UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA DE TECNOLOGÍAS Y SERVICIOS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Design and development of a diabetes and overweight characterization system using social media data mining

Luis Martín de Vidales Palomero
2019

TRABAJO DE FIN DE GRADO

Título:	Diseño y desarrollo de un sistema de caracterización de sobrepeso y diabetes utilizando minería de datos de redes sociales
Título (inglés):	Design and development of a diabetes and overweight characterization system using social media data mining
Autor:	Luis Martín de Vidales Palomero
Tutor:	Carlos Ángel Iglesias Fernández
Departamento:	Departamento de Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:	—
Vocal:	—
Secretario:	—
Suplente:	—

FECHA DE LECTURA:

CALIFICACIÓN:

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN

Departamento de Ingeniería de Sistemas Telemáticos
Grupo de Sistemas Inteligentes

TRABAJO DE FIN DE GRADO

Design and development of a diabetes and overweight characterization system using social media data mining

Junio 2019

Resumen

La nutrición y la comida son aspectos importantes en la vida de todos. La forma en la que comemos determina como nos sentimos, las enfermedades que podemos padecer en el futuro, e incluso habla del transfondo social del que procedemos, así como de nuestra cultura.

En este proyecto exploramos el uso de las redes sociales, y más en concreto de Twitter, como fuente de información nutricional. Los objetivos de este proyecto son el desarrollo y diseño de una herramienta capaz de llevar a cabo análisis nutricionales sobre la población, así como de un clasificador que sea capaz de distinguir entre comidas saludables y no saludables. Para completar estos objetivos el proyecto consta de las siguientes fases: captura, desarrollo del clasificador y construccion del sistema de análisis.

En la primera fase capturamos 19773 tweets durante 18 días en español y en catalán. Estos tweets contienen metadatos como la localización geográfica y los nombres de usuario y conformaron el dataset que empleamos en nuestra investigación.

En la segunda fase construimos un sistema para preprocesar los datos y convertirlos en una fuente de información para modelos predictivos y herramientas de visualización. Para ello extrajimos datos acerca de los nutrientes que contienen las comidas mencionadas en los tweets. Estos datos fueron empleados como recursos de aprendizaje para varios clasificadores distintos que entrenamos y evaluamos usando diversas métricas.

El clasificador tiene el objetivo de determinar si la comida mencionada es saludable o no. En esta tarea, el clasificador que ha mostrado un mejor desempeño de los que hemos desarrollado ha sido el que empleaba el algoritmo K Nearest Neighbors. Este clasificador alcanzó unos valores de precision de 0,93% de aciertos y una puntuación f1 de 0,93%

En la última fase desarrollamos un servicio de análisis de nutrición, que permite visualizar un análisis de la nutricion de una población en concreto, siendo capaz de filtrar los tweets según la comunidad Autónoma a la que pertenecen, el género del autor, la etiqueta de "saludable" resultado del clasificador y la hora de creación del tweet.

Palabras clave: Nutrición, Machine Learning, Big Data, Python, NLP, Sentimientos, Twitter, Análisis.

Abstract

Food and nutrition in general are important aspects of everybody's life. The way you eat determines how you feel, what illnesses you might suffer in the future and it even talks about your social background.

In this study, we use Twitter as a source of nutritional information, capturing tweets that talk about food. The goals of this project are designing and developing a tool capable of performing nutritional analysis over the population, as well as developing a classifier that distinguishes between healthy and unhealthy dishes. The project consists of the following stages: capture, classifier development and nutrition analysis system building.

In the first stage we collected tweets 19773 in Spanish and Catalan containing meta data such as geographic locations and user names, and conformed a corpus for our research.

In the second stage we built a system to preprocess these tweets transforming them into a source of information for predictive models and visualizations. With that purpose we extracted features as nutrients contained by the food mentioned in those tweets. This features were used as input for different classifiers which were evaluated using various metrics.

The classifier had to determine if the food mentioned in the tweet was healthy or not. The classifier that gave the best performance was the one implemented with the K Nearest Neighbors algorithm, reaching an accuracy and f1-score of 0,93 and 0,93 respectively.

In the final stage we developed a nutrition analysis service that allows you to visualize the analysis of the nutrition of a certain population being capable of filtering the tweets between Autonomous Communities, gender, the health label given by the classifier and the hour of creation.

Keywords: Nutrition, Machine Learning, Big Data, Python, NLP, Sentiments, Twitter, Analysis.

Agradecimientos

A mis padres por acompañarme a lo largo de este viaje y no desfallecer nunca.
A todos mis amigos, por contribuir a mi proceso formativo, y ayudarme a ser mejor persona.
A mi novia por no perder las ganas después de tanto tiempo, y animarme en las horas bajas.
A mi hermana por ser una inspiración constante y un ejemplo de esfuerzo y éxito.
A Carlos Ángel, mi tutor, por ser paciente, perfeccionista, y por darme la oportunidad de trabajar con él.
A mi abuelo, al cual le dedico este trabajo, intentaré ser yo también un buen ingeniero.
A mi hija, que me ha impulsado todos los días y ha sido mi motivación.

A todos vosotros muchas gracias, sin vosotros esto no sería posible.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Context

Fighting against obesity and overweight in all age ranges is one of the main problems that faces the population of most of the countries in the world [51]. To solve this problem, the best solution is adopting healthy habits such as exercising, not smoking, but above all, having a healthy and balanced nutrition [35].

When doing statistical research about nutrition there are different sources of information and methods for its capture such as telephone surveys, blood analysis etc [9, 48]. These sources and methods produce good results but they are either too generic and imprecise [7] or take too much time to be analyzed properly. For all this, it is necessary that the medical researchers and nutritionists are provided with new techniques for recovering and analyzing the information in a more effective way and with more precise data.

Nowadays the use of social media is a widespread practice, and it is a common habit for the users of social networks to post the food they are consuming, via pictures and text. In our case the main goal of this project is to developing new techniques for analyzing the nutrition of the population and to characterize if this nutrition is either healthy or unhealthy. Also it is interesting to analyze the depth of analysis that we can reach, out of the posts that

talk about food.

The social network we used as a source of information is Twitter. Twitter posts are limited to 280 characters, with the possibility of including images, videos and urls. Twitter in specific has shown to be a good source of information about health related topics such as insomnia [49] and even geographic distribution of healthy habits, diet, and exercise [33, 16].

To summarize, we believe that a phenomenon such as nutrition can be studied using the information obtained from a platform like Twitter, following the procedures explained throughout the rest of the document.

1.2 Project goals

The main objective in this project is the deployment of a classifier using machine learning technologies. This classifier has to be able to characterize the nutrition in posts of Twitter users as healthy or unhealthy. For this research we use tweets that are written either in Spanish or Catalan.

Another goal is investigating about nutrition and the reasons why food should be differentiated as healthy or unhealthy,in order to obtain the kind of nutrition that solves the problem of overweight and obesity. It is also an objective performing this analysis in the different Autonomous Communities of Spain separately, as well as in Spain altogether.

The previous goals culminate in the development of a tool for obtaining information and performing analysis about nutrition.

1.3 Project tasks

To achieve the goals that we have previously set, we have had to complete the following tasks:

- Study of the state of the art in relation to the machine learning technologies, as well as the use of social networks as a source of information for research in diverse areas of knowledge.

- Capture of a sufficient amount of tweets that talk about food for the creation of a dataset.

- Development of a classifier capable of classifying tweets as healthy or unhealthy following the next steps:
  - Preprocessing of the data and extraction of the characteristics that the classifier has to learn.
  - Division of the preprocessed data in a training set for the classifier and a test set to evaluate it.
  - Analysis of the results obtained after the tests performed
- Development of a nutrition analysis service, which includes a visualization interface to display the information and the analysis performed.

1.4 Structure of this document

The structure of the document is the following:

Chapter 1: Context in which the project is developed as well as a description of the goals and tasks to complete.

Chapter 2: State of the art and description of the technologies used in the project.

Chapter 3: Description of the process followed to deploy the classifier, as well as the metrics used for its evaluation and results obtained.

Chapter 4: Architecture of the nutrition analysis service, including the description of the different subsystems it has and the development and design of the visualization and analysis interface.

Chapter 5: Case of Study with a brief description of the results obtained with our tool.

Chapter 6: Conclusions reached in this project, problem faced and the suggestions of future work.

CHAPTER 2

Enabling Technologies

This chapter illustrates the state of the art of the technologies related to the theme of this project, as well as the technologies used throughout this project. Specifically after an introduction section on which we describe the state of the art, we will talk about the technologies used in successive sections.

2.1 State of the art

Following healthy habits is one of the main preoccupations nowadays. Healthy lifestyle is a concept that started to be important at the end of the 20th century, when the OMS defined it as a lifestyle that reduces the risk of premature death and suffering severe diseases [35]. Healthy habits should not be followed only for aesthetic reasons, they are associated with cardiovascular diseases and mortality [21]. Nowadays the non transmissible diseases are still the main cause of morbidity and mortality (70 percent), with cancer and cardiovascular diseases as the most significant. These diseases share the same risk factors, where a bad nutrition, the lack of exercise and sedentary lifestyle stand out between others such us stress, lack of sleeping, tobacco, alcohol and drug consumption [51]. Control of healthy weight, overweight in specific, as well as eating disorders are serious problems in actual society. Obesity is considered, after tobacco consumption, the second avoidable cause of illness and death. The OMS consider obesity has triplicated since 1975 [36]. The data presented is alarming. In 2016 more than 1,9 thousand million people over 18 years old

suffered overweight. Between them there were 650 million obese people. This means that 39 percent of the worlds adults where overweight in 2016 and 13 percent of them where obese. The magnitude of the problem is the reason why the OMS has cataloged overweight as a 'worldwide epidemic'.
Spain is in an intermediate position in that issue, as well as Madrid, where 47,2% of adults are overweight or obese, and 65% does not exercise [42]. However child obesity has one of the highest values in all Europe reaching the 19% [8]. Spanish population has stopped following their traditional nutrition pattern, Mediterranean diet, that between overweight teenagers and adults helps maintain metabolic health [47]. It has also been proven that at every age having a good physical condition contributes to metabolic health, even in overweight people [38]. That is why physical inactivity has become the 4th global risk factor of morbidity and mortality for every age and both genders [22], maintaining the lack of exercise levels stable since 2001 [11].
Data about obesity and sedentary lifestyle do not correspond with scientific evidence available or the effort made in the administration and societies to transmit the information to the population. It doesn't correspond either with the preoccupation of the citizens for following good nutrition and exercise habits. This could be due to the influence of the changes our society has had around habits with a verified effect on peoples health. Also consumer society has been a bad influence on the acquisition of healthy habits [6]. The factors that affect health the most are [6]:

- genetic risk factors
- biologic risk factors such as high blood pressure, dyslipidemia, mellitus diabetes, etc.
- behavioral risk factors such as diet, physical exercise, tobacco consumption, etc.

Behavioural factors have a decisive influence on healthy habits. Even though healthy habits are well-known (healthy diet and exercise), the reasons for choosing unhealthy habits depend on education [14], social influence [29, 12], and difficulties turning intentions into habits [34].
In addition to all these factors there is one that has not been deeply studied by nutrition and physical activity science, social networks. Social Networks have commonly been associated to teenager population, however in the last decades this habit has been adopted by more segments of the population. Even though this growth is exponential, the influence social networks may have over people is almost unknown.

Traditionally, nutrition science has used both qualitative (focal groups) and quantitative (surveys and interventions in control groups) methods to get to know the healthy habits

state of the subjects as well as the effectiveness of the interventions [9] using mixed methods that combine both approaches [48]. However, quantitative methods have shown precision problems, as the interviewed tend to be reluctant to report unhealthy food consumption [7]. Also they tend to overvalue their physical activity habits when they are interviewed. This is the main reason why we believe that the massive use of social networks provides a new source of primary data. A lot of the factors associated to lifestyle (number of daily meals, foods and drinks type, size of rations, eating at home or outside, eating in front of a screen, alone or accompanied, buying habits, going to the gym, for a walk, etc) are transmitted via social networks. Nevertheless, this social network data with big data characteristics hasn't been used in health science as in nutrition, even though there have been related studies suggesting how useful they would be in the implantation of healthy habits [4]. The usage of hybrid methods that combine quantitative data, qualitative data from surveys, and massive data comes as a new opportunity for obtaining quantitative data on a large scale [16]. The possibilities that give us using the tools provided by machine-learning for predicting the answer to interventions, as well as for proposing solutions that suit a subjects characteristics and behaviour are really interesting, and that's the reason why they should be implemented to join in the effort made for improving societies health [54].

