Online Food Recipe Title Semantics: Combining Nutrient Facts and Topics

Tomasz Kusmierczyk
NTNU, Trondheim, Norway
tomaszku@idi.ntnu.no

Kjetil Nørvåg
NTNU, Trondheim, Norway
noervaag@idi.ntnu.no

ABSTRACT

Dietary pattern analysis is an important research area, and recently the availability of rich resources in food-focused social networks has enabled new opportunities in that field. However, there is a little understanding of how online textual content is related to actual health factors, e.g., nutritional values. To contribute to this lack of knowledge, we present a novel approach to mine and model online food content by combining text topics with related nutrient facts. Our empirical analysis reveals a strong correlation between them and our experiments show the extent to which it is possible to predict nutrient facts from meal name.

Keywords

LDA, text regression, social media mining, online food recipe

1. INTRODUCTION

As fundamental concepts in our daily lives, food and dietary patterns are important research topics. However traditional ways of studying them are limited in scope and reach; the access to resources generated in social media have enabled new opportunities and research directions. In recent years online food-oriented communities have gained immense popularity and have moved this sphere of our lives on-line. Thus designing appropriate, specialized mining techniques may benefit not only in better understanding of online patterns but also in enabling new possibilities for practical applications.

The primary form of information that online users interact with is text. However in the context of food studies other factors play an equally important role. Key recipe characteristics that influence our health are nutrient facts, among which the most popular and important are food energy and quantities of fat, proteins, sugars, carbohydrates, cholesterol and sodium. Although several works studying those factors in the context of social media recently appeared, only few take into account textual content, and, moreover, mostly using external, predefined databases. One of the reasons for that is the lack of appropriate tools and techniques for effectively exploring these combined data dimensions. In our paper we address this gap.

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983897

Our first objective is to gain a deeper *understanding of associations between words and nutritional values*. Thus, the analytical part of our paper is devoted to studying correlations between them. In contrast to previous works, we rely only on short, free-form textual food titles provided by inexperienced social media users and do not require additional meta data such as a list of ingredients.

Another challenge this paper targets is to design and develop a model that would be helpful in mining online food content. Empirical observations provide useful hints that motivate design choices of such a model. We expect that using text in conjunction with additional information, i.e., nutrients, should provide noticeable improvement in many practical applications.

Two practical problems that we are particularly interested in are *identifying compact and meaningful food topics* and *predicting nutrient fact values* in various settings. One common setting in the context of online food platforms is when only some subset of nutrients is known. Another is when no nutritional values are known at all. For example we imagine a mobile application that can scan a meal name in a restaurant and then provide approximate nutritional values to help select food appropriately.

Our main contributions are: (i) a study of a large-scale online food community in terms of relations between nutritional values and textual descriptions (recipe names), (ii) the introduction of a new topic model combining text with several outputs (nutrient facts) regression, and (iii) an evaluation of our approach's efficacy in discovering recipe topics and predicting nutritional values.

2. RELATED WORK

Studying social media contents in order to discover food and dietary patterns and correlations is a relatively new research direction; so, few related studies exist in this context.

Some recent works study dietary preferences in social media from a geographic and temporal perspective. In the influential work by West et al. [11], recipe access logs were analyzed in order to observe seasonal and weekly trends in people's preferences towards nutritional values in different US regions. Similar observations were made by Said and Bellogin [8] when studying user rating behavior. Wagner et al. [9], recently investigated the dynamics of online food consumption. Finally, Kusmierczyk et al. [4] observed differences in nutrient value preferences from the perspective of both online recipe consumption and creation.

Recently, several researchers tried to exploit connections between textual content and dietary patterns. Abbar et al. [1] managed to build a model predicting county-wide obesity and diabetes statistics using food mentions in Twitter. Muller et al. [5] designed a system that calculates recipe nutritional values. In the proposed approach, the authors first match recipe ingredients and based on them then infer nutrient facts. A similar approach was taken by

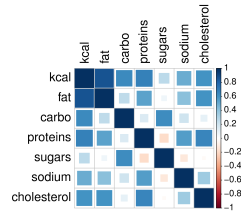


Figure 1: Correlations between recipe nutrients.

