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 MACHINE LEARNING SYSTEM FOR FAST FOOD PREVALENCE CHARACTERIZATION USING SOCIAL MEDIA MINING

DANIEL VERA NIETO
ENERO 2020

TRABAJO DE FIN DE GRADO

Título:	Diseño y desarrollo de un sistema de aprendizaje automático de caracterización de la comida rápida usando Social Media Mining
Título (inglés):	Design and development of a machine learning system for fast food prevalence characterization using Social Media Mining
Autor:	Daniel Vera Nieto
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 FIN DE GRADO

DESIGN AND DEVELOPMENT OF A MACHINE LEARNING SYSTEM FOR FAST FOOD PREVALENCE CHARACTERIZATION USING SOCIAL MEDIA MINING

Daniel Vera Nieto

Enero 2020

Resumen

Una creciente epidemia mundial de sobrepeso y obesidad se está apoderando de muchas partes del mundo. Entre los diversos factores que influyen en ello, los hábitos alimentarios son un elemento clave debido a que tienen un profundo impacto en la vida, la salud y el bienestar de los seres humanos. Además, varias enfermedades crónicas han sido relacionadas con el sobrepeso y la obesidad, las cuales se han asociado positivamente con el consumo de comida rápida.

Por lo tanto, el estudio de los hábitos alimentarios es importante tanto para la comprensión cultural como para el control de la salud pública. Tradicionalmente, los estudios a gran escala sobre el consumo de alimentos han utilizado cuestionarios y diarios para hacer un seguimiento de las actividades e ingestas diarias de sus participantes, lo cual puede ser intrusivo y costoso de llevar a cabo. En los últimos años las redes sociales se han convertido en una valiosa fuente de información para evaluar los hábitos, opiniones y decisiones tomadas por sus usuarios, por lo que los investigadores están examinando formas de utilizar los datos sociales para abordar cuestiones relacionadas con la salud.

Enmarcado en un proyecto de colaboración dentro del programa europeo Food4Health, nuestro trabajo pretende determinar si es posible caracterizar el consumo de comida rápida a través de la información de los mensajes publicados en Twitter. Para ello, hemos realizado un clasificador mediante técnicas de aprendizaje automático, disponibles con la biblioteca de Python Scikit-learn, capaz de clasificar los tweets de los usuarios que hacen referencia a la comida rápida. Además, hemos mejorado nuestro sistema con el análisis de las imágenes adjuntas a los tweets utilizando tanto modelos de clasificación de imágenes como de detección de objetos. Están basados en redes neuronales convolucionales enfocadas al reconocimiento de objetos, y se han utilizado técnicas de aprendizaje por transferencia para utilizar modelos pre-entrenados en nuestro objetivo, el reconocimiento de alimentos. Además, se ha realizado el análisis de sentimientos y emociones en los tweets para evaluar los sentimientos y emociones relacionados con la comida rápida.

Palabras clave: comida rápida, aprendizaje automático, clasificación de textos, Twitter, redes neurales convolucionales, análisis de sentimientos

Abstract

An escalating global epidemic of overweight and obesity is taking over many parts of the world. Among several factors that influence this, dietary habits are a key element due to they have a profound impact on human life, health and well-being. In addition, several chronic diseases have been linked to overweight and obesity, which have been found to be positively associated with eating fast-food.

Thus, the study of dietary habits is important for both cultural understanding and for monitoring public health. Traditionally, large-scale dietary studies of food consumption used questionnaires and food diaries to keep track of the daily activities of their participants, which can be intrusive and expensive to conduct. In last years social networks have become a valuable source of information to assess the habits, opinions and decisions taken by their users, so researches are examining ways to use social data to address health-related issues.

Framed in a collaboration project part of the Food4Health European program, our work aims to determine if it is possible to characterize fast-food consumption through the information of the messages posted on Twitter. For this purpose, we have made a classifier using machine learning techniques, availables with Scikit-learn Python library, capable of classifying tweets from users referring to fast-food. In addition, we have enriched our system with the analysis of the images attached to tweets using both image classification and object detection models. They are based on convolutional neural networks focused on object recognition, and transfer learning techniques have been used to use pre-trained models in our objective, to recognize food. Also, sentiment and emotion analysis have been performed on tweets in order to assess the sentiments and emotions related to fast-food.

This system has been tested in the Spanish case, assessing the fast-food prevalence in the country along with image and sentiment analysis to provide health-related information.

Keywords: fast-food, machine learning, text classification, Twitter, convolutional neural networks, sentiment analysis

Agradecimientos

Me gustaría expresar mis agradecimientos a la Universidad Politécnica de Madrid (UPM), en especial a mi escuela, la ETSIT (Escuela Técnica Superior de Ingenieros de Telecomunicación), que me ha aportado todas las herramientas necesarias para realizar este trabajo y me han permitido conocer a gente extraordinaria.

Mi gratitud al personal que forma el Grupo de Sistemas Inteligentes (GSI), por abrirme las puertas y apoyar mi trabajo. En especial, a mis compañeros, con los que he intercambiado ideas y buenos ratos.

También quiero expresar mi más sincero agradecimiento a mi tutor, Carlos Ángel Iglesias Fernández, cuya orientación, apoyo y colaboración ha sido clave para realizar este trabajo y me ha permitido adquirir nuevos conocimientos no solo a nivel técnico, si no a nivel personal.

Finalmente, quiero agradecer a mis padres el haberme apoyado contantemente durante toda esta etapa de mi vida que se cierra con la entrega de este trabajo, apoyo sin el cual no sería quien soy hoy.

Contents

Resumen **I**

Abstract **III**

Agradecimientos **V**

Contents **VII**

List of Figures **XI**

1 Introduction **1**

1.1 Context 1

1.2 Project goals 3

1.3 Structure of this document 4

2 State of Art & Enabling Technologies **5**

2.1 State of art 5

2.2 Enabling Technologies 7

2.2.1 Machine learning technologies 7

2.2.1.1 Scikit-learn 7

2.2.1.2 Image AI 8

2.2.1.3 Keras 8

2.2.1.4 Tensorflow 9

2.2.2 Data technologies 9

2.2.2.1 Pandas . . . 9
2.2.2.2 Geopandas . . . 9
2.2.2.3 Matplotlib . . . 10
2.2.3 Natural Language Processing (NLP) technologies . . . 10
2.2.3.1 Natural Language ToolKit (NLTK) . . . 10
2.2.3.2 GSITK . . . 10
2.2.3.3 Stanford Topic Modeling Toolbox (STMT) & Topbox . . . 11
2.2.4 Twitter developer platform . . . 11
2.2.5 Tweepy . . . 11
2.2.6 Senpy . . . 12

3 Architecture **13**

3.1 Introduction . . . 13
3.2 Retrieval Module . . . 15
3.2.1 Food Dictionary . . . 15
3.2.2 Geographical analysis . . . 16
3.2.3 Annotation . . . 17
3.2.3.1 Themes . . . 18
3.3 Analysis Module . . . 19
3.3.1 Tweet Classification . . . 19
3.3.1.1 Preprocessing . . . 20
3.3.1.2 Feature Engineering . . . 20
3.3.1.3 Classifiers . . . 23
3.3.1.4 Model Performance Evaluation . . . 26
3.3.1.5 Conclusion . . . 30
3.3.2 Image analysis . . . 33
3.3.2.1 Image classification . . . 33

3.3.2.2 Image object detection . . . 37
3.3.3 Sentiment and emotion analysis . . . 40

4 Case study: Fast-food in Spain **41**
4.1 Introduction . . . 41
4.2 System use on one tweet . . . 41
4.3 Spanish dataset . . . 43
4.4 Fast-food in Spain . . . 43
4.5 Image object detection . . . 45
4.6 Sentiment and Emotion analysis . . . 46

5 Conclusions and future work **47**
5.1 Conclusions . . . 47
5.2 Future work . . . 48

Appendix A Project impact **i**
A.1 Social impact . . . i
A.2 Economic impact . . . ii
A.3 Environmental impact . . . ii
A.4 Ethical implications . . . ii

Appendix B Project budget **iii**
B.1 Project structure . . . iii
B.2 Costs evaluation . . . v
B.2.1 Material resources . . . v
B.2.1.1 Software . . . v
B.2.1.2 Hardware . . . v
B.2.2 Human resources . . . vi

B.2.3 Taxes vi

B.3 Conclusion vi

Bibliography **vii**

List of Figures

3.1 System diagram 14

3.2 Map representing the tweets per country 16

3.3 Text classification 19

3.4 Classification process 20

3.5 Model training 23

3.6 Prediction process 24

3.7 Classification with ensembler 31

3.8 Neural network architecture 33

3.9 Image classification supports text classification 34

3.10 Model training process diagram 35

3.11 Transfer learning 36

3.12 Chocolate cake. Accuracy: 97.45% 36

3.13 Pizza. Accuracy: 99.35% 36

3.14 Sushi. Accuracy: 99.35% 37

3.15 Chessecake. Accuracy: 99.98% 37

3.16 RetinaNet architecture 38

3.17 Object detection example 38

3.18 Pizza 39

3.19 Sandwich 39

3.20 Sentiment and emotion analysis 40

4.1 System use on one tweet 42

4.2 Map representing the tweets per province 43

4.3 Fast-food wordcloud 44

4.4 Fast-food related keywords in Spain 45

4.5 Top detections in fast-food tweets 46

CHAPTER 1

Introduction

1.1 Context

Food has a profound impact on human life, health and wellbeing. In the late years, overweight and obesity have increased dramatically. World Health Organization estimates [1] there are more than 1.9 billion adults aged 18 years and older were overweight and, of these, over 650 million adults were obese. Worldwide obesity has nearly tripled since 1975, and overall, about 13% of the world's adult population were obese in 2016 [1].

Several chronic diseases have been linked to overweight and obesity, such as metabolic syndrome, type 2 diabetes, hypertension, coronary artery disease, cancer, osteoarthritis and infertility [2, 3]. This leads to medical care and other expenses associated with obesity costing up to almost 10% of health in United States, 2-3.5% of expenses in countries like Canada, Swiztzerland, Australia, France and Portugal [4]. In Spain, obesity costs up to 7% of national health care system expenses [4]. In addition, obese and overweighted people needs more health resources than people in normal weight [5].

Eating fast food was positively associated with a high-fat diet and Body Mass Index (BMI) [6] and is associated with higher energy and fat intake among adolescents [7]. Besides, positive correlation has been found between obesity and fast food restaurants [8] [9]. In

addition, consumption of fast food, which have high energy densities and glycemic loads, and exposes customers to excessive portion sizes, may be greatly contributing to and escalating the rates of overweight and obesity [10]. We observe there is a strong relation between an unhealthy diet and overweight and obesity risk.

Thus, the study of dietary habits is important for both cultural understanding and for monitoring public health. To best address this issue, public health awareness campaigns use data on dietary behavior across various segments population to tailor their messages to particular focus groups. Having detailed and accurate data on the cultural and individual behaviors that lead to unhealthy dietary habits is necessary for effective intervention programs.

Traditional studies focus on applying qualitative methods (focus groups) as well as quantitative ones (clinical trials and surveys). But the results during these decades are not enough. Until now, large-scale dietary studies of food consumption used questionnaires and food diaries to keep track of the daily activities of their participants, which can be intrusive and expensive to conduct.

World is changing and in last years social networks have become a valuable source of information to assess the habits, opinions and decisions taken by citizens, providing its users with a means of documenting the minutiae of their daily lives, including their dietary choices. Researchers have been examining ways to use social data to better understand and monitor public health problems in real-time. This growing area of research has been called infodemiology or infoveillance studies [11].

In our case, we are interested in determining if it is possible to characterize fast-food consumption through the information of the messages posted on a social network. Social media technology, such as Twitter, allows users to communicate with each other by sharing short messages of a maximum of 280 characters with the possibility of accompanying them with a photo, video, and/or link. Users can share their thoughts, feelings, and opinions on these social media platforms and, as a result, social media data may be used to provide real-time monitoring of behavioral outcomes that inform health behaviors. A unique aspect of social media data from Twitter is that the posts are public and possibly geotagged and thus, all internet users, including health researchers, can readily access these data. In addition, unique to Twitter is the use of hashtags (#) that allows a user to highlight and allow other users to follow relevant topics of interest. Given their high level of use, these sites collect an enormous amount of data (eg, over 500 million tweets per day on Twitter) [12].

