T.R.

GEBZE TECHNICAL UNIVERSITY

FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER ENGINEERING

PREDICTING CUISINE NAME FROM RECIPE INGREDIENTS

TUBA TOPRAK

SUPERVISOR
YRD. DOÇ. DR. BURCU YILMAZ

GEBZE
2022

T.R.
GEBZE TECHNICAL UNIVERSITY
FACULTY OF ENGINEERING
COMPUTER ENGINEERING DEPARTMENT

PREDICTING CUISINE NAME FROM RECIPE INGREDIENTS

TUBA TOPRAK

SUPERVISOR
YRD. DOÇ. DR. BURCU YILMAZ

2022
GEBZE

This study has been accepted as an Undergraduate Graduation Project in the Department of Computer Engineering on 31/08/2022 by the following jury.

JURY

Member
(Supervisor) : Yrd. Doç. Dr. Burcu YILMAZ

Member : Doç. Dr. Habil KALKAN

ABSTRACT

Over the years, people have tried to discover new ingredients and incorporate them into recipes or create new recipes. With the recipes, an international food culture has emerged. While neighboring countries were influenced by each other, some of them created their own cuisines.Therefore, it has become difficult to distinguish cuisines that are so similar to each other.One of the obvious relationships to explore is the relationship between ingredients and cuisines.This thesis describes deep learning and machine learning models applied to predict cuisine based on ingredients.

Keywords: cuisine, predict,ingredient.

ÖZET

Yıllar geçtikçe insanlar yeni malzemeler keşfetmeye ve bunları tariflere dahil etmeye veya yeni tarifler yaratmaya çalıştı. Tariflerle birlikte uluslararası bir yemek kültürü ortaya çıkmıştır. Komşu ülkeler birbirinden etkilenirken bazıları kendi mutfaklarını oluşturmuşlardır. Bu nedenle birbirine bu kadar benzeyen mutfakları ayırt etmek güçleşmiştir. Keşfedilmesi gereken bariz ilişkilerden biri de malzemeler ve mutfaklar arasındaki ilişkidir. Bu tez, malzemelere dayalı olarak mutfağı tahmin etmek için uygulanan derin öğrenme ve makine öğrenimi modellerini açıklar.

Anahtar Kelimeler: mutfak, içindekiler.

ACKNOWLEDGEMENT

Special thanks to my advisor, Burcu YILMAZ, for her patience, insightful comments, practical ideas, and powerful wisdom. I have a deep admiration and respect for her for helping me with encouraging me under all circumstances to complete this project.

Also, I would like to thank my family and friends for their endless support.

Tuba TOPRAK

LIST OF SYMBOLS AND ABBREVIATIONS

Symbol or Abbreviation	:	**Explanation**
NN	:	Neural Network
SVM	:	Support Vector Machine
ML	:	Machine Learning
NLP	:	Natural Language Processing
TF-IDF	:	Term Frequency–Inverse Document Frequency
BERT	:	Bidirectional Encoder Representations from Transformers

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. PROJECT DESCRIPTION

Food is an indispensable part of our lives. The most basic element with which one can identify a food item are its ingredients. Ingredients are the atomic components of food. Over the years, people have tried to explore new ingredients and incorporate them into recipes or produce new recipes all together. However, the choice of ingredients is characterized by geographical locality. One of the factors responsible for this behaviour could be the similarity in availability of an ingredient in a particular geographic region. This has resulted in the set of recipes being divided into geographic classes known as cuisines. One of the obvious relations that I would like to explore is the relation between ingredients and cuisines. It is quite apparent that availability and popularity are important factors influencing the choice of ingredients in a recipe. People in different regions have different taste preferences and hence tend to favor a particular set of ingredients in comparison to the other. Thus, there seems to be a strong co-relation between these two entities. In this project, I tried to develop a model to classify a recipe based on the ingredients it uses.

Figure 1.1: World Cuisine.

1.2. PROJECT PURPOSE

The aim of this project is to find out which culinary culture a dish belongs to with its ingredient list.

2. DATASET

The Yummly dataset used for the prediction task consists of 39,774 recipes. Each recipe is associated with a particular cuisine and a particular set of ingredients. Initial analysis of the data-set revealed a total of 20 different cuisines and 6714 different ingredients.Cuisine is the target variable.The Ingredients for every recipe is given as a list. Their are recipies from 20 different Cuisines. This means that the problem at hand is a multi-class classification.

