Contents

List of Tables

List of Figures

Abstract

There is a need for developers of consumer nutrition applications to accumulate food-related data. However, studies about the methods to assess the quality of food related data are scarce. This study lays a foundation for further research into quality of such data. The central part of the research is a way to solve merge problem that occurs when trying to merge multiple datasets into one with minimal number of duplicates. This study solves it using lexical similarity function. This study also proposes a range of metrics that can be used to gauge quality of a dataset and demonstrates their results.

Chapter 1

Introduction

Nowadays more and more problems are being solved by application of machine learning approaches. As the number of problems that seem to have their solution in machine learning increases so does the demand for data as without data machine learning approaches are ineffective. However, as many studies show [1]–[5], not any data is suitable for machine learning. Data of insufficient quality can introduce many problems, such as different biases [4], [5] or decrease in accuracy or precision [1]–[3]. To avoid such problems data science engineers commonly spend a great deal of effort to make sure that the data that they are using is of best possible quality. Studies on the quality of particular open datasets are not uncommon [6]. In addition, some companies even construct a complicated pipeline for particular data quality evaluation [7], [8]. While some researchers try to construct a generic approach that would work on different data [9].

The biggest challenge for data quality evaluation so far has been the intrinsic problem of quality – it's subjectiveness. Studies point out that, despite the existence of somewhat universal quality metrics e.g. the size of the dataset or labeling consistency [9], the quality of data is ultimately defined by the approach that will be used to process said data. As such it is a lot more reliable and efficient to evaluate data quality in a particular field

One of the fields that has seen an increase in interest because of machine learning is consumer nutrition. Consumer nutrition is a field that is concerned with what people eat and what are the implications of their diet. Despite many approaches existing in this field, one of the approaches has gained significant popularity amongst consumers. The approach involves analyzing the dishes or food items, evaluating their nutritional contents and making further recommendations aimed at making the diet more varied, healthy and balanced. There are various methods that can be used to achieve this, however there is one thing that they have in common – the need for data. Without "knowing" enough about the dishes it is quite obvious that one could not be efficient at searching for them or suggesting them.

The goal of this study is to suggest ways how development team of a consumer nutrition application could analyze datasets of food items to make sure that the data is of sufficient quality and is sufficiently new to include it into the data pool for the application. We achieve that goal by reviewing which methods are used to track quality in various fields and by noting down which methods are used in consumer nutrition applications to then define the standards for the data and suggest the metrics by which quality could be assessed.

Chapter 2

Literature Review

This chapter presents an overview of existing metrics of quality that are applied in various domains. When reviewing literature, we had several questions in mind. First, we needed to understand which methods are commonly used in consumer nutrition space for dish identification and recommendations. Second, we wanted to know what metrics are applied to datasets in general irrespective of their domains. Third, we wanted to know what specific concerns exist is some relevant specific domains. Fourth, we wanted to know whether there are some pre-existing methods for assessing dataset quality for accumulation of food related data. Answering these questions leads to the answer to the main question: which metrics could be applied to assess quality of datasets for consumer nutrition applications. We have searched for literature via Google Scholar using the following list of keywords: "Dataset, Quality, Dataset nutrition, Consumer Nutrition, Machine Learning". In total we have processed over 20 different articles.

I Methods Used in Consumer Nutrition

Since quality attributes heavily depend on the methods used in processing data it is important to learn what methods are used in the research field, for our case it is consumer nutrition. We focus on non-invasive approaches only.

A. Dish Identification

One of the common features of consumer nutrition applications is dish identification. This feature contains two main subproblems. The first one is single dish identification and it aims to solve arguably the more common case when a plate contains a single specific dish. However in Asia it is rather common to encounter multiple dishes plated together and this is the other subproblem – mixed dish identification.

1) Single dish identification: The single dish identification task has quite a diversity in terms of approaches. The main approach is leveraging power of deep networks made for object recognition [10], [11]. Other quite interesting proposals involve recognising dishes by history data [citation needed] or even by position utilizing closest restaurant menus [12], [13]. Some solutions focus on real-time dish recognition [14].

2) Mixed dish identification: Mixed dish identification, despite being arguably a more rare case than single dish identification, has also received a fair share of research attention. The most common methods utilized for this problem are R-CNNs and YOLO networks [15]–[17].

3) Text based dish retrieval: There is also a possibility for applications to utilize text based methods to find relevant dishes. Despite there being not many studies on this subject as these methods are quite common across many fields, the more common methods that can be used for such search are:

- Deep learning algorithms
- Clustering techniques (K-means)
- Various Natural Language Processing techniques

B. Nutrition Recommendation Systems

One of the most important aspects of consumer nutrition applications is, of course, recommendation systems. The goals of recommendation systems here commonly is to suggest dishes based on previous history of the user. According to systematic literature review [18], the most common methods used for dish recommendation systems are:

- Rule based systems
- Deep learning algorithms
- Genetic algorithms
- Decision Trees
- Domain ontology
- Clustering techniques (K-means)
- Various Natural Language Processing techniques

C. Implications from methods in Consumer Nutrition

There are various methods that are utilized for both dish identification and for recommendation system part of applications. The dish searching methods imply the necessity to have some metrics for quality of images, while text based searching implies the need to process the text as well. The variety of the recommendation system methods implies that for this study it is best to lay the overall groundwork for the quality instead of focusing down on a few particular methods. All of the approaches tend to use the same overall data: dish names, ingredients lists, nutritional information and maybe some pictures of the dish. This is the data that the study will be focused on.

II General quality metrics for dataset

It is a well-known fact that data quality influences any solution that will be built using data. Therefore, a lot of research was dedicated to some of the more common quality issues.

A. Data coverage

One of the possible quality issues is insufficient data coverage. Most of the datasets cannot account for all of the diversity that one can encounter in a particular issue, hence they are collected to be as representative of the diversity as possible. The quantification for the degree of representativeness is called coverage. Insufficient coverage can lead to various issues, for example [19] describe a case nicknamed “google gorilla” where an image recognition algorithm was trained on data that did not contain enough images of dark-skinned people which lead to misidentification of some African Americans as gorillas. In order to quantify it authors propose an algorithm based on Maximal Covered Patterns which can highlight underrepresented combinations of parameters in the data. They have demonstrated the correctness of their approach on COMPAS dataset. However, the algorithm's runtime drastically increases with the number of dimensions in the dataset showing an exponential trend, which may complicate its use in some of the bigger datasets. The other common way to test coverage is to measure coverage not based on the dataset, but on the model by, for example, utilizing metamorphic testing [20], [21]. Metamorphic testing was proposed as early as 1998 and is based on the idea that we could test an application by defining rough relations between input and the output. Even though it is rather hard to define the exact relations one could define some of the general rules e.g., with the decrease of particular samples the accuracy should decrease. That is exactly the approach that some of the researchers took and have found some degree of success [21]. However, this approach is also limited by the fact that the rules have to be defined manually and although there have been some attempts to produce rules automatically, they have their own limitations. For instance, authors of [20] have found an approach to generate effective rules for SVM classifiers, but it is unclear whether the same approach could be applied to other models.

B. Completeness

Another type of issue that commonly plagues data is completeness issues. Completeness shows how much data describes an object of interest. Issues that may make datasets less complete involve missing or null values. Incomplete datasets may pose a challenge as completing datasets can be fairly difficult and using an incomplete dataset may decrease accuracy of the models that are using it. However, since completeness may depend heavily on the domain and the approach, metrics for measuring it vary.