1.2 Project goals

The main objective of this project is develop a tool capable of gathering health indicators from social media, to be used in a collaboration project framed in the Food4Health European program [13]. For this purpose, we have made a classifier using machine learning techniques capable of classifying messages from users referring to fast-food , adding further insights generated from the messages and its related data (images) using Deep-Learning models. These messages are taken from the social network Twitter, selecting only Spanish-speaking users. In addition, we will carry out a geographical study of the subject focused in Spain. The results obtained from these objectives will help us to develop a system that allows us to obtain the characteristics of the tweets related to fast-food.

In order to achieve these objectives explained in the previous section, the following tasks have been carried out during the project:

1. T1: Study the state of the art in relation to Natural Language Processing and machine learning technologies. Specifically, its application on health related studies.
2. T2: Collection of tweets related to fast-food in order to create a dataset.
3. T3: Geo-location study of the tweets that form the dataset.
4. T4: Development of a first classifier capable of selecting tweets related to fast-food following the next steps:
   (a) Development of a preprocessing technique for each message to be analysed. In this preprocessing process, the characteristics that we want the classifier to learn must be extracted.
   (b) Dividing the processed data in two parts, with one part we will train the classifier and with the second part we will check if it performs its function satisfactorily.
   (c) Analysis of the results of the different machine learning algorithms.
5. T5: Development of a second classifier capable of distinguish the theme in fast-food tweets.
6. T6: Analysis of characteristic patterns of users who consume fast-food in Spain by province.

1.3 Structure of this document

In this section we provide a brief overview of the chapters included in this document. The structure is as follows:

- Introduction ***(Chapter 1).*** This chapter introduces the reader to the context where the project is developed in, giving a quick overview of the health issues related of fast-food, how classical health studies are made and how we pretend to develop a tool to make health studies easier. In addition, the goals of the project are presented along with the specific tasks we will perform.
- State of Art & Enabling Technologies ***(Chapter 2).*** This chapter offers the reader a review of all the previous knowledge and works that have inspired ours. It explains the current approaches to apply social media mining to health related studies. Also, a brief review of the main technologies that have made possible this project is included.
- Architecture ***(Chapter 3).*** This chapter describes the architecture of the project, from the captured tweets to all the outputs our system provides based on the texts and images of the tweets.
- Case study: Fast-food in Spain ***(Chapter 4).*** This chapter describes the system implementation in a particular case, where its validity is discussed and insights are extracted as a Proof-of-Concept
- Conclusions and future work ***(Chapter 5).*** This chapter details the achieved goals and outcomes. In addition, future work lines are described.
- Project impact ***(Appendix A).*** This appendix shows the social, economic, environmental impact as well as ethical implications
- Project budget ***(Appendix B).*** This appendix describes the necessary project budget regarding material and human resources needed, as well as taxes involved.

CHAPTER 2

State of Art & Enabling Technologies

2.1 State of art

Neighborhood data collection is expensive and time consuming, and then existing data are only available for certain time periods or certain areas. To solve this situation, using large scale data from social media has become a new trend among health researches since the global spread of this new way of communication. There is an incremental interest in using social media to track health issues such as flu [14], alcohol consumption [15] insomnia [16] and depression [17].

Social media mining techniques have been proved to work on social media networks of different types. One recent study involving the use of big data was performed by Achananuparp et al [18]. They conducted a research to discover whether journaling encourage healthy eating choices analyzing data from MyFitnesspal, a diet tracking social network. They used a dataset containing almost 100K unique food entries recorded over almost 2M meals. In addition, Schwartz et al. [19] developed predictive models of well-being over the language used in status of Facebook users. Additionally, they introduced a lexicon with annotated well-being data containing 10 categories and 1522 words that has been used in several researches.

But we could say Twitter is one of the biggest sources of information. Given the fact most of Twitter's content is public and its use is extended, it is not a surprise that Twitter has been used to study human behavior on a big scale. Twitter data can be used to analyze phenomena ranging from the number of people infected by the flu, to national elections, to tomorrow's stock prices. In addition, researchers have been able to track public behaviors, information and opinions about a broad range of topics, including those related to health issues [20]. Information generated via Twitter can be useful in the examination of beliefs, attitudes, and sentiment towards certain health topics (e.g., vaccines) [21] as well as health-related activities and health status. Views and activities described via social media can help shape perceived norms, attitudes, beliefs and subsequently, behaviors of people.

Focusing in the filed of nutrition research using media mining on Twitter, we find a recent study that analyzed a large scale conversational public health data set to characterize the general public's opinions regarding diabetes, diet, exercise and obesity [22], discovering the topics usually related to each keyword. The insights extracted could support public health experts and social scientists in better understanding common public opinions about these topics.

Augmenting traditional models based on demographic variables with Twitter-derived information improves predictive accuracy for several health-related statistics [23], including obesity, diabetes and access to healthy foods. Also, linguistic indicators were identified and correlated to different health-related statistics. This suggest that Twitter is a useful complementary source of information, not only detecting explicit mentions of the research interest, but processing the language characteristics.

As a representative work on Twitter mining, Yanai et al [24] proposed a system to collect food photos from Twitter monitoring the social media stream to find tweets with photos containing food-related keywords and apply a 2-steps classifier to label up to 100 different food classes. This allowed them to perform a spatio-temporal analysis discovering the most popular foods in each part of Japan at each moment of the year.

Food message and image analytics have been used to classify food types and sentiments towards those food terms in order to perform a geospatial analysis of tweets and map them onto the obesity prevalence map [25]. The outcome of this is a Big Data framework that can be used to reveal social food trends or sentiments in the obesity prevalence regions.

Using geotagged Twitter data, Nguyen et al were able to create neighborhood indicators for happiness, food and physical activities, finding that happy tweets making reference to healthy food and physical activity were less frequent in low-income locations from three counties [26]. Nguyen et al. also extended their research in another work to create zip code

level indicator of community happiness and social modeling of diet and physical activity from tweets, combining them with administrative data from the state of Utah to examine the relation between neighborhood characteristics and chronic disease [27].

In a similar way, Widener et al. implemented a data-mining framework to use geolocated Twitter data to explore the prevalence of healthy and unhealthy food across USA, performing sentiment analysis on them. This information was related to demographic indicators like low-income and low-access census tracts, finding a lower proportion of tweets about healthy foods with positive sentiment and a higher proportion of unhealthy tweets in general [28].

A recent study [29] explored to what extent it is possible to use twitter to get insights into dietary habits of a country. They used Twitter to monitor dietary habits at both national and personal scale, performing a large scale analysis of 210K Twitter users in the United States, tracking their 502M tweets. They enriched the collected data with demographic indicator such as income, education and gender; with interests from each user; and with the nutritional values of food. The results show that the foods mentioned in the daily tweets of users are predictive of obesity and diabetes statistics and that the calories tweeted are linked to interests and demographic indicators.

Thus, it has been proved the potential of using social media mining to address health issues, in special, by using Twitter platform. However, most of these studies have been performed in United States and there are few focusing on European countries. For this reason, the goal of this research is to find out to what extension it is possible to create an automatic analysis system similar to those presented but focusing on the specific case of Spain. That is, we pretend to develop an automatic system that allow us to perform province-level nutritional analysis in Spain focusing on a specific kind of food: the fast-food.

2.2 Enabling Technologies

2.2.1 Machine learning technologies

2.2.1.1 Scikit-learn

Scikit-learn [30] is an open source machine learning Python library built on NumPy, SciPy, and matplotlib that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection and evaluation, and many other utilities.

This project was created by David Cournapeau as a Google Summer of Code project. Later, Matthieu Brucher started work on this project as part of his thesis. Finally, INRIA assumed the leadership of the project and made the first distribution in 2010. Since then, they have published several releases with the help of a growing international community.

We have used this library to make some preprocessing steps; and to select and evaluate different models.

2.2.1.2 Image AI

ImageAI [31] is a python library that integrates state-of-the-art Deep Learning and Computer Vision capabilities using simple and few lines of code. It is built on Tensorflow, OpenCV [32] and Keras. It currently supports image prediction and training using 4 different Machine Learning algorithms trained on the ImageNet-1000 dataset. ImageAI also supports object detection, video detection and object tracking using RetinaNet, YOLOv3 and TinyYOLOv3 trained on COCO dataset. Also, it allows you to train custom models for performing detection and recognition of new objects. ImageAI is a project developed by Moses Olafenwa and John Olafenwa , the DeepQuest AI team.

We have used ImageAI to perform object detection on the images included in our tweet dataset. The object detection API provides the capability to detect, locate and identify 80 most common objects in everyday life in a picture using pre-trained models that were trained on the COCO Dataset [33]. The selected pre-trained model is RetinaNet [34], which has high performance and accuracy, but with longer detection time.

2.2.1.3 Keras

Keras [35] is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. It was developed with a focus on enabling fast experimentation. Being able to go from idea to result with the least possible delay is key to doing good research. It offers a higher-level, more intuitive set of abstractions that make it easy to develop deep learning models regardless of the computational backend used.

Keras was initially developed as part of the research effort of project ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System) in 2015 by François Chollet, a Software engineer and AI researcher at Google. In 2017, Google's TensorFlow team decided to support Keras in TensorFlow's core library.

The image classification algorithm we used to perform analysis on the images of our

tweet dataset is built on Keras.

2.2.1.4 Tensorflow

Tensorflow [36] is an end-to-end open source platform for machine learning. It provides a comprehensive, flexible ecosystem of tools, libraries, and community resources that lets researchers push the state-of-the-art in Machine Learning and developers easily build and deploy Machine Learning applications.

TensorFlow was originally developed by researchers and engineers working on the Google Brain team in 2011 within Google's Machine Intelligence Research organization to conduct machine learning and deep neural networks research and it was released on 2015.

It has been used as back-end for Keras in the image classification algorithm.

2.2.2 Data technologies

2.2.2.1 Pandas

Pandas [37] is an open-source Python library that provides high-performance, easy to use data structures and data analysis tools. It offers data structures and operations for manipulating numerical tables and time series. The principal features and functionalities that provide pandas library are data grouping, merging and querying operations as well as time series analysis, reading and writing tools in different file formats and data munging.

Development was started in 2008 by Wes McKinney to perform quantitative analysis on financial data. In 2015, pandas became a NumFOCUS sponsored project, a non-profit organization from United States which promotes open-source projects.

Pandas was used in this work to manage, clean and manipulate our tweets dataset.

2.2.2.2 Geopandas

Geopandas [38] is an open source project to make working with geospatial data in Python easier. It combines the capabilities of Pandas and Shapely [—7:online], providing spatial operations on geometric types extending the datatypes used by Pandas.

We have used Geopandas to easily work with geographic data with the goal of performing location analysis on our dataset.

2.2.2.3 Matplotlib

Matplotlib [39] is a Python library which provides 2D plotting capabilities to produce publication quality figures such as plots, histograms, power spectra, scatterplots and errorcharts with just a few lines of code.

It has been originally developed by John Hunter and has grown with the contribution of a large community, becoming one of the NumFocus sponsored projects.

We have used this library to produce the figures shown in this work. In particular, we have used Geoplot for creating maps easy taking advantage of its great compatibility with matplotlib. Geoplot [40] is a high-level Python geospatial plotting library which makes mapping easy.

2.2.3 Natural Language Processing (NLP) technologies

2.2.3.1 Natural Language ToolKit (NLTK)

NLTK [41] is a Python open source community-driven platform to work with human language data. It provides easy-to-use tools of text processing for classification, tokenization, stemming, tagging, parsing and semantic reasoning.

The NLTK project began at the University of Pennsylvania in 2001 when Steven Bird and Edward Loper agreed a plan for developing software infrastructure for NLP teaching that could be easily maintained over time.

NLTK library has been used in this work to preprocess the texts in our dataset and extract the features needed to fed the machine learning models.

2.2.3.2 GSITK

GSITK [42] is a library on top of scikit-learn that eases the development process on NLP machine learning driven projects. It manages datasets, features, classifiers and evaluation techniques, so that writing an evaluation pipeline results fast and simple.

GSITK is a tool developed by the Intelligent Systems Group at ETSIT, Universidad Politécnica de Madrid (UPM), which development started in 2017.

We have used this library mainly for text preprocessing.

2.2.3.3 Stanford Topic Modeling Toolbox (STMT) & Topbox

STMT [43] goal is providing easy-to-use topic modeling tools to perform analysis on datasets that have a substantial textual component. Some of its features are topic models (such as LDA, Labeled LDA, and PLDA) training and easy text manipulation.

The Stanford Topic Modeling Toolbox was written at the Stanford NLP group by: Daniel Ramage and Evan Rosen, first released in September 2009.

Since this toolbox is written in an old version of Scala, we have used Topbox [44], a small Python wrapper around the Stanford Topic Modeling Toolbox (STMT) that makes working with L-LDA a bit easier with no need to leave the Python environment.