As far as we know, the approach we are proposing for studying nutritional situation as well as intervention effectiveness has not been used. In the Intelligent Systems field there are various studies that have analyzed the potential of social networks for sleeping disorder detection [49], feeding [25], geographic distribution of healthy and unhealthy food [55], geographic distribution of healthy habits , diet, and exercise [33, 16].

There are some studies that confirm the usage of social networks as a useful information source about nutrition and health state [31, 32]. Twitter in specific has been used as an information source for a study about obesity in population related to characteristics of neighbourhoods [32], a study about the most consumed meals based on geographic situation of consumers [1], as well as more complex studies such as the analysis of negative or positive attitude observed in tweets related to obesity, diet, diabetes and exercise [46, 50].

It is also relevant getting to know and understanding the information sources the population uses for following certain habits. One of the actual problems is the lack of information related to nutritional practices [53] that lead to following diets or using remedies with high associated risk. Social network analysis will help to analyze the sources used [31] or detecting new trends or fake news [41].

Big Data technologies have changed the processing of the big volumes of information produced nowadays by the application of artificial intelligence techniques, as machine learning, semantic processing and natural language processing. These techniques are often applied to social networks due to the high volume of information generated by its users.

Various studies have successfully explored the use of social networks as a useful source of information referring to nutrition and health state [31, 32]. There are others that have analyzed the kind of content published by the users about their diet or their physical activity in social networks such as twitter [20, 50], combining qualitative analysis and surveys, that have determined that this publication helps them to maintain their healthy habits, thanks to the support received by their followers. This conclusion has an implication with the trans theoretical model of change of Prochaska and Diclemente [39], because the users focus more on maintaining the motivation rather than achieving the change objectives.

All this related work suggests Twitter, as well as other social networks, combined with Big Data technologies, as very promising resources for analyzing healthy habits.

2.2 Technologies used

The main objectives of this project are developing a tool capable of characterizing and classifying different messages from Twitter users and provide a visualization tool that provides a simpler way to analyze the data obtained. Therefore the technologies used in the project correspond to the implementation of artificial intelligence through machine learning technologies and the implementation of a visualization interface with the technologies involved in its deployment.

Machine Learning is the field with the objective of developing techniques to show computers how to learn. There are different approaches: supervised, unsupervised and semi supervised [37].

Our problem is a classification one, which means our learning process will be supervised. This approach starts by training the algorithm with a series of classified documents, and the algorithm produces a function that establishes a correspondence between the characteristics of each document and the desired classification. The output of this function determines the classification of a new document that the system is trying to classify.

The machine learning algorithms we use work through a series of accounting character-

istics that define each of the samples, which in our case are texts. This extraction of characteristics is known as feature extraction and will be performed with different analysis techniques.

2.2.1 Python libraries

The main functionality of the project has been programmed in Python. The following libraries written in this language, have provided us the tools to implement it:

- **Matplotlib:** Matplotlib [13] is a Python library used for plotting 2D which allows us to produce different representations such as histograms, heat maps, etc. from a source of data using a Python script. Every graphical representation shown in this project has been generated using directly or indirectly this library.

- **Numpy:** Numpy [52] is a Python library for scientific computing. Other libraries used in this project such as Scikit-learn need this library to execute their functions.

- **Pandas:** Pandas [30] is the Python library used in this project in the data analysis stage. This library implements data structures and data analysis tools with high performance. It also has the capability of merging, grouping and querying data structures.

  These data structures are **Series** and **DataFrames**. Series are one dimensional objects in which every element has its own index. DataFrames are two-dimensional labelled objects that have columns with potential different types.

  Pandas functions have been used for managing the data compiled from Twitter messages. With its functions the data has been exported to formats such as CSV, filtered based on the criteria imposed and reduced in order to save as much storage as possible.

- **CSV and JSON:** CSV and JSON libraries will be used for our datasets storage and interchange between systems. Pandas can interpret both formats, and write files in both of them.

- **GeoPandas:** GeoPandas [15] is another Python library related to Pandas that allows the user to represent geographic dataframes with coordinates, countries and all sorts of geographic data using matplotlib.

- **RegEx(re):** RegEx in Python is performed through a library named re. This library provides functions to obtain regular expressions included in the range of characters given as inputs.

- **Requests:** Requests is the Python library used for executing http requests to the external APIs used in the project such us the genderize API described later.
- **Scikit-learn:** Scikit-learn [37] is an open source Python library with machine learning algorithms already implemented. It uses other libraries such as Numpy, SciPy(Scientific Python) and Matplotlib. This library includes function to perform tasks such as:
  - **Classification:** deciding which category a certain object belongs to.
  - **Clustering:** automatic gathering of similar objects into sets.
  - **Dimensionality reduction:** reduction of the amount of random variables to consider.
  - **Model selection:** comparing, validating and choosing parameters and models.
  - **preprocessing:** feature extractions and normalization.
- **NLTK:** NLTK [23] is a set of Python libraries used for statistical natural language processing. This library allows us to extract measurable characteristics from different human-written texts. In our project this library has been used to perform the following tasks:
  - **Tokenization:** Tokenization is the task of breaking a certain text into parts, called tokens. For this purpose certain characters or words in the document that contain no information are discarded. Commonly punctuation and stop words (such as articles, prepositions, conjunctions, pronouns, etc.) that are words that have no meaning in themselves.
  - **Part Of Speech tagging:** this task consists of tagging each word of a text to its grammatical category.

2.2.2 Twitter API

The Twitter API [26] has been used in the data-extraction stage. This API provides the tools needed for filtering real-time tweets, read the profile of users and their followers(along with other data), identifying twitter applications and users who register using authentication and OAuth authorization.

The streaming API has been used for collecting a high volume of tweets related to food consumption with low latency.

In order to use the API inside of a Python project, the Python library **Tweepy** [40] has

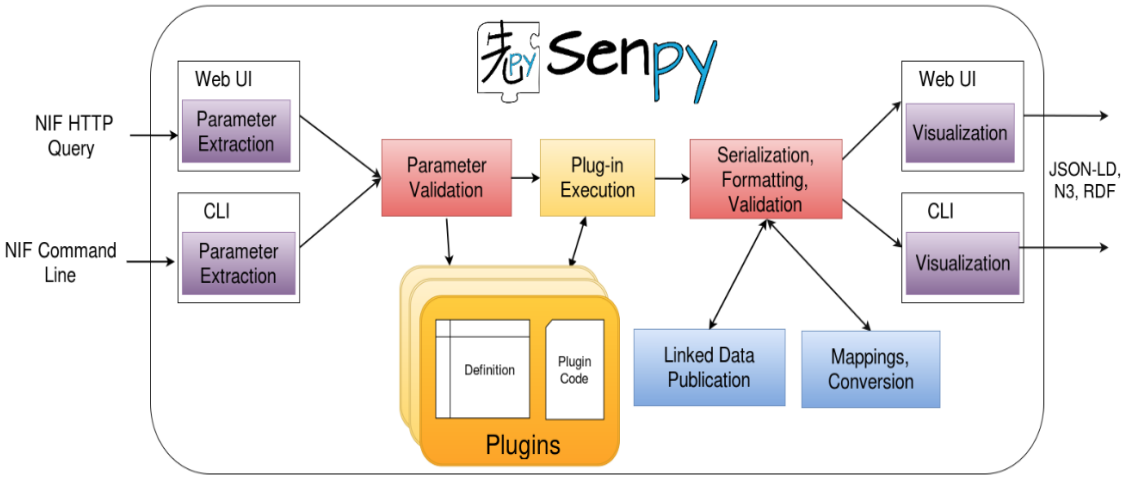

Figure 2.1: Senpy's Architecture

been used as a wrapper for communicating these two technologies. This library has performed the tasks required to authenticate, connect, disconnect with the Twitter API, and many others, during the usage of the streaming API.

2.2.3 Genderize API

The Genderize API [17] has been used in the data-analysis stage. This API provides the tools needed for extracting the gender of an individual given him/her name.

2.2.4 Senpy

Senpy [44] is a framework developed by the GSI at ETSIT-UPM that allows the user to develop sentiment and emotion analysis services. All the services developed in Senpy share the same API which allows the use of all of them in a simple and interchangeable way.

The architecture of Senpy consists mainly of:

- **Senpy Core:** where the service is built and where all the tasks of a web analysis service (data validation, user interaction, formatting, logging, etc.) are performed.
- **Senpy Plug-in:** where the classifiers and the analyzers of a determined service are deployed.

For our project we have used the already implemented plug-in sentiment140. This plug-in analyzes the text of every tweet captured giving as a result the sentiment with which every tweet was written.

2.2.5 ElasticSearch

Elasticsearch[1] is a real-time, open source, distributed full-text search and analysis engine. It is used by many large organizations around the world [10] and developed in Java. It can be accessed from a Restful web service interface and uses JSON documents to store and upload data.

It offers features such as:

- Distributed and Highly Available Search Engine.
- Various APIs (Restful HTTP API, Native Java API, between others).
- Reliable, Asynchronous Write Behind for long term persistence.
- Single document level operations are atomic, consistent, isolated and durable.

Every Elasticsearch instance in a virtual or physical machine is known as a **node**. There can be as many nodes supported by a single machine as the machines resources allow. In a node there can be one or multiple **indexes**. These indexes are a collection of **documents**, being those documents a collection of fields that compose an object. The documents are defined in JSON format with a unique identifier, the UID. The indexes are subdivided in **shards** and each shard contains the same properties as the index it comes from. Both indexes and shards can be replicated in **replicas** for improved availability and performance.

2.2.6 Sefarad

Sefarad [45] is an environment developed by GSI at ETSIT-UPM that extracts the data stored in Elasticsearch and provides tools to implement a visualization interface.

These are the main modules Sefarad consist of:

- **Visualization module:** Is the module that contains the main functionality of this environment. It is composed of Polymer Web Components that conform a dashboard. These web components perform the task of drawing the graphics with the information extracted from Elasticsearch.
- **ElasticSearch:** Every dashboard has an ElasticSearch node associated to it. This node contains the data represented by the dashboard in one or multiple indexes and represents the persistence layer.

[1] https://github.com/elastic/elasticsearch

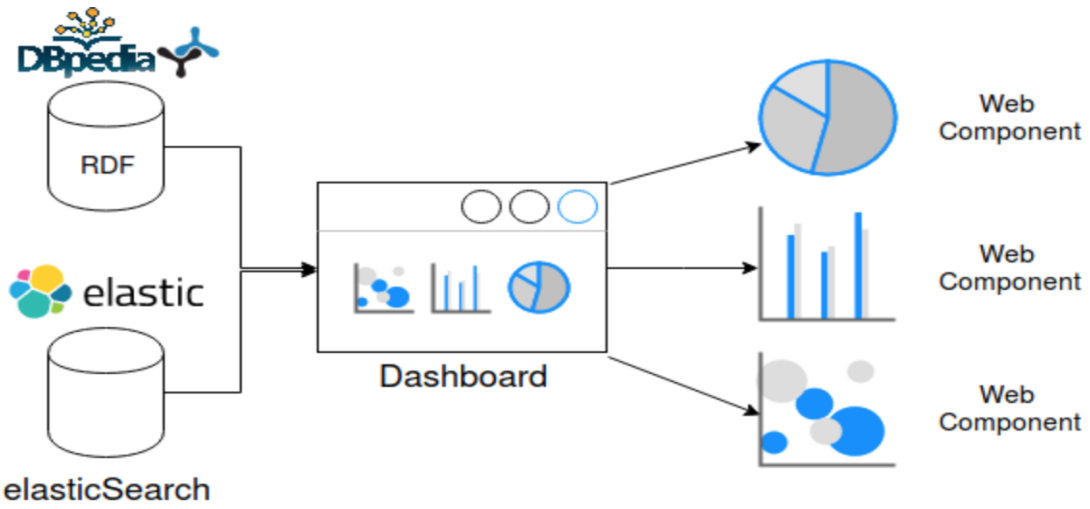

Figure 2.2: Sefarad's Architecture

Sefarad is also capable to recover data from other external sources, such as Fuseki or DBPedia.

CHAPTER 3

Machine learning model building and evaluation

3.1 Introduction

In this chapter we will review the process followed for the development of the classifier used in this project.

There will be a description of the formation of the dataset. This includes the capture process, the analysis of the information contained in the tweets and the labelling process. Later we will explain the feature extraction performed. These features extracted will be used as input for three different machine learning algorithms, which will be evaluated using various evaluation metrics. Then, based on the results of the evaluation, one of the algorithms will be chosen and discussed.

