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 experts in substantiation processes using NLP, namely 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. Achieving this goal requires a Named Entity Recognition (NER) model, a Relation Extraction (RE) model, and a Knowledge Graph creation process. This project focuses on the development of a Relation Extraction tool. To this end, a rule-based data annotation method is provided, as the annotated dataset originally available was small in size and contained a strong label imbalance. Also, a baseline machine learning approach is employed, followed by a state-of-the-art deep-learning approach, making use of the BERT model. These models aim at classifying the relation between a food entity and a phenotype entity within specific labels. The approaches employed are accompanied by several limitations, such as a small data size, a label imbalance in the data set leading to bias in the training, and 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 turn can improve regulatory and market conditions for manufacturers and consumers.**

I. Introduction

Diet plays a fundamental role in the development of diseases and health conditions in humans. It constitutes the individual's needs for energy and nutrients to develop and maintain their 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 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 has introduced the Food and Drug Administration (FDA), Japan has introduced the Nutrition Improvement Laws, and the European Union (EU) has 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 use of health claims in the food industry, more specifically in 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 referred to as a substantiation 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 a significant 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 the 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 of 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 process of reviewing food health claims at the EFSA? To answer this, a literature review is conducted first. 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 economic impact of regulations. Studies that employ similar methodologies, 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, which 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 refer to psychological and behavioral functions; and claims that relate to slimming, weight control, and appetite control. Claims that fall under Article 13.5 are any claims that are based on newly developed scientific evidence or that require access to proprietary data. Article 14 refers to claims that concern the reduction of a disease risk and any claims that concern children's health. Article 13 claims are only required to present sufficient references to relevant existing scientific evidence that supports the claim, while Articles 13.5 and 14 require 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 claims that are substantiated with a high level of scientific evidence, meaning they have high credibility. Qualified claims refer to claims that are substantiated with a low level of scientific evidence, having a lower credibility, and make use of emerging knowledge. [5] Claims submitted in the EU are handled by the European Commission at first. Claims that fall under Article 13 must reference sufficient scientific evidence that : emphasizes the need for evidence linking the food to a beneficial effect on human health; recognizes the usefulness of markers or intermediate effects in the biological process; and emphasizes that the effects are both biologically and statistically significant and meaningful. Claims that fall under Article 13.5 or Article 14 must contain scientific evidence obtained specifically for the product in question. The European Commission submits the claims to the EFSA, which conducts an evaluation based on the following questions : Has the product been sufficiently characterized? Is the claim beneficial for health? Has a cause-and-effect relationship been established? [5]. If the claim answers positively to these questions, it is 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 that states, suggests, or implies that a food has certain characteristics relating to its origin, nutritional properties, nature, production, processing, composition, or any other quality” [13]. From this definition, two types of claims arise : nutritional claims, which refer to any claim that states, suggests, or implies that the concerned food has a particular beneficial nutritional property due to the energy, nutrients, or other substances provided or not; and health claims, which refer to any claim that 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 communicating 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 at ensuring that ”the average consumer” can effectively understand the claim. Reference [4] published a paper on this subject in 2011, analyzing 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 resource 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 beliefs and how those relate to the claim. It follows 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, 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 analyzed the impacts of the 1924/2006 regulations on market innovation and found that they pose an important challenge to companies. Different factors can explain this : namely, as the regulation came into effect, companies were incentivized to revisit the ingredient 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 become less attractive 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, so that a sensical, useful, and structured presentation of the data can help obtain knowledge easier. Multiple approaches and methods exist for this, but those most 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 labeled to represent entities and edges are labeled to represent relations between entities. 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 tasks. 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 that 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 Electronic Medical Records (EMR) of a particular hospital in China. Their approach consists 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 for 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