C. Consistency

One more of the important issues is inconsistency. Consistent data should not contain duplicate records or contain surprising values in records in a sense of constraints. For instance, when dealing with a field called “date” one usually does not expect to find strings amongst numerical date values. Such irregularities are usually dealt with before constructing a model, however, when missed, they can cause various issues. Therefore, it would be a good practice to check a dataset against expected constraints e.g., constraints on format, type, duplicates or even in terms of labels.

D. Provenance

An additional issue that could arise with the data is its provenance. For instance, data from an untrustworthy source may contain hidden issues or biases that will be hard to detect. Questionable legality of the data may produce other trouble. In addition, if a researcher themselves produces a dataset, they need to put their processing pipeline under scrutiny, so as not to change the meaning of data. There have been several approaches proposed for tracking lineage [3, 6, 8] however tracking source and general trustworthiness of data remains a challenging and mostly manual task.

III Concerns specific to domains

Some of the domains may possess additional requirements to the data apart from the general ones. Here we limit the scope to only include some of the more relevant ones.

A. Computer vision domain

It is common knowledge that computer vision has many challenges and a lot of its challenges find their reflection in the datasets. For example, here are some of the challenges of image classification, a subfield of computer vision: Intra-class variation, scale variation, view-point variation, illumination variation, occlusion and background clutter. Intra-class variation refers to intrinsic variation between objects in the same class, for example butterflies that can have different colors or forms of wings. Scale variation is more specific and focuses on the same object having different sizes. View-port variation is about the same object being seen from different angles or positions. Illumination variation is when objects can exist under different illumination conditions and therefore have different color intensity. Occlusion happens when the object can be partially obstructed and background clutter describes a situation when the object can coexist on the picture with other (maybe irrelevant) objects. Some of these challenges can be mitigated via choice of an approach, for example, some of the techniques can be illumination- or scale-invariant. Other problems can be avoided by choosing a narrower scope e.g., view-port variation becomes irrelevant when objects are placed in a consistent position before the camera. However, all of these have to be a conscious decision pertaining to a specific problem and must be taken into account when evaluating a particular dataset.

B. Consumer nutrition applications domain

In consumer nutrition applications domain the main data-related concerns are usually concerns that relate to general lack of relevant, actual, accurate data. As there are only a handful of truly reputable sources of food-related information, a lot of data may be inaccurate or out-of-date by the time it is found. As such there are multiple companies dedicated to filtering and accumulating food-related data, however the data that such companies accumulate is commonly not open to use.

IV Quality metrics in use for food-related datasets

It has to be noted that there is a severe lack of research into quality of food related datasets. The keywords that one could reasonably expect to be in use for such studies are occupied by other fields. Overall, research into quality of various datasets is quite a recent trend and we believe that currently there is almost no research into the food-related data or at the very least it is incredibly hard to find. We were unable to find any research into food data related quality metrics in publicly searchable databases.

V Summary

To summarize, we have found some answers to our questions although many of them tend to be quite complex in their nature.

To begin with, there are various metrics that can be usually measured for just about any dataset such as coverage, completeness, consistency, and provenance however the approaches to measure them will depend on the particular field for which the dataset is used.

There are numerous concerns that can be raised for the domain of computer vision, however sometimes these concerns can be sidestepped or mitigated by the choice of an algorithm and as such we may need to provide some general quality metrics for the images such as whether the image actually shows something and is not just static noise, but there is no particular need to get in-depth on the images as some of the undesirable features may actually be necessary to simulate real-world scenarios.

As for the approaches in the consumer nutrition applications, there are lots of different methods that can be utilized. This makes it quite challenging to pick particular quality attributes to analyze. Despite this, it is very clear what kinds of data the methods need: they all do require food names, ingredients lists, nutritional information and this can be used as a basis for fundamental analysis.

As there is almost no research done on the topic, or at least it is difficult to find, we have designed our own quality metrics that can be used as a foundation for further research into this topic.

Chapter 3

Methodology

In this chapter we describe the steps that we have taken in order to approach the creation of some metrics that could help assess the data quality in the field of consumer nutrition and present some of the difficulties we have encountered and solutions that we have tried to cope with said difficulties.

As was stated previously, our end goal is to produce a metric or a set of metrics that can be measured on the datasets in the field of consumer nutrition. These metrics should be useful in assessment of quality for a given dataset and therefore should produce results that a human knowing about the quality of a given dataset could expect.

I Dataset formatting rules

First and foremost the most fundamental problem of the metrics has to be resolved. Whichever metric would be constructed they would all have the same downfall. Applying any metric to a dataset of unknown format is infeasible. Any metric will require some assumptions to be made regarding the dataset contents and/or format. For instance, if we decide to measure the size of images in the dataset it will require assumption that images exist in the dataset. Moreover, even the simplest ideas such as measuring the number of missing values requires an assumption about how the values are stored – are they stored in rows or in columns. The stronger the metric will be, the more assumptions about dataset formatting it will require and the stronger the assumptions will have to be. However formatting does not exist in a vacuum either. A lot of decisions on the formatting are made based on ease of further use. Considering this, we have identified existing types of nutrition datasets and possible use cases for them in order to come up with some standards that then could be used to apply some metrics.

We assume that the use case for the datasets is to create an application that will allow users to input the dish (either by taking a photo or by typing in a name) to then get some further dish recommendations along with some general nutrition advice in terms of macro- and micro- nutrients. Considering this use case we limit the datasets to the following information:

1. Dish name
2. Dish pictures
3. Dish nutrients
4. Dish ingredients

Please note that despite this, datasets can and commonly do contain more information. However we cannot take into account all of the information that could be contained within the dataset so we limit the scope of research to the most common and probably the most impactful information. Additionally we are not taking into account personalized user history as that is commonly built during the application lifecycle and usually is not available publicly sometimes due to legal or ethical concerns.

Then we define some standards. We define them based on the more common formats of the existing datasets. We require the dataset to be split into multiple different tables each containing dish name as the first column and other information of one type starting from the second column. For instance we could get a table with dish name and pictures or with dish name and nutrients, but there should not be a table with only pictures and nutrients or with name pictures and nutrients simultaneously. However several different nutrients can exist in one table. Other requirements to the data can be seen in Tab. I.

TABLE I
Requirements on formatting dish parameters

Parameter	**Requirements**
Dish name	Written in English
Dish pictures	Is in .jpeg format, is is separate directory per food item/class
Dish nutrients	Name is in English, value provided for serving of 100g, measured in micrograms, every nutrient is in a separate column
Dish ingredients	Written in English, ingredients are in a separate column, written with commas between ingredients, are alphabetically sorted

II Quality metrics

After the standards have been defined some metrics can be measured over the datasets. Previous research which examined the problems of so-called "dirty data" have proposed multiple frameworks for systematizing the requirements to the data. We follow the [22] article which proposes several quality dimensions and rule categories. We improved on that by slightly modifying both dimensions and categories to fit our purpose. Most notable of the changes is that we have removed Business Entity rules. All of the datasets have quite different formatting and the formatting transformations have to be done manually, so Business Entity rules in their original sense are not really applicable in our case. Additionally we have introduced Data Coverage rules as it is quite an important aspect of quality and it can produce quite a few useful metrics. Our modified quality dimensions and rule categories can be seen in Tab. II.

