Recipes Review
First Project for NLP Course, Winter 2022

Amir Ali
Warsaw University of Technology
amir.ali.stud@pw.edu.pl

Stanisław Matuszewski
Warsaw University of Technology
first2.last2@xxxxx.com

Jacek Czupyt
Warsaw University of Technology
jacek.czupyt@gmail.com

Supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

Food computing recently emerged as a stand-alone research field, in which various Natural Language Processing (NLP) are applied to the various stages of food production pipelines. In this project, we introduce a new dataset called FoodReview where we collect more than 18k+ comments which were given on different Recipes. The purpose here is to extract ingredients from different comments and For that, we will implement the Named Entity Recognition model which is widely used for text extraction.

1 Introduction

Food computing is an interdisciplinary field that focuses on the development of technology and systems for the processing, distribution, and consumption of food. It recently emerged as a stand-alone research field, in which artificial intelligence, deep learning, and data science methodologies are applied to the various stages of food production pipelines. Food computing may help end-users in maintaining healthy and nutritious diets by alerting them of high-caloric dishes and/or dishes containing allergens (0). Food computing can also involve the use of virtual and augmented reality to enhance the culinary experience, as well as the use of machine learning algorithms to create personalized nutrition plans and improve food safety.

Owing to that, food computing has currently a rapidly growing research field. Nowadays, Natural Language Processing (NLP), especially Natural Language Understanding(NLU), is playing an increasingly important role in the field of recognizing food entities.

In our project, we introduced a new dataset – called FoodReview, which consists of recipes with 18k+ real users comments from www.tasteofhome.com. Apart from building a dataset, we aimed at detecting all recipe ingredients inside of collected comments and based. We used Stanford and Spacy NER models in order to handle these tasks.

2 Literature review

While working on the project, we used available studies on the topic of food computing (0). The authors of the first similar to us undertook the task of preparing and processing a new data set of their own - TASTEset. It contained ingredient data from 700 recipes, a total of 13,000 entities. In further work, they compared the effectiveness of several NER models - BERT and LUKE. The main limitation encountered by the authors was that the amount of available data was too low.

In this paper (0), the authors undertook the task of classifying input recipes into one of the following classes: pork, nut grains, meat, gluten, fish, dairy, seafood, restricted diet, and egg or shrimp. To achieve this, the authors used the Random Forest Classifier model and compared the effect of different ways of training on their accuracy.

In this paper (0), the authors proposed a novel way to represent recipes and also experimented with them. They have implemented their model for the estimation of nutrition from cooking recipes, as well as for finding similar recipes in the dataset. They proposed to represent each recipe as a set of keywords belonging to three categories - ingredients, utensils, and processes. The models were learned on a dataset consisting of 118,171 recipes from RecipeDB. Their best model gave an F1 Score of greater than or equal to 0.95 across all datasets.

In the last paper (0), the authors used deep learning models - LSTM and GPT-2 models to generate the novel recipes using a list of ingredients as input. As in the previous work, the learning process took place on the RecipeDB collection. The project also includes a web application in which the user, after selecting a set of available ingredients, obtained the novel recipe created by the model. The application is publicly available at https://cosylab.iiitd.edu.in/ratatouille2/

3 Data Collection

We gathered our data from the well-known cooking recipe website tasteofhome.com, we downloaded the ingredients and comments left on the list of 100 most popular recipes presented by the site. We used the Selenium library to scrape the list of ingredients from each of the HTML pages. As for the comments, we found and used the hidden backend API used by the site, which avoided the trouble of loading all the comments on the HTML page, and gave us some additional data that would be otherwise difficult to scrape.

In total, we obtained 100 lists of ingredients and 18182 comments, gathered in two CSV files. The ingredients dataset simply contains name of the recipe, and scraped text with the list of ingredients. The comment dataset contains a bit more information, which could be useful for some machine learning problems, as well as for scraping additional related data from the site (figure 1).

4 Exploratory Data Analysis

The histogram in figure 1 shows the distribution of comments by the recipe ranking. As expected, the recipe's popularity is positively correlated with the number of comments and the data appears to be following some form of power law distribution. With only one exception, the number of comments per recipe never falls below 85.