3.2 Dataset

The dataset created for the project contains 8 million tweets captured from April 05, 2019 to April 23, 2019. The capture was performed using the Streaming Twitter API and the Tweepy library for Python. We used these technologies in order to capture tweets in real

time. The tweets captured were related to nutrition, and they had to meet the following characteristics:

- They must have been written either in Spanish or Catalan.
- They could not be re-tweets (a re.post of a tweet originally posted by a different user) because our interest focuses on the habits that the user follows, and supporting someone else habit don't necessarily mean that the user follows it.
- The tweets must contain words related with food, in order to perform a nutrition investigation having the Spanish population as a research subject. The words used for the captures were food terms such as pizza, salad, hamburger and also brand names as Coca Cola, Burger King, McDonald's etc.

3.2.1 Location

The first step of the analysis consists in obtaining the geographical origin of the tweets from the dataset. The countries in which a higher amount of tweets was captured are shown in Table 3.2.1. To determine the location of the tweets there were two main approaches:

Country	Tweets	Percentage
Argentina	35327	29,44%
Chile	6954	5,79%
Spain	19773	16,48%
Uruguay	4008	3,34%
Colombia	9162	7,63%

Table 3.1: Amount of tweets captured in Spanish and Catalan and percentage of the corpus per country.

Geolocating the tweets using the location field included in the data retrieved from the user that wrote the tweet. To do so, we would apply the Google Maps Geocoding API [1] which implements functions that allow to obtain the coordinates (40.416775,-3.703790) of a certain address (Madrid).

[1] https://developers.google.com/maps/documentation/geocoding/intro

Using the place field included in the metadata of the tweets that are already geotagged to a place to conform our dataset.

We have followed the second approach that consists on filtering the tweets between the ones that have the place field filled in and those who don't. As a result of applying the filter we got 119958 tweets that conform our dataset. The geolocated tweets represent 1.4% of the 8 million tweets captured.

A visual representation of the analysis carried out is provided in 3.2.1. Using the GeoPandas library we produced a heat map that represents the countries with more tweets written in them.

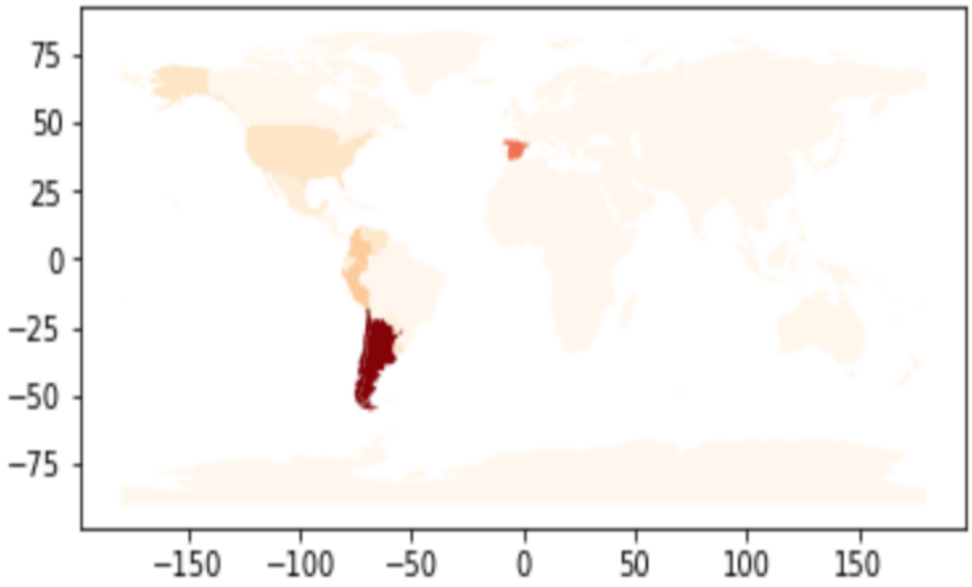

Figure 3.1: Map representing the countries in relation to the number of tweets with food words.

3.2.2 Semantic Mistakes

There are some terms in the capture filter that have different meanings depending on the context. This produces noise in our dataset, as there are tweets that are not actually speaking about food. Examples of different usages given to food terms can be seen in Table 3.2.2. To eliminate this noise as much as possible, we analyzed the list of food names we used in the capture, and checked all the different usages of each term. We came to the conclusion that, a food term was used with the appropriate meaning only when it was used as a common noun.

Text	Does it talk about food?
I love having wine (me encanta beber **vino**)	yes
he came to my party (él **vino** a mi fiesta)	no

Table 3.2: Examples of different usages of food terms.

Therefore, to know which tweets are talking about food and which aren't, we must perform a syntactical analysis over the text of the tweets.

The technologies used for this task was the nltk Python package along with the Stanford POS Tagger explained in Sect 2.2.1. With this pair of technologies we tokenized the texts from the tweets and obtained the syntactical categories of each word. We then kept only the tweets that were talking about food for a deeper analysis. The dataset included 9930 tweets that were actually talking about food. This represents a 50,21% of the tweets from Spain.

3.2.3 Feature extraction

Feature extraction consists of transforming text into numerical vectors that provide a simpler view of it to the automatic learning model. As the categorization of the tweet as healthy or unhealthy depends more on the characteristics of the food mentioned than other features that could be extracted from the text such as LDA topics, Ngrams between others.

The features that are relevant for our research are described below.

3.2.3.1 Nutrient features

To characterize if a tweet mentions healthy or unhealthy food it is important to know the different nutrients that the food mentioned in it has.

To extract those values we looked for each food term on the Spanish Food Composition Database [2] and obtained the amount of calories, carbohydrates, fats, proteins, cholesterol and fiber that every food in our list had.

[2] http://bedca.net/bdpub/

After that, we associated a value of each of these nutrients to every tweet based on the different food terms that appeared in it.

3.2.3.2 Gender

The goal of extracting the gender of the authors is to obtain the information necessary to perform statistical analysis based on demographic variables in order to find correlations between the gender and nutrition.

We have applied the methods implemented in the genderize API explained in Sect 2.2.3 to obtain the gender of the author based on its twitter user name.

With this approach we were able to extract the gender of 6903 tweet authors, which makes a 69,5% of the tweets that talk about food, being 4010 (58,1%) male and 2893 (41,9%) female. The authors gender we couldn't find out using the genderize API where labelled as "unspecified".

3.2.3.3 Sentiment

As well as with the gender, we have extracted the sentiment with which the tweets where written to study the relation between the sentiments and nutrition, performing statistical analysis over the tweets.

To determine the sentiments associated to every tweet we have used Senpy, a technology described in Sect 2.2.4 with which we have obtained the amount of tweets that showed a positive, negative and neutral sentiment.

With this approach we were able to extract a sentiment different than neutral in 2218 tweets being a 22,3% of the tweets that talk about food. This shows that when talking about food people tend to be neutral rather than positive or negative. Between these 2218 tweets there were 1995 (89,9%) positive tweets and 223 (10,1%) negative tweets. This denies the hypothesis that there is a relation between sentiment and nutrition, but shows that if somebody talks about food with sentiment associated with the text there is a higher probability that this sentiment will be positive.

3.2.4 Labelling

Our goal is to characterize the dietary choices in tweets as healthy or unhealthy from a nutritional perspective and to do so, we have used a supervised classifier that has had to be

trained. In this training process a set of already classified documents has to be necessarily used as input and to obtain it we have labelled the tweets as healthy or unhealthy from a nutrition perspective.

The approach followed on this classification considers the amount of each nutrient that every tweet has. Based on the amount of nutrients included in the traditional Mediterranean diet. This well-known traditional diet has proven to be balanced in what comes to nutrition [28] and effective against health problems such as obesity [5].

Along with that, we have taken into account the medical recommendations for a diet to fight against overweight [24] to establish thresholds on every nutrient above or below which the tweet should be labelled as unhealthy.

Based on those thresholds we have labelled the documents evaluating which tweets have a healthy nutrition by evaluating which thresholds had been surpassed.

3.2.5 Dataset Summary

In this section we revise the results of the processes described above, to give an overview of the dataset that is used as input for the classifier.

From the 8 million tweets that where captured, only a 1,4% were geolocalized. These tweets add up 119958 from which 19773 (16,48%) are localized in Spain. From these 19773 50,21% talk about food, and these 9930 tweets conform the dataset that is used as input for the classifier. In the following tables and figures we describe the content of this dataset.

Language	Number and percentage	Number of males	Number of females	Number of positives	Number of negatives
Spanish	9285 (93,5%)	3705 (37,3%)	2739 (27,5%)	1848 (18,6%)	222 (2,2%)
Catalan	645 (6,5%)	305 (3,1%)	154 (1,5%)	147 (1,4%)	1 (0,01%)

Table 3.3: Evaluation of the Nutrition Classifier with the first set of features.

After this analysis, we would like to do a nutritional summary of the dataset. In the following figure, we can see the amount of food terms that fall into the different food categories that we have taken into acount.

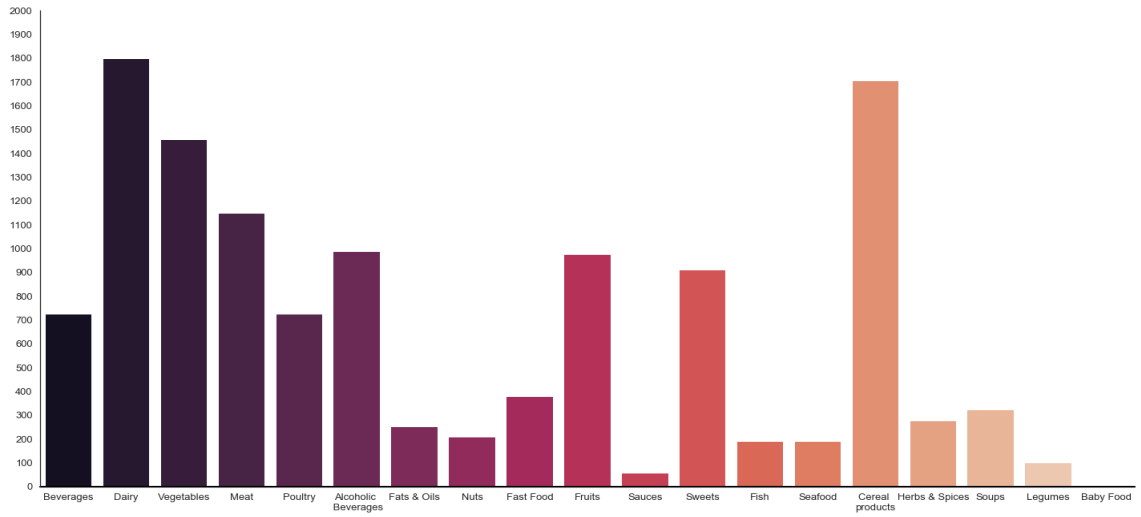

Figure 3.2: Amount of food terms mentioned of each food category.

3.3 Classifier

A classifier has been developed for the automation of the labelling task. This classifier tags tweets that meet the characteristics of the healthy or unhealthy tag.

This classifier executes the task of classifying texts. This task always has a similar workflow which consists of three different steps, shown in the figure below and described in the following sections.

Figure 3.3: Different steps in machine learning process.

3.3.1 Preprocessing

In this phase the raw text from the tweets is processed in order to make it easier to analyze. This includes eliminating punctuation, stop words, rare characters, emoticons and urls, as well as usernames of users mentioned in the tweets. This process is necessary because none

of these different items are useful for the classification task, and they hinder the feature extraction task.

The concept stop words must be explained to understand the reasons of their elimination. Stop words are words that provide no information about the content of the document, they are meaningless and are used only for syntactical purposes. This kind of words includes articles, pronouns, prepositions, connectors etc. The list of these words, along with the punctuation marks is provided by the NLTK library which has been explained in Sect 2.2.1.

We have performed this task before determining if a tweet was talking about food or not as described in Sect 3.2.2. To do so we have used the TweetTokenizer module of the NLTK library, That allows us to convert the document to a list of different tokens while taking into account the exclusive characteristics of a tweet such as the mention of a user or the use of hashtags.

3.3.2 Feature Extraction

As explained before in Sect 3.2.3 feature extraction consists of analyzing the preprocessed text and extract features that provide a simpler view of the text, easing the understanding of these texts to the machine.

There are multiple kinds of feature extraction techniques that can be used to extract these features from texts. Most of them focus on extracting the importance of words in a document and all the documents that conform the dataset. Other focus on extracting topics from the texts, considering every text as a mixture of different topics, belonging each word to one of those topics.