TABLE II
Quality dimensions and rules

Quality dimension	**Rule category**
Business attribute rules	Data formatting rules Data lineage indicators
Data coverage rules	Data variety indicators Data volume indicators Data recency indicators
Data validity rules	Data completeness rules Data correctness rules Data uniqueness rules Data consistency rules

TABLE II
Quality dimensions and rules

Quality dimension	**Rule category**
	Data accuracy rules Data precision rules
Data dependency rules	Attribute dependency rules

With the help of quality dimensions and rule categories we introduce our quality metrics specific to the domain of consumer nutrition. These metrics and their classification can be viewed in Tab. III.

TABLE III
Quality rules and metrics

Rule Category	**Data Quality Metric**
Data volume	Number of rows
Data variety	Percentage of novel records Number of pictures per dish Entropy of pictures
Data uniqueness	Percentage of duplicate dishes
Data accuracy	Ingredient consistency with ground truth Nutrient consistency with the ground truth
Data precision	Image quality Picture resolution
Data formatting	Uniformity of format
Data recency	Recency of data
Data completeness	Percentage of missing values in nutrients' columns

III Collecting Datasets on Nutrition

In order to assess the efficacy of the aforementioned metrics we decided to calculate them for some datasets of varying quality levels. That way we could see whether the metrics can be used to distinguish some of the lower quality datasets from the datasets of higher quality. However before this we had to collect some datasets to utilize them for testing.

We collected only publicly available datasets that were relevant to the field of consumer nutrition. We collected them from several sources, including, but not limited to: Kaggle, Data.World, USDA website, OpenFoodFacts and others. We have to note that despite multiple researchers repeatedly mentioning some of the datasets in their papers we have not managed to find publicly available sources for some of them. For instance, we could not find: ChinFood1000, Food-475, ChineseFoodNet, FooDI-ML. Despite this, we have managed to collect around 12 different datasets overall.

Further explanation of this step is available in section 4.1.

IV Merge problem

Once we collected the datasets we had to calculate the metrics. Some of them are quite straightforward in calculation. However around a third of the metrics are dependent on what is known as a Merge Problem. Merge problem is what happens when there is a need to understand that two records in one or multiple datasets actually describe the same entity. This understanding is highly domain-dependent as the notion of entities and ways to describe them vary wildly between domains. This reliance on the domain of the data is what makes the merge problem quite troublesome. There already exists some research both for the general case and for food data. For instance some researchers [23] propose a method that allows for a quite fast and reliable way to merge dataset records if a similarity function for two records is defined. Ways to define a similarity function for food are covered by [24] where researchers have tested word2vec and GloVe models against their carefully crafted lexical similarity function. During their testing their lexical similarity function seemed to perform better than other tested models. Additionally some researchers have proposed the AgriBERT [25] – BERT model that has been fine-tuned on the food and agricultural data. This could theoretically produce better results than manually crafted lexical similarity. However the researchers have not tested the performance of the AgriBERT in the application to the Merge problem and it seems that the resulting model was never published, so we had no way to test it ourselves.

Chapter 4

Implementation

In this chapter we describe how exactly we have implemented quality metrics mentioned in the Methodology chapter 3.2, what we have discovered in terms of implementation difficulties and how we have dealt with them.

I Dataset collection

First we had to collect some datasets in order to experiment on them with different ways to collect metrics. We have collected only publicly available datasets that included the following information:

- Food name & Ingredient information
- Food name & Nutrient contents
- Food name & Food Images

As it was identified in the Literature Review chapter, this information is integral to enabling the most common methods that are used by nutrition applications.

While collecting datasets we have encountered an unexpected problem: while there are quite a few public datasets of various quality, some datasets that are commonly mentioned in research papers are quite hard to either find or download. Some notable datasets that we were unable to find a way to download were:

- ChinFood1000
- CHINESEFOODNET
- Food-475
- FooDI-ML

This problem may have some serious implications both for the research in the Nutrition field and for consumer applications. We explore this further in the Discussion chapter.

Otherwise we have managed to find quite a bit of data both reputable and somewhat lacking. Notable public datasets include:

- USDA National Nutrient DB
- USDA Table Cooking Yields Meat
- nutrition5k
- Open Food Facts
- Nutritionix
- Food101
- MAFood121

A concern to note here is that most datasets are relevant to first-world such as USA or Europe, which may be not desirable for world-wide applications.

II Formatting datasets

Every dataset that we have collected had to be reformatted in order to bring it to the stardard that we have defined earlier in the Tab. I. While initially we set out with a desire to automate the process of computing metrics, the process of standardization is the issue that has stopped us from automating everything further down the line. All of the reformatting had to be done manually due to the large variability of the datasets that we collected. We were performing this adaptation process using python with various data processing libraries. Adaptation of a single dataset took anywhere between 30 minutes and 6 hours depending on the initial configuration of the dataset. The processes that took the most time for writing the processing were:

- Separating & grouping necessary information
- Identifying measurement units
- Adjusting nutrient values to be per 100g serving

A. Finding Duplicates

In order to calculate some metrics a way to identify similar records is required. To identify similar records across different datasets or even in one dataset one has to find a practical way to solve merge/push problem for the records in the dataset(s).

In order to solve merge/push problem there needs to be a function on two records that decides whether the records are sufficiently similar. As described in the Methodology chapter, we propose to implement a lexical similarity function that is adapted a bit to work better with food and ingredients. This function takes a string of words as an input and if the two strings contain sufficient number of similar words returns $True$, otherwise returns $False$. Implementation of proposed lexical similarity function is similar to intersection over union, but adapted to work with sentences and it is as follows:

1. Tokenize input

   Tokenization splits sentences into words. In case of food sentence is food name with possibly appended list of ingredients.

2. Tag tokens with part of speech

   Part of speech tagging is important as different parts of speech tend to convey different information in regards to food. For instance, nouns usually contain information about what is the dish made from, for example beef, tomatoes, or lettuce. The verbs convey how the ingredients were processed to obtain the dish, for instance, boiled, fried, or baked. The adjectives convey some properties of the dish e.g. crispy, sweet.

   Taking into account this difference in meaning, for the metric to be effective it needs to give more weight to the nouns as boiled beef and boiled potato are two drastically different dishes. However, the metric still needs to consider both verbs and adjectives, although allocate a bit less of a weight to them.

3. Lemmatize nouns, verbs, adjectives

   Lemmatization is a commonly known procedure to strip away ends of words in order to bring different forms of the same words to the same basic form.

Without lemmatization the next step would become too inaccurate.

4. Compute intersection over union for nouns and verbs with adjectives

   This and next step compute IoU for nouns and verbs with adjectives separately to maintain the priority of nouns over adjectives and verbs as discussed in tagging step.

   The details of how exactly the intersection over union is computed are available in [24].

5. Multiply the two computed IoU metrics together

   This step unites the metric computed for nouns and for verbs with adjectives back into a single metric.

In our implementation have performed tokenization using $nltk$ python library. We have also used this convienient library to remove the stopwords, to perform part of speech tagging, and to perform lemmatization.

Despite providing detailed enough instructions on how to compute the metrics [24] seems to omit how exactly it is further used to obtain matches. The metric by itself is not useful unless the thresholds for it can be identified.

We performed several experiments, but we could not identify a single threshold that would work reliably in all cases. First and foremost, this metric is heavily dependent on having input lengths. Thresholds need to be drastically different for sentences of 5 words and for sentences of 20 words. Moreover, the accuracy of matching sharply decreases with the decrease in number of words as information available is also reduced heavily. This decline needed to be addressed somehow and we settled on separating thresholds for these cases and we produced two thresholds: one for matching foods by name only and the other for matching foods based on names and ingredients. This allows some flexibility as not all datasets have ingredients, but almost all of them have at least names, but in case ingredients are present it is possible to match more reliably. We have produced the thresholds experimentally on several datasets, but as the datasets vary these thresholds may not work well on all of them.

