Towards Contextual Healthiness Classification of Food Items - A Linguistic Approach

Michael Wiegand and Dietrich Klakow
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
{Michael.Wiegand|Dietrich.Klakow}@lsv.uni-saarland.de

Abstract

We explore the feasibility of contextual healthiness classification of food items. We present a detailed analysis of the linguistic phenomena that need to be taken into consideration for this task based on a specially annotated corpus extracted from web forum entries. For automatic classification, we compare a supervised classifier and rule-based classification. Beyond linguistically motivated features that include sentiment information we also consider the prior healthiness of food items.

1 Introduction

Food plays a substantial part in each of our lives. With the growing health awareness in many parts of the population, there is consequently a high demand for the knowledge about healthiness of food. In view of the variety of both different types of food and nutritional aspects it does not come as a surprise that there is no comprehensive repository of that knowledge. Since, however, much of this information is preserved in natural language text, we assume that it is possible to acquire some of this knowledge automatically with the help of natural language processing (NLP).

In this paper, we take a first step towards this endeavour. We try to identify mentions that a food item is healthy (1) or unhealthy (2).

(1) There is not a healthy diet without a lot of fruits, vegetables and salads.
(2) The day already began unhealthy: I had a piece of cake for breakfast.

This task is a pre-requisite of more complex tasks, such as finding food items that are suitable for certain groups of people with a particular health condition (3) or identifying reasons for the healthiness or unhealthiness of particular food items (4).

(3) Vegetables are healthy, in particular, if you suffer from diabetes.
(4) Potatoes are healthy since they are actually low in calories.

The major problem of identifying some Is-Healthy or Is-Unhealthy relation is that the simple co-occurrence of a food item and the word healthy or unhealthy is not sufficiently predictive as shown in (5)-(7).

(5) Chocolate is not healthy.
(6) The industry says chocolate is healthy, but I guess this is just a marketing strategy.
(7) If chocolate is healthy, then I will run for the next presidential election.

We describe the contextual phenomena that underlie these cases and provide detailed statistics as to how often they occur in a typical text collection. From this analysis we derive features to be incorporated into a classifier.

Our experiments are carried out on German data. We believe, however, that our findings carry over to other languages since the aspects addressed in this work are (mostly) language universal. For the sake of general accessibility, all examples will be given as English translations.

To the best of our knowledge, this is the first work that addresses the classification of healthiness of food items using NLP.

2 Related Work

In the food domain, the most prominent research addresses ontology or thesaurus alignment (van Hage et al., 2010), a task in which concepts from different sources are related to each other. In this context, hyponymy relations (van Hage et al., 2005) and part-whole relations (van Hage et al., 2006) have been explored. More recently, Wiegand et al. (2012a) examined extraction methods for relations involved in customer advice in a supermarket. In Chahuneau et al. (2012), sentiment information has been related to food prices with the help of a large corpus consisting of restaurant menus and reviews.

In the health/medical domain, the majority of research focus on domain-specific relations involving entities, such as genes, proteins and
3 The Dataset

In order to generate a dataset for our experiments, we used a crawl of chefkoch.de\(^1\) (Wiegand et al., 2012a) consisting of 418,558 webpages of food-related forum entries. chefkoch.de is the largest German web portal for food-related issues.

While we are aware of the fact that the healthiness of food items is also discussed in scientific texts we think that the text analysis on social media serves its own purpose. The language in social media is much more accessible to the general population. Moreover, social media can be considered as an exclusive repository of popular wisdom containing, for example, home remedies.

3.1 Healthiness Markers & Food Items

As it is impractical for us to manually label the entire web corpus with healthiness information, we extracted for annotation sentences in which there is a healthiness marker and a mention of a food item. By healthiness marker, we understand an expression that conveys the property of being healthy. Apart from the word healthy itself, we came up with 17 further common expressions (e.g. nutritious, healthful or in good health). Since the word healthy covers more than 95% of the mentions of healthiness markers in our entire corpus, however, we decided to restrict our healthiness marker exclusively to mentions of that expression. Thus, our main focus in this classification task is the contextual disambiguation, i.e. the task to decide whether a specific co-occurrence of the expression healthy and some food item denotes a genuine Is-(Un)Healthy relation.