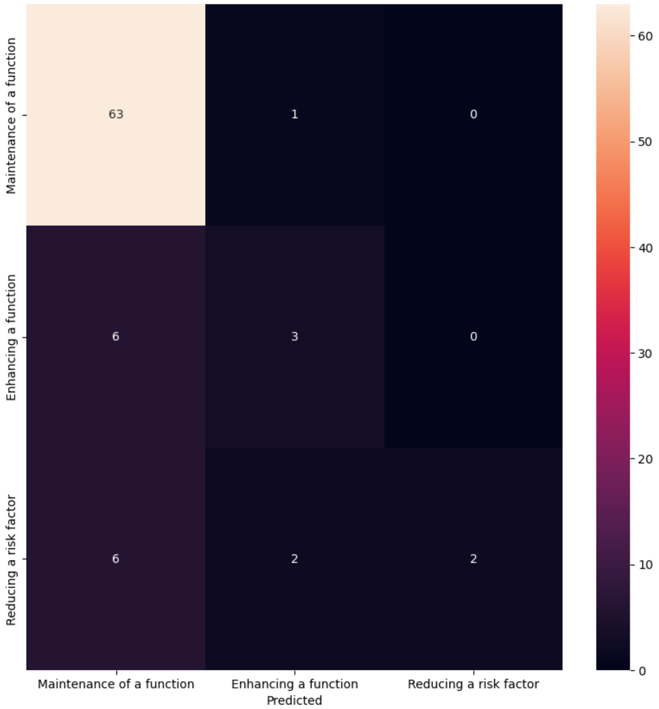
This research uses data from the EFSA, available online. It consists of 2563 claims, which relate to 1216 different foods and 523 different health aspects. Each claim contains information such as the food term, the health term, the claim, whether it was authorized or not, and other attributes less relevant 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 the 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. These authorized claims are annotated and are therefore initially used for training and evaluating the algorithms concerned. As the aim of this project is to create a tool that automatically creates knowledge graphs of food health claims, there are three main tasks at hand. The first task is to develop a NER model that will identify food entities or phenotype entities. 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-health relationships are identified: ”maintenance of a function”, ”enhancement of a function” and ”reduction of a risk factor”. These relationships are already annotated in the authorized claims dataset, therefore, the only pre-processing concerns entities of food and phenotype. Indeed, these entities are given in the columns ‘Food’ and ‘Phenotype’, but the values do not necessarily match 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 identifying 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 classifying 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 context, it is assumed 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. This problem is tackled with two distinct methods. The baseline method consists of a traditional machine learning approach, while the other method consists of making use of state-of-the-art deep learning technologies. A weak supervision approach for increasing the size of the dataset and decreasing the label imbalance is also presented.

A. Baseline method : Machine Learning

For this approach, entities in the text are replaced by an entity tag, and the sentence is vectorized using a TF-IDF vetorizer. The resulting vectors are then used to train a classification model using the library Scikit learn [15]. Different classification algorithms are considered for this task, namely a Logistic Regression classifier, a Mutlinomial Naïve-Bayes classifier, and a Linear Support Vector Classifier. The models were evaluated using K-fold cross-evaluation, with $K = 10$. The metrics employed to evaluate and select the models are accuracy, precision, and 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. However, it has relatively poor precision and recall, meaning it is quite limited in discriminating between different classes. This can be attributed to the training data used, which only consists of 235 entries, 80% of which are labeled ”Maintenance of a function”. This label imbalance creates a bias in the model, which can be seen in the confusion matrix below :



When summing the columns up, it is noticed that the model predicts the label ”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 others. These limitations can be addressed with a larger annotated dataset with a smaller imbalance in labels. The following subsection describes how this task is approached.

B. Rule-based annotation

The 235 data entries of authorized claims used in the previous subsection were manually annotated by domain experts. However, the EFSA dataset also contains a number of unauthorized claims, that were not manually annotated. Due to time and resource constraints, manual annotation of these data entries is not considered, instead, a rule-based annotation approach is used. Indeed, the library Snorkel [16] offers a weak supervision approach to data annotation. Through the manual identification of general heuristics that can somewhat identify the label of a data entry. The library then employs these heuristics to annotate data, resolving conflicts between the rules when possible and abstaining from annotating when necessary. Upon inspection of the labeled dataset, rules are inferred. Claims containing the words ”maintenance” or ”normal” are associated with the label ”Maintenance of a function. Claims containing the words ”improvement”, ”acceleration”, ”increase” and their declensions are associated with the label ”enhancing a function”. Claims containing the words ”protection”, ”reduction”, ”neutralization,” and their declensions are associated with the label ”Reducing a risk factor”.