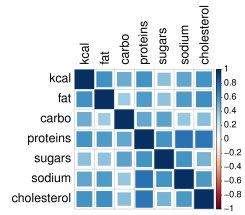


Figure 2: Correlations over words between info-gains corresponding to nutritional values.

De Choudhury et al. [2], who used the US Department of Agriculture National Nutrient Database to obtain nutritional and calorific information from Instagram food posts.

When the focus is on understanding social media patterns, LDA and topic models are the standard approach to identify textual clusters of interest, e.g., when studying topics around patterns in diet [2], relations between tweets and public health [7], or life satisfaction [3]. Although the method is popular not only in our context, out-of-the-box solutions are not always sufficient and proposing variants adjusted to specific domain is often necessary, as in [10]. For example, existing models extending LDA, e.g. [6], attach regression component to words whereas we link it on topic level.

3. DATA SET

Our study relies on the data retrieved from the largest English online food recipe platform, namely *allrecipes.com*.[1] The web site was crawled and archived in July 2015, and the data set contains more than 240 thousand recipes.

For each recipe, metadata with title and information about nutritional facts (per 100 g) are provided. Unfortunately, some nutrient values were missing for the majority of recipes, and all seven of the most important facts (i.e., kilocalories (denote *kcal*), *fat*, carbohydrates (denote *carbo*), *proteins*, *sugars*, *sodium*, *cholesterol*) were present in only 58 thousand recipes. Thus, our experiments focused on this subset.

Initially, recipe titles were arbitrary strings defined as free-form text by the users. Several standard text pre-processing, data cleaning steps were necessary. First, we filtered out punctuation, special characters, numbers and stop-words. Then, using a Porter stemmer, word forms were unified. Finally, we filtered out all words occurring less than 2 times in the corpus. This procedure resulted in a vocabulary containing of 4,679 unique words. However, only 1,578 words were used more than 10 times. The most popular words are 'chicken', appearing in 5,230 (9%) of titles, and 'salad', found in 3,728 (6%) of titles.

4. NUTRIENT FACTS AND TEXT CORRELATIONS

In our study we focus on relations between nutrient facts and associated textual content as one of our goals is to extend the understanding of how they depend on each other.

[1]http://allrecipes.com

Table 1: Top 5 words with the highest information gain for nutrient facts.

Nutrient Fact	Important Words
kcal	chicken, cooki, pie, dip, pasta
fat	cooki, pie, chicken, casserol, sausag
carbos	cake, pie, dip, pasta, cooki
proteins	chicken, cooki, cake, chocol, pork
sugars	cake, chocol, pie, cooki, appl
sodium	cooki, chicken, cake, chocol, casserol
cholesterol	chicken, cooki, shrimp, pork, egg

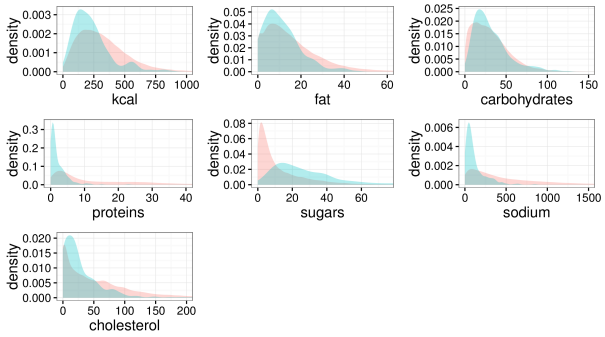


Figure 3: Comparison of nutrient fact distributions in recipes with (green) and without (red) a word 'frost' in the title.

First, we were interested in associations between nutrition values in the context of recipes produced by the studied online community. Figure 1 presents a Pearson correlation matrix between all 7 nutrient facts. We observe a cluster of strongly positively correlated values composed of almost all facts. The only two nutrition values that behave differently are sugars and carbohydrates. Sugars have almost no correlation to fat and cholesterol and are negatively correlated to proteins and sodium. Furthermore, the correlations between carbohydrates and fat, proteins and cholesterol are significantly weaker than for other nutrients. Some of these discoveries are surprising and demonstrate that online food content may be biased in an unexpected way: studies like ours are needed to gain better understanding of these online communities.