We have work with this wrapper to use L-LDA, a supervised topic modeling algorithm, in order to get features to train our topic classifier.

2.2.4 Twitter developer platform

Twitter developer platform [45] provides many API products, tools, and resources that enables to harness the power of Twitter's open, global, and real-time communication network. It is advertised as a tool to publish and analyze Tweets, optimize ads, and create unique customer experiences. Some of the functionalities it provides are searching tweets, create Twitter Ads campaigns, get engagement metrics, send direct messages, retrieve account activity or embedding tweets on other websites.

Twitter is recognized for having one of the most open and powerful developer APIs of any major technology company, becoming a reference implementation for public REST APIs since its first release in 2006.

In this work we made use of the Streaming API to download twitter messages in real time. It is useful for obtaining a high volume of tweets, which suits perfect for our goal of collecting a tweets dataset.

2.2.5 Tweepy

Tweepy [46] is an easy-to-use Python library for accessing the Twitter API. It provides all the functionalities needed to exploit the different Twitter APIs, including the Streaming API. Tweepy makes it easier to use the twitter Streaming API by handling authentication, connection, creating and destroying the session, reading incoming messages, and partially routing messages.

We have made use of this library to easily utilize the Twitter Streaming API.

2.2.6 Senpy

Senpy [47] is a framework for sentiment and emotion analysis services. Its goal is to produce analysis services that are interchangeable and fully interoperable. All services built using Senpy share a common interface, which allow users to use them simply by pointing to a different URL or changing a parameter.

The development started in 2014 was carried by the Intelligent Systems Group at ETSIT, UPM, as part of the european Mixed Emotions project.

Senpy has been used to perform sentiment and emotion analysis on the tweets to carry further analysis.

CHAPTER 3

Architecture

3.1 Introduction

In this chapter, we cover the design phase of this project, as well as implementation details involving its architecture. Firstly, we present an overview of the project, divided into two main modules: Retrieval and Analysis Module. This is intended to offer the reader a general view of this project architecture. After that, we present each module separately and in much more depth explaining the details of their sub-modules.

A general overview of the system is depicted in Figure 3.1. First, the Retrieval Module (Section 3.2) collects data from Twitter and prepares it to be used by the Analysis Module. Then, the Analysis Module (Section3.3) process each tweet in the Tweet Classification module (Section 3.3.1) to predict if it is related to fast-food. After that, for those tweets related to fast-food, images are analyzed in the Image Analysis module (Section 3.3.2) and Sentiment and Emotion Analysis is performed (Section 3.3.3).

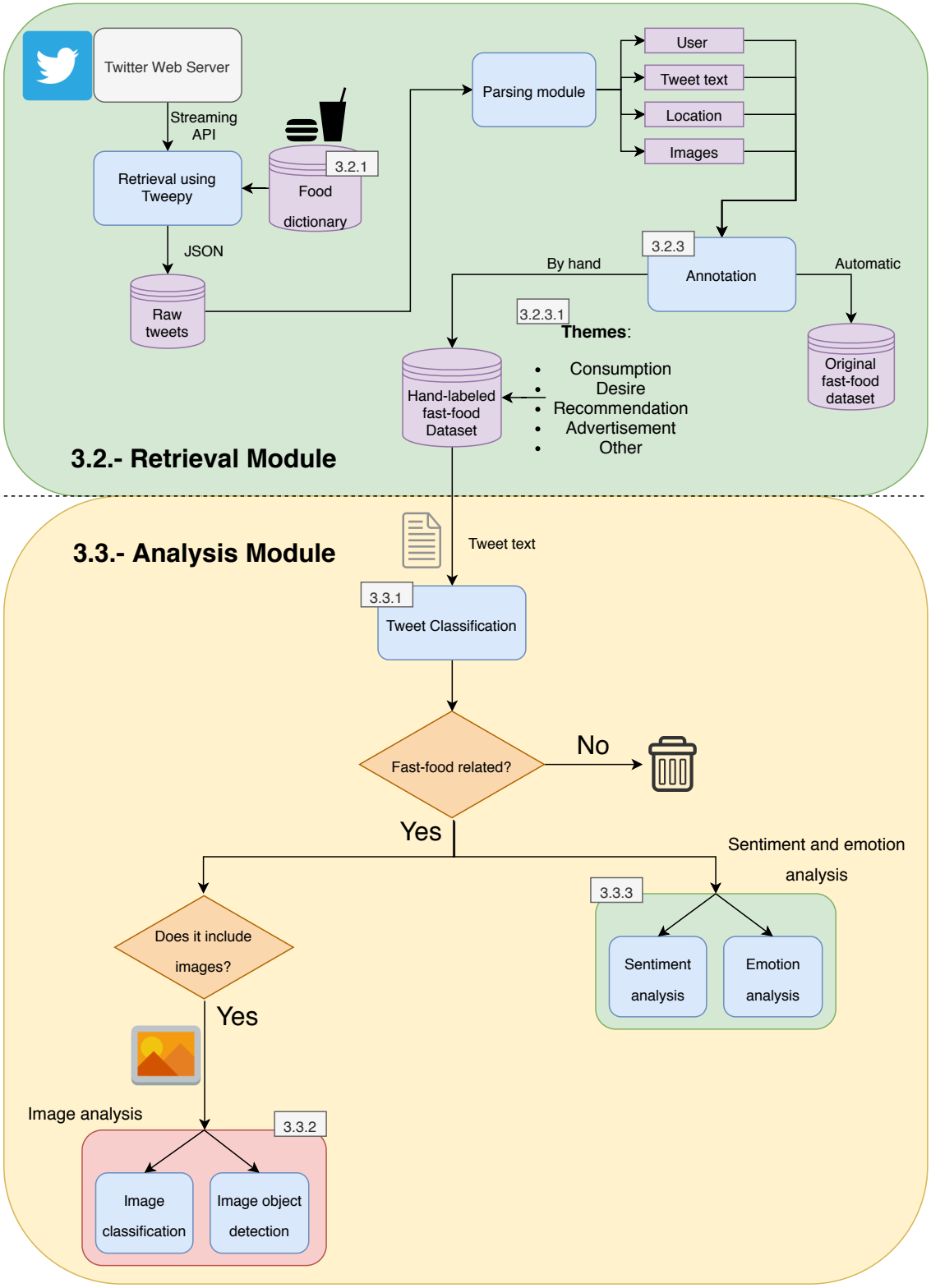

Figure 3.1: System diagram

3.2 Retrieval Module

This section describes the main parts of the Retrieval Module, from tweets collection to dataset annotation after processing the raw tweets received from Twitter.

The first task is to create a tweets dataset: the tweets were captured from November 7th, 2019 to December 15th, 2019. The capture was performed using the Streaming Twitter API and the Tweepy library, tools that allow us to continuously collect a random 1% sample of publicly available tweets matching our criteria. The tweets had to meet the following criteria:

- The tweets language must be in Spanish
- They have to contain a word of our food-keywords dictionary (see Section 3.2.1, since we are looking for food-related tweets.
- They can not be re-tweets (a re-post of a tweet originally posted by a different user). Therefore we focus only on original tweets.

The collected sample dataset matching this criteria contained a total number of 1.341.198 tweets containing fast-food keywords. The tweets, collected in JSON format, are parsed to extract the attributes we are interested on: User, Text, Location and Images. Note that this attributes are not directly available and some preprocessing has been needed to be able to obtain them. Also, a geographical analysis have been performed on this dataset in section 3.2.2 using the location extracted from tweets.

In addition, we have collected a dataset of tweets from Spain without specific keywords to use it as a “control dataset", in order to use it as negative sample for the classification task. This might introduce noise to the classification task since it is possible to have captured tweets talking about food in this dataset, but we have considered it insignificant. We will refer to this dataset as “generic dataset".

3.2.1 Food Dictionary

We handcrafted a dictionary of unhealthy foods based on the same criteria used by Nguyen et al [27], adapting it to the Spanish food culture. For the unhealthy food list, we add some of the foods used in the previously mentioned work that are considered relevant to the Spanish case, the top fast food restaurants in Spain [48], the most consumed alcoholic beverages in Spain [49] as well as well-known beer brands in Spain. The lists is presented

in Table 3.1. We have chosen these fast-food related keywords since, in most of the cases, they are good representatives of fast-foods and unhealthy eating.

Unhealthy foods list

'Cocacola', 'Coca cola', 'Pepsi', '100 montaditos', 'McDonalds', 'Burguer King', 'Taco-Bell', 'Telepizza', 'Dominos', 'Dominos Pizza', 'Starbucks', '#Rodilla', 'Pans & company', 'KFC', 'Fanta', 'Estrella Galicia', 'Cruzcampo', 'Mahou', 'Estrella Damn', 'Amstel', 'bacon', 'pastel', 'tarta', 'galletas', 'bebida energética', 'perrito caliente', 'helado', 'pizza', 'frito', 'cerveza', 'hamburguesa', 'kebab', 'refresco', 'fastfood', 'vodka', 'whiskey', 'ron', 'tequila', 'sandwich', 'ginebra', 'anís', 'brandy','nuggets'

Table 3.1: Food keywords

3.2.2 Geographical analysis

The first task we performed on this dataset was locating the geographical origin of the captured tweets. From the whole dataset, only 53786 tweets were geolocated. This is a 4.01% of the dataset. The tweets origins are depicted in the Figure 3.2. The top countries with more tweets related to food are shown in Table 3.2.

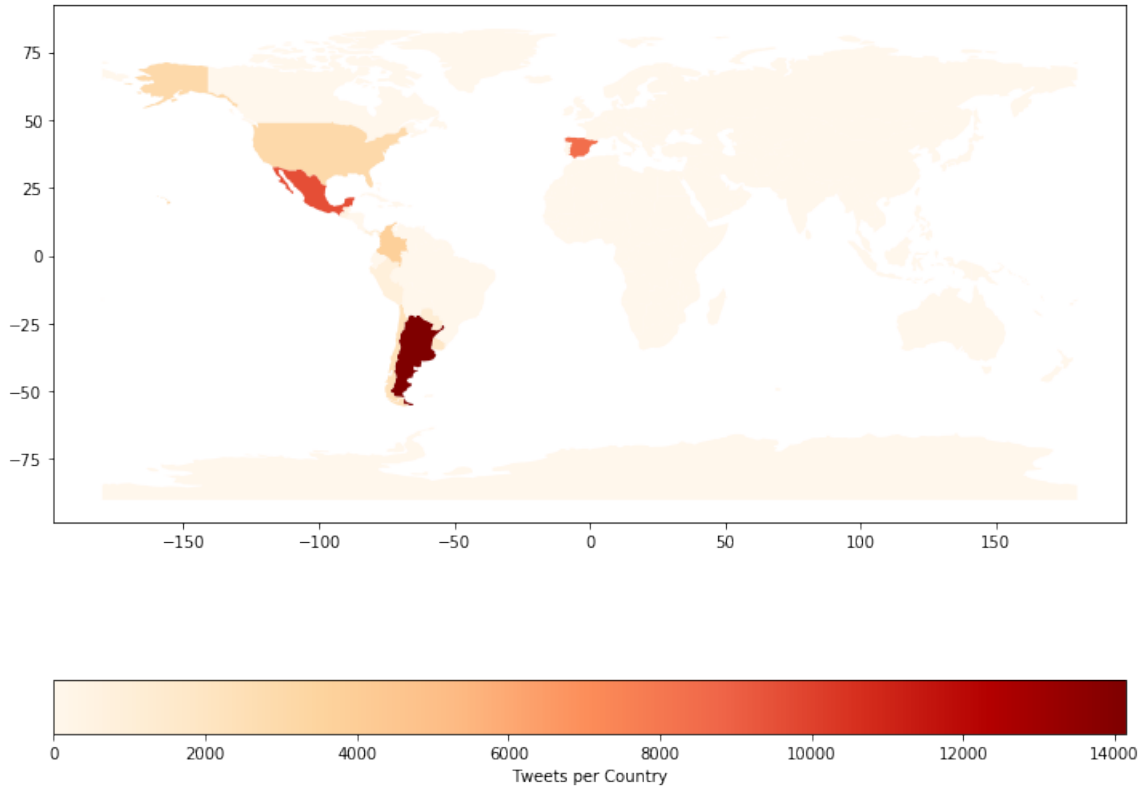

Figure 3.2: Map representing the tweets per country

Argentina	14146
México	9573
España	8431
Colombia	3864
Estados Unidos	3062
Chile	2403
Uruguay	1841
Paraguay	1174
Panamá	959
Perú	869

Table 3.2: Top-10 countries by number of tweets

3.2.3 Annotation

Our main goal is to detect and characterize tweets from users who talk about fast food as a supervised classification task. We need a sample of labelled tweets for this process. We have followed two different approaches to label our dataset in order to perform the classification task.