C. State-of-the-art approach : Deep Learning

As seen in the literature review, the state-of-the-art approach for NLP tasks consists of employing deep learning models. This is done through the usage and fine tuning of existing pre-trained language models given the task and the available data. Namely, the use of the BERT model [17] is investigated. BERT stands for Bidirectional Encoder Representations from Transformers, and is pre-trained on a book corpus and a Wikipedia corpus. The base model contains around 110 million parameters and is used for a wide range of NLP tasks in the scientific community. The model is instantiated, and fine-tuned using a train-validation-test split of proportions 80-10-10 respectively.

IV. Results

A. Rule-based annotation

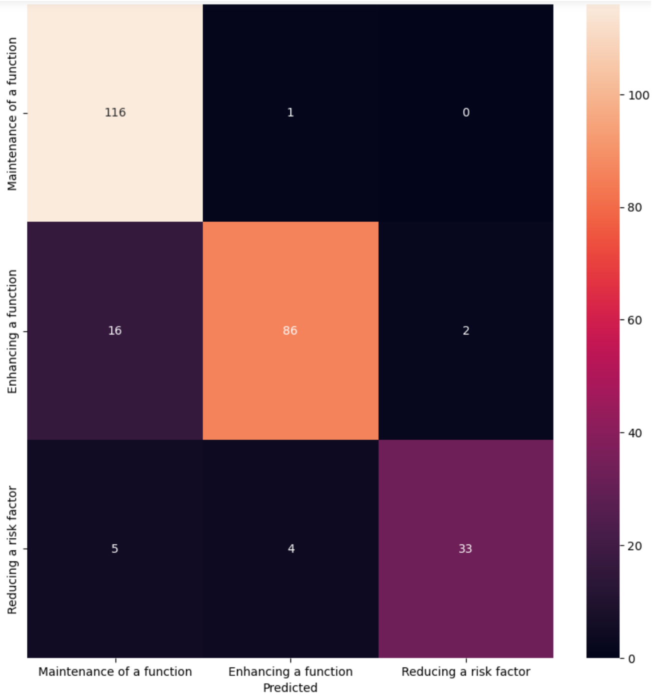
The rules manually derived from the inspection of the labeled data set are used to label data with the Snorkel library [16]. Out of 1491 unauthorized claims from the EFSA dataset, applying the set of rules resulted in 544 new labels. The program abstained from annotating other entries due to conflicts and inconsistencies. From these newly labeled entries, 205 were labeled as ”maintenance of a function”. 255 entries were labeled as ”enhancing a function”, and 84 were labeled as ”reducing a risk factor”. These claim-label pairs are merged with existing annotated data. Initially, there were 235 annotated entries, and adding the 544 newly annotated entries, bringing the total size of the dataset to 779 entries, which represents a $218,72\%$ increase in data size. The resulting set is composed of around 48% of ”Maintenance of a function” labels, around 38% of ”Enhancing a function label” and around 14% of ”Reducing a risk factor” labels. Therefore, the impacts of the limitations due to label imbalance are decreased.

B. Baseline method : Machine Learning

After the increase in data size, the machine learning approach is reinvestigated : Linear Support Vector Machine, Multinomial Naïve Bayes and Logistic Regression models are trained and evaluated using a 10-fold cross-evaluation process. Below is a table that describes the performance of these models in this process.