2.1. Cuisine

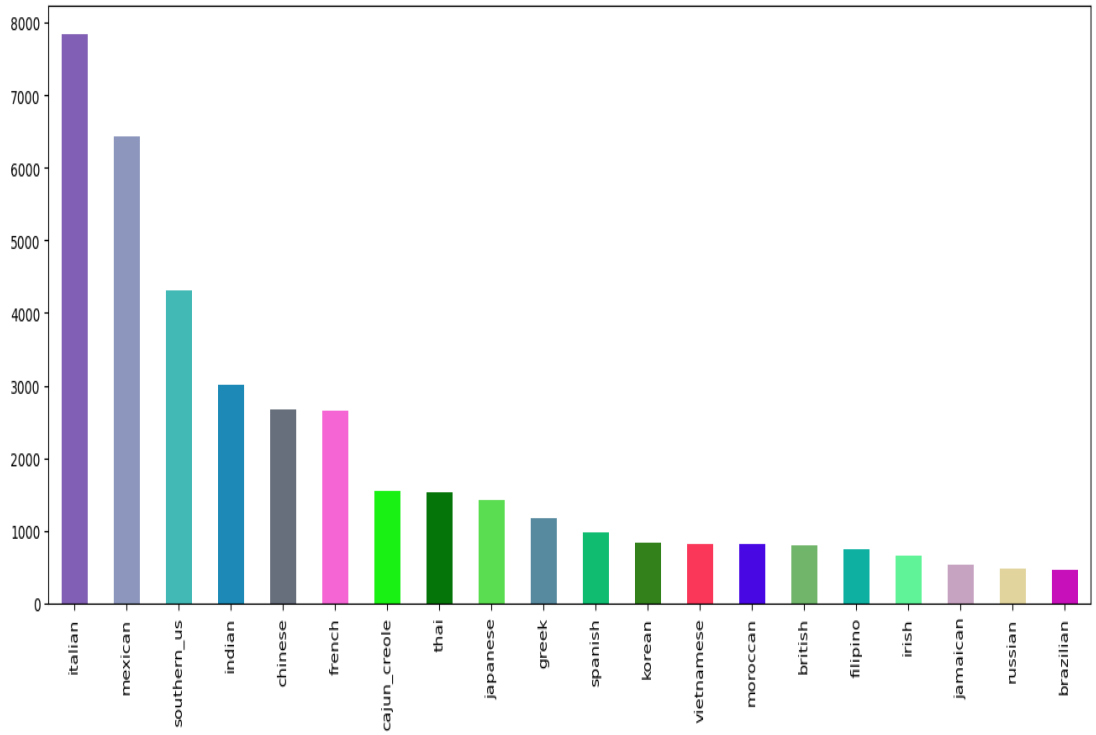

Figure 2.1: : Counts of all cuisines.

From the plot of label distribution, we observe that the most common category in our sample is the Italian cuisine, followed by the Mexican. The least represented cuisines are the Irish, Jamaican, Russian and Brazilian - counting for only 6% of our training sample of recipes.

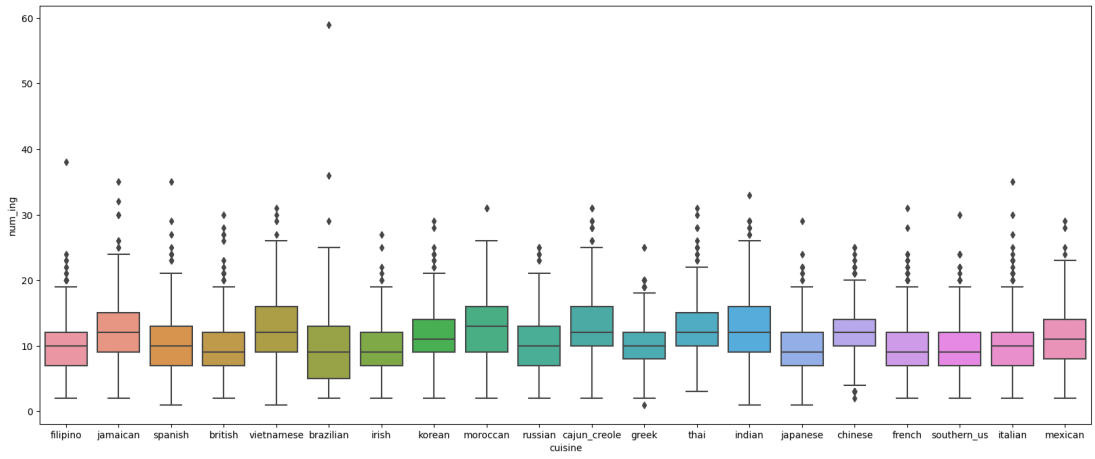

Figure 2.2: :Box plots of recipe length distributions.