The food items for which we extract co-occurrences with the healthiness marker healthy (Table 7) will henceforth be referred to as target food items. In order to obtain a suitable list of items for our experiments, we manually compiled a list of frequently occurring types of food.

3.2 “Unhealthy” vs. “Not Healthy”

In order to obtain instances that express an Is-Unhealthy relation, we exclusively consider negated instances of the Is-Healthy relation (8). We also experimented with a dataset with mentions of the word unhealthy (paired with our target food items) to extract instances such as (9).

Using the same target food items, the unhealthy-dataset is, however, less than 14% of the size of the healthy-dataset. We also found that instances of the Is-Unhealthy-relation are not easier to detect on the unhealthy-dataset, since the unhealthy-dataset produced much poorer classifiers for detecting Is-Unhealthy relations than the healthy-dataset using negations as a proxy.

4 Annotation

Our final dataset comprises 2,440 instances, where each instance consists of a sentence with the co-occurrence of some food item and the word healthy accompanied by the two sentences immediately preceding and the two sentences immediately following it.

The dataset was manually annotated by two German native speakers. On 4 target food items (this corresponds to 574 target sentences)\(^2\) we measured an inter-annotation agreement of Cohen’s $\kappa = 0.7374$ (Landis and Koch, 1977) which should be sufficiently high for our experiments.

The annotators had to choose from a rich set of category labels that particularly divide the negative examples (i.e. those cases in which the co-occurrence of the target food item and healthy neither expresses an Is-Healthy nor an Is-Unhealthy relation) into different categories.

In the following, we describe the different category labels. Their distribution is shown in Table 1.

4.1 Is-Healthy Relation (HLTH)

This class describes instances in which there holds an Is-Healthy relation between the mention of healthy and the target food item (10).

Table 1 shows that less than 20% of the co-occurrences of the target food item and healthy express this relation. This may already indicate that its extraction is difficult.

\(^1\)www.chefkoch.de

\(^2\)This is the only part of the dataset which was annotated by both annotators in parallel.
<table>
<thead>
<tr>
<th>Type</th>
<th>Abbrev.</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is-Healthy</td>
<td>HLTH</td>
<td>488</td>
<td>20.00</td>
</tr>
<tr>
<td>Is-Unhealthy</td>
<td>UNHLTH</td>
<td>171</td>
<td>7.01</td>
</tr>
<tr>
<td>OTHER:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Relation</td>
<td>NOREL</td>
<td>788</td>
<td>32.30</td>
</tr>
<tr>
<td>Restricted Relation</td>
<td>RESTR</td>
<td>312</td>
<td>12.79</td>
</tr>
<tr>
<td>Unspecified Intersection</td>
<td>INTERS</td>
<td>198</td>
<td>8.11</td>
</tr>
<tr>
<td>Embedding</td>
<td>EMB</td>
<td>157</td>
<td>6.43</td>
</tr>
<tr>
<td>Comparison Relation</td>
<td>COMP</td>
<td>121</td>
<td>4.96</td>
</tr>
<tr>
<td>Unsupported Claim</td>
<td>CLAIM</td>
<td>87</td>
<td>3.57</td>
</tr>
<tr>
<td>Other Sense</td>
<td>SENSE</td>
<td>77</td>
<td>3.16</td>
</tr>
<tr>
<td>Irony</td>
<td>IRO</td>
<td>25</td>
<td>1.02</td>
</tr>
<tr>
<td>Question</td>
<td>Q</td>
<td>16</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the different (linguistic) phenomena.

4.2 Is-Unhealthy Relation (UNHLTH)

We already stated in §3.2 that we consider negated instances (11) as instances for the Is-Unhealthy relation. We have a fairly broad notion of negation, e.g. (12) and (13) will also be assigned to this category. These partial negations are at least as frequent as full negations (11). However, we assume that the latter are often employed only as a means of being polite even though the speaker’s intention is that of a full negation. The fact that we also observed fewer mentions of unhealthy co-occurring with a target food item than negated mentions of healthy would be in line with this theory (unhealthy is usually perceived to be more intense/blunter than not healthy).