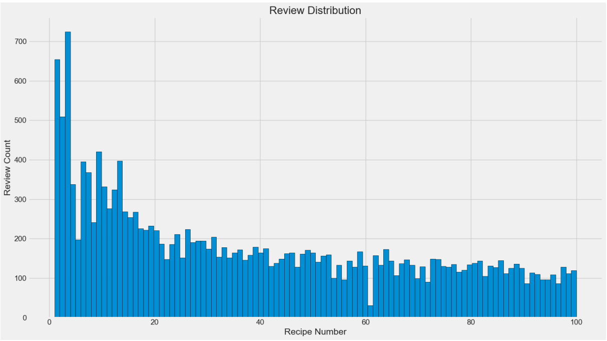


Figure 1: Food Recipes Distribution

Column name	Description
recipe_name	name of the recipe the comment was posted on
recipe_number	placement of the recipe on the top 100 recipes list
recipe_code	unique id of the recipe used by the site
comment_id	unique id of the comment
user_id	unique id of the user who left he comment
user_name	name of the user
user_reputation	internal score of the site, roughly roughly quantifying the past behaviour of the user
create_at	time at which the comment was posted as unix timestamp
reply_count	number of replies to the comment
thumbs_up	number of up-votes the comment has received
thumbs_down	number of down-votes the comment has received
stars	the score on a 1 to 5 scale that the user gave to the recipe. A score of 0 means that no score was given
best_score	score of the comment, likely used by the site the help determine the order the comments appear in
text	the text content of the comment

Table 1: Columns of the comments dataset

The figures 2 and 3 show the comment's character count and word count distributions respectively. Most comments are rather short, averaging around 40 words or 200 characters. There are no zero-length comments, as those are impossible to post one the site, but there are a few single character comments. The longest comment has 556 words and 2742 characters.

The figure 4 shows the distribution of the review scores given by the comments, on a range from 1 to 5 stars, where 0 represents a lack of review. As can be seen, the vast majority of the reviews gave the maximum score. This is to be expected, not only because online reviews of anything are in general mostly positive, but also because this represents the list of the most popular recipes. However the extreme proportion of positive reviews is

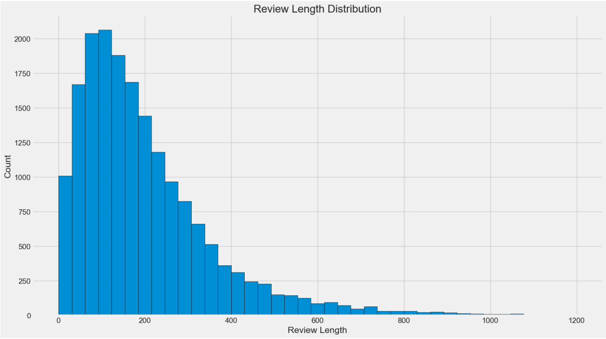


Figure 2: Comment Character Count Distribution

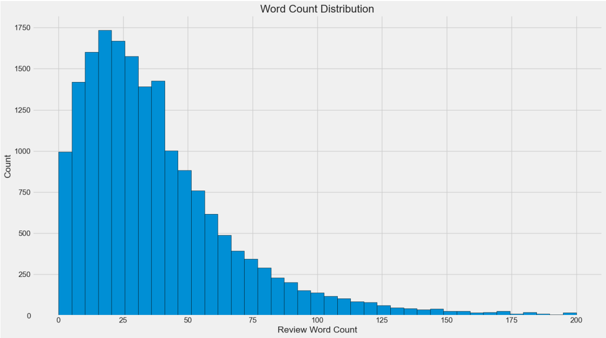


Figure 3: Comment Word Count Distribution

still a little surprising, totaling to an average score of 4.73 stars, but given our use of this data, this was not a problem.

It's very important to understand the context of the text, especially when working on Text Extraction, because every word's meaning is important. In order to better understand the sentiment of the text, we used the TextBlob library to find the estimated sentiment polarity of the comments. The results are present in figure 5. The polarity score ranges between -1 and 1, where positive values represent a positive sentiment and negative values represent a negative sentiment. The graph shows that more than 88% of the comments have a positive sentiment. However the majority of them to be less extreme then could have been expected from the ratings distribution, bringing the mean value to only 0.36,

5 Data Preprocessing

Data Preprocessing is an important role to build a model. If we see our text we found special case letters, digits, Punctuation, and repeat words with different forms. So, to handle these stuff we will implement the followings:

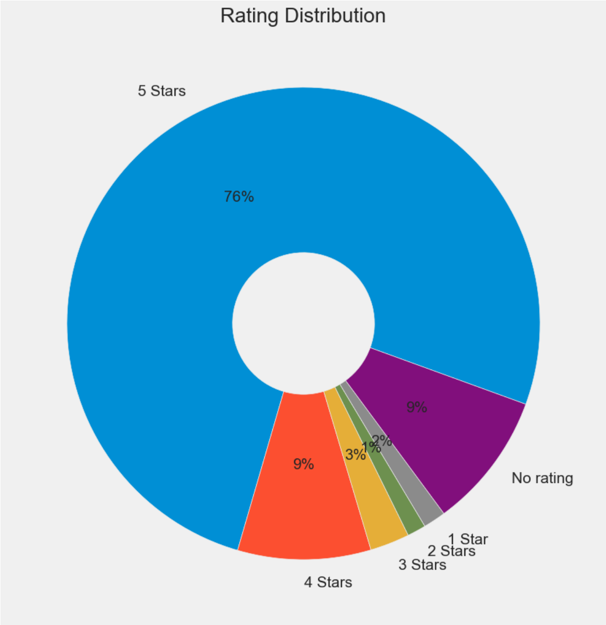


Figure 4: Review Rating Distribution

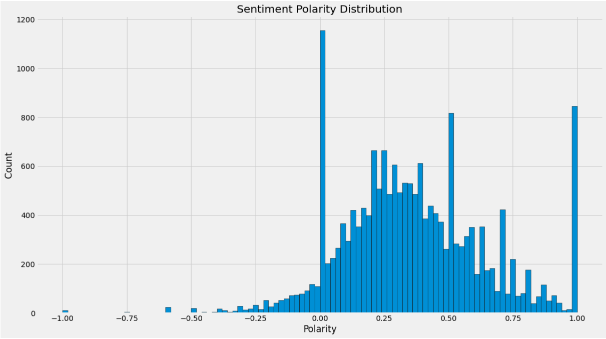


Figure 5: Sentiment Polarity Distribution

5.1 Regular Expression

Regular Expression is a key role to clean the text. By using Regular Expressions we will remove the punctuation and special case letters like @%$*.

5.2 Stop Word Removal

As we deal with textual data of recipe review. If we see our target as extracting the ingredients from the text then words like A, The, IS, ARE, etc don't necessary. So, it's very import remove such words which have no meaning.

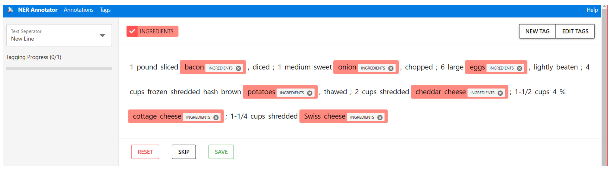
5.3 Lemmatization

Lemmatization is a process where we can derive words into root words. For instance, both "tomatoes" and "Tomato" are output as "tomato". The purpose is to categorize into 1 word.

5.4 NER Annotations

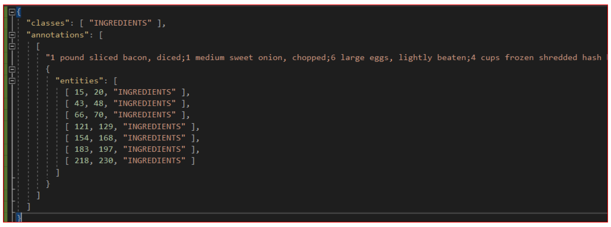
In order to further work with NER models, we needed to add annotations to the data indicating parts of the descriptions containing key words - in our case, ingredients. Due to the relatively small size of the dataset (100 recipes), we decided to use a manual process to add annotations. The choice of this method is also associated with the highest quality of added annotations i.e without any errors possibly made by ML algorithms.

For the data annotations needed for NER models, we used the dedicated NER Text Annotator application available online(0). It allows the user to easily mark parts of the text and assign them to the appropriate tag category. In our case, the only tag added to texts was *INGREDIENTS*. At the end of the process, the application generates a dedicated, easily compatible with the NER Models classes JSON file.