The Moroccan cuisine seems to have the longest recipes on average compared to all the rest cuisines in our sample.The opposite phenomenon is observed for Irish, British, French and Southern US cuisine.There exist outliers in all cuisines.Recipes part of the European cuisine tend to be with average length or shorter compared to the rest of the sample.

2.2. Ingredients

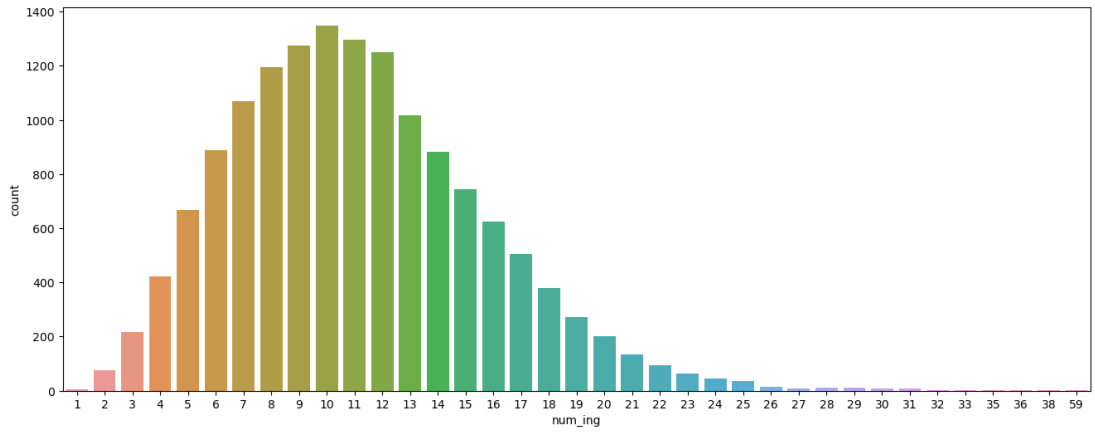

Figure 2.3: : number of ingredients in recipes.

A kitchen can often be known for its unique ingredients. Top 3 ingredients with each cuisine in the dataset:

• Brazilian: cachaca, acai

• British: stilton cheese, suet

- Cajun Creole: Cajun seasoning, andouille sausage
- Chinese: Shaoxing wine, Chinese five-spice powder
- Filipino: lumpia wrappers, calamansi
- French: Gruyere cheese, Cognac
- Greek: feta cheese, Greek seasoning
- Indian: garam masala, ground turmeric
- Irish: Irish whisky, Guinness
- Italian: parmesan cheese, ricotta cheese
- Jamaican: scotch bonnet chiles, jerk seasoning
- Japanese: mirin, sake
- Korean: Gochujang, kimchi
- Mexican: corn tortillas, salsa
- Moroccan: couscous, preserved lemon
- Russian: beets, buckwheat flour
- Southern US: buttermilk, grits
- Spanish: chorizo, serrano ham
- Thai: red curry paste, fish sauce
- Vietnamese: fish sauce, rice paper

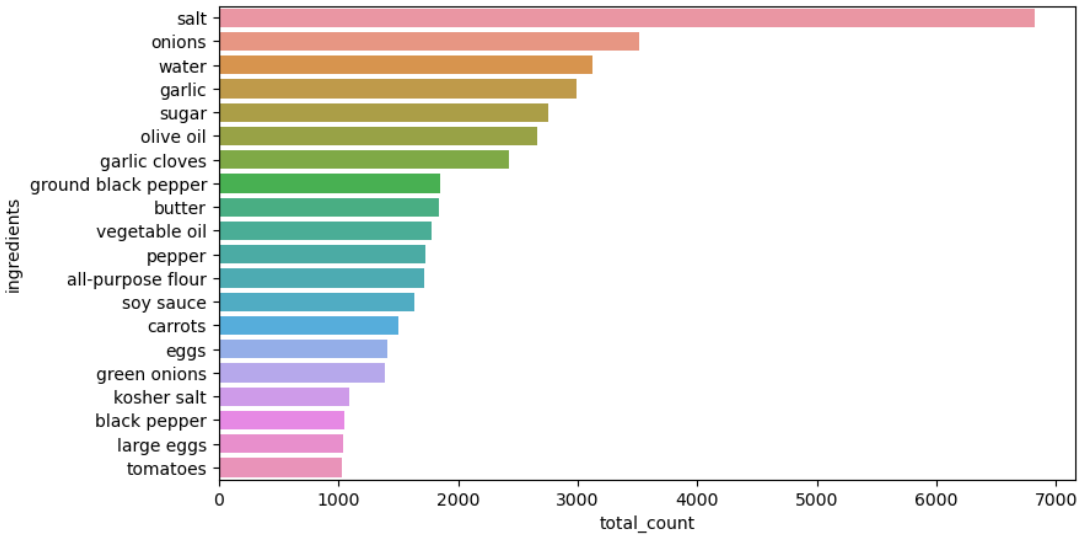

Figure 2.4: :Most Common Ingredients in the whole dataset.