Nevertheless, in our case of study it doesn't make sense to extract the importance of the different words in a text or topics associated as we are categorizing tweets in terms of nutrition, and none of these features are actually useful to determining if a tweet is healthy or not.

We have focused instead on the extraction of the nutrients that are associated to the food terms mentioned in every tweet, being this the core of our feature extraction task. These features are numeric values and categorical variables explained in Sections 3.2.3.1, 3.2.3.2 and 3.2.3.3 that will be used as input for the machine learning algorithm.

3.3.3 Classification

This last step of the text classification task comprehends the training of a classifier used the features extracted in the previous step. To implement this classifier there are multiple automatic learning models that can be used to train a final model. The classifier that we aim to develop has to be able to classify unknown texts through supervised learning.

The Scikit-learn library described in Sect 2.2.1. provides different types of classification algorithms that are already implemented, with functions that perform the training and prediction tasks.

Each of the algorithms have different parameters that can be tuned to obtain a higher success rate on the prediction task. These parameters (called hyper-parameters) can't be tuned randomly, as they determine the success in our classification.

To choose the appropriate hyper-parameters we have used the GridSearchCV module, that allows you to know the hyper-parameters that give the best results to a classifier when classifying a dataset with certain features. Essentially, this module is used to optimize the algorithm in order to obtain the maximum success rate.

For measuring the level of success of the feature extraction and the classification tasks we have used the cross validation technique(K fold). This technique divides the data into k different sets, and goes through the sets k times, using each set as the training data, and the k-1 remaining sets as test data. This gives us k different test cases that are averaged to obtain an estimation of the results of the level of success in our model.

This technique allows us to check the results of the classification algorithms we are going to use whilst exploiting our dataset to the fullest and without incurring an over fitting problem.

In this project the following classification algorithms that are provided by the scikit-learn library have been tested:

- **Multinomial Naive-Bayes**: This kind of classifier implements the naive Bayes algorithm for multinomial models. This algorithm uses a multinomial distribution that normally requires integer feature counts. However, in practice, fractional counts may also work. Scikit-learn has implemented a classifier of this type called MultinomialNB which is the one we have used.

- **Linear Support Vector Classification (LinearSVC)**: This classifier belongs to the supervised learning Support Vector Machine (SVM) set for classification detection. It is an implementation of the SVC classifier with a linear kernel but implemented in terms of liblinear rather than libsvm to scale better to large numbers of samples. Again, it is provided by scikit-learn under the name LinearSVC.

- **K-Nearest Neighbours(KNN)**:A k-nearest-neighbour algorithm, often abbreviated k-nn, is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in. The k-nearest-neighbor is an example of a "lazy learner" algorithm, meaning that it does not build a model using the training set until a query of the data set is performed. We will use the classifier implemented in the scikit-learn library called KNeighboursClassifier.

3.3.3.1 Evaluation metrics

Once we have already implemented the workflow of the classifiers we have to check their performance in the classification task, measuring the success rates in order to deploy them or continue looking for a better combination of feature extraction techniques or other classifiers.

There are standard metrics that provide objective information on the amount and kind of mistakes that the classifiers make when deciding the category in which a tweet belongs.

These metrics use various concepts that have to be defined prior to explaining them. The concepts are the following:

- **True Positives (TP)**: the classifier obtains a true positive when a tweet is correctly predicted with a positive label (positives that are predicted as positives).

- **True Negatives (TN)**: the classifier obtains a true negative when a tweet is correctly predicted with a negative label (negatives that are predicted as negatives).

- **False Positives (FP)**: the classifier obtains a false positive when a tweet is incorrectly predicted as positive (negatives that are predicted as positives).

- **False Negatives (FN)**: the classifier obtains a false negative when a tweet is incorrectly predicted as negative (positives that are predicted as negatives).

These different categories allow us to calculate the metrics used to evaluate the performance of the classifiers. With the amount of tweets that fall into each of those categories we can

then calculate these metrics:

- **Accuracy**: this metric relates the amount of correct predictions to the total number of prediction, obtaining the percentage of predictions that have been done correctly. This is calculated using the following expression.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

- **Precision (P)**: this metric relates the amount of correct positive predictions to the total of positive predictions, obtaining the percentage of positive predictions that have been done correctly. This is calculated using the following expression.

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

- **Recall (R)**: this metric relates the amount of correct positive predictions to the total positive expected values, obtaining the percentage of positive values that have been predicted out of the expected. This is calculated using the following expression.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

- **F1 score**: It is the harmonic mean between precision and recall. It reaches its best value when the result equals 1 and the worst when it equals 0. To obtain this metric we use the following expression.

$$F1score = 2 * \frac{R * P}{R + P} \tag{3.4}$$

With these metrics we will evaluate the performance of the classifier while detecting healthy nutrition tweets.

3.3.4 Nutrition Classifier

This classifier has the task of labelling the tweets with the health label defined in Sect 3.2.4.

The first thing to do in the process was obtaining the sample of classified tweets for its use in the training process. This sample was obtained by a handmade labelling process, which gave a sample of 500 labelled tweets.

250 were labelled as “healthy” and the other half was labelled as “unhealthy”. We selected the same amount of “healthy” and “unhealthy” tweets in order to avoid a balancing problem.

Once we had the sample needed to train the classifier, we proceeded to perform the tasks of the classification workflow.

Preprocessing was done using the Tweet Tokenizer of NLTK and removing punctuation, stop words etc as explained in Sect 3.3.1.

In the Feature Extraction phase we performed the extraction of the nutrient values, the gender and the sentiments associated to each tweet as explained in Sections 3.2.3.1, 3.2.3.2 and 3.2.3.3.

With these tasks done, we continue with the classification process by training and testing the three different classifiers we will be evaluating. For that purpose with the GridSearchCV module we tuned the following hyper-parameters on each of the classifiers.

- **Multinomial Naive Bayes Parameters:** the parameter we tuned for the MultinomialNB classifier is the alpha parameter. This is an additive Laplace/Lidstone smoothing parameter that ranges between 0 and 1.
- **Linear Support Vector Classification Parameters:** the parameters we tuned for the LinearSVC classifier are the penalty parameter of the error term (C), the tolerance for stopping criteria (tol) and the norm used in penalization (penalty).
- **K Nearest Neighbours Parameters:** the parameters we tuned for the KNeighboursClassifier are the number of neighbours(n_neighbours), the power parameter for the Minkowski metric (p) and the algorithm used to compute the nearest neighbours (algorithm).

To evaluate the performance of the classifier we have used the metrics explained earlier along with the K Fold cross validation with K=5.

It is important to note that we have used two different sets of features as input for the classifiers. First we have used as features the nutrient values of each tweet, and then the nutrient values and also the gender of the author.

Even though the gender has been extracted for statistical purposes to perform a nutritional analysis over the dataset, we have chosen to do things this way to see if using gender as a feature helps the classifier achieve better results than using only the other set of features.

As we can see in Table 3.3.4 the KNeighboursClassifier performs outstandingly better than Linear SVC and MultinomialNB using as features the nutrient values assigned to the tweet. It gives much higher values in all the metrics defined before given the optimal hyper-parameters that were extracted using the GridSearchCV module.

Classifier	Hyper-params	Accuracy	Precision	Recall	F1 score
MultinomialNB	alpha = 0,0001	0.65 (+/- 0.04)	0,68	0,65	0,64
LinearSVC	C = 3 tol = 0,8	0.66 (+/- 0.16)	0,7	0,66	0,64
KNeighboursClassifier	n_neighbours = 1 p = 2 algorithm = 'ball_tree'	0.93 (+/- 0.05)	0,92	0,92	0,92

Table 3.4: Evaluation of the Nutrition Classifier with the first set of features.

When we use the second set of features, as we can see in Table 3.3.4 the evaluation metrics show a slightly better performance in MultinomialNB and KNeighboursClassifier and a higher accuracy in LinearSVC but lower recall and F1 score. This means that, although in Linear SVC the gender has improved the accuracy, the number of tweets that should have been predicted as positive but have been predicted as negative has increased. This also harms the F1 score as it depends on the recall as much as on the precision metric.

Classifier	Hyper-params	Accuracy	Precision	Recall	F1 score
MultinomialNB	alpha = 0,0001	0.66 (+/- 0.04)	0,68	0,65	0,65
LinearSVC	C = 9 tol = 0,01	0.70 (+/- 0.05)	0,7	0,65	0,62
KNeighboursClassifier	n_neighbours = 1 p = 2 algorithm = 'ball_tree'	0.93 (+/- 0.05)	0,93	0,93	0,93

Table 3.5: Evaluation of the Nutrition Classifier with the second set of features.

As we have said before, with these results in hand we can conclude that the gender of the author is not a useful feature for our classification task. We also determined the

classifier that gives the best performance. This classifier is the KNeighboursClassifier and it is the one we have chosen to deploy.

CHAPTER 4

Nutrition Analysis Service

4.1 Introduction

In this chapter, we cover the development of a visualization service that we have implemented as a nutrition analysis tool to analyze the data obtained from Twitter posts related to food. The technologies used for its development have already been described in Chapter 2

First we explain the architecture in which this project is based, presenting the different systems that conform it and the relationship between those systems. Second we describe every system of the architecture with more detail, as well as the function it performs within it.

4.2 Architecture

In this section we present the global architecture of the project, defining the different subsystems that include this architecture. We can divide the architecture into the following subsystems:

- **Capture system:** This is the subsystem that performs the capture of the Twitter posts that include food names.
- **Analysis system:** This subsystem is the one responsible for the analysis of the

posts. For this task different technologies described in Chapter 2 are used to classify the tweets between the ones that actually talk about food and the ones that doesn't, as well as to obtain the location, nutritional information, gender, and sentiment.

- **Persistence system:** This is the subsystem that stores and searches the tweets. It has been implemented using the ElasticSearch search engine described in section 2.2.5.
- **Visualization system:** This system is a server to visualize the analyzed data. We have used Sefarad environment, defined in Section 2.2.6.

This architecture defines a workflow that has been performed manually but could be performed automatically if an orchestrator was developed to monitor and manage the results of every subsystem. The capture and analysis systems have already been explained in Chapter 3. The remaining subsystems are explained in detail in the following sections. An overview of the whole architecture is given in Figure C.1.

4.3 Capture system

As mentioned before, this is the subsystem that captures the tweets that contain any food name. It is developed in Python and uses the Twitter Streaming API as explained in Sect 3.2.

This search gives as a result a JSON file containing the tweets that have been captured every hour. This JSON files are the input files of the Analysis system.

4.4 Analysis system

This is the system in charge of the analysis of the tweets. To do so, we have used various technologies described in Chapter 2. The whole process is described in Chapter 3 and gives as a result a JSON file where each entry is a tweet of the dataset captured.

4.5 Persistence System

This system store the information that will be used later in the visualization system, and has as an input the results from the Analysis system.

As mentioned before, we will be using ElasticSearch for this task, creating an index to store and access our data.

To store the tweets in ElasticSearch, we developed a Python script that uses the ElasticSearch Python module to index the tweets.

The data has been introduced to ElasticSearch in JSON format. We have used only one index named "tweetsnutweetion" with 5 shards for its storage. The JSON file has the following structure:

Field	Content
Autonomous Community	Autonomous Community in which the tweet is located.
Country	Country in which the tweet is located.
created_at	Hour of the day in which the tweet was published.
user	user name of the author.
text	Text content in UTF-8 format.
lat	latitude of the place in which the tweet was written.
long	longitude of the place in which the tweet was written.
Calories	Amount of Calories contained by the foods in the tweet in Kcal.
Carbohydrates	Amount of Carbohydrates contained by the foods in the tweet in g.
Cholesterol	Amount of Cholesterol contained by the foods in the tweet in mg.
Fats	Amount of Fats contained by the foods in the tweet in g.
Fiber	Amount of Fiber contained by the foods in the tweet in g.
Proteins	Amount of Proteins contained by the foods in the tweet in g.
Gender	Gender of the author.
Sentiment	This field contains the result of the sentiment140 plug-in.
Health	This field contains the result of the Health classifier.

Table 4.1: Fields of each document indexed in ElasticSearch.

4.6 Visualization system

The last system in our workflow is a visualizing tool for the visualization of the data stored in Elasticsearch. It has been implemented using the Sefarad environment, a tool developed by the Intelligent System Groups.

Specifically, we have designed a dashboard composed of a series of widgets described in the following sections.

Our dashboard performs API REST requests to our Elasticsearch index in order to obtain the desired documents based on some filters imposed in the query. These requests are called aggregations in Elasticsearch and allow you to group and extract statistics from documents based on the value of one of the fields, that is used as a filter. Another advantage of this type of request is that it allows you to concatenate several aggregations and obtain the result of both in a single response.

