natural Language Processing (NLP) methods, namely automated feature extraction to assist domain experts improves the yield of human efforts in such review processes. Therefore this study aims at developing such a tool specific to the review process concerning food health claims conducted by the EFSA. This entails the research question : How can automated feature extraction assist the substantiation proccess of reviewing food health claims at the EFSA? To answer this, first a literature review is conducted. Additional background knowledge on the food health claim regulation 1924/2006 is given, followed by an analysis of the consumer perception of food health claims, as well as the economical impact of regulations. Studies that employ similar methodology such as NLP and Knowledge Graphs are also reviewed. A description of the methodology employed is conducted, describing the 3 steps of this method : Named Entity Recognition (NER),

Relation Extraction (RE), and Knowledge graph construction. The results are then presented, followed by a discussion of their limitations and implications.

II. Litterature review

A. Regulations

The term functional food, first introduced in Japan in the 1980s, refers to processed foods containing ingredients that aid specific body functions in addition to being nutritious [2]. As functional foods rose in popularity, there came a need for regulation, as companies spent millions of dollars in research and development (RnD) of such functional foods [7], and therefore a large number of unsubstantiated claims flooded the market. In 1963, the United Nations (UN) developed the Codex Alimentarius, a collection of guidelines and internationally recognized standards relating to food safety and production [13]. Following this, in the 1990s, the USA and Japan were the first countries to implement a regulatory framework concerning food safety, followed later on by the EU, who in 2006 introduced the 1924/2006 regulations. These regulations distinguish between three main, types of claims : Article 13, Article 13.5 and Article 14. Article 13 refers to claims that concern the role of a nutrient in growth development or maintenance of body functions, claims that refers to psychological and behavioural functions and claims that relates to slimming, weight control and appetite control. Claims that fall under Article 13.5 are any claim that are based on newly developed scientific evidence or which requires access to proprietary data. Article 14 refers to claims that concern the reduction of a disease risk, and any claim that concern children's health. Article 13 claims are only required to present sufficient references to relevant existing scientific evidence that support the claim, while Article 13.5 and 14 requires the submission of an extensive scientific dossier.

This means that claims that are not based on any type of scientific evidence are systematically denied, which contrasts with regulations in the USA and Japan. Indeed, these countries make a distinction between unqualified and qualified claims. Unqualified claims refer to claim that are substantiated with a high level of scientific evidence, meaning they have a high credibility. Qualified claims refer to claims that are substantiated with a low level of scientific evidence, having a lower credibility and making use of emerging knowledge. [5] Claims submitted in the EU are handled by the European Commission at first. Claims that fall under the Article 13 must reference sufficient scientific evidence that : emphasise the need for evidence linking the food to a beneficial effect on human health, recognises the usefulness of markers or intermediate effects in the biological process and emphasises that the effects are both biologically and statistically significant and meaningful. Claims that fall under Article 13.5 or Article 14 must contain scientifical evidence obtained specifically for the product in question. The European Commission submits the claims to the EFSA, who conduct an evaluation based on the following questions : Has the product been sufficiently characterised? Is the claim effect beneficial for health? Has a cause and effect relationship been established? [5]. If the claim answers positively to these questions they are added to the EFSA register of food health claims, which is the data set used in this project.

B. Food Health Claims

The Codex Alimentarius defines food health claims as "any representation, which states, suggests or implies that a food has certain characteristics relating to its origin, nutritional properties, its nature, its production , its processing or its composition or any other quality" [13]. From this definition, two type of claims arise : nutritional claims, which refer to any claim which states, suggests or implies that the concerned food has a particular beneficial nutritional property due to the energy/nutrients/other substances (not) provided; and health claims, which refer to any claim which states, suggests or implies that a relationship exists between food and health.