Here are the popular ingredients in every kitchen. This gives an idea of the materials that make up an integral part of the kitchen. As you can see, Salt is at the top of the list in almost every kitchen. That's why ingredients like salt and onions don't help define the cuisine. Every kitchen should have ingredients that are rarely found in other kitchens. That's why this content has been deleted.

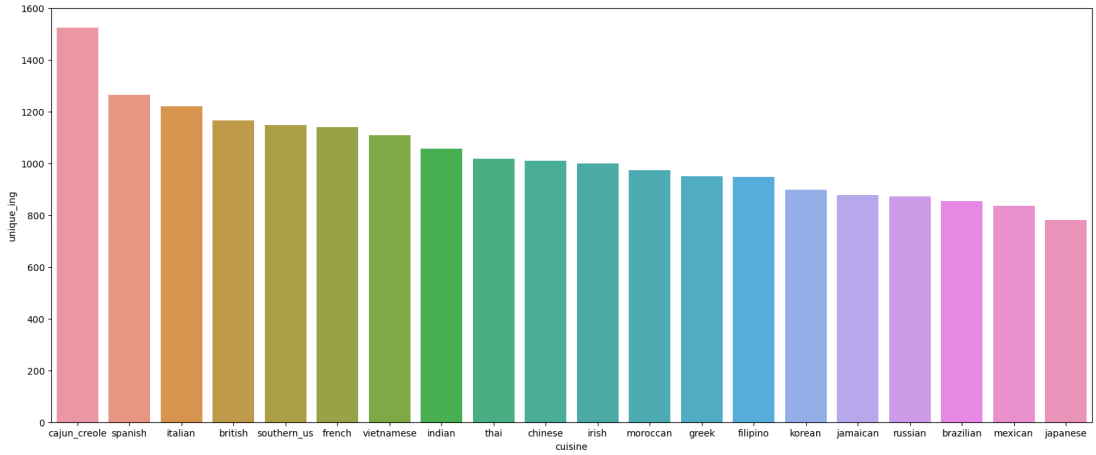

Figure 2.5: :Unique ingredients in cuisines

From the bar chart above it can be seen that kitchens with more examples in the training example need not be associated with more content that represents them. It turned out that French cuisine, which makes up 6.65% of the training sample, has more variability in content than Indian cuisine (this observation is unexpected as Indians use many spices in their recipes).

3. SYSTEM ARCHITECTURE

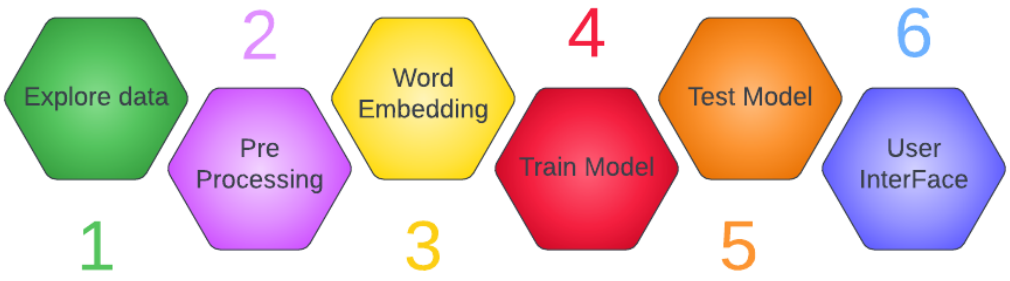

Figure 3.1: :System Architecture

3.1. Pre- Processing

After observing the dataset, I pondered upon the mining techniques that could be applied to this dataset. I explored the possibilities of application of regression, dimensionality reduction and found that only classification could be applied to this dataset.

Since the data is limited for estimating cuisines based only on ingredients, I used some prior knowledge about recipes. E.g. an instance containing flour, butter and sugar would have a high probability of having eggs in it. This not only presents a pattern with respect to ingredient duplets and triplets but also opens up a lot of possibilities for exploration despite the size limitation of the dataset.

The problem of multi-class text classification required cleaning and adjustment of data according to needs and I worked on a general framework for achieving that. Input and output of data was done through Pandas library. All the text was converted into lowercase.