4.3 Other Relations

Apart from the two target relations, we observe the following other relationships:

4.3.1 Restricted Relation (RESTR)

This category describes cases in which the Is-Healthy relation holds provided some additional condition is fulfilled. Typical conditions address a special kind of preparing the target food item (14) or make quantitative restrictions as to the amount of the target food item to be consumed (15). As such, one cannot infer from restricted relations to general properties of food items.

4.3.2 Unspecified Intersection (INTERS)

In relation extraction, syntactic relatedness between the candidate entities of a relation is usually considered an important cue (Zhou et al., 2005; Mintz et al., 2009). In particular, the specific type of syntactic relation needs to be considered. If in our task healthy is an attributive adjective of the target food item (16), this is not an indication of a genuine Is-Healthy relation that we are looking for. With this construction, one usually refers to all those entities that share the two properties (intersection) of being the target food item and being healthy. This case is different from both HLTH (17) and RESTR (18).

(16) I usually buy the healthy fat.
(17) Fat is healthy.
(18) I usually buy the healthy fat, the one that contains a high degree of unsaturated fatty acids.

4.3.3 Comparison Relation (COMP)

If the target food item is compared with another food item with regard to their healthiness status (19) & (20), one cannot conclude anything regarding the absolute healthiness of the target food item. This is due to the fact that a comparison assumes healthiness as a (continuous) scale rather than a binary (discrete) property. It determines the positions of the two food items relative to each other on that particular scale.

(19) Honey is healthier than chocolate. (target food item: honey)
(20) Honey is as healthy as chocolate. (target food item: honey)
4.3.4 Unsupported Claim (CLAIM)
In our initial data analysis, we found frequent cases in which the author of a forum entry reports a (controversial) statement regarding the healthiness status of a particular food item. These claims are often used as a means of starting a discussion about that issue (21).

(21) Some people claim that chocolate is healthy. What do you make of it?

If it is not possible to infer from such reported statement that the reported view is shared by the author (and we found that this is true for many reported statements), we tag it as CLAIM.

4.3.5 Question (Q)
There may also be cases in which the Is-(Un)Healthy relation is embedded in a question (22).

(22) Is chocolate healthy?

4.3.6 Irony (IRO)
Irony (23) is a figure of speech that can frequently be observed in user-generated text (Tsur et al., 2010). With a proportion of less than 1\%, this, however, does not apply for the forum entries that comprise our data collection.

(23) Everyone knows that sweets are healthy, in particular, chocolate with its many calories even makes you lose weight.

4.3.7 Embedding (EMB)
In addition to the previous categories CLAIM and IRO, there exist other ways of embedding the healthiness relation into a context so that the general validity of it is discarded. We introduce a common label for all those other remaining types that include, for instance, modal embedding (24) or irrealis construction (25).

(24) Honey could be healthy.
(25) If chocolate were healthy, people eating it wouldn’t put on so much weight.

4.3.8 Other Sense (SENSE)
Both the target food item and the German healthiness cue gesund are (potentially) ambiguous expressions. For instance, gesund can be part of several multiword expressions, such as gesunder Menschenverstand (engl. common sense).

4.3.9 No Relation (NOREL)
While in all previously discussed cases the target food item and healthy are somehow related, there are cases in which the co-occurrence is merely coincidental (26).

(26) Tomatoes are very healthy and they can be ideally served on bread. (target food item: bread)

On our dataset, this is the most frequent label.

5 Feature Design
All features we use are summarized in Table 2 along examples. Apart from bag of words (word), we use following features:

5.1 Linguistic Features
The linguistic features are mainly derived from our quantitative data analysis in §4. Given the limited space of this paper, we will only point out some special properties.

The first group of (linguistic) features (Table 2) is designed to detect some relationship between target food item and healthy. The co-occurrence within the same clause is usually a good predictor. There are three features to establish this property: clause, boundary and otherFood.