Once a claim is approved by the EFSA, there is still a challenge in the communication of this claim to the consumer of the product concerned. Namely, additional regulations apply to the methods of disclosure and communication of this claim to consumers. These regulations are aimed to ensure that "the average consumer" can effectively understand the claim. Reference [4] published a paper on this subject in 2011, analysing the effectiveness of this regulation. They have found that there is an important risk of consumer misunderstanding. This is because consumers are faced with time and resources constraints, because they experience many biases when exposed to claims, because the acquisition process is influenced by individual motivation and ability, and because the understanding of a food health claim is based largely on the consumer's prior belief and how those relate to the claim. It entails from these factors that addressing claims to the "average consumer" is a limited approach as targeting specific consumer groups would be more effective. Indeed, they have found that highly educated consumers, more specifically women are the most likely to respond positively to a health claim. They have also determined that elderly people are more likely to respond positively to a health claim than younger people. Such findings confirm that there is an important effort to be made by companies to successfully communicate a claim to consumers.

Reference [14] have analysed the impacts of the 1924/2006 regulations on the market innovation, and have found that it poses an important challenge to companies. Different factors can explain this : namely, as the regulation came into effect, companies were incentivized to revisit the ingredients list of their existing products to "cover" ingredients not supported by a health claim. The regulation is also deemed to lack transparency in regards to the process of scientific substantiation. Companies also redirect their efforts towards effectively communicating claims to consumers and away from innovation. Limited financial and R&D resources also come into play. Because of these factors, health claims as a route of innovation has reduced attractiveness to companies. Investment in research is perceived as risky as there is a low probability of actually obtaining a health claim. Therefore

companies may focus on other aspects of a product instead of scientific substantiation, such as sustainability and traceability. A cause for positive change towards innovation in the industry would be a reduced cost of processing health claims, which is the aim of this project.

C. Natural Language Processing and Knowledge Graphs in medicine

In an increasingly digitized world, vast amounts of data are collected constantly. However, there is a need to make sense of the raw unstructured data, such that a sensical, useful and structured presentation of the data can help obtain knowledge easier. Multiple approaches and methods exist for this, but those more specifically and often used in the field of medical research are NLP and Knowledge Graphs. A Knowledge graph is a knowledge base structured as a graph, in which vertices are labelled to represent entities and edges are labeled to represent relations. Indeed, Reference [11] have found that, along with Optical Character Recognition (OCR), NLP is widely used as a tool to improve the efficiency of medical chart review processes.

A study conducted with similar methods to this project is that conducted by Reference [10], which aims at creating a knowledge graph to enhance fraud, waste and abuse (FWA) detection in the Chinese healthcare system. Their approach consisted of creating an annotated corpus with medical data concerning drug labels, disease information from medical textbooks and information from medical examinations. This corpus was employed to train and test an NER model, as well as a relation extraction model to extract information from their unstructured corpus of data. They used entity linking to build the knowledge graph, which yielded a 70% detection rate of claims suspected of FWA. This means that the model can improve the efficiency of claim processing as it can assist domain experts in their task. Although the scope is wider, the approach employed is similar to that of this project, making use of NLP and Knowledge graphs to assist domain experts in their task.

Another study which highlights the use of such technologies is conducted by Reference [8] , who have used NLP to develop a tool that automatically builds a medical knowledge graph given large numbers of Electronical Medical Records (EMR) of a particular hospital in China. Their approach consist of a novel method of defining a Knowledge Graph fact as quadruplets instead of triplets. They employed deep-learning techniques to perform entity recognition, entity normalization and relation extraction to structure the data in usage. They employed graph cleaning, related-entity ranking and graph embedding to create the knowledge graph. The result is a Knowledge Graph containing 22508 different entities and 579094 quadruplets. It can be successfully used in many practical applications such as clinical decision support and information retrieval, which can be used by domain experts to improve the efficiency of their efforts. The approaches employed have a wider scope than in this project, as there is no need for a quadruplet structure in the knowledge graph, and the entity normalization process is overlooked. However it is still relevant to this project as it makes similar uses of NLP and Knowledge graphs to assist domain experts.