Firstly, we have supposed all captured tweets captured with the fast-food keywords are talking about food. This might have introduced some noise in our dataset due to polysemic words or errors during the capture, but we have considered this to bee a minor issue. We took a sample of 50K tweets from the gathered fast-food dataset. The advantage of this approach is we quickly have a big labeled-tweets dataset with minor work that we can use for classification. We added to this dataset a random sample from the generic dataset of 50K tweets, forming a labeled dataset of a total of 100K tweets with fast-food and non-categorized tweets. We will refer to the dataset result of this approach as original fast-food dataset.

Secondly, we have performed our own labelling process distinguishing between tweets talking about fast-food and the rest of the tweets. Despite we tracked tweets by a specific set of keywords, there are some tweets not related to food due to usernames containing any

of our keywords mainly.

Tweet	Fast-food
“Hoy fuimos a tomar un helado con Sacha [...]"	Yes
“Esta debería ser mi vista en estos momentos"	No

Table 3.3: Example of tweets.

We annotated by hand a random sample of 500 tweets, where 385 were related to fast-food somehow, that is, the 77% of the labeled tweets. Since this dataset is imbalanced (there are more tweets related to fast-food than generic tweets), we added a random sample of 270 tweets from the generic dataset, ending with a balanced dataset of 770 tweets. We will refer to the dataset result of this approach as hand-labeled fast-food dataset.

3.2.3.1 Themes

Next task is to group the tweets dataset by topic, allowing us to better characterize the messages and thus draw better conclusions. We took into consideration the most common topics we detected while inspecting the hand-labeled fast-food dataset. The defined topics are:

- Consumption: User express consumption of fast-food.
- Desire: User express desire or intention of consuming fast-food.
- Recommendation: User makes a recommendation of fast-food.
- Advertisement: User advertises fast-food. This topic is usually related to corporate accounts.
- Others: Tweets about fast-food not belonging to previous categories.

Once we have defined the topics, we performed the labelling of hand-labeled fastfood dataset. It can be seen that most of the tweets does not belong to any category (30.4%) and a significant amount of captured tweets is not related to fast-food at all (23%). There are 90 tweets where the user express consumption of fast-food (18%), 82 where the user express desire (16.4%) 33 where the user makes a recommendation (6.6%) and 28 where there is any kind of advertisement (5.6%).

These percentages show that we have a dataset where only 46.6% of tweets will be useful to develop an automatic theme classifier. In addition, the considerably amount of tweets considered to be not related to fast-food shows the limitation in the use of original fast-food dataset.

3.3 Analysis Module

This section describes the different components of the Analysis Module. First, text classification is performed on tweets texts (Section 3.3.1). Then, to those tweets classified as related to fast-food, images are analyzed (Section 3.3.2) as well as sentiments and emotions associated to tweets are obtained (Section 3.3.3).

3.3.1 Tweet Classification

Text classification, also known as text categorization or text tagging is the task of assigning predefined tags to unlabeled texts using a Text Classification Model (Figure 3.3). This is one of the natural language processing (NLP) applications in different use cases.

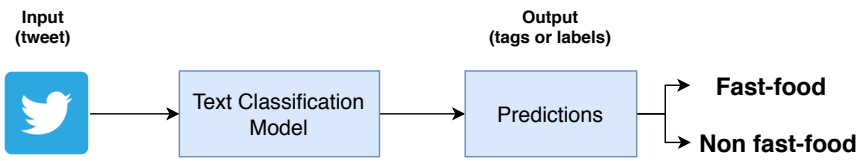

Figure 3.3: Text classification

In our case, we have developed different classification models for the automatic labelling task. First, we have developed a classifier able to distinguish when a tweet is talking about fast-food. Then, we want to discover how it address the topic, i.e., classify the tweet in themes, resulting in our second classifier.

The workflow to develop both classifiers is the same, that has been automated using a pipeline from the scikit-learn library. The steps used in the pipeline are Preprocessing, Feature Engineering and Classification. After that, we have performed the Model Evaluation of each classifier.

Figure 3.4: Classification process

3.3.1.1 Preprocessing

Prior to feature engineering, we made a small preprocess on the raw tweets that are delivered as JSON objects by the Twitter Stream API. The captured tweets are transformed to a pandas dataframe, retweets are filtered, the text to analyze is extracted from the "text" field or the "full_text" field if the latter exists and geolocation information is extracted.

Then text cleaning is performed to raw text to ensure no distortion is introduced to the model. Special characters such as "/n" are removed, the text is downcasted and punctuation signs are removed.

Then, we have used the module for preprocessing tweets from the GSITK library. This module allow to gather the important information and characteristics from the tweet and to generalize specific unimportant information such as urls and usernames changing them by the tag "<URL>" or "<USER>" respectively.

The result of this preprocessing is also tokenized using the TweetTokenizer module from NLTK library, which converts the text to a list of tokens taking into account the characteristics of the tweets.

In addition, we have removed stopwords. These are meaningless words that do not provide further information about the text. The list of words include articles, pronouns, prepositions, auxiliary verbs, etc. The dictionary used as stopwords is provided by the NLTK library. We have added to this list the tags generated with the preprocess module from GSITK.

Finally, the tokens are stemmed, this is, reduced derived words to their word stem or root form. We have used the Porter Stemmer algorithm provided by NLTK to perform this.

3.3.1.2 Feature Engineering

Feature engineering is the process of transforming data into features to act as inputs for machine learning models such that good quality features help in improving the model performance. When dealing with text data, there are several ways of obtaining features that

represent the data. We will cover the methods we have used thanks to scikit-learn package, which provides a library of transformers that allow us to perform the text feature extraction easily.

- **Lexical and Part-Of-Speech (POS) Features**

  We have used a custom transformer to extract lexical features such as the length of the text and the number of sentences it has using the sentence tokenizer from NTLK package, which split a text into its sentences.

  A similar approach has been used to extract POS features, we have created a transformer which is able to process a text and attach a part of speech tag to each word using the pos-tag module from NLTK.

  Therefore, after these two transformers we obtained features regarding the length of the texts, the number of sentences they have and the lexical categories (or POS) of the words in the texts.

- **TF-IDF term weighting**

  This transformer provided by scikit-learn produces a matrix of vectors that represents the relative importance of a term in the document and the dataset.

  This importance is calculated the Term Frequency (TF) and the Inverse Document Frequency (IDF), where TF represents how many times a term appears in a text and IDF the inverse of the number of texts in the corpus containing the term.

  The goal of TF-IDF is to reduce the impact of words that occur very frequently, thus carrying very little meaningful information compared to features that occur in a small fraction of the training corpus. The importance score increases proportionally to the number of times a word appears in a text and is adjusted by the number of texts containing that word. In addition, the length of the texts is taking into account by normalizing the TF term.

  This transformer produces useful features, however, it does not take into account any word order dependence and hence it cannot capture phrases and multi-word expressions.

- **N-grams**

  This transformer is implemented using modules also provided by scikit-learn and the result is also a matrix of TF-IDF vector, but now occurrences of N consecutive words are counted.

N-grams is the combination of N terms together and it allows to take into account the word order dependence, extracting characteristics from the context of each word. First, the N-grams are extracted from each texts, obtaining a matrix of vectors with the number of times each N-gram appears in each text. Then, this matrix is converted using the TF-IDF transformer to another matrix representing the relative importance of each N-gram.

- **Latent Dirichlet Allocation (LDA)**

  Latent Dirichlet Allocation is a probabilistic model for collections of discrete dataset such as text corpora. It is also used for discovering abstract topics from a collection of documents. This algorithm considers that the corpus is a collection of documents, a document is a sequence of words and there are a determined number of topics in the corpus, where each document is a mixture of topics. Then, the probability of a document belonging to a topic is calculated based on the probability of each word of that document in an unsupervised manner.

  We have implemented this algorithm using the corresponding module from scikit-learn to obtain features related to the common patterns of each kind of texts.

- **Labeled LDA (L-LDA)**

  This variation of LDA allows to use the algorithm for supervised learning purposes since it allows to indicate not only the number of topics but a set of labeled documents too.

  Apart from that, the algorithm works on the same principle as LDA and allows to determine the topic of a document using the probability of the words belonging to each topic. To obtain the words belonging to each topic we have used the Stanford Topic Modeling Toolbox. The result is a dictionary with the words in the dataset and the probability of belonging to each of the topics.

  We have implemented a transformer using this information to extract the probability of a document belonging to each of the themes, which will be useful for the topic classification.

- **Feature preparation**

  We have used the Feature union module of scikit-learn library to apply the list of transformer objects presented above in parallel to the input data and then concatenate the results. This is very useful to combine several feature extraction mechanisms into a single transformer.

The features used in the original fast-food dataset are Lexical features and the features obtained from TF-IDF and N-grams of two words.

However, these features were not enough to characterize the hand-labeled fast-food dataset, so in this case we have also used features obtained using the LDA algorithm. For the theme classifier, we have used the same features except LDA that was interchanged with L-LDA.

3.3.1.3 Classifiers

Once we have obtained the features from our initial data, the text from tweets, we are able to train a machine learning model to perform the classification of tweets. In our case we have used supervised machine learning models, this is, we give the model both the features and the correct label or tag for each test in order to allow the model to “learn" which features are related to which label. This process is known as Model training and the result is a trained model able to classify data similar to the one it has learned from (Figure 3.5).

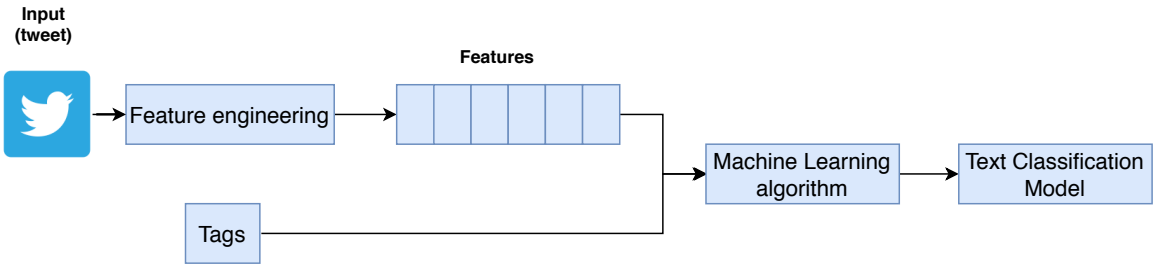

Figure 3.5: Model training

Most of the times, to train a model the dataset is splitted in train set and test set. The reason of this is to not use the same data to both train and evaluate a model, this way we avoid the model learning how to classify only the data we fed into it and not generalizing the relations between features and the corresponding class. This problem is known as overfitting and it occurs when our model does not generalize well from our training data to unseen data.

The trained machine learning model can make predictions after the same feature extractor is used to convert unknown raw text to features that will be used to get predictions on labels (Figure 3.6).

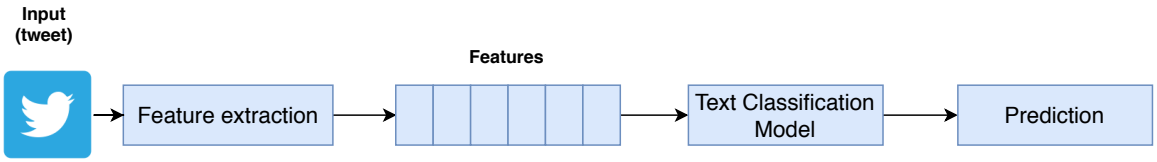

Figure 3.6: Prediction process

We have tested several machine learning classification algorithms provided by scikit-learn. Each of them has different hyper-parameters, parameters that are not directly learnt within our model or estimator. Then, a previous task to perform the classification is to select the best parameters for our case or fine-tune the model. Fine-tuning has been done using the GridSearchCV, a module from scikit-learn. This module generates candidates from a grid of parameter values specified, it fits on our dataset all the possible combinations of parameter values, they are evaluated and the best combination is retained. The parameters of the estimator used are optimized by cross-validated grid-search over a parameter grid. This way we obtain the best model hyper-parameters for our use case.