We already pointed out in §4.3.2 that not only syntactic relatedness between healthy and the target food item as such but also the specific syntactic relation plays a decisive role for this task. The two most common relations are that healthy is a predicative adjective (of the target food item), which is usually indicative of HLTH, and that healthy is an attributive adjective (of the target food item), which is usually indicative of INTERS (on our dataset in more than 90\% of the instances labeled with INTERS this is the case). This is reflected by the two features predRel and attrRel (and the backoff features pred and attr). An additional feature attrFood captures a special construction in which healthy as an attributive adjective actually denotes HLTH instead of INTERS.

For the conditional healthiness RESTR (§ 4.3.1), we found two predominant subcategories of restrictions: restrictions with regard to the quantity with which the target food item should be consumed (quant) and references to a specific subtype of the target food item, which we want to capture with a few precise surface patterns (spec) and a feature that checks whether the target food item precedes an attributive adjective (attrNoH).

Table 2 also contains features to detect various contextual embeddings (opHolder, question, irrealis, modal and irony). opHolder is to detect cases of CLAIM. We assume once some opinion holder other than the author of the forum post (i.e. 1st person pronoun) is identified, there is a CLAIM.

We also investigate whether healthiness correlates with sentiment. For instance, if the author promotes the healthiness of some food item, does this also coincide with positive sentiment (e.g.
tasty, good etc.)? Our features positive/negative polar check for the presence of polar expressions.

5.2 Knowledge-based Features using a Healthiness Lexicon

We also incorporate features referring to the prior knowledge of healthiness of food items. We use a lexicon introduced in Wiegand et al. (2012b) which covers approximately 3000 food items, and we refer to it as healthiness lexicon. Each food item is specified as being either healthy or unhealthy in that lexicon. The healthiness judgment has been carried out based on the general nutrient content of each food item. A detailed description of the annotation scheme and annotation agreement can be found in Wiegand et al. (2012b).

The specific features derived from that lexical resource are listed in Table 2. They are divided into two groups. prior describes the prior healthiness of the target food item. Since our task is to determine the contextual healthiness, the usage of such a feature is legitimate. The contextual healthiness need not to coincide with the prior healthiness. For instance, in (27), chocolate is described as a healthy food item even though it is a priori considered unhealthy.

(27) Chocolate is healthy as it’s high in magnesium and provides vitamin E.

We use this knowledge as a baseline. If we cannot exceed the classification performance of prior (alone), then acquiring the knowledge of healthiness with the help of NLP is hardly effective.

priorCont describes the prior healthiness status of neighbouring food items in the given context.

6 Rule-based Classification

We also examine rule-based classifiers since they can be built without any training data. Each classifier is defined by a (large) conjunction of linguistic features. Features indicating a class other than the target class are used as negated features in that conjunction. The rule-based classifiers only consider features where a positive or negative correlation towards the target class is (more or less) obvious. Table 3 shows the rule-based classifiers for each of our classes. For HLTH, it basically states that healthy has to be a predicative adjective of the target food item (predRel), and the target food item and healthy have to appear within the same clause (or there is no boundary sign between them). After that, a long list of negated features follows: quant, spec and attrNoH, for example, are negated because they are typical cues for REST. The remaining features are negated since they are either indicative of UNHLTH, COMP, EMB, CLAIM, SENSE, IRO or Q. The classifier for UNHLTH only differs from HLTH in that either of the negation cues, i.e. negTarget or negHealth, has to be present.

7 Experiments

In this section we present the results on automatic classification.

7.1 Classification of Individual Utterances

In this subsection, we evaluate the performance of the different feature sets on sentence-level classification using supervised learning and rule-based classification. We investigate the detection of the two classes HLTH (§4.1) and UNHLTH (§4.2). Each instance to be classified is a sentence in which there is a co-occurrence of a target food item and a mention of healthy along its respective context sentences. The dataset was parsed using the Stanford Parser (Rafferty and Manning, 2008). We carry out a 5-fold cross-validation on our manually labeled dataset. As a supervised classifier, we use Support Vector Machines (SVMlight (Joachims, 1999) with a linear kernel).