Initial approaches involved the use of NLTK library So this part was done only for the ingredient lists present in the dataset. In this section, the following was done: The unbalanced dataset has been balanced.
Remove outliers
Convert to lowercase
Remove hyphen

Remove numbers
Remove words which consist of less than 2 characters
Remove units
Remove accents
Lemmatize

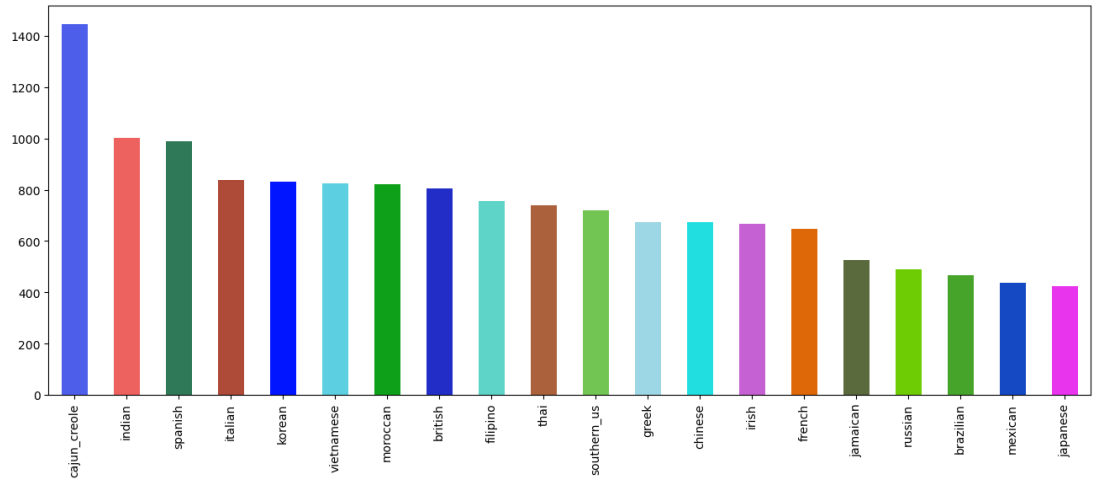

Figure 3.2: :Balance Dataset

3.1.1. Special Characters

The Special Characters have come from ingredients from a specific company or from ingredients which come in packaging: ”Bertolli® Alfredo Sauce” , ”Progresso™ Chicken Broth”, ”green bell pepper, slice”, ”half & half”, ”asafetida (powder)”, ”Spring! Water”

```
In [22]: ' '.join(sorted([char for char in set(' '.join(raw_ingredients)) if re.findall('[^A-Za-z]', char)]))

Out[22]: "  ! % & ' ( ) , - . / 0 1 2 3 4 5 6 7 8 9 ® â ç è é í î ú ' € ™"
```

Figure 3.3: :Special Characters

3.1.2. Upper Cases

Company names and Region Names may be capitalized as they are used in the content, so it has been checked and convert lowercase.

```
In [23]: list(set([ingredient for ingredient in raw_ingredients if re.findall
         ('[A-Z]+', ingredient)]))[:5]

Out[23]: ['English toffee bits',
          'Tuttorosso Diced Tomatoes',
          'Fisher Pecan Halves',
          'Jell-O Gelatin Dessert',
          "Frank's® RedHot® Original Cayenne Pepper Sauce"]
```

Figure 3.4: :Upper Cases

3.1.3. Apostrophes

```
In [24]: list(set([ingredient for ingredient in raw_ingredients if ''' in ingre
         dient]))

Out[24]: ['sheep's milk cheese', 'Zatarain's Jambalaya Mix', 'Breakstone's S
         our Cream']
```

Figure 3.5: :Apostrophes

3.1.4. Hyphens

Hyphens replaced with empty string.

```
In [25]:
list(set([ingredient for ingredient in raw_ingredients if re.findall
('-', ingredient)]))[:5]

Out[25]:
['chinese five-spice powder',
 'free-range chickens',
 'soft-boiled egg',
 'Jell-O Gelatin Dessert',
 'veal demi-glace']
```

Figure 3.6: :Hyphens

3.1.5. Numbers

Numbers indicate amount or density. Quantities can be a factor in defining the cuisine, but the number of ingredients containing the numbers is only 40, so their number has been removed.

```
In [26]:
temp_ing = list(set([ingredient for ingredient in raw_ingredients if r
e.findall('[0-9]', ingredient)]))
temp_ing[:6]

Out[26]:
['33% less sodium ham',
 '33% less sodium cooked ham',
 '2% milk shredded mozzarella cheese',
 '95% lean ground beef',
 'v8',
 'mexican style 4 cheese blend']
```

Figure 3.7: :Numbers

3.1.6. Units

Removed all units from ingredients as they would not be useful to the model.