The responses from Elasticsearch will be used by the widgets to represent the data in the most representative way. These widgets have been developed using the JavaScript Polymer library. With Elascticsearch's aggregations, widgets allow users to filter and view only tweets that meet multiple characteristics at the same time.

As an example, the dashboard is going to be able to display only the tweet statistics from female users, from the Autonomous Community of Madrid, that express positive sentiment.

This way of displaying the data makes our visualization system a powerful tool for the Analysis of nutrition in a specific location, time, gender, etc. The widgets that make up our visualization dashboard are:

4.6.1 Widgets

4.6.1.1 Autonomous Communities

This widget has been done with the google-chart component implemented in Sefarad, which takes the ElasticSearch data to elaborate a chart that displays an aggregation of tweets based on the autonomous community they belong to.

The chart is a column diagram and takes as input parameters the following fields:

- **data:** JSON with all the documents received from ElasticSearch.
- **field:** field of each document that will be displayed in the chart. For this certain widget the value of this field will be "AutonomousCommunity".
- **type:** Google Chart type. The value of this field in this widget will be "column".
- **filters:** list of aggregation filters for ElasticSearch queries.
- **options:** field that includes the title of the chart.
- **optionsbi:** field that includes the settings of the colours to be displayed, as well as other customization settings.
- **cols:** array with the labels of each axis.

This widget performs a filter aggregation to the ElasticSearch index when the user clicks on one of the Autonomous Communities, allowing the user to visualize only the tweets of a certain Autonomous Community.

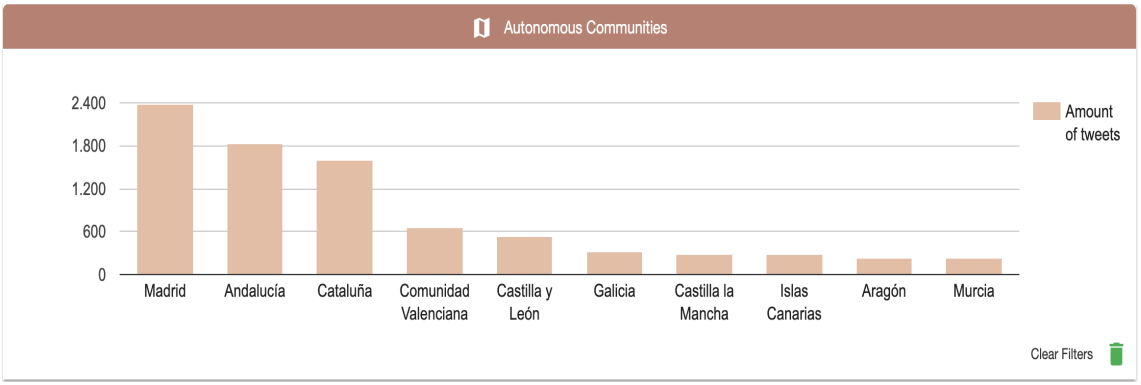

Figure 4.1: Autonomous Communities widget

4.6.1.2 Health widget

This widget has been also made with the google charts component but with "pie" type, and it displays the percentage of tweets that have been labelled with a healthy nutrition or unhealthy nutrition. It also extracts the data from the ElasticSearch index but using the field "Health" for the aggregation.

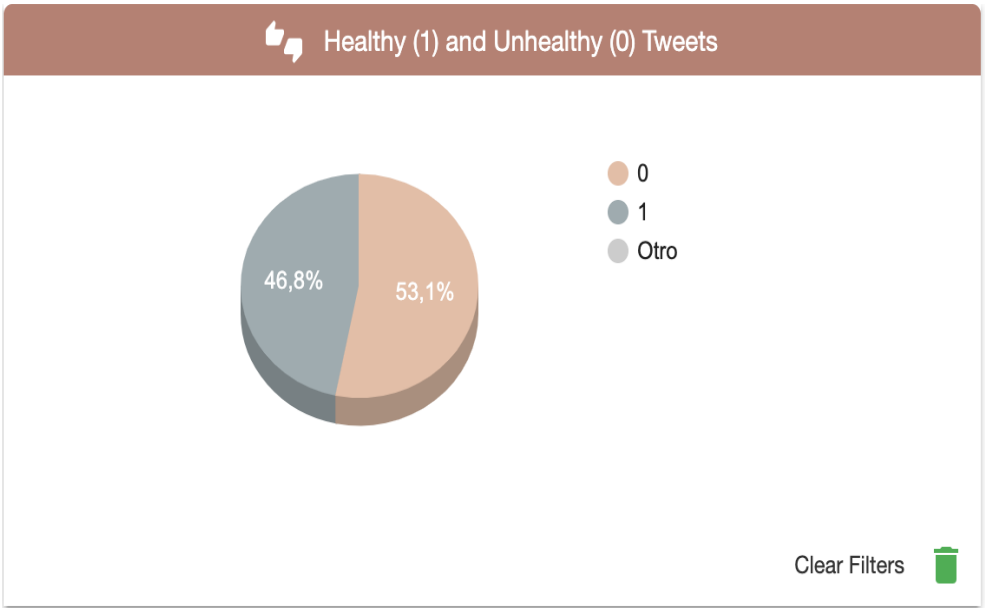

Figure 4.2: Health widget

4.6.1.3 Number charts

This type of charts has been used for the implementation of three widgets that classify the tweets as written by males, by females or written by individuals whose gender is unspecified.

These widgets give the amount of tweets that fall into each category, and receives as an input the following parameters:

- **data:** JSON with all the documents received from ElasticSearch.
- **filters:** list of aggregation filters for ElasticSearch queries.
- **aggkey:** field of the documents used to categorize.
- **icon:** logo corresponding to each of the categories.
- **title:** field that includes the title of the chart.
- **object:** value of the field used as the aggkey parameter.

These widgets perform a filter aggregation to the ElasticSearch index when the user clicks on one of the number charts, allowing the user to visualize only the tweets of a certain category.

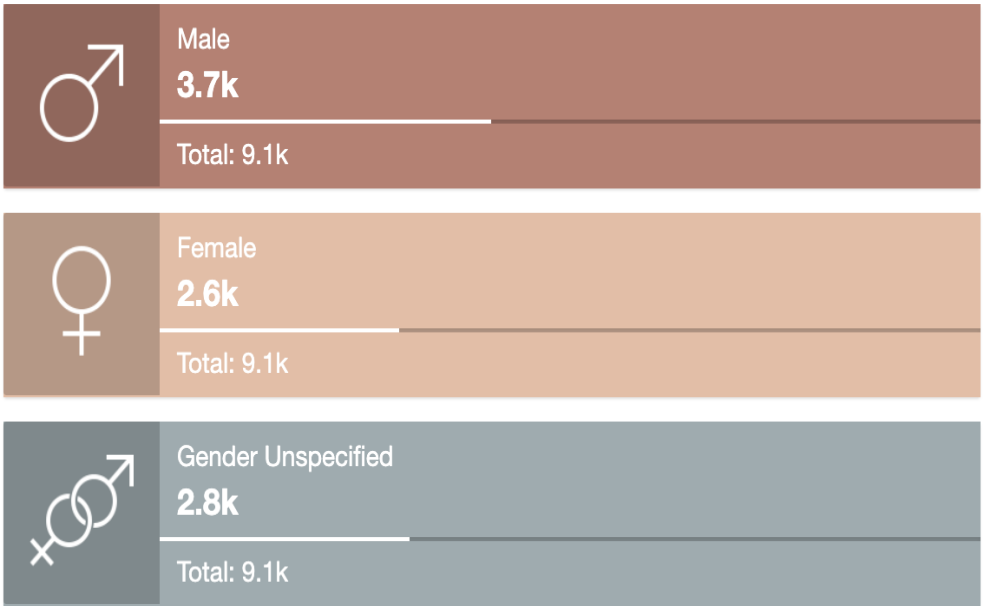

Figure 4.3: Number chart widget

4.6.1.4 Time of creation widget

This widget has been done using the google chart, but using the "area" type and performing a previous function that adds the tweets that have been written in each hour of the day.

This widget allows the user to know how many tweets have been written at every hour of the day, which gives a significant insight in nutrition analysis, allowing the user to know the importance that the population of a certain territory, with a certain gender gives to each meal of the day.

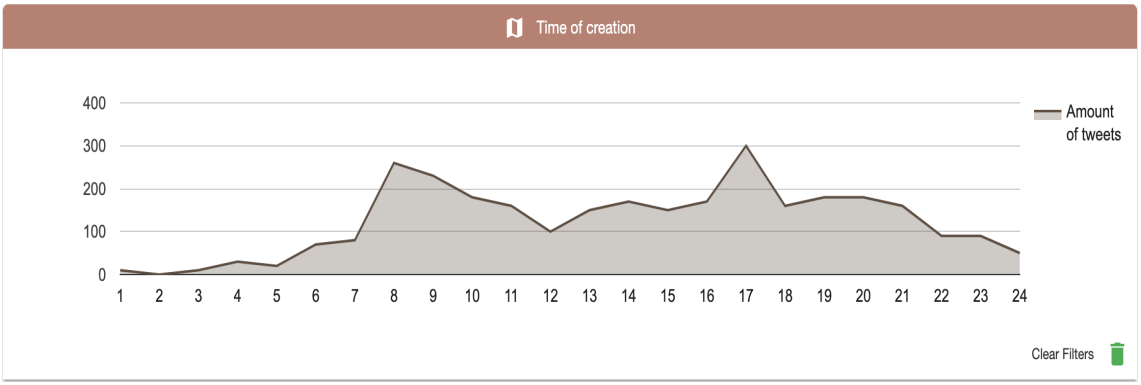

Figure 4.4: Time of creation widget

4.6.1.5 Nutrient mean widget

This widget has also been done using the google chart with the "column" type, but using as input the nutrient mean of the tweets that meet the filters established. It gives the mean of calories, carbohydrates, fats, proteins, cholesterol and fiber of the tweets, giving an insight of the nutrient mean of a certain territory, gender.

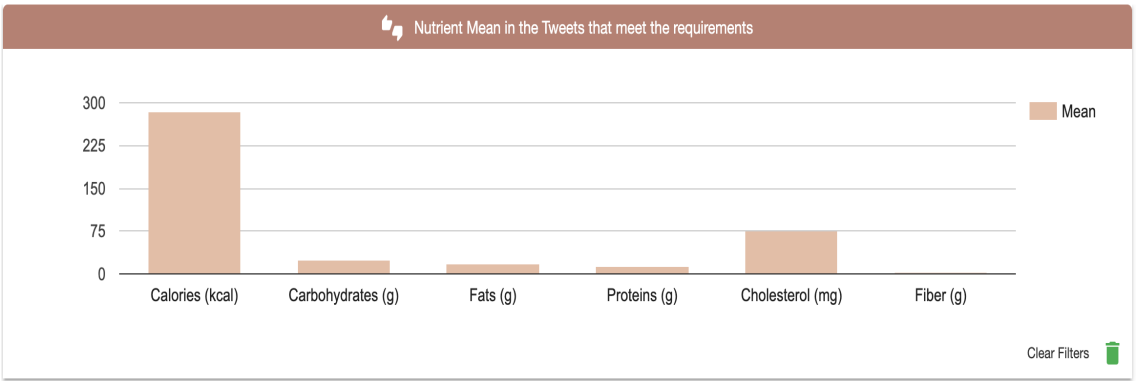

Figure 4.5: Nutrient mean widget

4.6.1.6 Geolocation widget

This widget has been made using the happy-map Sefarad component and displays the location of the tweets analyzed on a map having only the data as an input, being this data the latitude and longitude of the tweets organized as a tuple list.

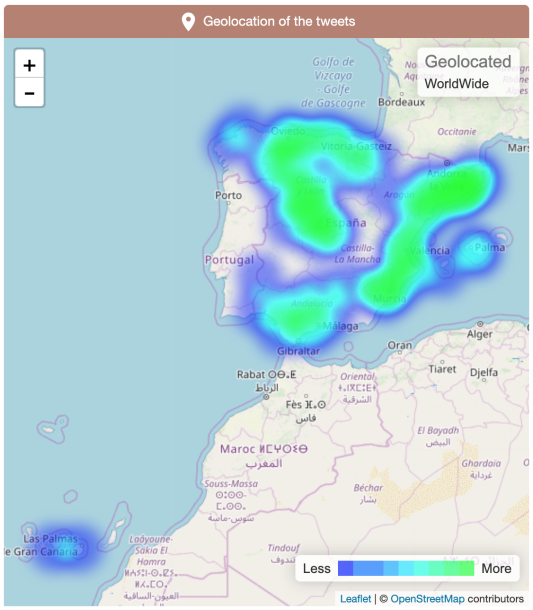

Figure 4.6: Geolocation widget

4.6.1.7 List of tweets widget

This widget has been made using the tweet-chart component, which allows you to view a list of tweets stored in ElasticSearch. The tweets displayed show the sentiment expressed in the text with the background colour, being green the positive sentiment, red the negative and peach the neutral. If the tweet has been labelled as healthy nutrition it shows broccoli as a logo and if it has been labelled as unhealthy nutrition it shows a hamburger as a logo.