However in order to produce the thresholds we needed to apply the metric to the dataset. Considering that it is a pairwise metric, running pairwise comparisons on datasets bigger than ten thousand records is not useful as it is unlikely to be computed in reasonable time. To speed up the computation the most sane approach is to reduce the number of times is is necessary to compute the pairwise metric. One way to do it that we employed is clustering. At its best, clustering reduces hundreds of thousands of records to clusters of a few hundred records. Computing pairwise metric on a hundred records is decently inexpensive and having lots of clusters is not a problem.

We wanted to avoid setting the number of clusters by hand as it implies assumptions about the dataset, which is some cases may produce misleading results. To keep the algorithm general we employed $DBSCAN$ clustering algorithm implemented in $sklearn$ python library. In order for the clustering to function we settled on using BERT embeddings produced from the input sentences. While BERT may not "know" enough about food domain to use only it to match foods, its embeddings may be helpful enough to provide a sense of general direction that the food is in lexically. And this is more than enough for the clustering as the goal of it is to merely reduce the pool for more time-consuming calculations.

Introducing clustering into the picture however added but another variable to optimize for. Instead of the number of clusters $DBSCAN$ algorithm requires minimal distance as its variable. This is a very important variable as it practically indirectly controls the size and number of clusters. If set too low $DBSCAN$ be-

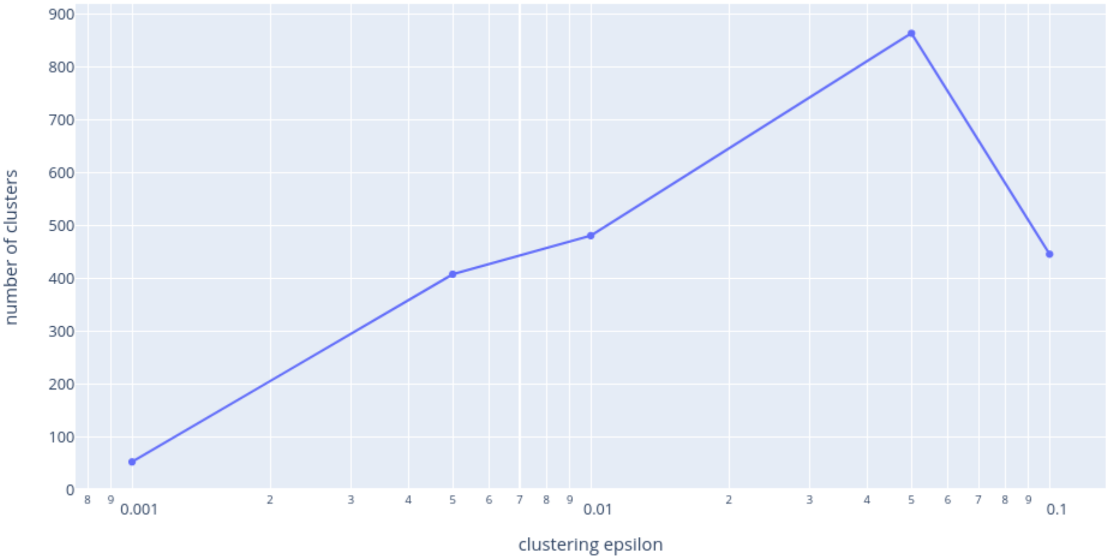

Fig. 1. Number of clusters in a sample of OpenFoodFacts dataset against clustering epsilon for DBSCAN

comes too hard to compute and if set too high it does not produce useful clusters. Completely empirically we settled on 0.01 as the value for the $DBSCAN$ ε as it produced adequate number 1 of adequately sized clusters 2 and also resulted is quite fast computations 3.

After we got clustering working and having established the ε value we have began experimenting for the threshold of lexical similarity. We sampled several similarity values carefully observing the number of resulting matches and manually checking the quality of some of them. After this procedure we settled on 0.5 for matching by names only and 0.8 for the matches using ingredients in addition to names. These values produced adequate matches that were quite hard (or impossible) to distinguish even for humans. The experiments additionally utilized the knowledge about the overall number of matches, usually checking the number and quality of matches of datasets with themselves (supposed duplicates) 4, 5.

During manual tuning of the thresholds for lexical similarity we noticed that for the aforementioned values matches were quite rigid and often having some

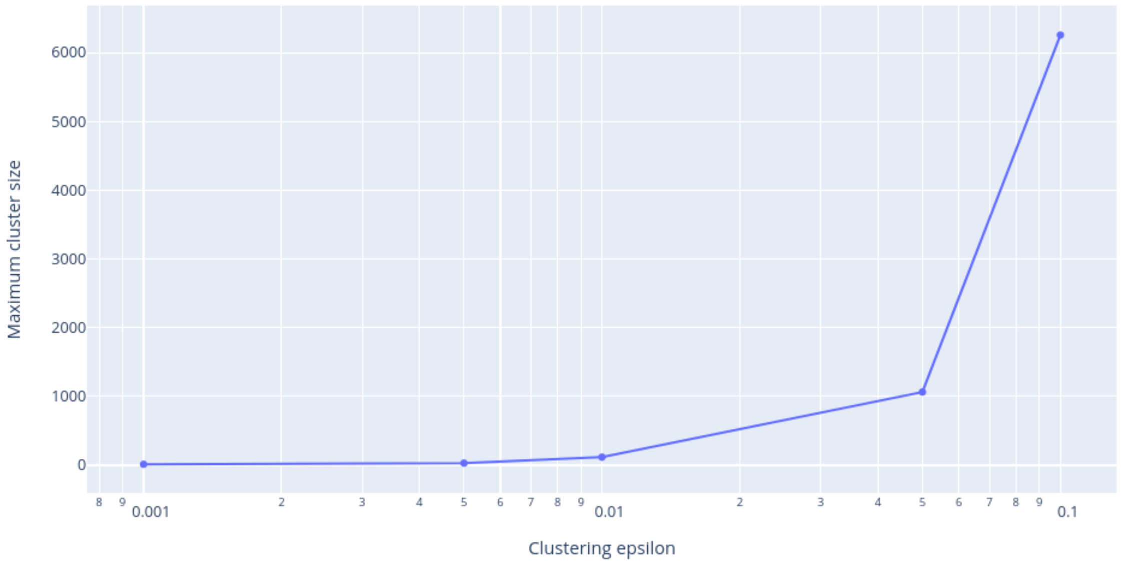

Fig. 2. Maximum size of clusters in a sample of OpenFoodFacts dataset against clustering epsilon for DBSCAN