In [28]:
```
units = ['inch', 'oz', 'lb', 'ounc', '%'] # ounc is a misspelling of ou
nce?

@interact(unit=units)
def f(unit):
    ingredients_df = pd.DataFrame([ingredient for ingredient in raw_in
gredients if unit in ingredient], columns=['ingredient'])
    return ingredients_df.groupby(['ingredient']).size().reset_index(n
ame='count').sort_values(['count'], ascending=False)
```

	ingredient	count
0	kinchay	3
1	pork chops, 1 inch thick	2

Figure 3.8: :Units

3.1.7. Regions

It was searched whether the name of the region was found in the name of the ingredients, For example:-Greek Yogurt. If found it will help this model and a model will outperform it.

In [30]:
```
d['american']
```

Out[30]:
```
['american cheese slices',
 'american cheese slices',
 'american cheese slices',
 'american cheese food']
```

Figure 3.9: :Regions

3.2. Word Embedding

TF-IDF is a statistical measure used to determine the mathematical significance of words in documents. The vectorization process is similar to One Hot Encoding. The value corresponding to the word is assigned a TF-IDF value instead of 1. The TF-IDF value is obtained by multiplying the TF and IDF values.In order for computers to perform mathematical operations, their ingredeints must be converted to numeric values. Text data converted into numeric data using tf-idf and Bert.

3.3. Models

3.3.1. SVC:

A multi-class Support Vector Classifier was used to classify the recipes. Like Logistic Regression, a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. The performance of the model was best observed for a value of C = 100.

```
Out[51]: OneVsRestClassifier(estimator=SVC(C=100, coef0=1, decision_function_shape='ovo',
                                           gamma=1),
                             n_jobs=4)
```

Figure 3.10: :SVC Model

3.3.2. Neural Network:

3.3.3. Neural Network with TF-IDF:

It is a model inspired by the human brain and nervous system. Neural Network is a structure established in layers. The first layer is called the input and the last layer is called the output. The layers in the middle are called Hidden Layers. The digitized information using TF-IDF was run in the model below and the result was obtained.

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 2229)]            0

 dense (Dense)               (None, 80)                178400

 batch_normalization (BatchN  (None, 80)               320
 ormalization)

 activation (Activation)     (None, 80)                0

 dropout (Dropout)           (None, 80)                0

 dense_1 (Dense)             (None, 20)                1620

 batch_normalization_1 (Batc  (None, 20)               80
 hNormalization)

 activation_1 (Activation)   (None, 20)                0

=================================================================
```

Figure 3.11: :Neural Network Model with TF-IDF

3.3.4. Neural Network with Bert:

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing NLP pre-training developed by Google.Since it is a pre-trained model, it has been tried to get successful results.

```
Model: "model"
__________________________________________________________________________________________________
 Layer (type)                   Output Shape         Param #     Connected to
==================================================================================================
 text (InputLayer)              [(None,)]            0           []

 preprocessing (KerasLayer)     {'input_mask': (Non  0           ['text[0][0]']
                                e, 128),
                                 'input_word_ids':
                                (None, 128),
                                 'input_type_ids':
                                (None, 128)}

 BERT_encoder (KerasLayer)      {'pooled_output': (  41373185    ['preprocessing[0][0]',
                                None, 512),                       'preprocessing[0][1]',
                                 'default': (None,                'preprocessing[0][2]']
                                512),
                                 'sequence_output':
                                 (None, 128, 512),
                                 'encoder_outputs':
                                 [(None, 128, 512),
                                 (None, 128, 512),
                                 (None, 128, 512),
                                 (None, 128, 512),
                                 (None, 128, 512),
                                 (None, 128, 512),
                                 (None, 128, 512),
                                 (None, 128, 512)]}

 dropout (Dropout)              (None, 512)          0           ['BERT_encoder[0][9]']

 classifier (Dense)             (None, 20)           10260       ['dropout[0][0]']