III. Methods

This research uses data from the EFSA, available online. It consists of 2563 claims, which relate to 1216 different foods and 523 different health aspects impacted. Each claim contains information such as the food term, the health term, the claim, whether it was authorised or not, and other attributes less relevant to this research such as the regulations specifications, supporting evidence and so on. Upon inspection, many claims are difficult to use in this case as they contain some data entry malpractices, namely use of different languages, spelling mistakes, synonyms, abbreviations and so on. It is also noted that out of all claims submitted to the EFSA, only 235 were authorized. Therefore, the dataset is reduced only to those authorized, which are easily usable. As the aim of this project is to create a tool that automatically creates knowledge graphs of food health claims, there are three main task at hand. The first task is to develop a NER model that will identify mentions of entities of food or phenotype. The second task is to develop a Relation Extraction (RE) model that will automatically qualify the relationship between the identified food and phenotype. The final task is to combine the NER and the RE model and use them to create a knowledge graph. Reference [9] developed an NER model specific to the EFSA dataset used in this project. The model achieved a 90% accuracy when identifying entities of foods and phenotypes on a test set. This project takes into consideration the model developed and therefore focuses on the second task : developing a relation extraction model. From all the authorized claims, three different food-heatlh relationships are identified: ”maintenance of a function”, ”enhancement of a function” and ”reduction of a risk factor”. These relationships are already annotated in the dataset, therefore the only pre-processing concerns entities of food and phenotype. Indeed, these entities are given in the columns ‘Food’ and ‘Phenotype’, however, the values do not necessarily match to those in the claims, as often synonyms, abbreviations or spelling mistakes are present when referring to the food or the phenotype. This is done manually through a dictionary that maps all such synonyms to a single entity. Relation Extraction : The task of relation extraction consists of identify the relationship between two given entities in a string of text. For example, in the sentence “Biotin contributes to normal macronutrient metabolism”, given the food entity ‘Biotin’ and the phenotype ‘normal macronutrient metabolism’, a relation extraction model aims at identifying the relationship “contributes to” and classify it in one of the three different relationships mentioned earlier, thus creating a triplet Subject, predicate, object : (Biotin, contributes to, normal macronutrient metabolism). Given the problem in occurence, all data entries are guaranteed to contain a relationship within three possible labels. Therefore the method used in this project consists of reducing this problem to a 3-class classification problem. After replacing the entities in the text by an entity

tag, the sentence is vetorized using a TF-IDF vetorizer. The resulting vectors are then used to train a classification model using the library Scikit learn.

Different classification algortihms are considered for this task namely a Logistic Regression classifier, a Mutlinomial Naïve-Bayes classifier and a Linear Support Vector Classifier. The models were evaluate using K-fold cross evaluation, with $K = 10$. The metrics employed to evaluate and selection the models are accuracy, precision, recall. Below are the testing scores for the models and the metrics.

Metric	LinearSVC	MultinomialNB	Logistic Regression
Accuracy :	0.816667	0.808152	0.799819
Precision :	0.586813	0.434486	0.433934
Recall :	0.503899	0.438889	0.435380

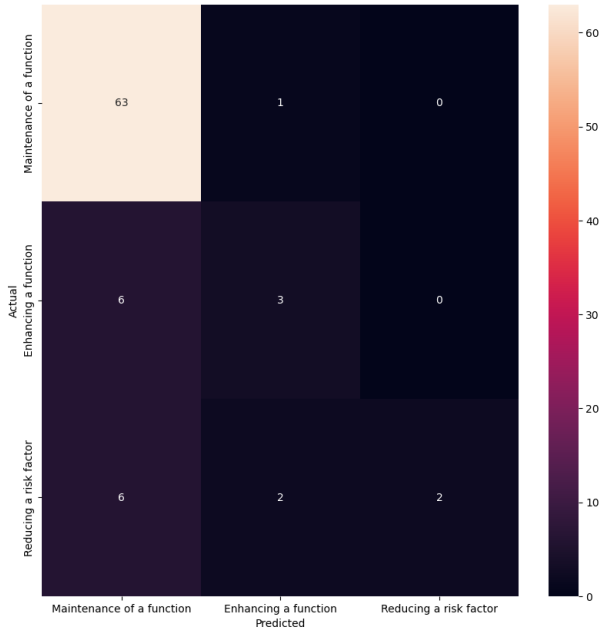
As we can see in the table above, the Linear Support Vector Classifier outperforms the other classifiers on every metric used.