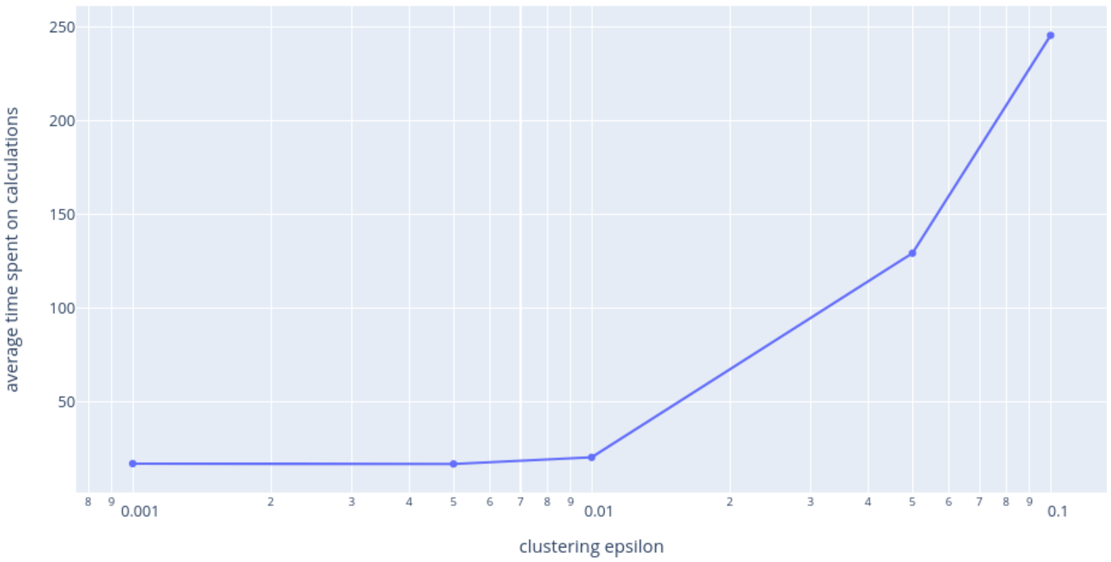

Fig. 3. Average time spent for clustering a sample of OpenFoodFacts dataset against clustering epsilon for DBSCAN

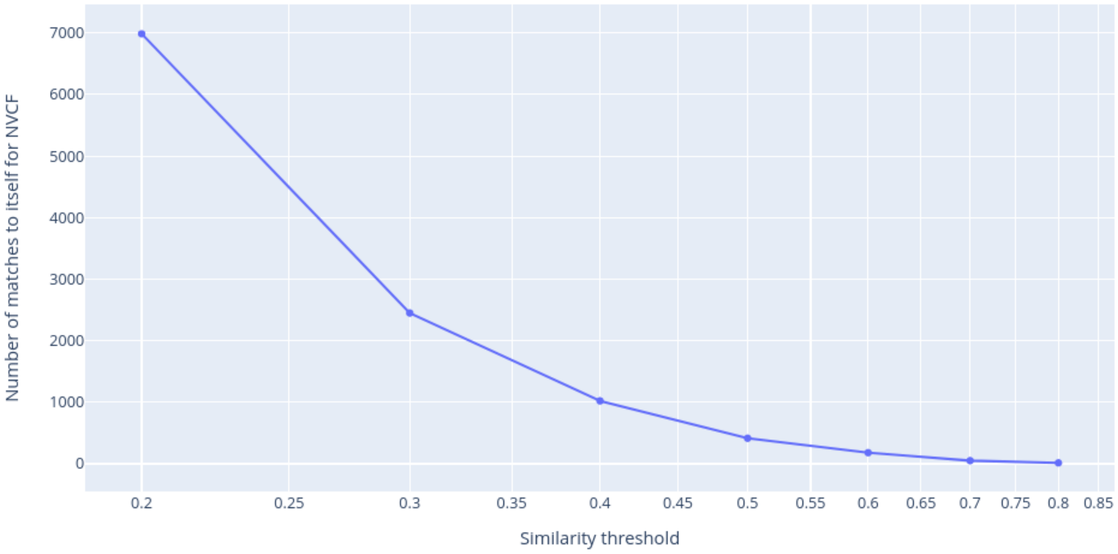

Fig. 4. Number of matches to itself (duplicates) for Nutritional values for common foods and products dataset per given similarity threshold

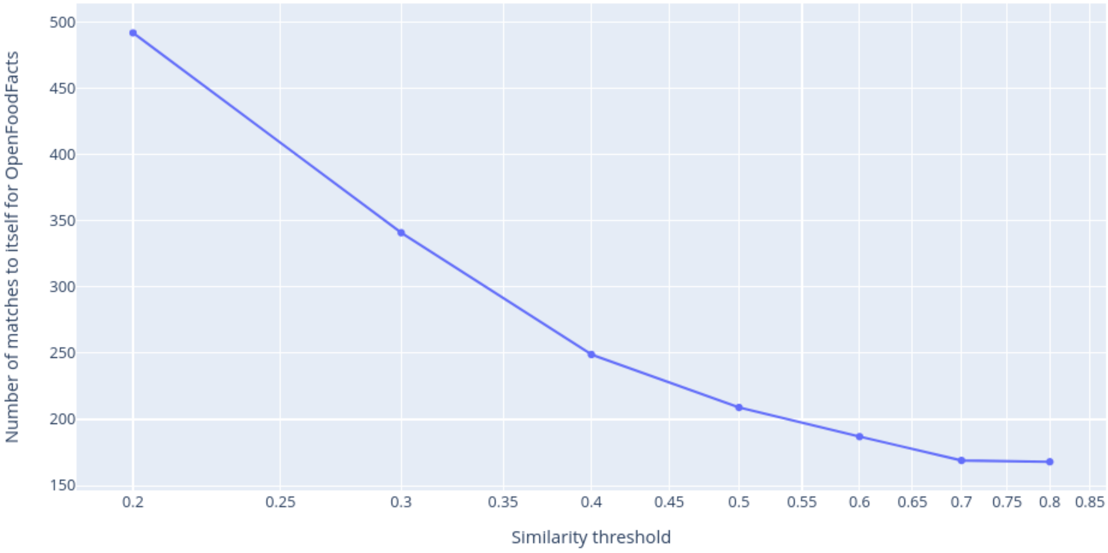

Fig. 5. Number of matches to itself (duplicates) for a sample of OpenFoodFacts dataset per given similarity threshold

strange additional adjectives or having the same ingredient with a different name was enough to throw the matcher off. We wanted to also have some more relaxed constraints in case some specific words were added to the names, descriptions or ingredients that would not be easily matched. We settled on 0.4 and 0.7 for the names and names with ingredients matching respectively.

III Calculation of metrics

The central part of our effort is, of course, metric calculation. This section will outline all of the metrics that we calculated along with how exactly we decided to do it.

A. Missing values

One of the stronger quality requirements to the most datasets is, of course, completeness. It is a first metric we implemented and it is for a reason. During experimentation it turned out that despite the fact that a dataset can contain lots of missing values they may just so happen to be concentrated in a relatively few parameters. Additionally, those few parameters, while being nice to have, may not be necessarily be essential for most purposes. Therefore it turns out to be extremely important to not only know how many values across all dataset are missing, but also which parameters are contributing to the missing values. It might be that discarding most offending parameters may make the core data that remains much more reliable.

B. Record recency

Some datasets happened to have a date for the last time each record was updated. It brought an entire dimension to the analysis as it allowed us to calculate quantiles, median and standard deviation for the dates in the dataset and see across all the dataset whether the information it contains is recent enough.

For the cases when only the date when the whole dataset was last updated was known no calculations had to be applied. Instead we just recorded the time that has passed since the latest update in years as usually it is enough to judge the recency of the dataset.

C. Number of similar records between new and known data

When accumulating data into a large data pool it is quite important not to accidentally include the same data two or more times as that can cause issues. For instance, if a particular set of dishes was duplicated several times it may skew further analysis techniques like clustering and make all of the data more biased towards this dishes.

We used lexical similarity metric described above to determine how many similar records are there in the new data when compared against known data. As was also stated above, we also differentiate between matches with strict constraints and matches with more relaxed constraints as their goals are fundamentally different. Stricter matches tend to show existence of almost identical records, while relaxed matches are more about having similar records.