To measure the influence of words on nutritional values we applied information gain. As an attribute feature we used either a presence or an absence of the word in the recipe title. Table 1 presents the most influential words. We noticed a high overlap between top correlated words, e.g., the word 'cooki' (cookie) is important for all nutrients. Furthermore, we observed that information gains of words were similar between nutrient facts. Figure 2 presents the info-gain correlation measured over words. For each word we calculated information gain to all nutrient values. Then we measured Spearman correlation over words between info-gains corresponding respectively to kcal, fat etc. Observed correlations are very high, showing that the same words are important for all nutrient facts. For example, we found 13 words that are present in top-100 lists of all nutrients. Some of these words are meal names (e.g., 'lasagna'), some represent ingredients (e.g., 'cranberries') but also more surprising observations were made. For example, the word 'frost' was identified as having a very high information gain. Figure 3 compares nutrient value distributions in recipes with and without 'frost' in the title. We observe, for example, that sugars are biased towards higher and proteins towards lower values.

5. COMBINING NUTRIENT FACTS AND TEXT TOPICS

The empirical analysis in Section 4 revealed a strong correlation among nutritional values and between nutritional values and particular words. Furthermore, we noticed that similar words are

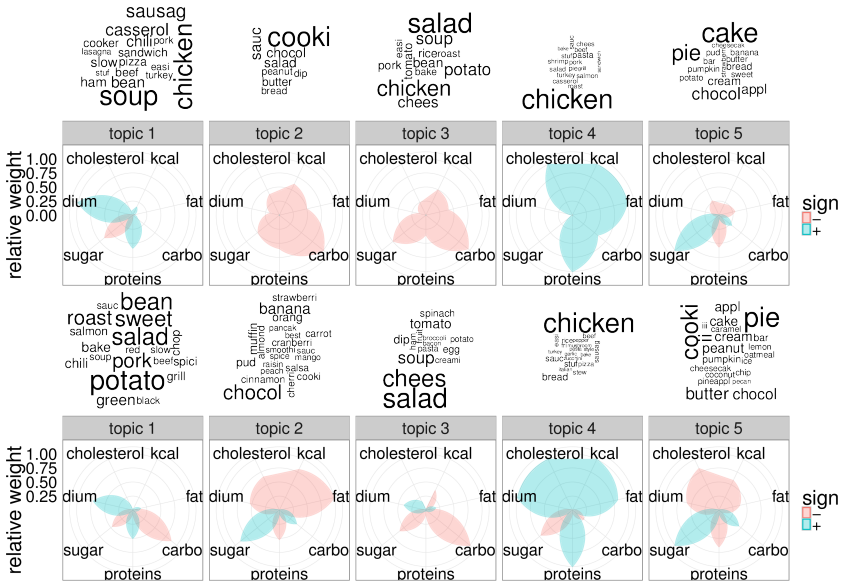
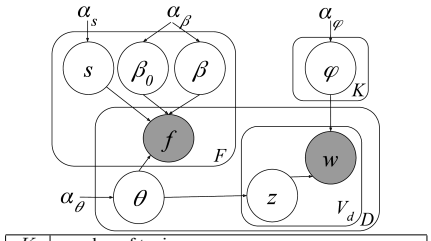


Figure 5: Comparison of topics found by standard LDA (bottom) and our model (top).



K	number of topics
D	number of documents (recipes)
F	number of outputs (nutrients)
V	number of unique words
V_d	number of words in document (recipe) d
θ	the multinomial distribution of topics for recipe d
φ	word distributions for topics
z	topic assignment for word w in document d
w	observed word from document d
f_{id}	observed i-th nutrient for document (recipe) d
β_i	i-th nutrient vector of topic weights
β_0^i	i-th nutrient bias
s_i	i-th nutrient standard deviation

Figure 4: LDA with built-in multi-output linear regression extension.

associated with all nutrient facts. Taking those observations together we can conclude that it is possible to model text topics in conjunction with related nutrient facts to provide an interpretable and low-dimensional representation of such content. Our approach to combining nutrients and topic modeling is to propose a model that uses the latent topic space to explain both observed words and nutritional values.

Among latent topic models, the most popular is *latent dirichlet allocation* (LDA). Unlike simple clustering methods, LDA allows text documents to exhibit multiple topics, i.e., each document is assigned a distribution of topics from which word topics and words are later drawn. We adopt LDA by extending it with multiple linear regression components.