It takes data as its only input being data a JSON file with all the documents retrieved from ElasticSearch.

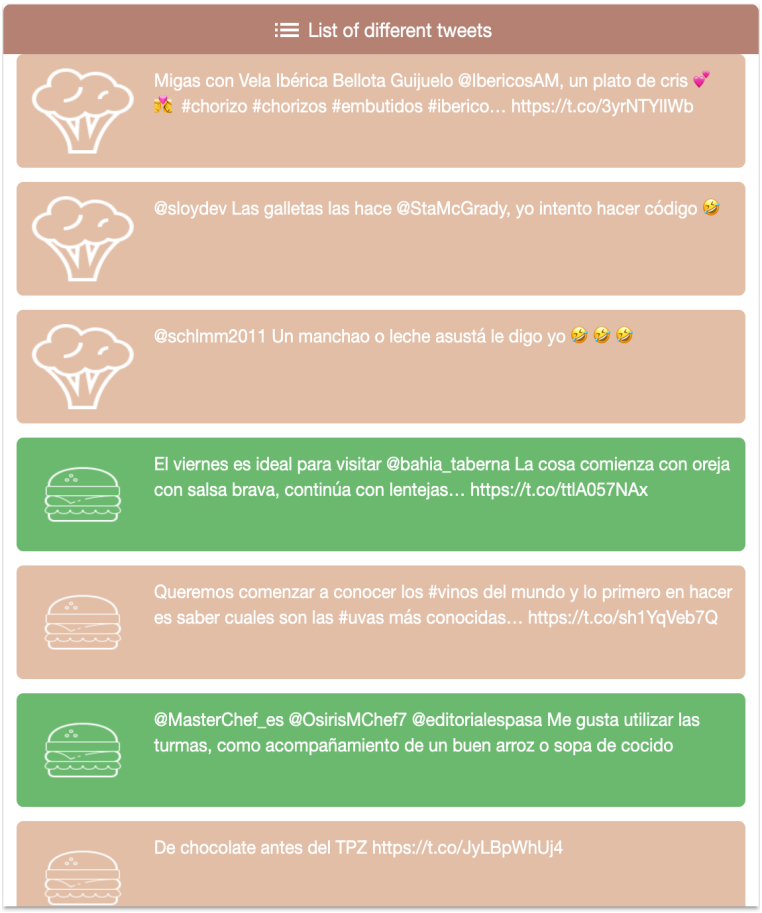

Figure 4.7: List of tweets widget

4.6.2 Dashboard

The widgets united conform a dashboard that can be used as an analysis and visualization tool for nutrition investigation over a period of time and having as a research target the population of a certain territory.

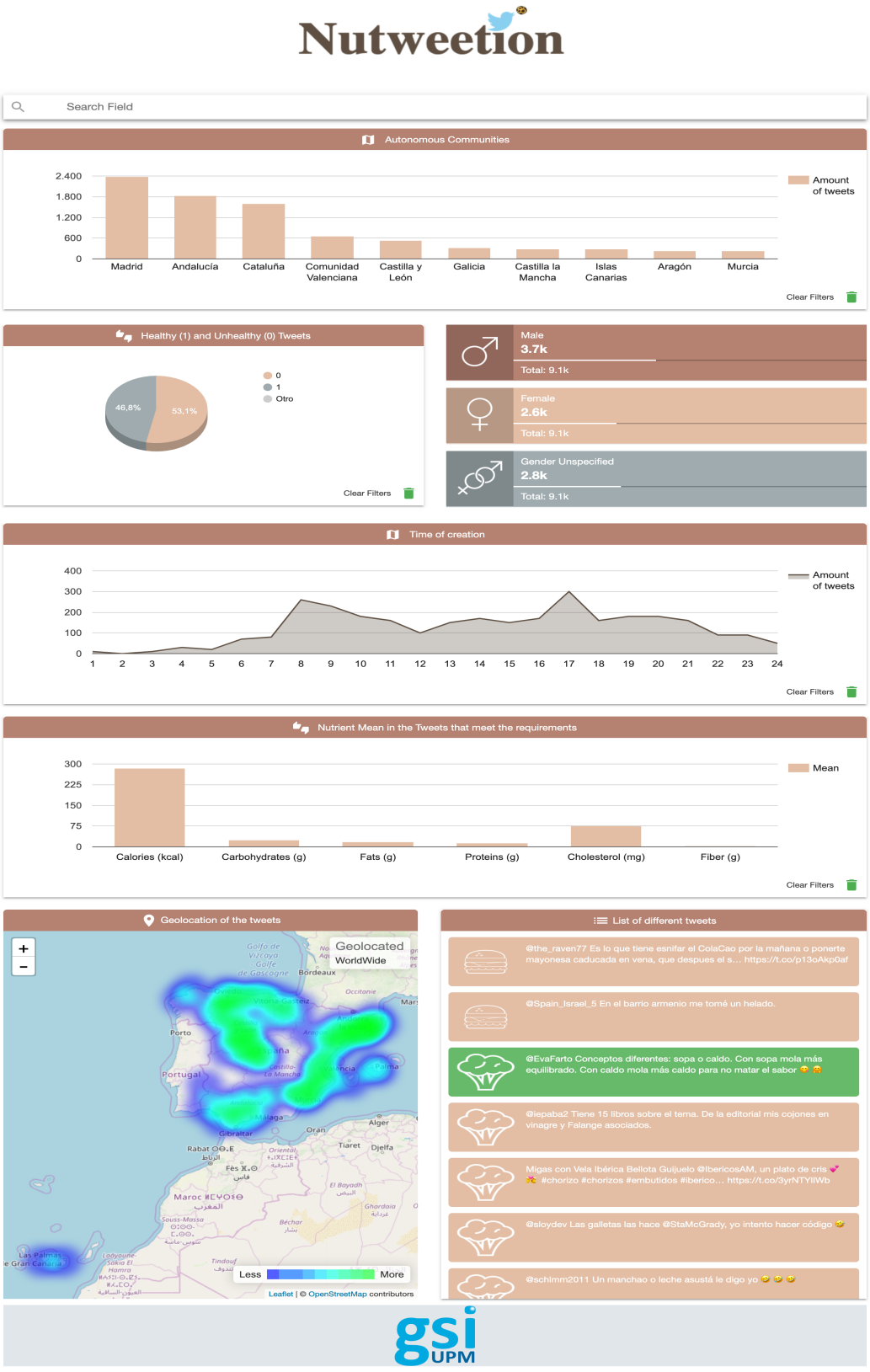

Figure 4.8: Nutweetion Dashboard

CHAPTER 5

Case study

5.1 Introduction

In this chapter we are going to describe a selected use case. This description will cover the advantages that the tool developed can offer while performing statistical research about nutrition over a certain population group.

5.2 Selected Use Case

The selected use case is a nutrition research over the population of Spain. To do this kind of statistical research the methods used tend to be telephone surveys about the nutrition of a sample of the population that is chosen randomly.

While this approach tend to obtain more standard and generic information, our tool allows the user to gather specific information about the food that is being consumed by the population of every Autonomous Community over a period of time that can be as long as the storage and processing resources of the user can handle.

Not only we know the food that is actually being consumed, but also the nutrients associated to that food, allowing the user to perform exhaustive analysis such as obtaining

the nutrient means in the food consumed by a group of population using the "Nutrient Mean" widget described in Sect 4.6.1.5.

This tool also provides methods to know the time of the day in which the most or the least nutrients are consumed and the mean values of these nutrients using the widget mentioned before along with the "Time of Creation" widget explained in Sect 4.6.1.4.

With the different filters implemented, we can establish nutritional profiles between the different Autonomous Communities by performing t-tests using the nutrient values of the tweets as an input, and not only between Autonomous Communities, also between the males and females of a certain Autonomous Community, and even between the different meals consumed during the day of the whole country.

We can also determine if the percentage of population of a certain Autonomous Community, gender and in a certain hour of the day that is having a healthy or unhealthy meal. Also, with the classifier implemented we can monitor the nutrition of a population to estimate overweight rates and establish if a certain population is at risk of overweight based on the nutrition they have.

As we have been describing throughout the whole section, the main advantage of this tool is that, with the big data analytic done, we can gather information to perform a much deeper analysis than the traditional methods used for obtaining this kind of information. That along with the classifier developed for the detection of healthy or unhealthy nutrition and the visualization system allows the user of the tool to analyze the nutrition in a glance.

We have performed an example analysis of the information to obtain nutritional profiles of the population of different areas of Spain. With the nutritional content of the tweets in hand, we proceeded to analyze that data and to do so we divided Spain in 5 different regions containing the following Autonomous Communities:

- North of Spain (**N**): Galicia, Asturias, Cantabria, Pais Vasco and Navarra.
- East of Spain (**E**): Catalonia, Comunidad Valenciana, Aragon and the Balearic Islands.
- South of Spain (**S**): Andalucia, Ceuta, Melilla, Murcia and the Canary Islands.
- Center except Madrid (**C**): Extremadura, Castilla la Mancha, Castilla y León and La Rioja.

- Madrid (**M**).

We analyzed the tweets collected from these regions, and performed t-tests between each region with the rest in order to understand the nutritional differences between them reported by the tweets nutritional content. The test measures whether the average (expected) value differs significantly across samples. If we observe a large p-value, for example larger than 0.05 or 0.1, then we cannot reject the null hypothesis of identical average scores. If the p-value is smaller than the threshold, e.g. 1%, 5% or 10%, then we reject the null hypothesis of equal averages. On the other hand, the F-value is the calculated t-statistic.

The results of these t-tests are shown in table 1. We considered variations in nutrients intake as significant if the p value in the t-tests performed was less than 0,01. As it is shown in the table, there are significant variations between Madrid and the north, the east and the south in terms of energy, being Madrid the area that consumes food with higher caloric content in all cases. Also Madrid has a higher consumption of fats than the north and of proteins than the east and south.

Even though there are only significant variations between Madrid and the north, east and south area, we can see slight differences between regions. For example, in the south region we can see that compared with the north the nutritional intake is lower in all nutrients, while compared with the east and the center region the consumption of fats and fiber is higher in the north. In the east region we can observe that it has a higher carbohydrates consumption than the south and the center region, but less than the north and Madrid, while in terms of fiber consumption the east region has the higher intake above all regions. The center region shows no significant variations with any region, but we can observe that its consumption of energy and proteins is the highest of all regions excluding Madrid, and its consumption of cholesterol is the highest of all regions.

Summarizing, even though Spanish nutrition is more or less similar between all regions, we can detect slight differences between each region's nutrition through tweets. A conclusion that we can extract from the t-tests performed is that Madrid shows a more nutritive diet than the other regions analyzed. This might be because of the gentrification that Madrid has suffered through the years which increases the demand of food in the area. However, an excess of nutrients such as carbohydrates and fats are not beneficial for the fight against overweight, and we can see how the amount of this nutrients is the highest of all regions. This might be due to fast food chains and the stressing life of the metropolitan areas that gives little or no time for cooking your own meals.