D. Number of similar records in the new data

We also calculated number of similar records in the data itself so as to find possible duplicates. Looking a bit ahead, this does tend to identify foods with different serving sizes as duplicates, but in the event that standardization of all the data is performed this may be desirable.

E. Image metrics

In this section we list a number of metrics pertaining to images in the dataset and their quality attributes.

1) Images per food item: There is a lot of variation in the datasets that have food pictures. Some of them are more focused towards use in computer vision specifically, others are less focused on it and offer pictures just for completeness and modality benefits. In order to differentiate between these dataset types we calculated average the number of images per food item.

2) Image resolution: Datasets with pictures can have quite a drastic variance in their resolution. In order to help discover this we calculated average and 80th percentile of image resolutions.

3) Image entropy: Entropy, roughly speaking, is a measure of how chaotic the image is. If the image is too chaotic – it is hard to understand what is depicted, but if the image does not have any entropy at all maybe it was generated by a bad algorithm. In order to detect these potential problems we measured average entropy, top and bottom 10 percentiles of entropy.

4) Subjective image quality: Apart from entropy and resolution it is desirable to assess the overall quality of the image. While the exact definition of quality for images is quite hard to pin down, machine learning approaches are there to help. We used a pre-trained SVM wrapped in an easy-to-use python library to assess subjective image quality. For a dataset we calculated average subjective image quality and recorded the number of pictures with subjective quality less than 0. Subjective quality being less than 0 usually indicated some artifacts or even that the image itself was missing.

5) Ingredient nutrient consistency with the ground truth: Unfortunately, we were unable to create any meaningful way to check ingredient and nutrient consistency with the ground truth. This is due to several factors: first, as we obtain matches some of the matches may be false positives; second, measurement inaccuracies and ingredient variance introduce too much noise into the measurements of nutrients to be able to tell anything by looking at only two values; third, ingredients commonly have several names and for their accurate matching there needs to be a way to identify synonyms. All of these problems could potentially be remedied by having a sufficiently large and trustworthy statistic per every dish, which is currently quite hard, if not entirely impossible to find. As technology and our knowledge of the world progresses and as data accumulates there may be a way to meaningfully check these consistencies.

Chapter 5

Results and Discussion

In this chapter we demonstrate the results that we have obtained by applying metrics we have devised to datasets of varying quality. We also provide our insight into dataset quality and some directions for further research.

I Test design

In order to assess the usefulness of of metrics we needed to identify which cases we want to cover with the metrics. We have identified the following needs in terms of evaluation:

1. How do metrics behave when applied to large well-established datasets?
2. How do metrics behave when applied to small, niche datasets? Is it different from large, popular datasets?
3. Are metrics able to identify dataset subsamples?

If we are able to answer these questions with some degree of certainty, then our metrics are definitely useful in analyzing new datasets as a human using them will theoretically be able to distinguish subsamples that are known and datasets of potentially questionable quality, which is the goal of the study.

To answer these questions we have devised the following tests:

1. Apply metrics to a well-established, popular dataset
2. Apply the metrics to some smaller, less reputable datasets comparing them to the well established dataset in terms of contents and metric values
3. Apply the metrics a scrambled subsample of a popular dataset that we have checked in step 1 against the dataset itself

While the tests themselves are quite straightforward, what may not be that straightforward is which datasets to apply the metrics to. We described a way we collected the datasets in the relevant Methodology section. Due to the time costs associated with processing datasets, as every dataset has unique formatting that needs to be pre-processed into described standard form, we settled on using only a few datasets, which were the closest to our dataset standart. The datasets that we have chosen for evaluation purposes can be observed in Tab. IV.

TABLE IV
Datasets chosen for evaluation

Dataset	**Dataset Type**
Sample of OpenFoodFacts dataset	Well-established
Nutritional Values for Common foods and produce	Small, niche
Food101	Well-established, images only
MAFood121	Well-established, images only
Epicurious Scraped	Small, niche, mainly images

One can notice that we do not use full OpenFoodFacts dataset (can be shortened to OFF for convenience) instead using only an imageless sample from it. We decided to use a sample instead of the full dataset because processing and storing the entire OpenFoodFacts database is too much to handle for our small scale infrastructure. However the size difference between this sample and other datasets is still big enough for us to judge whether the metrics are effective or not.

Another thing that may warrant some explanation is the presence of image-only datasets. It turns out, that our assumption that the datasets are usually mixed is not entirely true. There are some mixed datasets, but most datasets either focus on textual component or on image component. As such we have decided that it would be best to include some datasets focused primarily on images into our evaluation.

For the comparison with a sample we used a random, uniformly sampled subsample of OpenFoodFacts. We decided to produce two variants of it. In the first variant we have removed some percentage of words from the names of foods. This is designed to test resilience of matching metric to omission or underspecification. In the second version we have modified the sample switching some percentage of words randomly between the food names of the records. This was designed to test how resilient or sensitive the matching metric is to not omission, but change in the food name. In both cases we have compared these modified samples against a clean sample from OpenFoodFacts that had slightly different scope that the modified one in order to make sure the results are resilient to additional irrelevant records.

II Results of metric application

After applying metrics to the datasets listed in the Tab. IV we have obtained results that can be seen in Tab. V – VII.

TABLE V
Calculated metrics for text datasets

Metric	**Open Food Facts**	**Nutritional Values for Common foods and produce**	**Epicurious Scraped**
Number of records	356027	8789	13501
Number of parameters	163	77	5
Number of parameters with missing values	161	1	5
Number of parameters with <1% missing values	11	76	5
Number of parameters with <10% missing values	14	76	5
Number of parameters with <20% missing values	18	76	5
Number of parameters with >80% missing values	109	0	0
Number of parameters with >95% missing values	88	0	0
Number of parameters with >99% missing values	75	0	0
Soft matches to reference	-	0 (to OFF)	0 (to OFF)

Hard matches to reference	-	0 (to OFF)	0 (to OFF)
Soft duplicates	32289	2866	0
Hard duplicates	5025	1454	0
Number of matching food parameters to reference	-	32	0
Mean time since last updated in years	6	4	2
STD for time since last updated in years	0.54	-	-

TABLE VI
Matches for subsamples of OpenFoodFacts with omission

Metrics	**OFF**	**20% removed**	**40% removed**
Number of records	80000	10000	10000
Soft matches to reference	-	2510	2213
Hard matches to reference	-	2450	2197

TABLE VII
Matches for augmented subsamples of OpenFoodFacts

Metrics	**OFF**	**5% augm.**	**10% augm.**	**15% augm.**	**20% augm.**	**40% augm.**
Number of records	80000	20000	20000	20000	20000	20000
Soft matches to reference	-	102	98	80	37	0
Hard matches to reference	-	18	19	14	5	0

TABLE VIII
Calculated metrics for image based datasets

Metrics	**Food101**	**MAFood121**	**Epicurious Scraped**
Number of records	101000	21175	13501
Number of parameters	3		5
Soft matches to reference	-	0	0
Hard matches to reference	-	0	0
Soft duplicates	-	42	0
Hard duplicates	-	42	0
Average number of images per food item	1000	177.2	1
90% best image resolution	384x384x3	512x370x3	274x169x3
Average image entropy	7.59	7.74	7.30
Average subjective image quality	18.09	13.5613	1.9055
Number of Images with negative subjective quality	2132	1081	1503

When evaluating out metrics we were interested mainly in 2 things. First, whether the metric demonstrated some difference for different datasets. Second, whether this difference can be ascribed in some way to the quality level and whether some assumptions about the dataset quality can be made based on the metric value. Below we point out and discuss the most notable parts of the obtained results.