For each class, we train a binary classifier where positive instances represent the class to be extracted while negative instances are the remaining instances of the entire dataset (§4).

7.1.1 Comparison of Various Feature Sets

Table 4 lists the results for various feature sets that we experimented with. take-all is an unsupervised baseline that considers all instances of our dataset as positive instances (of the class which is examined, i.e. HLTH or UNHLTH). In other words, this baseline indicates how well the mere co-occurrence of healthy and the target food item predicts either of our two classes. Our second

Table 3: Rule-based classifiers based on linguistic features (Table 2).
Table 2: Description of the feature set; the set contains several cue word lists, in order to avoid overfitting, we either translated existing resources from English or used diverse web-resources that are not related to our dataset.
Table 4: Comparison of different feature sets.

<table>
<thead>
<tr>
<th>Features</th>
<th>HLTH</th>
<th>UNHLTH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Rec</td>
</tr>
<tr>
<td>take-all (baseline 1)</td>
<td>20.3</td>
<td>100.0</td>
</tr>
<tr>
<td>prior (baseline 2)</td>
<td>28.0</td>
<td>87.3</td>
</tr>
<tr>
<td>priorCont</td>
<td>21.2</td>
<td>96.9</td>
</tr>
<tr>
<td>prior+priorCont</td>
<td>28.0</td>
<td>86.9</td>
</tr>
<tr>
<td>word</td>
<td>35.9</td>
<td>66.5</td>
</tr>
<tr>
<td>linguistic</td>
<td>38.3</td>
<td>66.1</td>
</tr>
<tr>
<td>word+linguistic</td>
<td>40.2</td>
<td>63.6</td>
</tr>
<tr>
<td>word+prior</td>
<td>38.1</td>
<td>70.1</td>
</tr>
<tr>
<td>word+priorCont</td>
<td>35.0</td>
<td>65.3</td>
</tr>
<tr>
<td>word+prior+priorCont</td>
<td>37.4</td>
<td>70.8</td>
</tr>
<tr>
<td>word+linguistic+prior</td>
<td>41.4</td>
<td>64.3</td>
</tr>
<tr>
<td>word+linguistic+priorCont</td>
<td>44.1</td>
<td>68.3</td>
</tr>
<tr>
<td>all features</td>
<td>44.5</td>
<td>69.3</td>
</tr>
</tbody>
</table>

significantly better than word* at p < 0.1 ** at p < 0.05; better than word+linguistic* at p < 0.05; better than word+prior at p < 0.05 (paired t-test)

Table 5: List of the best subset of linguistic features (Table 2) for each individual class.

7.1.2 Inspection of Linguistic Features

Table 5 shows the best performing feature subset using a best-first forward selection as implemented in Weka (Witten and Frank, 2005). The table shows that diverse features are important including features to detect restricted relations (§4.3.1) (i.e. attrNoH) or comparisons (i.e. comp), features to distinguish predicative from attributive adjectives for the detection of unspecified intersection (§4.3.2) (i.e. predRel and attrRel), various features to determine contextual embedding (i.e. opHolder, irrealis and negHealth) and sentiment information (i.e. negative polarEC).

7.1.3 Detecting Anti-Prior Healthiness

We now take a closer look at anti-prior instances which are utterances in which the relation expressed is opposite to the relation that one would a priori assume, e.g. chocolate is healthy instead of chocolate is unhealthy. In our gold standard, we identified these instances with the help of the actual (manually assigned) label and our healthiness lexicon (§5.2).\(^4\) Such instances may be very interesting to extract, even though they are rare (15% on HLTH and UNHTLH). Previously, supervised classifiers with word+prior produced similar performance as classifiers with word+linguistic (Table 4). Since linguistic features are fairly expensive to produce, the prior knowledge of healthiness seems an attractive alternative. But this is misleading. Table 6 displays the recall (by supervised classification) on only anti-prior instances and shows that the usage of prior which, in isolation, would detect none of these instances, gives a much lower recall than linguistic when added to word. Therefore, word+linguistic would be the preferable feature set if one had to choose between word+prior and word+linguistic.