Linear regression is a well established statistical technique where dependent variables are modeled as a weighted sum of explanatory variables and bias. Each input is assigned a weight that measures its influence on the output, e.g., close to 0 weight means no influence. In our case dependent variables are nutritional values and regression is repeated as many times as there are outputs. Explanatory variables are LDA topic distributions per document (recipe). In contrast to a standard approach where LDA clustering is done before and separated from regression, our approach combines both models and learns topic distributions in a way that they express both text clusters and all nutrient facts well.

The detailed graphical representation of our approach is presented in Figure 4. LDA is extended by adding F regressions (top left corner) where the i-th output ($i \in 1..F$) for the d-th ($d \in 1..D$) recipe is drawn from the normal distribution: $f_{id} \sim \mathcal{N}(\beta_i^T \theta_d + \beta_0^i, s_i)$, where β_i is a K-dimensional vector of topic weights for the i-th output, β_0^i is a bias and the output's error is modeled by the standard deviation s_i. The model has several parameters. α_θ and α_φ parameterize respectively the document topic distributions (Dirichlet) and the topic word distributions (Dirichlet). In both cases we used popular heuristics, i.e., $\alpha_\theta = 0.1$ and $\alpha_\varphi = 0.1$. Linear regression weights were assigned the following priors: $\beta \sim \mathcal{N}(0, \alpha_\beta)$ where we used uninformative parametrization with standard deviation $\alpha_\beta = 100$. Similarly for $s = \frac{1}{\sqrt{\tau}}$ where precision $\tau \sim gamma(\alpha_s^{shape}, \alpha_s^{rate})$ we used $\alpha_s^{shape} = 3$, $\alpha_s^{rate} = 3$.

6. EXPERIMENTAL EVALUATION

We prototyped our model using a black box Gibbs sampler, namely *JAGS*.[2] To study its behavior and performance we compare it to standard LDA using the popular *gensim*[3] implementation. First, we examine the clustering results for recipe topics identification. Then, we observe performance for the nutritional values prediction task.

6.1 Recipe Topics Identification

Figure 5 compares recipe title topics found by our model (top) and by standard LDA (bottom) along with associated nutrient facts weights. For visualization purposes we chose $K = 5$. For each topic we present 7 weights coupled to respective nutrients. In the first case (for our model) we simply present empirical expected values of β_i obtained for each of topics. However, standard LDA provides only textual topics and to obtain nutritional value weights in the second case, we applied linear regression separately. In the figure positive values are represented with a green and negative with

[2]http://mcmc-jags.sourceforge.net/
[3]https://radimrehurek.com/gensim/

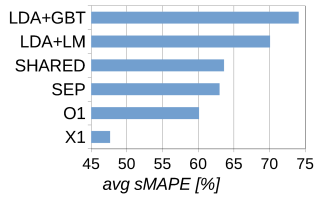


Figure 6: Nutrients prediction average performance.

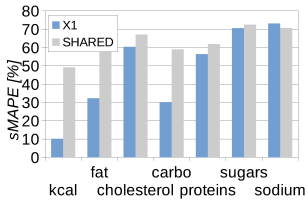


Figure 7: Prediction improvement when additional outputs values are included.

a red color. Although absolute values are not very informative and only relative proportions between topics are interpretable, weights were scaled, i.e., each weight was divided by the (absolute) maximum over topics (per nutrient fact).

The topics obtained by the two models differ noticeably both in terms of text and nutrient fact weights. Topics obtained by our model more consistently influence nutritional values, i.e., they either add or subtract from all outputs (for example topics 2 and 3 have only negative and topic 4 only positive weights), whereas standard LDA topics mix more often positive and negative influences. Furthermore, topics presented in the first row (our model) are more discriminative in selecting which nutrient facts they influence, i.e., they have more zeros. For example, only topic 1 contributes significantly to the output value of sodium. In Section 4 we observed that this nutrient correlates differently from the others and therefore our model separated it to a devoted topic. On the other hand, in the second row (standard LDA) almost all topics influence all outputs.

Topics discovered by our method differ also in terms of textual content. In general the first row in Figure 5 contains slightly more focused topics. For example, topic 1 is focused only on 'chicken' and 'soup' whereas in topic 1 in the second row there are 6 words of almost equal importance. Furthermore, we observe some differences in selecting and combining words. For example, our model combined 'cake' with 'pie', while the standard LDA combined 'pie' with 'cooki' (cookie) and 'cake' is not important for any topic.