1 pound sliced bacon INGREDIENTS , diced ; 1 medium sweet onion INGREDIENTS , chopped ; 6 large eggs INGREDIENTS , lightly beaten ; 4 cups frozen shredded hash brown potatoes INGREDIENTS , thawed ; 2 cups shredded cheddar cheese INGREDIENTS ; 1-1/2 cups 4 % cottage cheese INGREDIENTS ; 1-1/4 cups shredded Swiss cheese INGREDIENTS

Figure 6: Adding anotations - NER Anontator interface



```
{
  "classes": [ "INGREDIENTS" ],
  "annotations": [
    [
      "1 pound sliced bacon, diced;1 medium sweet onion, chopped;6 large eggs, lightly beaten;4 cups frozen shredded hash
      {
        "entities": [
          [ 15, 20, "INGREDIENTS" ],
          [ 43, 48, "INGREDIENTS" ],
          [ 66, 70, "INGREDIENTS" ],
          [ 121, 129, "INGREDIENTS" ],
          [ 154, 168, "INGREDIENTS" ],
          [ 183, 197, "INGREDIENTS" ],
          [ 218, 230, "INGREDIENTS" ]
        ]
      }
    ]
  ]
}
```

Figure 7: Adding annotations - an output JSON file

6 Methodology

In this project, we going to implement the Named Entity Recognition(NER) model which is widely used for information Extraction. Our aim is to extract the ingredient from Recipes Reviews.

In this project, we will implement the following techniques:

6.1 NER Model with Spacy 3.0

Here, the model uses a pre-defined set of rules for information extraction. Mainly for two purposes pattern-based and content base rules. This type of model is relatively simple to make and understand, however it is unable to process complicated expressions. Our main target is to use this technique to extract those ingredients that are connected with color. like green chili, etc.

The models are initially trained on labeled training data - the training sample is processed by the model, then its prediction is compared to the correct label using a loss function, and the model's weights are adjusted accordingly using backpropagation. The model's accuracy is then verified using labeled evaluation data, which was not used during training. We can then use the system to predict the labels for new data.

When preparing the training dataset, It is important to make sure it is large and varied enough to prevent overfitting - a situation where the model memorizes the training data instead of finding a general pattern. It is also crucial to ensure that the training data is derived from the same or similar distribution to the data one plans to process. For example, a model trained on English data will not be able to process text, and a model trained on reddit will likely not perform well when processing legal documents.

To help implement such a model, We use SpaCy - an advanced open-source library designed to facilitate commercial-grade natural language processing. It contains powerful tools that help in processing high volumes of text and building systems for various forms of information extraction and understanding.

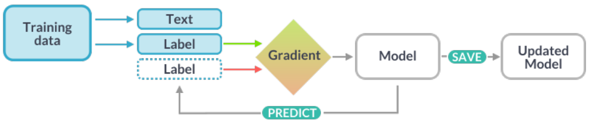


Figure 8: Machine Learning Based NER model(0)

7 Results

We experiment with different hyperparameters of the NER spacy model. The experiments are like Learning rate, Hidden Layer and Activation Function. The maximum F1 score that we are 0.67. Below you can see the table with the results of each experiment

Model	F1 Score	Precision	Recall
Learning Rate	0.67	0.71	0.69
Hidden Layer	0.52	0.58	0.54
Activation Function	0.59	0.67	0.60

8 Conclusion

We collected a very new dataset with crucial information about Recipe reviews and Ingredients. We build our own custom model to extract Ingredients from Review.

But this data was not enough to train the model perfectly. So, we need more annotated data to improve the model's accuracy.

9 Future Work

Based on the Result and Limited Dataset the model is not perfectly Trained. So, for the Future Project, we are planning to used the TASTEset dataset (0) and will test different predefined Models like Stanford NER, BER, and Transformers model and will compare the results.

References

N. Gilal, M. Agus, J. Schneider, and K. Al-Thelaya, "Slowdeepfood: a food computing framework for regional gastronomy," 11 2021.

A. Wróblewska, A. Kaliska, M. Pawłowski, D. Wiśniewski, W. Sosnowski, and A. Ławrynowicz, "Tasteset – recipe dataset and food entities recognition benchmark," 2022.

S. Gashi, E. Di Lascio, and S. Santini, *Multi-class Multi-label Classification for Cooking Activity Recognition*, pp. 75–89. Singapore: Springer Singapore, 2021.

N. Diwan, D. Batra, and G. Bagler, "A named entity based approach to model recipes," in *2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW)*, pp. 88–93, 2020.

M. Goel, P. Chakraborty, V. Ponnaganti, M. Khan, S. Tatipamala, A. Saini, and G. Bagler, "Ratatouille: A tool for novel recipe generation," in *2022 IEEE 38th International Conference on Data Engineering Workshops (ICDEW)*, pp. 107–110, 2022.

"Ner text annotator."

D. Campos, S. Matos, and J. L. Oliveira, "Biomedical named entity recognition: A survey of machine-learning tools," in *Theory and Applications for Advanced Text Mining* (S. Sakurai, ed.), ch. 8, Rijeka: IntechOpen, 2012.