\(^4\)Whenever HLTH co-occurs with prior unhealthiness (according to the healthiness lexicon) or UNHLTH co-occurs with prior healthiness, there is an anti-prior instance.
Table 6: Recall on anti-prior instances.

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>word+prior</th>
<th>word+linguistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>17.2</td>
<td>54.6</td>
</tr>
</tbody>
</table>

7.2 Aggregate Classification

Finally, we automatically rank food items according to healthiness based on the aggregate of text mentions. Ideally, the ranking should separate healthy from unhealthy food items. We want to know whether our text corpus and contextual classification, one can actually approximate a correct prior healthiness. Aggregate classification means that we make a healthiness prediction for a specific food item based on all text mentions of that food item co-occurring with the word healthy. It may be easier to achieve a robust aggregate classification than a robust individual classification. This is because in aggregate-based tasks, there is a certain degree of redundancy contained in the data, as instances of a group of utterances (belonging to the same food item) may often comprise similar information. For such classifiers, one should focus on a higher precision since a reasonable recall is enabled by the redundancy in the data.

Our baseline RAW is completely unsupervised and does not include any linguistic processing. We use the Pointwise Mutual Information (PMI) which is estimated on our large web corpus ($\S3$).

\[
PMI(food\ item,\ healthy) = \log \frac{P(food\ item,\ healthy)}{P(food\ item)P(healthy)} \tag{1}
\]

For the automatic classification, we consider LEARN which uses the output of the supervised classifier comprising the features word+linguistic (we must exclude the feature prior as this would include the knowledge we want to predict automatically in this experiment)$^6$ while RB is the output of the rule-based classifier we presented in $\S6$ (which does not contain prior as a feature either).

In order to convert the classifications of individual utterances for a target food item (by LEARN and RB) to one ranking score (according to which we rank all the target food items), we simply compute the ratio between instances predicted to be healthy and those predicted to be unhealthy:

\[
score_{LEARN/RB}(food\ item) = \frac{\#HLTH_{predicted}(food\ item)}{\#UNHLTH_{predicted}(food\ item)} \tag{2}
\]

Table 7: Aggregate ranking; green denotes (actual) healthy items, red (actual) unhealthy items.

where $\#HLTH_{predicted}(food\ item)$ are the number of instances the classifier predicts the label HLTH for the target food item while $\#UNHLTH_{predicted}(food\ item)$ are the number of instances labeled as UNHLTH, respectively.

Table 7 shows the results of the three rankings. The actual labels are derived from the healthiness lexicon ($\S5.2$). The table clearly shows that the ranking produced by RAW contains most errors. fat is the second most highly ranked food item. This can be explained by the high proportion of INTERS ($\S4.3.2$) among the co-occurrences of fat and healthy (almost 50%). LEARN and RB produce a better ranking, thus proving that a contextual (linguistic) analysis is helpful for this task. RB also outperforms LEARN presumably because of its much higher precision (as measured for individual classification in Table 4: 53.4% vs. 40.2% for HLTH and 45.0% vs. 40.9% for UNHLTH).

8 Conclusion

We presented a first step towards contextual healthiness classification of food items. For this task, we introduced a new annotation scheme. Our annotation revealed that many different linguistic phenomena are involved. Thus, this problem can be considered an interesting task for NLP. We demonstrated that a linguistic analysis is not only necessary for classifying individual utterances but also for ranking food items based on an aggregate of text mentions.

Acknowledgements

This work was performed in the context of the Software-Cluster project EMERGENT. Michael Wiegand was funded by the German Federal Ministry of Education and Research (BMBF) under grant no. “01IC10S01”. The authors would like to thank Stephanie Köser and Eva Lasarczyk for annotating the dataset presented in this paper. We would also like to thank Benjamin Roth for interesting discussions.

For $P(food\ item,\ healthy)$, we consider all sentences in which the target food item and healthy co-occur.

$^6$We train for each target food item a classifier using only the instances with the other target food items as training data.
References