Regions		Energy	Carbohydrates	Fat	Protein	Cholesterol	Fiber
MxN	F	2.966951	-0.030104	2.656508	0.729242	0.336523	1.033802
	p	0.003028	0.975985	0.007931	0.465902	0.736496	0.301300
MxE	F	2.905178	1.020297	0.880439	2.658364	0.372669	-0.678222
	p	0.003685	0.307633	0.378660	0.007875	0.709408	0.497660
MxS	F	3.993844	2.530750	2.494618	2.711566	1.243066	0.188121
	p	0.000065	0.011410	0.012639	0.006718	0.213898	0.850788
MxC	F	1.857332	1.386774	1.994307	0.356575	-0.830541	0.824343
	p	0.063342	0.165592	0.046190	0.721429	0.406285	0.409797
SxE	F	0.287851	-1.780104	0.965457	-1.189524	-0.526516	0.895681
	p	0.773477	0.075146	0.334383	0.234314	0.598562	0.370485
SxN	F	-0.692941	-1.481986	-1.4037621	-0.040914	-0.772736	-0.866050
	p	0.488376	0.138402	0.160447	0.967365	0.439712	0.386501
SxC	F	-1.169684	-0.502590	0.094791	-1.745093	-1.701796	0.676812
	p	0.242202	0.615281	0.924485	0.081051	0.088877	0.498566
ExC	F	-0.486151	0.610882	1.118159	-1.687564	-0.997894	1.354641
	p	0.626887	0.541314	0.263569	0.091576	0.318393	0.175612
ExN	F	0.692295	-0.719033	1.753584	-1.138687	0.045528	1.530133
	p	0.488797	0.472167	0.079587	0.254910	0.963688	0.126072
NxC	F	-1.145966	1.065137	-0.778092	-0.355892	-0.895300	-0.241374
	p	0.251950	0.286946	0.436609	0.721959	0.370737	0.809290

Table 5.1: Example of t-test performed between the different nutrition values of the dataset given certain regions.

CHAPTER 6

Conclusions and future work

6.1 Introduction

In this chapter we will explain the conclusions that we have reached after the study. We will also talk about the goals achieved, the problems we have faced in achieving these goals and how we have solved them. Finally, we will present the lines of future work that we have thought for the project.

6.2 Conclusions

In this section we will revise the conclusions that we have reached at the end of this project.

The first conclusion reached is that, as this system uses tweets written in Spanish as input, it can be applied to every Spanish-speaking country with little modifications to change the regions of study. As it can be seen in Sect 3.2.1 Argentina, Spain, Colombia and Chile report a high number of tweets that can be used for research.

In the case of the Spanish Twitter user community, knowing the penetration of this service in the country[1], the percentage of tweets that have the place field filled in, and the number

[1] https://es.statista.com/temas/3595/twitter-en-espana/

of users captured during the dataset formation process, we have estimated that 12,31% of the Spanish users write at least one tweet every month with a food term included in it. This equals to a number of 603.190 users.

The labelling process we have carried out has allowed us to know that 53,1% of the tweets handled talk about food that is unhealthy and the remaining 46,9% talk about healthy food.

We have also been able to extract nutritional profiles of the Spanish population based on the nutrient values present on each tweet, obtaining for example that in the south of Spain, compared with the north the nutritional intake is lower in all nutrients, while compared with the east and the center region the consumption of fats and fiber is higher in the north.

Another conclusion extracted throughout the realization of this project is that sentiment implied in tweets is not relevant for our classification task, as normally when talking about food no sentiments are implied. Nevertheless, this information is very useful when used for marketing purposes rather than medical research, as knowing which food terms produce positive sentiment in population is a useful piece of information for the food industry. We must then conclude that sentiments are not useful when talking about food for this projects target.

Regarding the user's activity of the users that conform our dataset, we have been able to determine that in Spain the population gives the adequate importance to each of the different meals of the day. We have observed that since 8 am the users start to have higher activity than during the night, and that this activity is maintained during the rest of the day until 11 pm when it starts to reduce, showing the highest activity at 8 pm at dinner time.

Finally we must highlight the performance of the classifier developed in this project. This classifier has performed outstandingly well with a dataset that has not a large size, which indicates that with a bigger dataset the performance could be even better.

We have also observed that even though gender is not related to the nutrient amount included in the tweets, its use as a feature for the classifier doesn't imply worse performance in the prediction task.

Using both sets of features, the best results have been given to us by the KNeighboursClassifier which evaluation was described in Sect 3.3.3.1 for both sets. The second set of features

that includes the nutrients and the gender has performed slightly better according to the precision, recall and F1 score. These results in the evaluation mean that including the gender as a feature reduces the amount of false positives and negatives.

Also, as noted in Chapter 5 one of the conclusions of our project is that nutrition in Spain is very similar among regions, but in the south region we can see that compared with the north the nutritional intake is lower in all nutrients, while compared with the east and the center region the consumption of fats and fiber is higher in the north. In the east region we can observe that it has a higher carbohydrates consumption than the south and the center region, but less than the north and Madrid, while in terms of fiber consumption the east region has the higher intake above all regions. The center region shows no significant variations with any region, but we can observe that its consumption of energy and proteins is the highest of all regions excluding Madrid, and its consumption of cholesterol is the highest of all regions.

6.3 Achieved goals

In this section we will explain the goals that we have accomplished in this project.

- **Formation of a dataset of food tweets from different countries.**
  The first goal we achieved was capturing messages posted by Twitter users using the Python Tweepy library along with the Twitter Streaming API and form a dataset with them.

- **Geographical analysis of the dataset.**
  From the dataset, we determine the origin of the food tweets getting to know the Spanish-speaking corresponding to each of them. Specifically, in the case of Spain, we were able to estimate the percentage of users that post food tweets in the Spanish user community, and distribute those tweets between the different Autonomous Communities.

- **Development of a nutrition tweet classifier.**
  As the dataset that we have used for the classifier has been built by us, we have had to design preprocessing and feature extraction techniques that are useful for our classification task. These techniques include NLP techniques for preprocessing and nutritional research for the extraction of characteristics. Finally, we have used machine learning algorithms for classification.

- **Development of a nutrition analysis tool for twitter.**

This last goal is the exploitation of all the goals fulfilled. The tool developed uses the classifier, along with the information extracted from the dataset to analyze the nutrition of the population and present it with an interactive visualization interface with which the analysis of the data extracted is simpler, and conclusions can be extracted in a glance.

6.4 Problems faced

While fulfilling the different tasks included in this project we have had to face several problems, and overcome the difficulties to achieve the goals mentioned before. The problems encountered are the following:

- **Ignorance about the different technologies used in the project.**
  The first problem encountered was the ignorance about most of the libraries and technologies used throughout the project. The first phase of the project consisted in research of the state-of-the-art technologies used to carry out tasks as ours, and once the technologies and methodology were decided, learn how to use these technologies.

- **Formation of the dataset.**
  For the formation of a dataset that was representative of the Spanish population we had to consider the inclusion of the different languages that are spoken in Spain. When considering this we concluded that only Catalan could be included as Euskera and Galician were not supported by the preprocessing technologies used.

6.5 Future work

In this last section of the chapter we will explain the different improvements that could be implemented in this project.

- **Automatizing the process.**
  The work flow of this tool could be automatized using an orchestrator such as Luigi[2] to manage the inputs and outputs of the different systems and supervise the status of the different processes.

- **Adding features of the tool.**
  The tool developed has a specific target that is the analysis of the nutrition of a certain population. However it could be used also for other purposes such as knowing which

[2]https://github.com/spotify/luigi

dishes are more popular in the Twitter community. Also by including another capture filter we could analyze the popularity of different restaurants of a certain area.

- **Adding Instagram as a source of information.**
  Although twitter has a great user community in Spain, Instagram has proven to be one of the most used social networks nowadays between almost all groups of population. Its inclusion in our tool as a source of data would increase the amount of information captured greatly.

APPENDIX A

Cost of the System

A.1 Introduction

In this appendix we are going cover the costs of the project by making an adequate budget for its realization. The economical costs included in this budget will be explained in the following sections.

A.2 Physical Resources

The costs of the physical devices necessary for the development of this project is mainly made up of a computer whose minimum requirements allow the development and deployment of the system.

The resources needed in this computer may be very varied, so as an example we present the features of the computer where the project has been developed:

- CPU: Intel Core i5 3.2 GHz x 4
- RAM: 8GB

- Disk: 128GB

Nowadays a computer with similar characteristics has a price of approximately 700 euros.

A.3 Human Resources

In this section we are going to cover the part of the budget that consists of the cost of the human resources needed to develop the system and for its maintenance.

We must first estimate the salary of a person to carry out the development of a project like ours. This is, a student of telecommunications engineering. Knowing this, we can base the salary of this student on the 360 hours of dedication stipulated in the UPM Collaboration Scholarship, through which similar projects are carried out, with an amount of 1,725 euros.

Also, we must consider the salary of a person in charge of software system maintenance and execution. Professional profiles that are able to execute these tasks are Telecommunication Engineers or Computer Engineers with knowledge of machine learning and NLP. The salary of a worker with this profile is approximately 24.000 euros per year.

A.4 Licences

This section considers the cost that correspond to the software licences necessary for the tools used in the development and deployment of the system built in this project.

As all the software used in this project is open-source, no licences cost must be taken into account.

A.5 Taxes

There is a possibility that the final product is sold to a company that is interested in its acquisition. In that scenario the sale is subject to a tax of 15% of the price of the product as defined in Statute 4/2008 of Spanish law.

APPENDIX B

Impact of this project

B.1 Introduction

In this appendix we will give our analysis of the social, economic, environmental, ethical and professional impacts related to the realization of this project.

B.2 Social impact

Social networks have become the most used online platforms on the Internet. Projects ours allow to perform an analysis of the great amount of content that is shared daily to draw conclusions about the population.

In our case, the tool developed and the study on nutrition, will give researchers on this subject or who carry out projects with social networks to use and start from the resources created (dataset, tools, etc.) and from the conclusions made. This system can be seen as the early stages of a system capable of measuring overweight and obesity rates on the population which would bring about a significant improvement on the characterization of the population's health.

Another group that benefits from our project are companies whose business models deals

with production or vending of food, because this project extract precise information on the interests and habits that their potential customers share.

B.3 Economic Impact

In this section the possible economic impacts that the usage of our project may produce on the companies and researchers will be covered.

From the point of view of the companies and researchers, using our system allows you to capture, analyze and visualize information in a way that allows to reduce by a considerable amount the costs required to perform each of the tasks we have just mentioned.

B.4 Environmental Impact

This section defines the main environmental impact that a system such as ours can produce.

Computers and other information technology infrastructures consume significant amounts of electricity, contributing to greenhouse gas emission due to the production of this electricity.

Also, it is important to take into account the energy required for the cooling system necessarily associated with this equipment, which is the second main reason for the consumption of a high amount of energy.

The main environmental impact of this project will then be the high energy consumption of the server where it is deployed.

B.5 Ethical Implications

In this section we will talk about the ethical implications of our a project.

The first ethical problem we face, is related to the economic impact (Sect. B.3) because our system allows companies to reduce costs, by reducing the human resources needed to perform the tasks that our tool completes.

We believe that systems like ours change the labour landscape, because even though that the system that we have developed could mean the reduction of the work force needed to

perform some tasks, it creates other tasks as maintenance tasks, analysis tasks and execution tasks.

Another ethical issue is due to the use of Twitter data to carry out research such as the one in this project. Twitters privacy policy clearly indicates that users consent to the collection, transfer, and storage of data that is public, while each user has the ability to change their account's privacy settings. This project only analyzed tweets that were available publicly.

APPENDIX C

Architecture of the nutrition analysis service

This appendix includes a diagram of the architecture explained in Chapter 4. It shows the different subsystems included in the architecture, as well as the inputs and outputs of each system.

As we can see the capture system gives as output a dataset with the tweets captured, that are then filtered and preprocessed in the analysis system, giving as an output a JSON file that includes all the features extracted along with other relevant characteristics. This JSON file is stored in the persistence system which communicates with the visualization system through queries to obtain the information from the JSON files and display the charts included in it.