IV. Results

After investigation of which model performs better at this task, the Linear Support Vector Machine classifier was selected. The model is trained using a training split of the data and evaluated using a testing split. This results in a classifier that achieves an accuracy of 81.92% on classifying sentences that contains a food entity and a phenotype entity into : "Enhancing a function", "Maintenance of a function" or "Reducing a risk factor".

V. Discussion

While the model developed achieves a relatively high accuracy, there are several limitations to it. The first limitation comes from the imbalance in data labels present in the training data. Indeed, the label "Maintenance of a function" represents 78% of entries, the label "Enhancing a function" represents 12% of entries and the label "Reducing a risk factor" represents 10%. This imbalance creates an important bias in the algorithm which can be observed in the confusion matrix below. It describes the test performance of the model selected and trained as described in the previous section.



When summing the columns up, we notice that the model predicts "Maintenance of a function" around 90% of the time. This means that the model is quite limited in its capacities, as it is heavily biased towards a particular label at the detriment of the others. Other limitations arise from the guiding assumptions made. Indeed, it is assumed that the input sentence is initially treated by a NER system trained for this task, meaning that the sentence systematically contains a "Food" entity and a "Phenotype" entity, and that the position of these entities is known. It is assumed that every input sentence is guaranteed to contain a relationship, and finally, it is assumed that all food-health relationships can be classified within one of those categories : "Maintenance of a function", "Enhancing a function" and "Reducing a risk factor".

Furthermore, there are different approaches that can lead to an improved performance or similar performance less dependent on initial assumptions. Simply, increasing the size of the data used will lead to more robust training which can achieve better performance. Reducing the imbalance in the data label distribution can also help the model perform better and eliminate some of the bias present toward the label "Maintenance of a function". Currently, it is investigated to use a weak supervision learning approach to pursue this line of research, by crafting logic-based rules to automatically annotate the data. This could be done through Snorkel Flow. Finally, the ideal approach to such a problem is making use of deep learning algorithms applied to NLP. Currently, the use of the pyTorch library is also currently investigated to pursue this line of research.

VI. Conclusions

In conclusion, this project aimed to develop a tool that automates the creation a knowledge graph of food-health relationships using natural language processing (NLP). The research focused on the review process of food health claims conducted by the European Food Safety Authority (EFSA)

and explored how automated feature extraction can assist in the substantiation process. A literature review is provided, which gave insights into the regulations concerning food health claims, the challenges in communicating these claims to consumers, and the impact of regulations on the market and innovation. The use of NLP and knowledge graphs in the medical field is also examined, highlighting their potential in improving efficiency and assisting domain experts. The methodology involved the utilization of data from the EFSA, specifically the authorized food health claims dataset. Following the development of a NER model in another similar project, which indentifies food and phenotype entities, a relation extraction (RE) model was created to determine the relationship between these entities. The combination of these models enables possibility of developing a method to automate the construction of a knowledge graph. The findings of this research contribute to addressing the resource-intensive process of reviewing food health claims. By automating certain tasks using NLP , domain experts can improve their efficiency and accelerate the substantiation process. The presented tool has the potential to reduce the cost and time required for evaluating claims, benefiting both the regulatory authorities and the food industry.

However, it is important to acknowledge the limitations of this study. The dataset used was limited to authorized claims, and further research could explore the inclusion of a wider range of claims, including those that were not authorized. Additionally, the NER and RE models can be further enhanced to improve their accuracy and performance using a deep learning approach.