Now we will provide a breve description of the classification models we have used in this work:

- **Naive Bayes.** Naive Bayes methods are algorithms based on applying Bayes' theorem, which describes the probability of an event based on prior knowledge of conditions that might be related to the event. These kind of algorithms make the "naive" assumption of conditional independence between every pair of features given the value of the class variable. The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of the conditional probability. Despite of their assumptions, these classifiers have worked well in different use cases and they have significant advantages such as the small amount of data they require or the high speed of the training process. For our problem, we have tested two kinds of classifiers of this family:

  - **Multinomial Naive Bayes.** This classifier is used for multinomial models and is suitable for classification with discrete features like word counts. The performance of this method is not expected to perform very good since it is thought to be used in multi-class problems.

  - **Bernoulli Naive Bayes.** This model is used for data that is distributed according multivariate bernoulli distributions. The main difference from multinomial NB is this one explicitly penalize the non-occurrence of a feature that is indicator

of a class, where the multinomial variant would simply ignore this fact. This model might perform better on datasets with shorter documents, which is really interesting for our case. This method is not expected to perform very good since all our features are not binary.

- **K-Nearest Neighbors.** This type of classification is computed from the instances of training data by simply majority vote of the K nearest neighbors of each point, this is, a point is assigned to the class which has the most representatives nearest to the point. For high-dimensional parameter spaces, this method becomes less effective due to the so-called "curse of dimensionality. When the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance like this one. As we generate thousands of features from the text, this method is not expected to perform well.

- **Logistic Regression.** Despite its name, is a model used for classification rather than regression. Logistic regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.

- **Support Vector Classification.** Support vector machines are supervised learning algorithms with interesting advantages for our case study: they are effective in high dimensional spaces, they are memory efficient and very versatile since they can be used in regression and classification problems and can easily handle multi-class variables. The classifiergenerates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane(MMH) that best divides the dataset into classes. SVM finds an optimal hyperplane which helps in classifying new data points.

- **Decision Tree.** A supervised learning method which predicts the value of a target variable by learning simple decision rules inferred from the data features. The probability that a given record belong to an option of the decision rule is calculated iteratively until the record is classified in the most likely class. Some of the advantages of this algorithm that are of interest for our case are that it is simple to understand and to interpret since trees can be visualised and it requires little data preparation.

- **Random Forest.** A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the

predictive accuracy and control over-fitting. That is, it consists of a large number of decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

- **ExtraTrees Classifier.** Similar to the random forest classifier, this classifier is a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The main difference is the splits use random variables to be computed. A random subset of candidate features is used, but instead of looking for the most discriminative thresholds, thresholds are drawn at random for each candidate feature and the best of these randomly-generated thresholds is picked as the splitting rule.

We have implemented a pipeline which take the raw text of the tweets, preprocess it, perform the feature extraction and then feed this data into each of the classifiers. The results will be discussed in next section.

3.3.1.4 Model Performance Evaluation

Already implemented the workflow of the classifiers, we need now a way to select the best classifier for our case. To accomplish this goal we will use different metrics that will be explained in this section.

Let us define some concepts before:

- True Positive (TP): Label which was predicted Positive (in our scenario “fast-food") and is actually Positive (i.e. belong to Positive “fast-food" Class).
- True Negative (TN): Label which was predicted Negative (in our scenario “non fast-food") and is actually Negative (i.e. belong to Negative “non fast-food" Class).
- False Positive (FP): Label which was predicted as Positive, but is actually Negative, or in simple words the sample wrongly predicted as “fast-food" by our Model, but is actually “non fast-food".
- False Negatives: Labels which was predicted as Negative, but is actually Positive ("fast-food” predicted as “non fast-food").

In addition, it is important to note we have used K-fold cross validation (CV) to compute the metrics. In CV the training set is split into k smaller sets or “folds". Then the model is trained using K-1 of the folds as training data and the resulting model is validated on the remaining part of the data (i.e., it is used as a test set to compute a performance measure such as accuracy). The performance measure reported by k-fold cross-validation is then the average of the values computed in the loop.

The metrics we have used to evaluate the performance are Recall (Rc), Precision (Pr), Accuracy (Ac) and F1 Score (i.e., F1).

- Precision. Measures the percentage of true results among the total number of positive predictions.
$$Precision = \frac{TP}{TP + FP}$$
- Accuracy. Proportion of correct predictions among the total number of cases examined.
$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$
- Recall. Measures the percentage of actual positives that is correctly classified
$$Recall = \frac{TP}{TP + FN}$$
- F1 score. Harmonic mean of precision and Recall. It will be low if either precision or recall are low.
$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Once we have defined the metrics we are able to evaluate the performance of each model.

- **Fast-food classifier**

  The goal of this classifier is to label fast-food related tweets. For this purpose, we have trained our models with two different datasets described in subsection 3.2.3. The main reason of using the second dataset is to develop a more precise classifier.

  First, we used the original fastfood dataset. The dataset is compounded of 100K tweets divided equally in positive and negative classes. The raw text is preprocessed and features are extracted using the lexical transformer, TF-IDF and bigrams (N-grams of two words) as explained in subsection 3.3.1.2.

  The metrics for each of the classification models are presented in Table 3.4. The metrics has been computed using K-fold cross validation with K=5.

Classifier	**Hyper-params**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Multinomial NB	alpha: 0.1	0.87 (+/- 0.00)	0.84 (+/- 0.01)	0.92 (+/- 0.01)	0.87 (+/- 0.01)
Bernoulli NB	-	0.91 (+/- 0.01)	0.94 (+/- 0.01)	0.88 (+/- 0.01)	0.91 (+/- 0.01)
K-NN	metric: 'euclidean' n_neighbors: 19 weights: 'distance'	0.58 (+/- 0.01)	0.68 (+/- 0.02)	0.31 (+/- 0.00)	0.42 (+/- 0.01)
Logistic Regression	C: 1000.0 penalty': 'l2'	0.92 (+/- 0.01)	0.97 (+/- 0.01)	0.87 (+/- 0.01)	0.92 (+/- 0.01)
Decision Tree	max_depth: None	0.92 (+/- 0.01)	0.94 (+/- 0.00)	0.89 (+/- 0.01)	0.91 (+/- 0.00)
Random Forest	criterion: 'gini' max_depth': 8 max_features: 'auto' n_estimators': 500	0.86 (+/- 0.03)	0.94 (+/- 0.04)	0.74 (+/- 0.12)	0.82 (+/- 0.06)

Table 3.4: Classifiers performance on original fastfood dataset

The machine learning algorithms with the best performance are Logistic Regression and Decision Tree.

In order to get a better performance and more accurate results, hand-labeled fastfood dataset was used. The dataset is compounded of 770 labeled-by-hand tweets divided equally in positive and negative classes. The raw text is preprocessed and features are extracted using the lexical transformer, TF-IDF, N-grams of up to three words and LDA as explained in subsection 3.3.1.2.

The motivation of using this handmade dataset is that in the first dataset some tweets labeled as positive were not really related to fast-food. The advantage of this dataset is that has been manually created and supervised by an human, so we have higher confidence in the labeled as positive records. However, the disadvantage of this dataset is the extremely smaller amount of data than the previous dataset.

The metrics for each of the classification models are presented in Table 3.5. The metrics has been computed using K-fold cross validation with K=5.

Classifier	Hyper-params	Accuracy	Precision	Recall	F1-score
Multinomial NB	alpha: 0.4	0.75 (+/- 0.12)	0.73 (+/- 0.23)	0.85 (+/- 0.20)	0.77 (+/- 0.09)
Bernoulli NB	-	0.74 (+/- 0.06)	0.88 (+/- 0.31)	0.59 (+/- 0.22)	0.69 (+/- 0.05)
K-NN	algorithm: 'ball_tree' n_neighbors: 3 p: 2	0.57 (+/- 0.09)	0.58 (+/- 0.12)	0.56 (+/- 0.07)	0.57 (+/- 0.06)
Logistic Regression	C: 13 penalty: 'l2' tol: 0.0001	0.85 (+/- 0.02)	0.83 (+/- 0.09)	0.88 (+/- 0.08)	0.85 (+/- 0.02)
SVC	C: 1 gamma: 1 kernel: linear	0.76 (+/- 0.05)	0.95 (+/- 0.12)	0.55 (+/- 0.11)	0.70 (+/- 0.07)
Decision Tree	max_depth: 6	0.77 (+/- 0.06)	0.99 (+/- 0.03)	0.54 (+/- 0.10)	0.70 (+/- 0.07)
Random Forest	n_estimators: 33	0.87 (+/- 0.08)	0.92 (+/- 0.06)	0.81 (+/- 0.14)	0.84 (+/- 0.11)
ExtraTrees	-	0.77 (+/- 0.05)	0.94 (+/- 0.10)	0.58 (+/- 0.11)	0.71 (+/- 0.06)

Table 3.5: Classifiers performance on hand-labeled fastfood dataset

The machine learning algorithm with the best performance is Random Forest.

- **Theme classifier**

  The goal of this classifier is to classify the tweets previously labeled as “fast-food" into the topics defined in section 3.2.3.1.

  From the hand-labeled fast-food dataset, that was labeled by hand, we observe that there are 90 tweets where the user express consumption of fast-food (18%), 82 where the user express desire (16.4%) 33 where the user makes a recommendation (6.6%) and 28 where there is any kind of advertisement (5.6%). This is a dataset with 233 categorized tweets in total divided in four topics where tweets are imbalanced distributed. To make the classifying task easier, we have grouped “Recommendation" and “Ads" categories with “Other", having a final dataset divided on three classes: “Consumption" (90 tweets), “Desire" (82 tweets) and “Other" (213).

  Once again, the raw text is preprocessed and features are extracted using the lexical transformer, TF-IDF, N-grams of up to three words and L-LDA as explained in subsection 3.3.1.2.

The metrics for each of the classification models are presented in Table 3.6. The metrics has been computed using K-fold cross validation with K=5. Due to the obtained results, only Multinomial Naive Bayes, Logistic Regression and Random Forest models, Support Vector Classifier and K-Nearest Neighbors are presented.

Classifier	**Hyper-params**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Multinomial NB	alpha: 0.1	0.59 (+/- 0.04)	0.62 (+/- 0.13)	0.59 (+/- 0.04)	0.51 (+/- 0.05)
Logistic Regression	C: 7 penalty: 'l2' tol: 0.0001	0.61 (+/- 0.05)	0.60 (+/- 0.06)	0.61 (+/- 0.05)	0.58 (+/- 0.05)
Random Forest	n_estimators: 9	0.62 (+/- 0.05)	0.65 (+/- 0.22)	0.62 (+/- 0.06)	0.59 (+/- 0.06)
SVC	C: 1 gamma: 9.9995e-07 kernel: 'linear' probability: True	0.58 (+/- 0.04)	0.57 (+/- 0.08)	0.58 (+/- 0.04)	0.55 (+/- 0.04)
K-Nearest Neighbors	n_neighbors: 24 p: 1	0.53 (+/- 0.05)	0.38 (+/- 0.18)	0.53 (+/- 0.05)	0.40 (+/- 0.06)

Table 3.6: Topic classifiers performance

The obtained results of these classifiers are extremely inferior to our expectations. For this reason, topic classification can not be included in the tweet characterization system. At this moment, the main hypotheses explaining these poor results are the lack of a more defined criteria to label the tweets on each category, the small amount of data available and the fact that the dataset is imbalanced. Further analysis is needed to achieve better results and other classifiers might perform better.

3.3.1.5 Conclusion

Despite we accomplished great results on original fast-food dataset, we gathered a new dataset. We have been able to achieve good performance on hand-labeled fast-food dataset, where target tweets were labeled by hand and, thus, provides a more accurate representation of the reality. The best model is Random Forest, which achieved a 87% (+/-0.08) of accuracy, nearly followed by the Logistic Regression model with an accuracy of 85% (+/-0.02). Not as good as the previous models, Extra Tress model achieved notorious performance also with an accuracy of 77%(+/-0.05). However, these favorable results make contrast with the poor

performance of the topic classifier models. For this reason, the latter is not included on the system.

Therefore, the text classification system will combine the predictions of the best three models, presented above, in order to improve robustness over a single estimator, architecture known as Ensembler (depicted in Figure 3.7). Two different implementations are considered, both of them implemented the Voting Classifier provided by scikit-learn. First, Majority or Hard Voting, in which the predicted class label for a particular sample is the class label that represents the majority of the class labels predicted by each individual classifier. Second, Weighted Average Probabilities or Soft Voting, returns the class label as argmax of the sum of predicted probabilities, that is, the maximum of the averaged weighted probabilities of the models. This is interesting since specific weights can be assigned to each classifier. In our case, we have assigned a weight of 1.5 to the Random Forest and Logistic Regression classifiers; and a weight of 1 to the Extra Trees classifier. This way we give greater confidence to the prediction of the first two estimators.