Metric	LinearSVC	MultinomialNB	Logistic Regression
Accuracy :	0.867928	0.801189	0.851928
Precision :	0.899914	0.842384	0.888604
Recall :	0.837566	0.742423	0.799692

Again, it is observed that the Linear Support Vector Machine model outperforms the other models on all metrics. It is also observed that the increase in data size for cross validation resulted in a 5.1261% increase in accuracy, a $31,3101\%$ increase in precision, and a 33.3667% increase in recall in this model. These results are quite significant, as the precision and recall scores have considerably increased. This means that the model is now more efficient at discriminating between labels and has a more varied output. This can be observed through the confusion matrix below :



Again, by summing up the columns, we observe that out of 263 testing entries, the model predicts the label ”Maintenance of a function” 137 times, the label ”Enhancing a function” 91 times, and the label ”Reducing a risk factor” 35 times. This shows how the model achieves better precision and recall as it makes more varied predictions than when trained on the original labeled data.

C. State-of-the-art approach : Deep Learning

The model was meant to be trained for at least 8 epochs with a learning rate of $1 * 10^-6$. However, due to time and implementation constraints, the training process was not able to terminate, as it was interrupted after 6 epochs. In this shorter than planned training process, the training accuracy iteratively improved and reached up to $85,6\%$. This almost matches the testing performance of the Linear SVC model, but it is believed that this model could reach a training accuracy above 90% with sufficient training. It is also estimated that such a model would outperform the LinearSVC model on testing metrics such as accuracy, precision, and recall. Efforts are currently being pursued to successfully train this model for longer than what has been achieved.

V. Discussion

Initially, this project was subject to a strong limitation that came from the small size of the available data and the imbalance in data labels. Indeed, the original annotated dataset only contained 235 entries, 78% of which were labeled as ”Maintenance of a function”, 12% of which were labeled ”Enhancing a function” represents of entries and 10% of which were labeled ”Reducing a risk factor”. This imbalance created an important bias in the initial machine learning algorithms employed. This limitation is addressed through an increase in the size of annotated data. The annotation process is done through Snorkel, a library that uses rule-based learning to annotate data entries. The total size of the data is now 779 entries, which represents a $218,72\%$ increase in data size. The resulting set is composed of around 48% of ”Maintenance of a function” labels, around 38% of ”Enhancing a function label” and around 14% of ”Reducing a risk factor” labels. This means that the original label imbalance is strongly reduced. However, another limitation arises from employing this rule-based approach to annotate data : indeed, contrary to data annotated by human experts, there is no guarantee for the correctness of the labels.

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 improved or similar performance that is less dependent on initial assumptions. Simply, increased annotated data size will lead to more robust training, which can achieve better performance. Further reducing the imbalance in the data label distribution can also help the model perform better and eliminate some of the bias.

Finally, it can be argued that a limitation that comes with employing deep learning lies in the difficulty of inspecting the classification process, as BERT is a black-box algorithm, which has poor explainability and could restrict the validity of its predictions, specifically if those are used in the legal domain. Another significant limitation of this project arises from the unfinished training process for the deep-learning model. This also means that the model was not evaluated.

VI. Conclusion

In conclusion, this project aimed to develop a tool that automates the creation of knowledge graphs 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 gives 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 for improving efficiency and assisting domain experts. The methodology involved the utilization of data from the EFSA, specifically the authorized food health claims dataset, as well as additional annotated data obtained through a rule-based annotation process developed through the Snorkel library. 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. For this, baseline machine learning models and trained, evaluated and compared. A state-of-the-art deep-learning approach is also described. The use of these models enables the 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, which were small in number and contained a label imbalance. Further research could explore the inclusion of a wider range of claims. Additionally, several contextual assumptions are made to simplify the relation extraction process. The assumptions are limitaitons as they reduce to scope of application of the models provided. Furthermore, the models can be further enhanced to improve their accuracy and performance by performing additional training.

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] Pedregosa, F. (2012, January 2). Scikit-learn: Machine Learning in Python. arXiv.org. https://arxiv.org/abs/1201.0490

[16] Ratner, A., Bach, S., Ehrenberg, H. R., Fries, J. A., Wu, S., & Ré, C. (2017). Snorkel. Proceedings of the VLDB Endowment, 11(3), 269–282. https://doi.org/10.14778/3157794.3157797

[17] Devlin, J. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. https://arxiv.org/abs/1810.04805

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