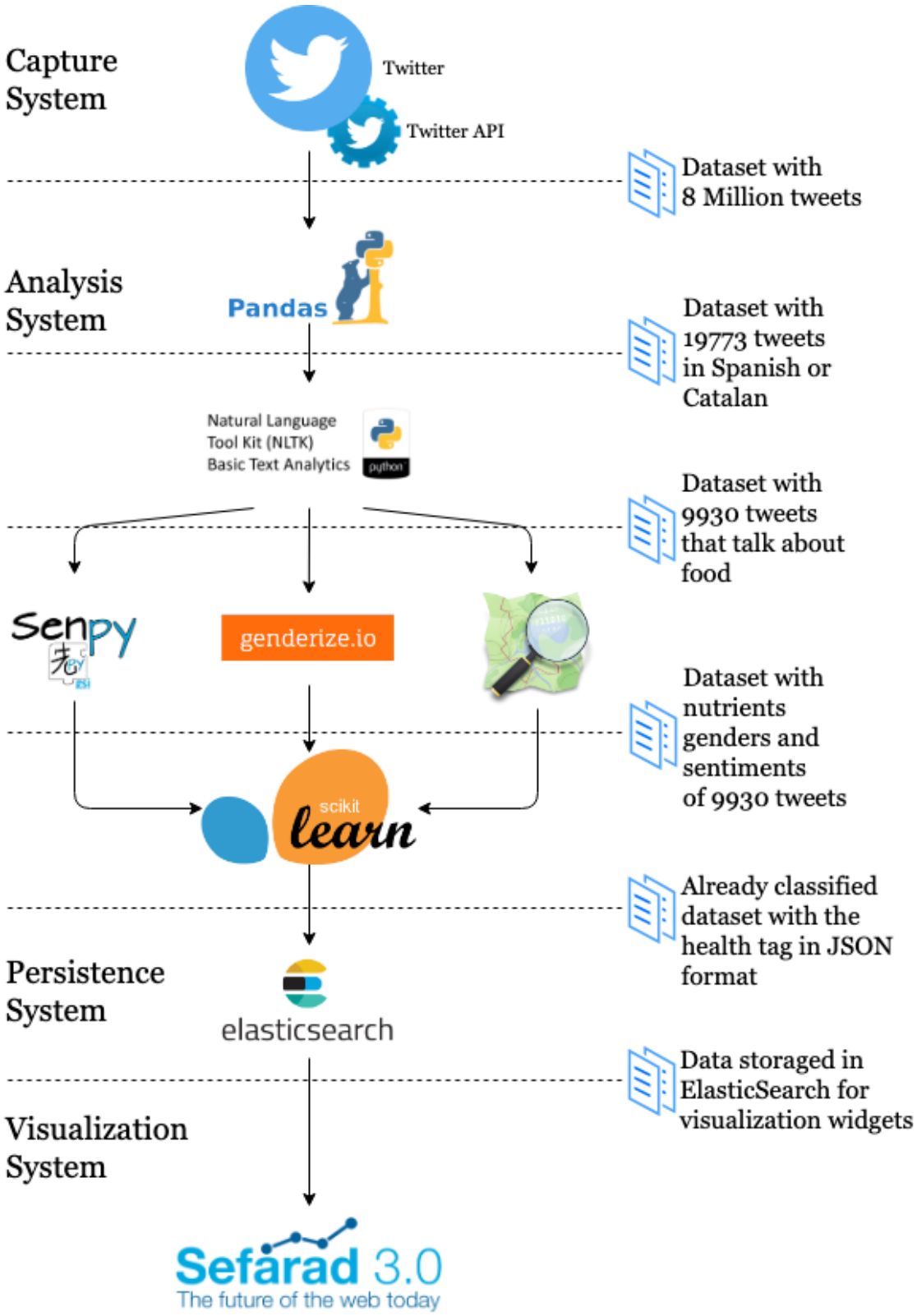

Figure C.1: Nutrition Analysis Service Architecture

Bibliography

[1] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. You tweet what you eat: Studying food consumption through twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3197–3206. ACM, 2015.

[2] Oscar Araque. Design and Implementation of an Event Rules Web Editor. Trabajo fin de grado, Universidad Politécnica de Madrid, ETSI Telecomunicación, July 2014.

[3] Lide Arenaza, Inge Huybrechts, Francisco B Ortega, Jonatan R Ruiz, Stefaan De Henauw, Yannis Manios, Ascensión Marcos, Cristina Julián, Kurt Widhalm, Gloria Bueno, et al. Adherence to the mediterranean diet in metabolically healthy and unhealthy overweight and obese european adolescents: the helena study. *European journal of nutrition*, pages 1–9, 2018.

[4] Albert Bandura. Health promotion by social cognitive means. *Health education & behavior*, 31(2):143–164, 2004.

[5] Genevieve Buckland, A Bach, and L Serra. Eficacia de la dieta mediterránea en la prevención de la obesidad. una revisión de la bibliografía. *Rev Esp Obes*, 6(6):329–39, 2008.

[6] Jose M Castellano, Jagat Narula, Javier Castillo, and Valentín Fuster. Promoting cardiovascular health worldwide: strategies, challenges, and opportunities. *Revista Española de Cardiología (English Edition)*, 67(9):724–730, 2014.

[7] Adrian Cook, Jane Pryer, and Prakash Shetty. The problem of accuracy in dietary surveys. analysis of the over 65 uk national diet and nutrition survey. *Journal of Epidemiology & Community Health*, 54(8):611–616, 2000.

[8] Revista Española de Cardio Velasco Elsoting cardiovascular health worldwide. Los niños de España, entre los más obesos de Europa. *La Vanguardia*, 2018.

[9] Rosalind S Gibson. *Nutritional assessment: a laboratory manual*. Oxford university press, 1993.

[10] Clinton Gormley and Zachary Tong. *Elasticsearch: The definitive guide: A distributed real-time search and analytics engine.* " O'Reilly Media, Inc.", 2015.

[11] Regina Guthold, Gretchen A Stevens, Leanne M Riley, and Fiona C Bull. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1· 9 million participants. *The Lancet Global Health*, 6(10):e1077–e1086, 2018.

[12] Suzanne Higgs and Jason Thomas. Social influences on eating. *Current Opinion in Behavioral Sciences*, 9:1–6, 2016.

[13] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90, 2007.

[14] Lars Johansson, Dag S Thelle, Kari Solvoll, Gunn-Elin Aa Bjørneboe, and Christian A Drevon. Healthy dietary habits in relation to social determinants and lifestyle factors. *British Journal of Nutrition*, 81(3):211–220, 1999.

[15] K Jordahl. Geopandas: Python tools for geographic data. *URL: https://github. com/geopandas/geopandas*, 2014.

[16] Dmytro Karamshuk, Frances Shaw, Julie Brownlie, and Nishanth Sastry. Bridging big data and qualitative methods in the social sciences: A case study of twitter responses to high profile deaths by suicide. *Online Social Networks and Media*, 1:33–43, 2017.

[17] Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 53–54. International World Wide Web Conferences Steering Committee, 2016.

[18] Ranjitha Kashyap and Ani Nahapetian. Tweet analysis for user health monitoring. In *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*, pages 348–351. IEEE, 2014.

[19] M Kebbe, S Damanhoury, N Browne, MP Dyson, T-LF McHugh, and GDC Ball. Barriers to and enablers of healthy lifestyle behaviours in adolescents with obesity: a scoping review and stakeholder consultation. *Obesity Reviews*, 18(12):1439–1453, 2017.

[20] Logan Kendall, Andrea Hartzler, Predrag Klasnja, and Wanda Pratt. Descriptive analysis of physical activity conversations on twitter. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 1555–1560. ACM, 2011.

[21] Dana E King, Arch G Mainous III, Mark Carnemolla, and Charles J Everett. Adherence to healthy lifestyle habits in us adults, 1988-2006. *The American journal of medicine*, 122(6):528–534, 2009.

[22] Harold W Kohl 3rd, Cora Lynn Craig, Estelle Victoria Lambert, Shigeru Inoue, Jasem Ramadan Alkandari, Grit Leetongin, Sonja Kahlmeier, Lancet Physical Activity Series Working Group, et al. The pandemic of physical inactivity: global action for public health. *The lancet*, 380(9838):294–305, 2012.

[23] Edward Loper and Steven Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[24] MJL Lozano and Alfonso Soto González. Actualización en obesidad. *Cad Aten Primaria*, 17(2):101–7, 2010.

[25] James E Lubben. Assessing social networks among elderly populations. *Family & Community Health: The Journal of Health Promotion & Maintenance*, 1988.

[26] Kevin Makice. *Twitter API: Up and running: Learn how to build applications with the Twitter API.* " O'Reilly Media, Inc.", 2009.

[27] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[28] F Márquez-Sandoval, M Bulló, B Vizmanos, P Casas-Agustench, and J Salas-Salvadó. Un patrón de alimentación saludable: la dieta mediterránea tradicional. *Antropo*, 16:11–22, 2008.

[29] Wm Alex McIntosh. Food and nutrition as social problems. In *Sociologies of Food and Nutrition*, pages 215–234. Springer, 1996.

[30] Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython.* " O'Reilly Media, Inc.", 2012.

[31] Yelena Mejova. Information sources and needs in the obesity and diabetes twitter discourse. In *Proceedings of the 2018 International Conference on Digital Health*, pages 21–29. ACM, 2018.

[32] Quynh C Nguyen, Kimberly D Brunisholz, Weijun Yu, Matt McCullough, Heidi A Hanson, Michelle L Litchman, Feifei Li, Yuan Wan, James A VanDerslice, Ming Wen, et al. Twitter-derived neighborhood characteristics associated with obesity and diabetes. *Scientific reports*, 7(1):16425, 2017.

[33] Quynh C Nguyen, Suraj Kath, Hsien-Wen Meng, Dapeng Li, Ken R Smith, James A VanDerslice, Ming Wen, and Feifei Li. Leveraging geotagged twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography*, 73:77–88, 2016.

[34] MC Onwezen, J Van't Riet, H Dagevos, SJ Sijtsema, and HM Snoek. Snacking now or later? individual differences in following intentions or habits explained by time perspective. *Appetite*, 107:144–151, 2016.

[35] World Health Organization et al. Healthy living: what is a healthy lifestyle? Technical report, Copenhagen: WHO Regional Office for Europe, 1999.

[36] World Health Organization et al. http://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. 2018.

[37] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[38] JR Pharr, Courtney A Coughenour, and TJ Bungum. An assessment of the relationship of physical activity, obesity, and chronic diseases/conditions between active/obese and sedentary/normal weight american women in a national sample. *Public health*, 156:117–123, 2018.

[39] James O. Prochaska and Waine F. Velicer. The transtheoretical model of health behavior change. *American journal of health promotion*, 1997.

[40] Joshua Roesslein. tweepy documentation. *Online] http://tweepy. readthedocs. io/en/v3*, 5, 2009.

[41] Sylvia B Rowe and Nick Alexander. Food and nutrition science communications: behind the curtain. *Nutrition Today*, 52(3):151–154, 2017.

[42] Esther Sánchez. El 47% de los madrileños tiene sobrepeso u obesidad. *El País*, 2018.

[43] J. Fernando Sánchez-Rada. Design and Implementation of an Agent Architecture Based on Web Hooks. Master's thesis, ETSIT-UPM, 2012.

[44] J Fernando Sánchez-Rada, Carlos A Iglesias, Ignacio Corcuera, and Oscar Araque. Senpy: A pragmatic linked sentiment analysis framework. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 735–742. IEEE, 2016.

[45] J. Fernando Sánchez-Rada, Alberto Pascual-Saavedra, Enrique Conde-Sánchez, and Carlos A. Iglesias. A Big Linked Data Toolkit for Social Media Analysis and Visualization based on W3C Web Components. In Henderik A. Proper Claudio A. Ardagna Dumitru Roman Robert Meersman Hervé Panetto, Christophe Debruyne, editor, *On the Move to Meaningful Internet Systems. OTM 2018 Conferences. Part II*, volume 11230 of *LNCS*, pages 498–515, Valletta, Malta, October 2018. Springer-Verlag.

[46] George Shaw Jr and Amir Karami. Computational content analysis of negative tweets for obesity, diet, diabetes, and exercise. *Proceedings of the Association for Information Science and Technology*, 54(1):357–365, 2017.

[47] Norbert Stefan, Hans-Ulrich Häring, and Matthias B Schulze. Metabolically healthy obesity: the low-hanging fruit in obesity treatment? *The lancet Diabetes & endocrinology*, 6(3):249–258, 2018.

[48] Leslie O Strolla, Kim M Gans, and Patricia M Risica. Using qualitative and quantitative formative research to develop tailored nutrition intervention materials for a diverse low-income audience. *Health education research*, 21(4):465–476, 2005.

[49] Daniel Suárez Souto, Óscar Araque, and Carlos Ángel Iglesias. How well do spaniards sleep? analysis of sleep disorders based on twitter mining. 2018.

[50] Rannie Teodoro and Mor Naaman. Fitter with twitter: Understanding personal health and fitness activity in social media. *ICWSM*, 2013:611–620, 2013.

[51] Alexander Manuel Valle Flórez, Oscar Fernando Romero Puerto, Nancy Carolina Valero, and Alejandra Vergara. Enfermedades no transmisibles. 2018.

[52] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22, 2011.

[53] Brian Wansink. Position of the american dietetic association: food and nutrition misinformation. 2005.

[54] Gary M Weiss, Jessica L Timko, Catherine M Gallagher, Kenichi Yoneda, and Andrew J Schreiber. Smartwatch-based activity recognition: A machine learning approach. In *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on*, pages 426–429. IEEE, 2016.

[55] Michael J Widener and Wenwen Li. Using geolocated twitter data to monitor the prevalence of healthy and unhealthy food references across the us. *Applied Geography*, 54:189–197, 2014.