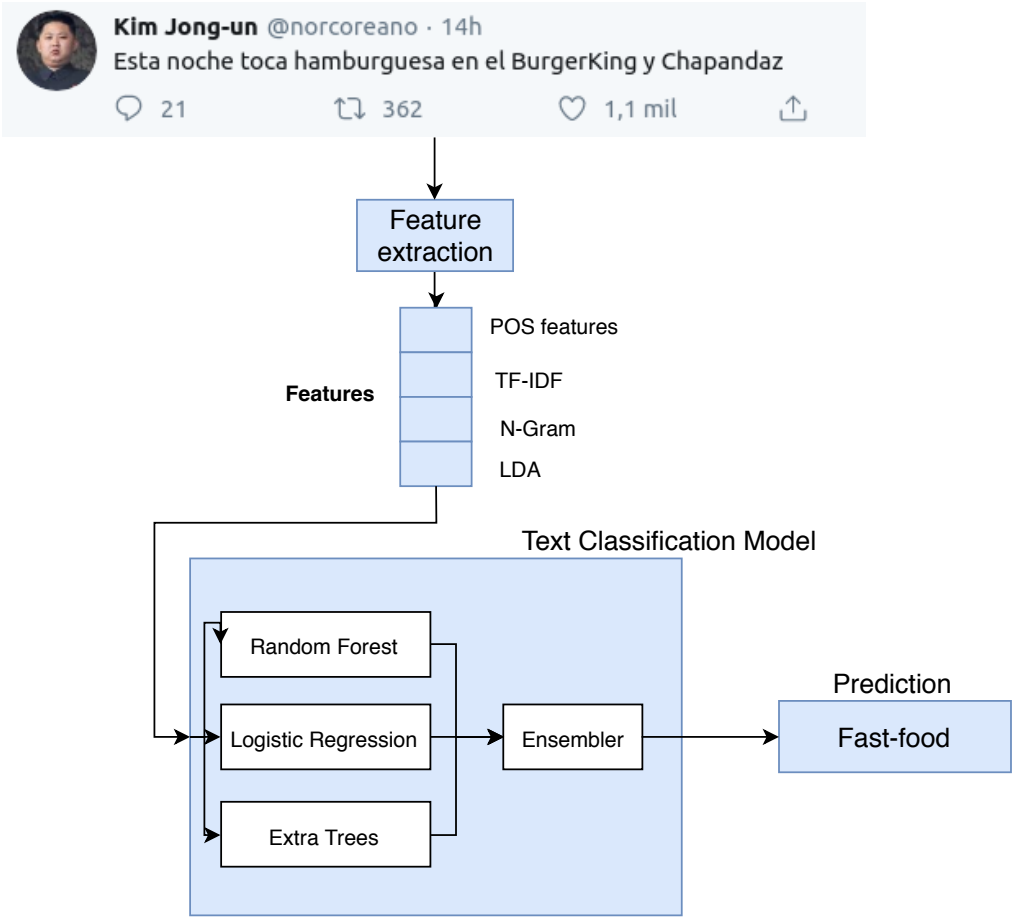

Figure 3.7: Classification with ensembler

Both ensemblers performance are presented in Table 3.7, where we can observe both ensemble methods have similar results. Thus, we choose the soft-voting ensemble arbitrarily.

Classifier	**Weights**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Hard Voting	-	0.85 (+/- 0.01)	0.85 (+/- 0.11)	0.87 (+/- 0.13)	0.86 (+/- 0.04)
Soft voting	RF: 1.5 LR: 1.5 ET: 1	0.86 (+/- 0.02)	0.85 (+/- 0.11)	0.87 (+/- 0.13)	0.86 (+/- 0.04)

Table 3.7: Ensemblers performance

3.3.2 Image analysis

Tweets are not only public messages of up to 280 characters, they also can contain polls, videos and images. In this section we present the work aimed to obtain further information from the tweets applying different computer vision techniques on the images contained in the tweets of our dataset. The main goals of this are to support the fast-food classifiers decision and to add extra information to better characterize fast-food related posts.

3.3.2.1 Image classification

Image classification is one of the tasks in computer vision with greater grow in the last years. The objective is classify an image according to its visual content. For example, an image classification algorithm may be designed to tell if an image contains a food or not, assuming that only one food is present in the image.

Traditional approaches on food recognition were based on classical image features and machine learning techniques, achieving low-mid accuracy results for datasets of 50-101 classes. In recent years, the grow of deep-learning (machine learning methods based on artificial neural networks, see Figure 3.8) have led to an enormous development of artificial intelligence field, including computer vision and, thus, image classification. Convolutional Neural Networks, a type of deep-learning architecture, have achieved the most significant technological advances for image classification and recognition tasks. In particular, works using this kind of networks have achieved accuracy above 79% for datasets of 101 classes [50].

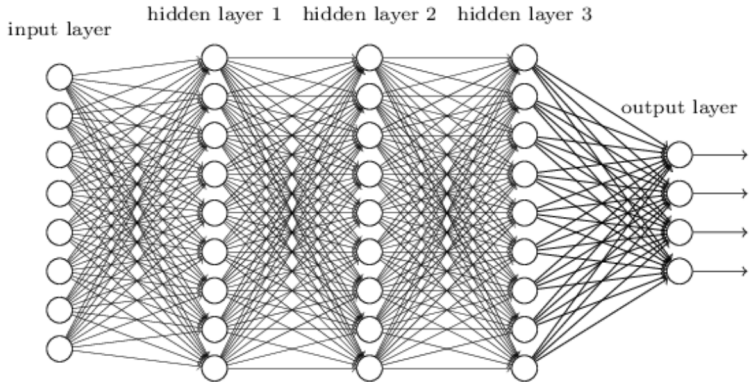

Figure 3.8: Neural network architecture

However, there is still no solution offering high accuracy. Food recognition is a difficult task due to the great variety and complexity of food, what makes it hard even for the human eye to recognize very similar foods. Also, the several different ways of food plating add an extra level of difficulty to the task. In addition, most of the times various food appear on the same plate.

Taking into account the limitations that this methods have nowadays, we have implemented an image classification algorithm to our fast-food characterization system following the work carried by Alberto Sanchez [51]. The main goal of this is to support the decision made by the fast-food classifier, this is, assess if the image associated to the tweet supports the conclusion of the text classifier in order to have a more robust system (Figure 3.9). Also, it could be useful to know if tweets not classified as fast-food related could have images of fast-food. We leave this premise as further work on the system.

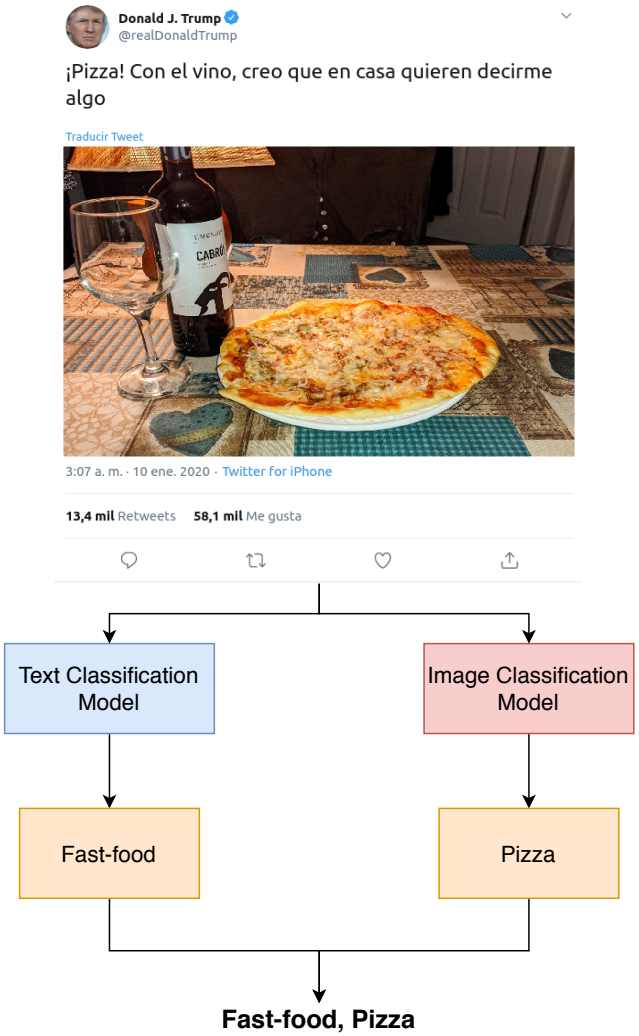

Figure 3.9: Image classification supports text classification

The image classification model we have developed is based on transfer learning with convolutional neural networks trained on object recognition. It has been constructed using Keras, a high-level API built on Tensorflow for deep-learning research. Model evaluation has been done using a popular food dataset, ETHZ Food-101 [52]. This dataset contains 101 food categories, with 101K images. For each class, 250 manually reviewed test images are provided as well as 750 training images.

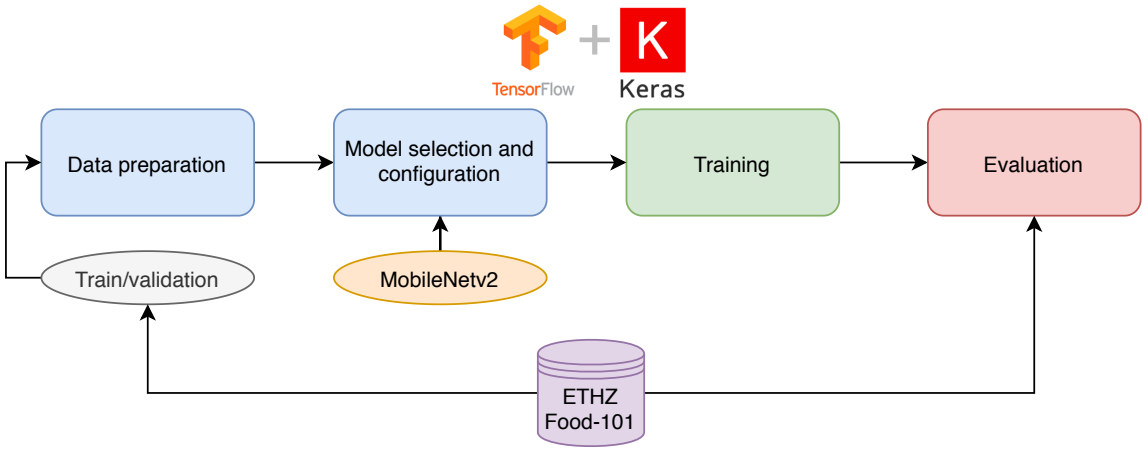

Figure 3.10: Model training process diagram

Creating a model from scratch for image classification can be very complicated due to the vast time and computing resources needed. For this reason, there are techniques to transfer the knowledge of pre-trained models to your own classification problem. Transfer learning is a computer vision method where a model developed for a task is reused as the starting point for a model on a second task, allowing to create accurate classification models in a short time. This requires the use of pre-trained models trained on a large dataset as the starting point to be able to solve a similar problem. In our case, the pre-trained model we have used is MobileNetV2 [53], loaded with the weights of the ImageNet dataset [54].

MobileNetV2 is a small model, with low latency and power, designed for mobile devices that improves the state of the art performance of mobile models on multiple tasks and benchmarks as well as across a spectrum of different model sizes. It is very useful to meet the computational restrictions of some use cases because of its lightness, allowing to obtain in little time trained models with good performance. MobileNetV2 with respect to its previous version is 30% faster obtaining the same results.

After selecting the architecture, the convolutional base of the model is loaded without including the part of the classifier to be able to adjust it to the desired requirements. Then

the model is trained on the food dataset to obtain a classification model capable of distinguish food of the 101 classes (Figure 3.11). Once trained the model, it can be used directly for getting a list of probabilities that belong to classes, and therefore, the food category to which the image belongs as the maximum probability category.