A. Number of records and its influence on quality

It is sort of expected that the bigger a given dataset has, the better it is suited for further usage due to the volume of data. However that may not be totally true as according to our results from Tab. V it is ascertainable that smaller, less popular datasets with fewer records and parameters may be a more complete and consistent source of truth than giant datasets that are constructed by ordinary users. Therefore while the number of records by itself can be a good indicator of how much data one could expect to get by adapting a given dataset into the data pool, it should not be utilized alone.

B. Missing values

Missing values are commonly dreaded in the data science industry as it is a typical source of problems. However during the evaluation process it became very clear that only the number of missing values is not at all representative of the scale of problems. A dataset may have 50% of records with missing values, but all of those missing values may all be present in only one problematic parameter. As such it is very important to know how critical the situation with missing values per parameters is. Thus we have included the number of parameters with different percentages of missing values. This along with the number of records already paints a picture of how much actual data one can expect from a dataset.

C. Matches to reference dataset: meaning for quality; Duplicates

During the evaluation phase we expected that due to sheer size of our OpenFoodFacts sample most datasets would end up with at least a few matches to it, however that was not what we have observed. As can be seen in the Tab. V, the number of matches to OpenFoodFacts for the two datasets tested was 0. While it is possible that many foods did not match due to similar foods not being present in OpenFoodFacts sample that we have utilized we have a few other hypotheses that could explain it.

First hypothesis is that foods did not match due to too many different details being provided in different datasets. To put it in other way, if a food name is overspecified in two different ways these ways may not match even if to humans they have the same meaning. We explore this idea a bit further and provide a bit more evidence for this hypothesis further in 5.2.6.

Second hypothesis that could explain the observed absence of matches is that the steps done for prepossessing of food names or for clustering them are insufficient. It stands to reason that either nltk python library or BERT encoder do not have enough contextual knowledge to be able to process difficult, niche words that can be present in food names. This could influence the results of matching in an undesirable way, but confirming this hypothesis requires further testing.

Despite this, we believe that this metric performs quite well in situations where data accumulation is needed. This metric does demonstrate a surprising robustness to omissions, something that we talk in part 5.2.5, and therefore can be utilized to detect samples from previously seen datasets, preventing information duplication.

As for duplicates, they are quite common to be found in many datasets, but not for the reasons that may be expected. Duplicates happen when the matching metric finds foods that are only a few word short of each other. This may produce quite good results, but also may have some side effects such as getting two flavours of the same commercial product recognised as duplicates of each other. Another thing that is may happen to be recognised as a duplicate is when two records only differ by serving size that has somehow made its way into the name of the product. This may be perceived as a false positive, but may also be recognised as a quality problem, as such it is important to look at what exactly matched as a duplicated before making decisions on quality. Despite this even the number of duplicates can provide some understanding of can be expected from the dataset.

As could be expected, soft duplicates are usually more prone to producing false positives than hard duplicates. This can be clearly observed on the OpenFoodFacts dataset where there are quite a lot of soft duplicates, but more than 6 times less hard duplicates. A thing of note here is that for smaller dataset of Nutritional Values for Common foods and produce the difference in number between hard and soft duplicates is not as drastic as for OpenFoodFacts. We attribute this to OpenFoodFacts simply containing much more variety of foods which may make the soft duplicates metric less reliable.

D. Time when records were last updated

While time when the dataset was last collected/updated could be important we have identified a few weaknesses of this metric. To begin with, this metric may not be accurately reported by creators of the dataset or reported at all, which makes it hard to judge just any dataset based on it. What is more, this metric may not be too relevant for a random dataset. Some datasets, like OpenFoodFacts contain food information for commercial, processed foods that can be bought in the supermarkets. For these kinds of foods it is important to have accurate, recent data. However for datasets that report on general food items like meat, fruits or vegetables or complex meals that are prepared either at home or in restaurants from these ingredients the information may not need to be recent at all as values for nutrition of these items drift quite slowly. In spite of this, we decided that in this particular case abundance of information is not hazardous and may inform better decisions and as such the metric stayed in the resulting code base.

E. Omissions and their influence on matching

When testing OpenFoodFacts against a sample of it, we have found it to be surprising just how well the matching metric performed on samples that had some information omitted from them. Even when 40% of words from the food names were removed the matches to the parent dataset clearly showed that the records are not totally new and contain at least some known information. We believe this to be an important advantage to the way we match records as it ensures that if the food names were to be cleaned up from some words the metric is still able to identify this.

F. Augmentation and its influence on matching

Testing matches in an environment of augmented samples has shown a weakness in matching. As can be seen in VII, even small augmentation of food names can bring down the accuracy of matching and thus the number of matches substantially. However when the augmentation is small the metric is still able to identify some records that can be traced to the parent dataset. This, combined this absolute lack of matches between different datasets that can be seen in Tab. V gives some ground to believe that any number of matches with our metric should be grounds for suspicion.

G. Image entropy

Image entropy is a hard metric to interpret. We believe this metric to be an indicator of overall health of the images. For instance, if entropy is unusually low it may mean that the images are somehow either generated or may be damaged somehow. Overall, we did not have such extremes among our dataset and as such cannot confirm or deny its effectiveness as a quality metric.

H. Subjective image quality

Subjective image quality has proven to be a vital metrics when assessing image-based datasets. As can be seen in Tab. VIII, subjective image quality definitely has some degree of correlation with the status of the dataset as Food101 and MAFood121 have far better scores than the dataset that was scraped from a website. It can also be noted from the table that we have created a counter for images with negative quality. This is done due to a curious effect. When assessing the scraped dataset we have found out that images with negative quality had visible artifacts and sometimes did not have the food on them at all. Since this correlation of negative quality with obvious visual artifacts was discovered we have decided to separate and hightlight both the number of images with negative quality and the names of images themselves for easier inspection.

Chapter 6

Conclusion

I Conclusion

To briefly summarize what is described above, we identified a research gap that is the lack of research into quality of food-related datasets. We designed the metrics that can reflect the quality of food related dataset. Then we confirmed that the metrics listed in Tab. V, Tab. VIII do indeed function and produce relevant results. We confirmed some of the advantages of the metrics and noted some potential downfalls that are to be avoided when utilizing them.

Overall we believe that this study addresses the research gap that is the total lack of research into quality of food-related datasets. We believe this study lays a groundwork for further research and provides helpful information about quality metrics that can be utilized by any person seeking to build a robust consumer nutrition application.

To summarize our contributions:

- We have created a dataset standard for food-related data
- We have identified and produced code for helpful metrics for assessing

dataset quality

- We have verified that metrics do indeed produce results and identified some of their advantages and weaknesses

The dataset standard can be observed in Tab. I and provides some idea as to what universal format may be for the data that is sourced for the application.

List of implemented quality metrics can be seen in Tab. V, Tab. VIII; the code for them can be found on github. These metrics have proven to be quite useful in analyzing quality of the dataset, although some careful consideration of the raw numbers is strictly necessary.

II Directions for further research

As the lack of research in this field is quite severe, there are multiple directions that futher research in this field can take.

First idea for further research is to develop a more robust record matching system. This robust matching system likely needs to be powered by machine learning techniques and as such needs a dataset of matches. Constructing a dataset of matches can be partially done using matching metric from this study.