==================================================================================================
```

Figure 3.12: :Neural Network Model with BERT

3.4. User Interface

The interface is created in Python through the Tkinter library. There is a textbox for the user to enter the ingredient list, a button to predict the entered list, and an output box for displaying the result.

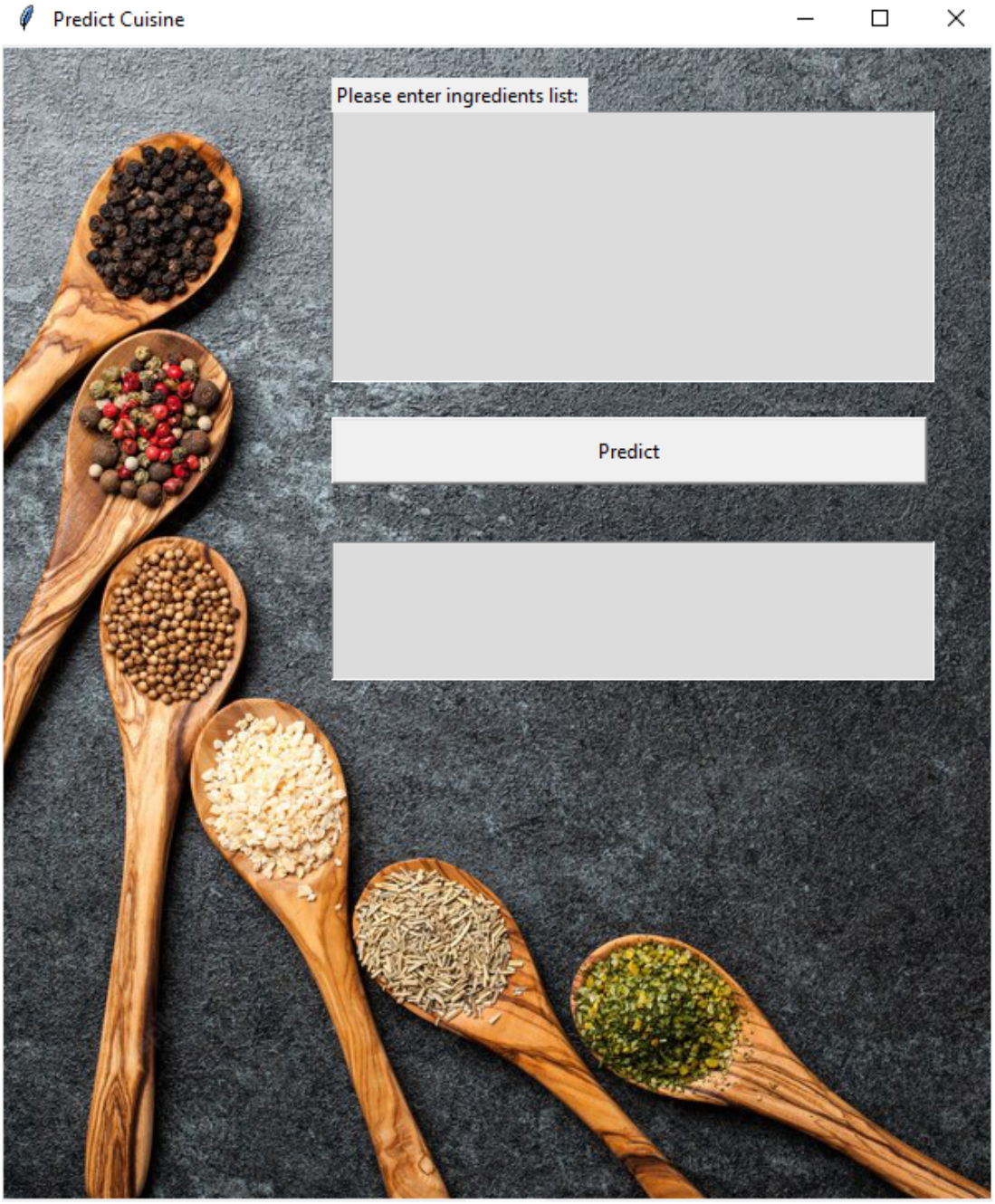

Figure 3.13: :User Interface

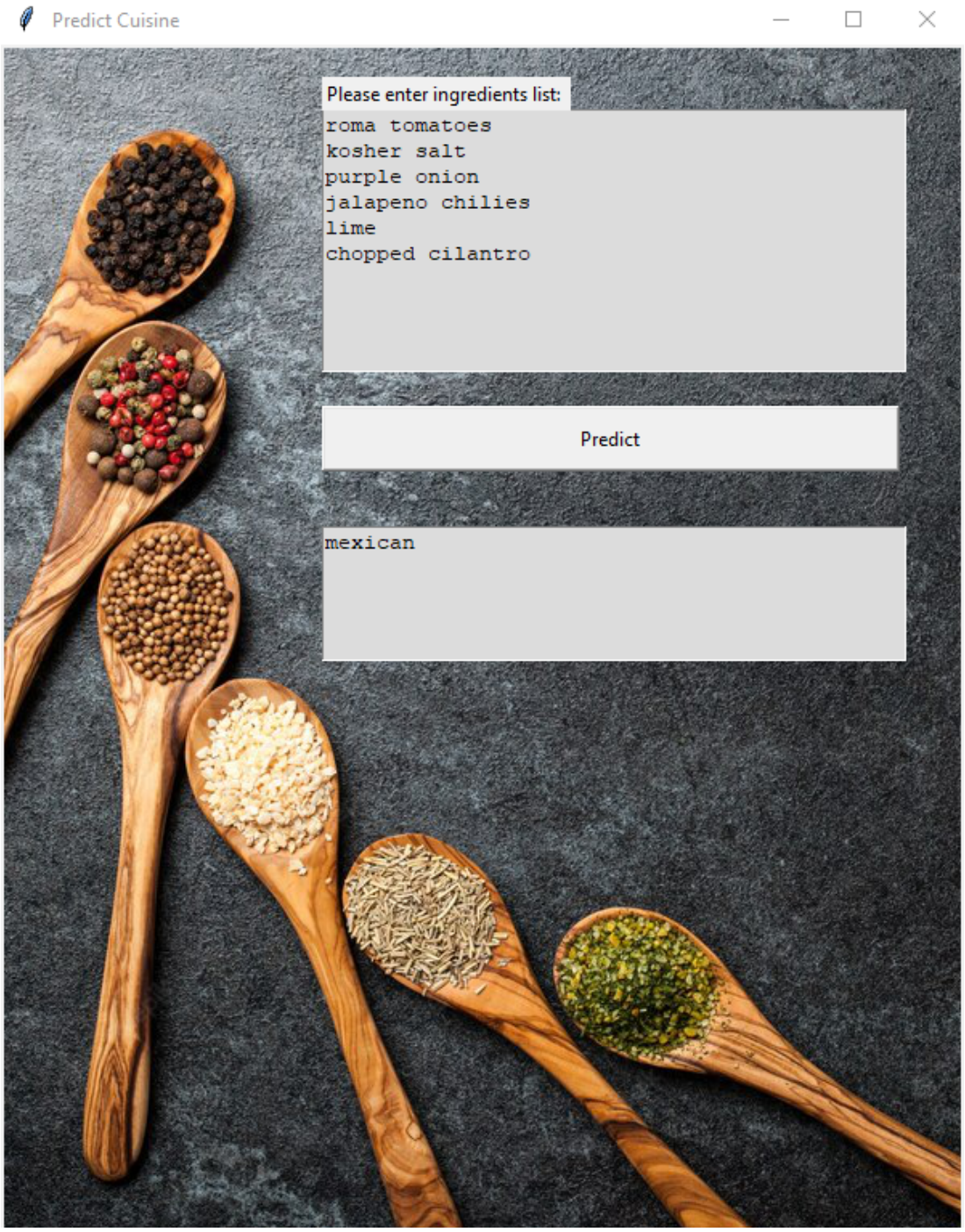

Figure 3.14: :User Interface

4. CONCLUSION

My approach to the problem of identifying the cuisines based on the ingredients has been broken into three parts :
First approach machine learning:
I have taken the data set and used it trainour tf-idf library and finally applied Supervised learning using SVC to classify.In this way, an accuracy of 80 % was obtained.

Second approach deep learning:
The inputs converted to digital by applying ift-idf are trained in the neural network model. The validation accuracy rate in this model is 71%. Val loss was not a good approach as it was 91%.

Third approach deep learning:
I tried the pre-trained bert approach as my results from the second approach were unsatisfactory. The accuracy rate I get from this model is 70%.

5. BIBLIOGRAPHY

1 - Yummly-Kaggle. 2015. Kaggle – What's Cooking? (2015). Retrieved November 28, 2015 from https://www.kaggle.com/c/whats-cooking

2- Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830,2011.

3 - Böhning, Dankmar. "Multinomial logistic regression algorithm." Annals of the Institute of Statistical Mathematics 44.1 (1992): 197-200.APA

4- Smola, A. J., Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and computing, 14, 199-222

5-H. Su, M.-K. Shan, T.-W. Lin, J. Chang, and C.-T. Li. Automatic recipe cuisine classification by ingredients. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, pages 565–570. ACM, 2014.

6- Han Su, Ting-Wei Lin, Cheng-Te Li, Man-Kwan Shan, and Janet Chang. 2014. Automatic recipe cuisine classification by ingredients. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct). ACM, New York, NY, USA, 565-570. DOI=http://dx.doi.org/10.1145/2638728.2641335

7- A Review on Word Embedding Techniques for Text Classification,Book cover Innovative Data Communication Technologies and Application pp 267–281Cite as

8- A Review on Word Embedding Techniques for Text Classification S. Selva Birunda,R. Kanniga Devi