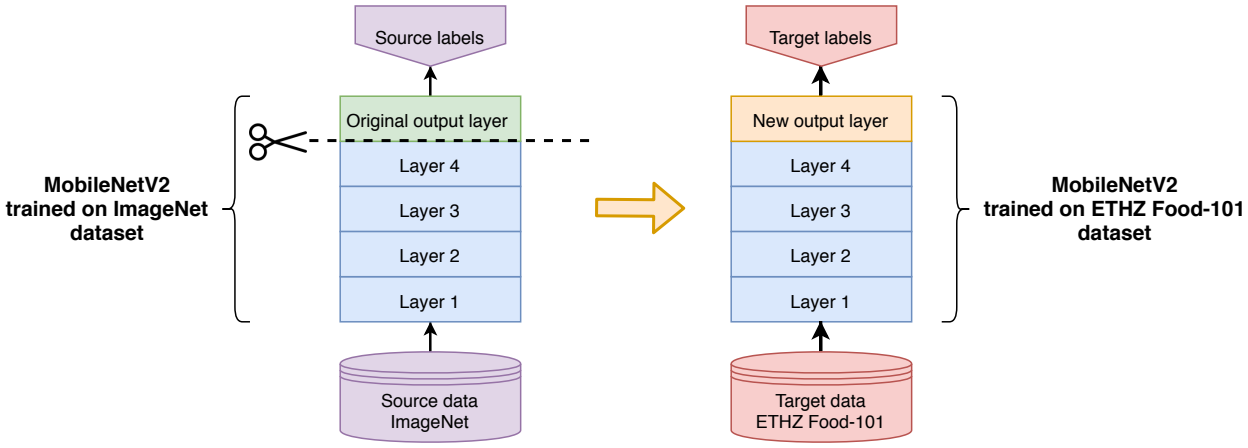

Figure 3.11: Transfer learning

The model evaluation metrics are presented in Table 3.8. To illustrate the model results, some images are shown from Figure 3.12 to Figure 3.15 along with the predicted label and its accuracy score.

Model	Accuracy	Precision	Recall	F1-score
MobileNetv2	0.817	0.820	0.820	0.820

Table 3.8: MobileNetV2 performance on ETHZ Food-101 dataset

Figure 3.12: Chocolate cake. Accuracy: 97.45%

Figure 3.13: Pizza. Accuracy: 99.35%

Figure 3.14: Sushi. Accuracy: 99.35%

Figure 3.15: Chessecake. Accuracy: 99.98%

Although it appears we achieved great results in this task, we have not include this module in the characterization system. Our image classification model works great to classify food between different classes, however, it does not work that great when a non-food image is predicted, predicting chocolate cake label for a person or pizza when a empty plate is shown. This limitation was among our expectations, considering we would need first a classification model able to distinguish which images correspond to food and which not. Anyway, this is a proof-of-concept and further work could be done to include this module in the system.

3.3.2.2 Image object detection

Object detection is a computer vision technique that allows us to identify and locate objects in an image. It can be used to count objects in a scene and determine and track their precise locations, all while accurately labeling them. It is widely used in computer vision tasks such as activity recognition, face detection, face recognition, video object co-segmentation and tracking objects.

As well as image classification, this is a classical machine-learning task which have obtained great result in the last years with the grow of deep-learning models. Nowadays, the

state-of-the-art models are RetinaNet [55] and YOLOv3 [56]. RetinaNet is a model developed by Facebook AI Research (FAIR) implemented with a recurrent convolutional neural network R-CNN, which requires multiple neural networks (Figure 3.16). On the other hand, YOLOv3 apply a single neural network to the full image, dividing the image into regions and predicting bounding boxes and probabilities for each region.

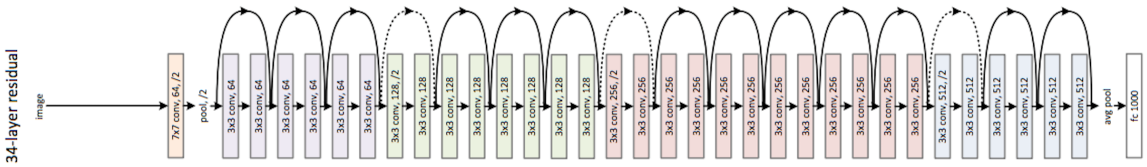

Figure 3.16: RetinaNet architecture

We have implemented the object detection algorithm using the Python library ImageAI, which provides both object recognition algorithms. While YOLOv3 has moderate performance and accuracy with a moderate detection time, RetinaNet has high performance and accuracy, but with longer detection time. We have chosen RetinaNet due to its better performance. The goal of performing object detection on the images of our dataset is to add extra information that is related to fast-food tweets such as which objects are more usually related to fast-food, as well as confirm which is food in the tweet (Figure 3.17.

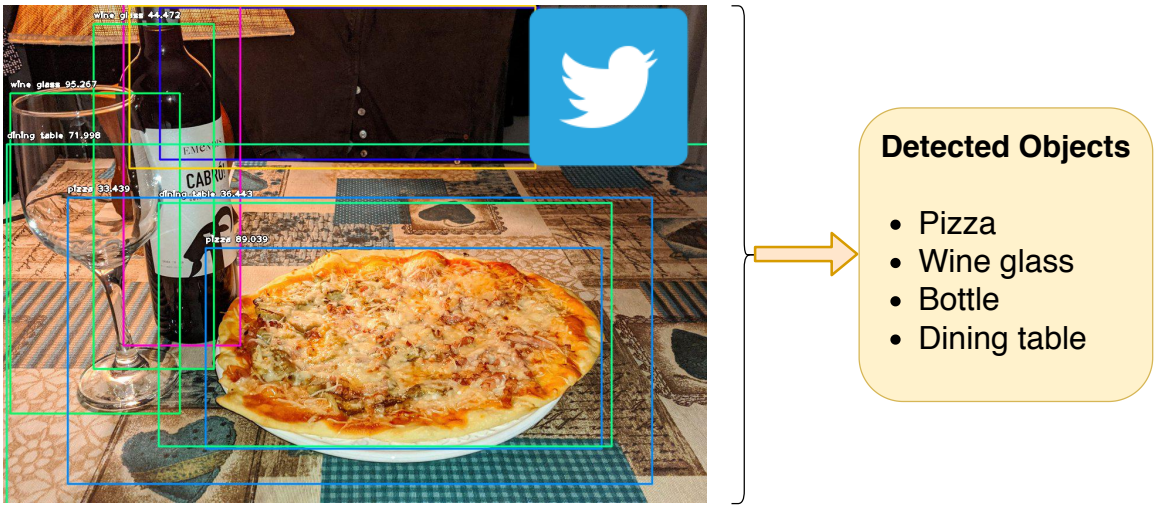

Figure 3.17: Object detection example

Using a few lines of code and without the complexity of setting up the neural network, ImageAI library allowed us to implement RetinaNet algorithm and detect up to 80 different kinds of objects, including different foods such as pizza or cake and food related objects like

bottle or cup. Some examples of the results are shown in next figures.

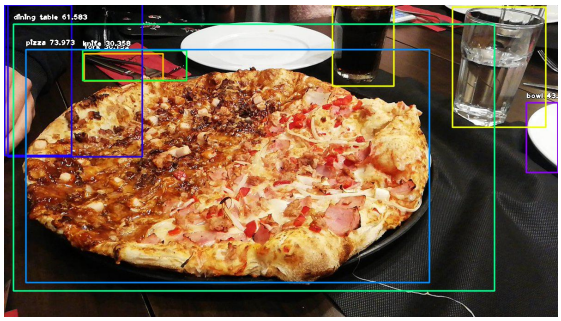

Figure 3.18: Pizza

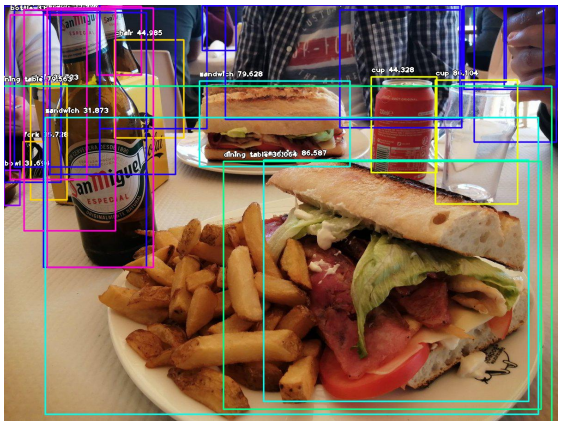

Figure 3.19: Sandwich

In order to better characterize what is shown in the images, only detections with 50% or higher accuracy have been taken into account. Using this kind of algorithms adds interesting extra information about the context of the tweets. Further work could be done training a custom algorithm to detect different kind of foods.

3.3.3 Sentiment and emotion analysis

This section describes the sentiment and emotion analysis module. Our objective here is to assess what kind of relation users from a region have with fast-food. To fulfill this goal, we have implemented this module using Senpy, a simple framework to build sentiment and emotion analysis services (Figure 3.3.3). Methods for both sentiment and emotion analysis are usually based on lexicons associated to each sentiment or emotion.

Figure 3.20: Sentiment and emotion analysis

Senpy allowed us to easily use Sentiment140 [57] algorithm. Sentiment140 perform automatic sentiment classification of tweets. These messages are classified as either positive, negative or neutral, allowing to discover the sentiment of a brand, product, or topic on Twitter.

In a similar way, we implemented DepecheMood [58] to assess the emotion associated to each tweet. It is a lexicon for emotion analysis obtained using unsupervised learning algorithms that allow us to associate different emotions to each tweets. For our purpose, we have associated to each tweet only the emotion with higher probability. The possible emotions associated to each tweet are: awe, amusement, anger, annoyance, indifference, joy, negative-fear and sadness.

CHAPTER 4

Case study: Fast-food in Spain

4.1 Introduction

The main objective of this chapter is to illustrate the use of our characterizing system and to test its capabilities to provide information in the Spanish case.

Thus, in this chapter we are going to describe the results of applying our system to the Spanish case. This description will cover the main system features, and its main purpose is to show the capabilities and potential of our system.

4.2 System use on one tweet

This section describes an example of how our system is applied on one tweet to extract information. The system is depicted in Figure 4.1. From the tweet, the text is analysed in order to evaluate if it is related to fast-food or not. As it is, sentiment and emotion analysis is performed, finding that the tweet is positive and it is associated with joy. Then, the image of the tweet is analysed: first, the classification model reports there is pizza in the image, supporting this way the fast-food label. This is corroborated by the object detection model,

which also identifies a pizza at the same time it detects a wine glass, a bottle and a dining table.

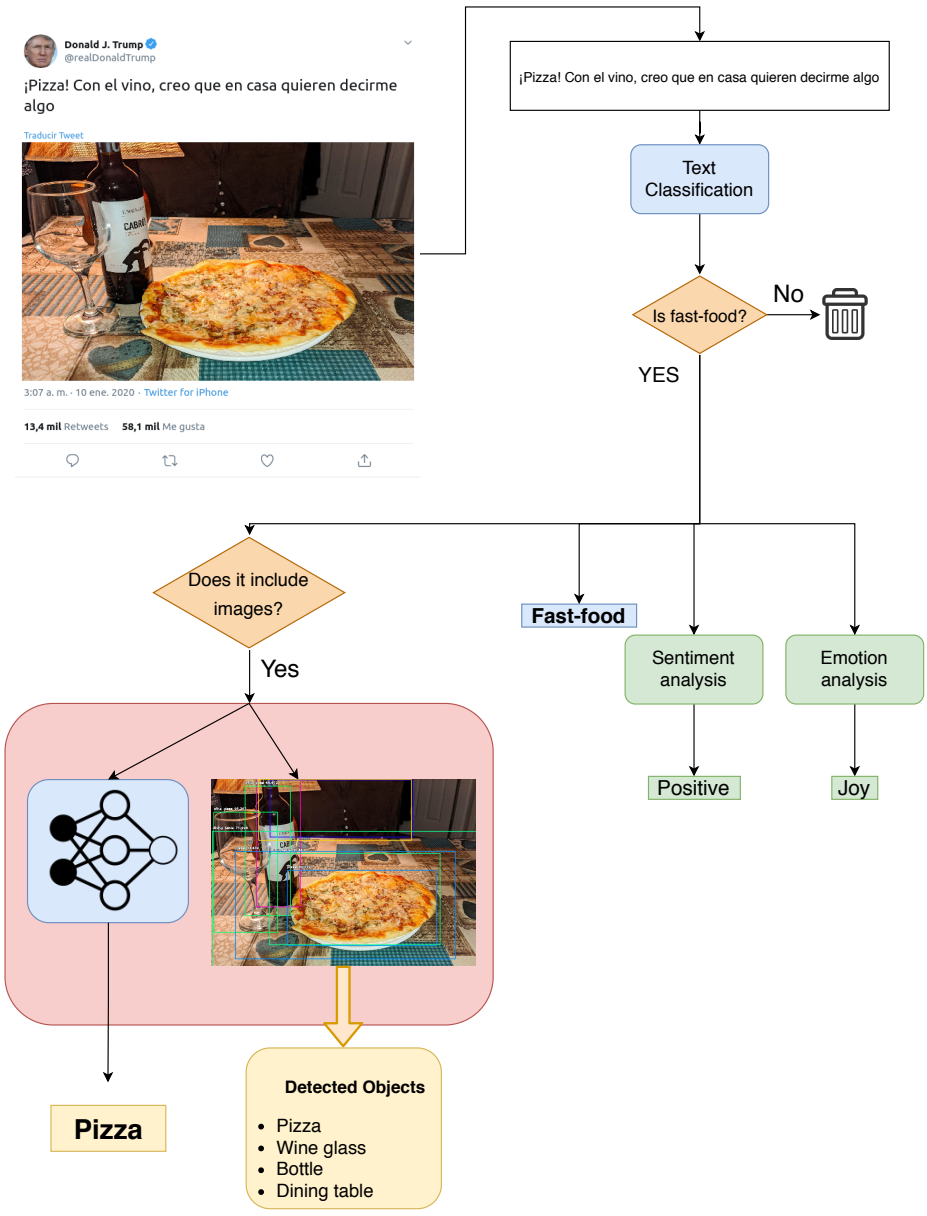

Figure 4.1: System use on one tweet

This way the system have allowed us to know the tweet is related to fast-food, specifically with pizza. Also, information from the image allowed us to extract information about the context, such as the person is in a dining table with wine. Finally, the tweet was associated with a positive sentiment and joy.