6.2 Nutritional Values Prediction

Predicting nutritional values of an unknown recipe can be helpful and applied in various settings depending on circumstances. We simulate three of them: when we have no information about the recipe (apart from its title), when one of the nutrients is known (denote O1) and when all but one nutrient are known (denote X1). Furthermore, in the first setting we consider two variants: with latent representation shared between all outputs (denote SHARED) and with a separate model for each of outputs (denote SEP). In all cases we set $K = 30$, as it is a popular heuristic value. Our model we compare to LDA with *linear regression* (LDA+LM) and with *gradient boosted regression trees*[4] (LDA+GBT).

For evaluation purposes we split the data set randomly into training (80%) and test (20%) subsets. Predictions $\hat{f_{id}}$ for unseen recipes we obtained by taking empirical expected values of samples from the model with fixed (from the training phase) word distribution φ, weights β and deviations s (precisions τ). Prediction quality we measure using *symmetric Mean Absolute Percentage Error*: $sMAPE_i = \frac{2}{|test|} \sum_{d \in test} \frac{|f_{id} - \hat{f_{id}}|}{|f_{id}| + |\hat{f_{id}}|}$, a measure that has much better statistical properties than *MAPE*. Using percentage error allows us to compare results for different outputs and to average over outputs $i \in 1..F$ without favoring any of them: $avg\ sMAPE = \frac{1}{F} \sum_i sMAPE_i$.

Figure 6 compares prediction performance of our model (in several variants) to the baselines. We observe a significant improvement over LDA+GBT and LDA+LM for our approaches. Surprisingly there is not much difference between the model with the latent representation shared between outputs (SHARED) and using the separate representation for each of the outputs (SEP), showing that the same topics can be successfully used for many outputs. On the other hand, we observe a huge improvement when outputs are known additionally to the title (O1 and X1). However the boost is different for different outputs. Figure 7 presents the performance results split by nutrients. For fat, sugars and sodium, differences are negligible, but when other nutritional values are known, energy value (kcal) can be determined almost with no error. Although this result is not surprising it shows that our model makes proper use of additional information if provided.

[4]https://cran.r-project.org/web/packages/gbm/gbm.pdf

7. CONCLUSIONS

In this paper, we extended our understanding of the relation between textual content and associated nutritional values in the context of online food communities. Exploiting observations from the analytical part we designed a novel topic model that exploits both correlations between outputs and words. The performance of our method we evaluated in two practical tasks, i.e., compact topics identification and multiple outputs prediction. Furthermore we exhibited the extent to which including regression information in the training process helps in selecting the appropriate compact representation.

Acknowledgments. We thank Christoph Trattner for creating the *allrecipes.com* dataset and sharing it with us.

8. REFERENCES

[1] S. Abbar, Y. Mejova, and I. Weber. You tweet what you eat: Studying food consumption through Twitter. In *Proc. of CHI*, 2015.

[2] M. De Choudhury and S. S. Sharma. Characterizing dietary choices, nutrition, and language in food deserts via social media. In *Proc. of CSCW*, 2016.

[3] H. A. Schwartz et al. Characterizing geographic variation in well-being using tweets. In *Proc. of ICWSM*, 2013.

[4] T. Kusmierczyk, C. Trattner, and K. Nørvåg. Temporality in online food recipe consumption and production. In *Proc. of WWW*, 2015.

[5] M. Muller et al. Ingredient matching to determine the nutritional properties of internet-sourced recipes. In *Proc. of PervasiveHealth*, 2012.

[6] D. M. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proc. of UAI*, 2008.

[7] M. J. Paul and M. Dredze. You are what you tweet: Analyzing Twitter for public health. In *Proc. of ICWSM*, 2011.

[8] A. Said and A. Bellogín. You are what you eat! Tracking health through recipe interactions. In *Proc. of RSWeb*, 2014.

[9] C. Wagner, P. Singer, and M. Strohmaier. The nature and evolution of online food preferences. *EPJ Data Science*, 3(1):1–22, 2014.

[10] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proc. of SIGKDD*, 2011.

[11] R. West et al. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. In *Proc. of WWW*, 2013.