Second idea is to create a dataset that would contain nutritional information of multiple instances of one food. For example, it would contain nutritional information of 20 identical in recipe tomato soups. Such a dataset could be then utilized to check the sanity of records in new datasets based on statistical techniques.

Third possible direction is to develop a more inclusive dataset standard and try applying metrics and techniques from this study to a more diverse group of datasets. For instance, it would be great to include some asian food datasets in the study.

Bibliography cited

[1] C. Castro, “What’s Wrong with Machine Bias,” en, *Ergo, an Open Access Journal of Philosophy*, vol. 6, no. 20201214, Jul. 2019, ISSN: 2330-4014. DOI: 10.3998/ergo.12405314.0006.015. [Online]. Available: http://hdl.handle.net/2027/spo.12405314.0006.015 (visited on 05/21/2023).

[2] J. Gu and D. Oelke, *Understanding Bias in Machine Learning*, arXiv:1909.01866 [cs, stat], Sep. 2019. [Online]. Available: http://arxiv.org/abs/1909.01866 (visited on 05/21/2023).

[3] S. Corbett-Davies and S. Goel, “The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning,” 2018, Publisher: arXiv Version Number: 2. DOI: 10.48550/ARXIV.1808.00023. [Online]. Available: https://arxiv.org/abs/1808.00023 (visited on 05/21/2023).

[4] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, “Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data,” en, *JAMA Internal Medicine*, vol. 178, no. 11, p. 1544, Nov. 2018, ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2018.3763. [Online]. Available: http://archinte.jamanetwork.com/

article.aspx?doi=10.1001/jamainternmed.2018.3763 (visited on 05/21/2023).

[5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," en, *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aax2342. [Online]. Available: https://www.science.org/doi/10.1126/science.aax2342 (visited on 05/21/2023).

[6] W. H. Clark and A. J. Michaels, "Quantifying Dataset Quality in Radio Frequency Machine Learning," in *MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM)*, 2021, pp. 384–389. DOI: 10.1109/MILCOM52596.2021.9652987.

[7] S. Picard, C. Chapdelaine, C. Cappi, *et al.*, "Ensuring Dataset Quality for Machine Learning Certification," in *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, 2020, pp. 275–282. DOI: 10.1109/ISSREW51248.2020.00085.

[8] L. Cenci, M. Galli, G. Palumbo, L. Sapia, C. Santella, and C. Albinet, "Describing the Quality Assessment Workflow Designed for DEM Products Distributed Via the Copernicus Programme. Case Study: The Absolute Vertical Accuracy of the Copernicus DEM Dataset in Spain," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, Brussels, Belgium: IEEE, Jul. 2021, pp. 6143–6146, ISBN: 978-1-66540-369-6. DOI: 10.1109/IGARSS47720.2021.9554393. [Online]. Available: https://ieeexplore.ieee.org/document/9554393/ (visited on 05/21/2023).

[9] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, “The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards,” 2018, Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.1805.03677. [Online]. Available: https://arxiv.org/abs/1805.03677 (visited on 03/14/2023).

[10] A. Meyers, N. Johnston, V. Rathod, *et al.*, “Im2Calories: Towards an automated mobile vision food diary,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1233–1241.

[11] S. Cheng, B. Chu, B. Zhong, *et al.*, “DRNet: Towards fast, accurate and practical dish recognition,” en, *Science China Technological Sciences*, vol. 64, no. 12, pp. 2651–2661, Dec. 2021, ISSN: 1674-7321, 1869-1900. DOI: 10.1007/s11431-021-1903-4. [Online]. Available: https://link.springer.com/10.1007/s11431-021-1903-4 (visited on 05/21/2023).

[12] R. Xu, L. Herranz, S. Jiang, S. Wang, X. Song, and R. Jain, “Geolocalized Modeling for Dish Recognition,” *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1187–1199, Aug. 2015, ISSN: 1520-9210, 1941-0077. DOI: 10.1109/TMM.2015.2438717. [Online]. Available: http://ieeexplore.ieee.org/document/7114316/ (visited on 05/21/2023).

[13] L. Herranz, S. Jiang, and R. Xu, “Modeling Restaurant Context for Food Recognition,” *IEEE Transactions on Multimedia*, vol. 19, no. 2, pp. 430–440, Feb. 2017, ISSN: 1520-9210, 1941-0077. DOI: 10.1109/TMM.2016.2614861. [Online]. Available: http://ieeexplore.ieee.org/document/7581115/ (visited on 05/21/2023).

[14] Y. Kawano and K. Yanai, "FoodCam: A real-time food recognition system on a smartphone," en, *Multimedia Tools and Applications*, vol. 74, no. 14, pp. 5263–5287, Jul. 2015, ISSN: 1380-7501, 1573-7721. DOI: 10.1007/s11042-014-2000-8. [Online]. Available: http://link.springer.com/10.1007/s11042-014-2000-8 (visited on 05/21/2023).

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[16] T. Ege and K. Yanai, "Estimating food calories for multiple-dish food photos," in *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, IEEE, 2017, pp. 646–651.

[17] L. Deng, J. Chen, Q. Sun, *et al.*, "Mixed-dish Recognition with Contextual Relation Networks," en, in *Proceedings of the 27th ACM International Conference on Multimedia*, Nice France: ACM, Oct. 2019, pp. 112–120, ISBN: 978-1-4503-6889-6. DOI: 10.1145/3343031.3351147. [Online]. Available: https://dl.acm.org/doi/10.1145/3343031.3351147 (visited on 03/14/2023).

[18] S. Abhari, R. Safdari, L. Azadbakht, *et al.*, "A systematic review of nutrition recommendation systems: With focus on technical aspects," *Journal of biomedical physics & engineering*, vol. 9, no. 6, p. 591, 2019, Publisher: Shiraz University of Medical Sciences.

[19] J. Vincent, *Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech*, en-US, Jan. 2018. [Online]. Available: https://www.theverge.com/2018/1/12/16882408/google-

racist-gorillas-photo-recognition-algorithm-ai (visited on 05/21/2023).

[20] T. Y. Chen, S. C. Cheung, and S. M. Yiu, "Metamorphic Testing: A New Approach for Generating Next Test Cases," 2020, Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.2002.12543. [Online]. Available: https://arxiv.org/abs/2002.12543 (visited on 03/14/2023).

[21] X. Xie, J. W. Ho, C. Murphy, G. Kaiser, B. Xu, and T. Y. Chen, "Testing and validating machine learning classifiers by metamorphic testing," en, *Journal of Systems and Software*, vol. 84, no. 4, pp. 544–558, Apr. 2011, ISSN: 01641212. DOI: 10.1016/j.jss.2010.11.920. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0164121210003213 (visited on 03/14/2023).

[22] L. Li, T. Peng, and J. Kennedy, "A rule based taxonomy of dirty data," *GSTF Journal on Computing (JoC)*, vol. 1, no. 2, 2014.

[23] M. A. Hernández and S. J. Stolfo, "The merge/purge problem for large databases," *ACM Sigmod Record*, vol. 24, no. 2, pp. 127–138, 1995, Publisher: ACM New York, NY, USA.

[24] G. Popovski, G. Ispirova, N. Hadzi-Kotarova, E. Valencic, T. Eftimov, and B. Korousic-Seljak, "Food Data Integration by using Heuristics based on Lexical and Semantic Similarities.," in *HEALTHINF*, 2020, pp. 208–216.

[25] S. Rezayi, Z. Liu, Z. Wu, *et al.*, "Agribert: Knowledge-infused agricultural language models for matching food and nutrition," IJCAI, 2022.