Overall, this research demonstrates the potential of NLP and knowledge graphs in assisting domain experts in reviewing food health claims. The automated feature extraction tool developed in this study provides a foundation for future research and advancements in the field of food-health relationships. It is hoped that this work will contribute to the development of more efficient evaluation processes on the impact of food on human health, ultimately leading to improved consumer awareness and healthier food choices.

REFERENCES

[1] Wild, C. P. (2005). Complementing the Genome with an "Exposome": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. Cancer Epidemiology, Biomarkers & Prevention, 14(8), 1847–1850. https://doi.org/10.1158/1055-9965.epi-05-0456 in press

[2] Kaur, S., & Das, M. (2011). Functional foods: An overview. Food Science and Biotechnology, 20(4), 861–875. https://doi.org/10.1007/s10068-011-0121-7 in press

[3] Choi, W. S., & Kim, H. S. (2011). Health Claims for Food Products Advertised on Korean Television and Their Regulation: A Content Analysis. Journal of Health Communication. https://doi.org/10.1080/10810730.2011.561911 in press

[4] Nocella, G., & Kennedy, O. B. (2012). Food health claims – What consumers understand. Food Policy, 37(5), 571–580. https://doi.org/10.1016/j.foodpol.2012.06.001 in press

[5] Lalor, F., & Wall, P. D. (2011). Health claims regulations. British Food Journal, 113(2), 298–313. https://doi.org/10.1108/00070701111105358

[6] Brookes, G. (2010). Economic Impact Assessment of the European Union (EU)'s Nutrition & Health Claims Regulation on the EU food supplement sector and market. European Health Claims Alliance (ECHA). in press

[7] Niva, M. (2007). 'All foods affect health': Understandings of functional foods and healthy eating among health-oriented Finns. Appetite, 48(3), 384–393. https://doi.org/10.1016/j.appet.2006.10.006 in press

[8] Li, L., Wang, P., Yan, J., Wang, Y., Li, S., Jiang, J., Sun, Z., Tang, B., Chang, T., Wang, S., & Liu, Y. (2020). Real-world data medical knowledge graph: construction and applications. Artificial Intelligence in Medicine, 103, 101817. https://doi.org/10.1016/j.artmed.2020.101817 in press

[9] Schulte, L. (2022). Knowledge Extraction from EU Food Health Claim Data using NLP, Maastricht University unpublished

[10] Sun, H., Wang, S., Zhu, W., He, Y., Zhang, S., Xu, X., Hou, L., Li, J., Ni, Y., & Xie, G. (2020). Medical Knowledge Graph to Enhance Fraud, Waste, and Abuse Detection on Claim Data: Model Development and Performance Evaluation. JMIR Medical Informatics, 8(7), e17653. https://doi.org/10.2196/17653 in press

[11] Straub, L., Gagne, J. J., Maro, J. C., Nguyen, M., Beaulieu, N., Brown, J. R., Kennedy, A., Johnson, M., Wright, A., Zhou, L., & Wang, S. V. (2019). Evaluation of Use of Technologies to Facilitate Medical Chart Review. Drug Safety, 42(9), 1071–1080. https://doi.org/10.1007/s40264-019-00838-x in press

[12] European Food Safety Authority (2022, November 29). Eu Register on nutrition and health claims [Data set]. https://food.ec.europa.eu/safety/labelling-and-nutrition/nutrition-and-health-claims/eu-register-health-claims

[13] Codex Alimentarius : https://www.fao.org/fao-who-codexalimentarius/en/

[14] Bröring, S., Khedkar, S., & Ciliberti, S. (2017). Reviewing the Nutrition and Health Claims Regulation (EC) No. 1924/2006: What do we know about its challenges and potential impact on innovation? International Journal of Food Sciences and Nutrition, 68(1), 1–9. https://doi.org/10.1080/09637486.2016.1212816 in press

[15] Code : https://github.com/AlexisMayet/ThesisCode.git