4.3 Spanish dataset

We focus the scope of our study on tweets from Spain, a total of 8431 tweets from 8087 different users. Since Twitter has a total of 4.9M users from Spain [59], we conclude we collected tweets from 0.16% of users in Spain.

Figure 3.2 shows a map with the number of tweets located per province. As it could be expected, the provinces with more population are also the ones with more tweets. This is the case of Madrid, Barcelona, Sevilla or Valencia.

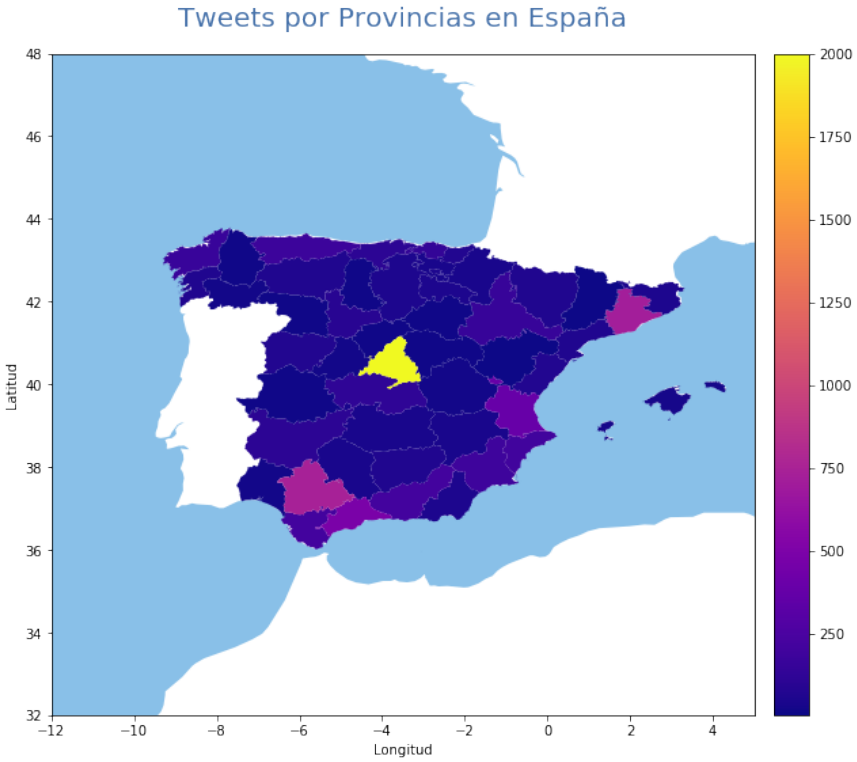

Figure 4.2: Map representing the tweets per province

4.4 Fast-food in Spain

As we want to find out the prevalence of tweets talking about fast food in the Spanish user community, we have used the text classification model described in Section 3.3.1 on the Spanish dataset. Out of the 8431 tweets, 5604 were labeled as fast-food related, i.e., a 66.5% of the dataset. This result supports again our previous decision of hand-labeling and creating dataset-02. We have determined the percentage of users who have written at least one tweet in the period of study talking about fast food as the number of users who wrote

tweets labeled as fast-food and the total number of users in Spain: a 0.11% of users in Spain have ever written about fast food.

Also, we can find the most common words related to fast-food, an useful indicator to know how fast-food is talked about and what feeling is more common when talking about it. Most common words are shown in Figure 4.3, where words with greater importance are bigger. We can observe the most common foods, as well as other words that could be considered as positive relation indicators such as “quiero“ (I want) and “sí“ (yes).

Figure 4.3: Fast-food wordcloud

In addition, we can compute the number of times each term in our fast-food keywords dictionary appear in the Spanish dataset. This is shown in Figure 4.4. We observe the most common keywords are “cerveza“ (beer), “pizza“ (pizza) and “tarta“(cake). The first brand that appears is Coca Cola, far more common than the rest of brands.

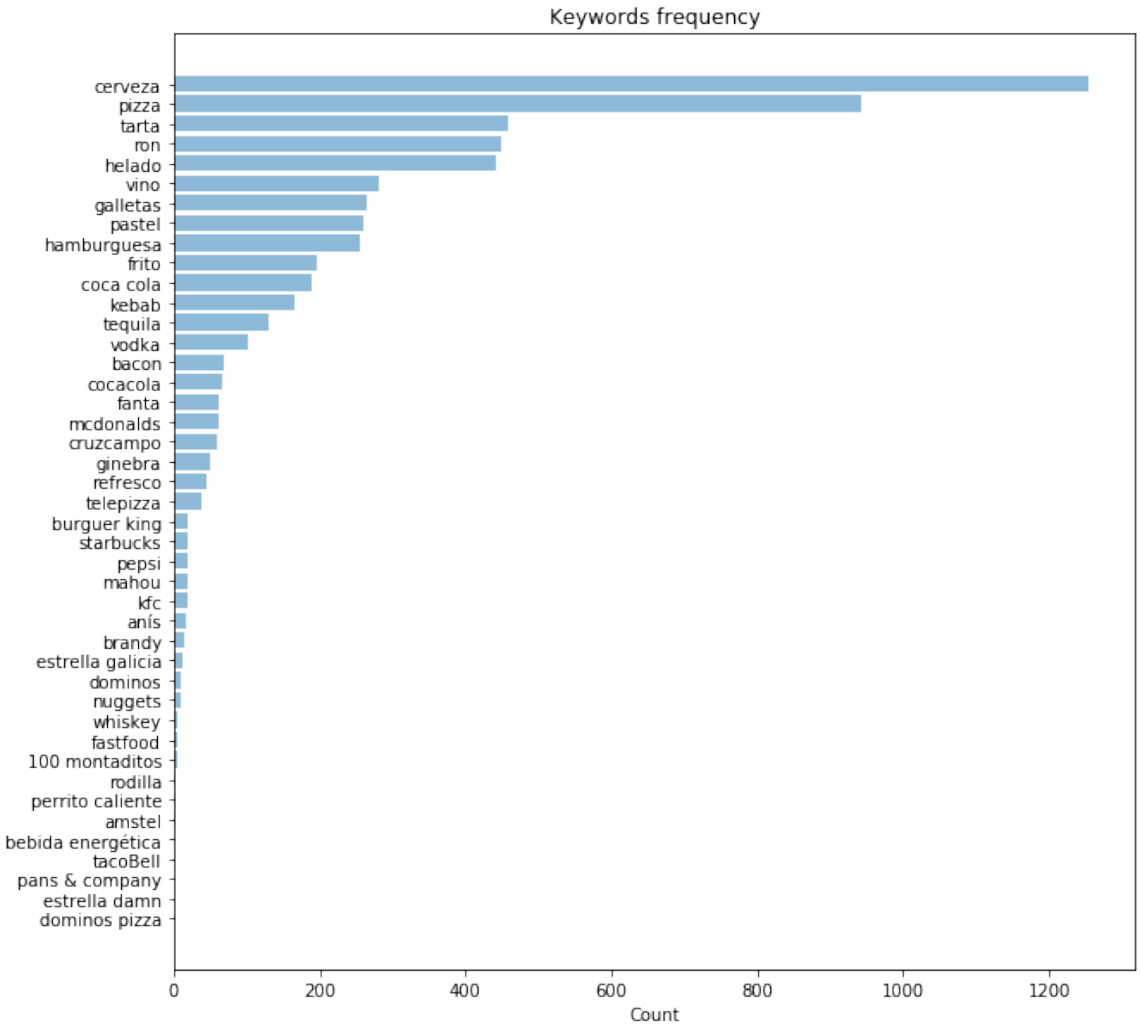

Figure 4.4: Fast-food related keywords in Spain

4.5 Image object detection

We use the image object detection algorithm to extract further information about the context from the images attached to tweets. From the 5604 tweets labeled as fast-food by our model, only 235 had images attached. This fact makes us see this module as a proof-of-concept instead of a valid representation of the reality since we do not have a representative sample.

We computed object detection on the images attached to the 235 tweets and related them to the location of the tweets. The most common detections are shown in Figure 4.5, where we observe fast-food related tweets are more commonly associated with these things: person, bottle, dining table, cup, pizza, sandwich, cake and wine glass. It is important to note that other objects with less detections have been discarded due to they might be errors

(i.e., mouse or skate-board detections).

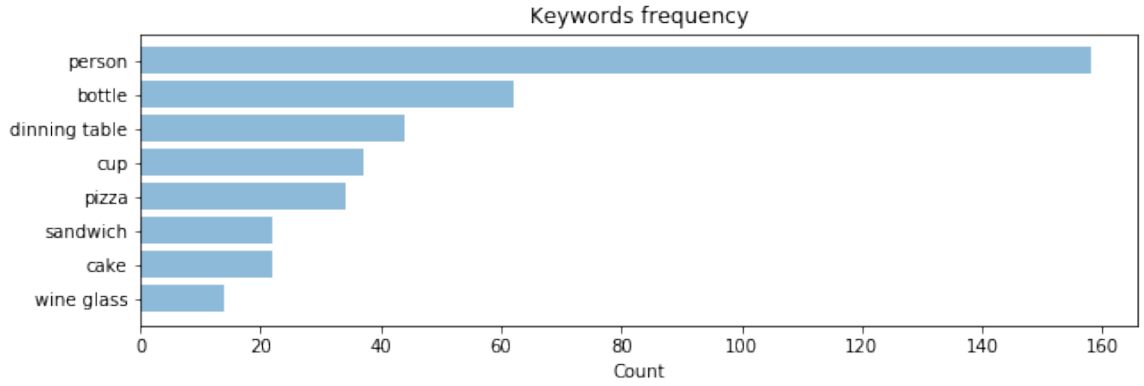

Figure 4.5: Top detections in fast-food tweets

4.6 Sentiment and Emotion analysis

The goal of this kind of analysis is to obtain province level indicators of sentiment and emotion related to fast-food. From the Spanish dataset of 5604 tweets, the 71.9% has been labeled as neutral (4033 tweets). The 24.8% is considered Positive (1392 tweets) and the last 3.2% Negative (179 tweets). This is consistent with studies that reveal fast-food is commonly associated with positive sentiments [60].

In a similar way we can perform emotion analysis. The emotions related to the Spanish dataset are shown in Table 4.1. We observe most of the tweets are related to positive emotions such as awe or amusement, but a significant number is related to negative ones such as anger or annoyance.

Emotion	**Count**	**Emotion**	**Count**
awe	2062	indifference	253
amusement	1409	joy	147
anger	786	negative-fear	65
annoyance	692	sadness	54

Table 4.1: Emotions related to fast-food tweets in Spain

CHAPTER 5

Conclusions and future work

In this chapter we will describe the conclusions extracted from this project, and the thoughts about future work, as well as a description of the achieved goals.

5.1 Conclusions

In this study we have provided a methodology to characterize fast-food patterns from province-level Twitter data. First, we found that the Spanish-speaking countries with higher proportion of fast-food related tweets are Argentina, Mexico and Spain. In the case of the Spanish Twitter user community, we have determined that 0.11% of users in Spain have written at least one tweet related to fast-food during the dataset capture. Second, the hand-labelling process has demonstrated to offer more precise results than automatic labelling based on keywords. However, it is difficult to obtain a big dataset labelling by hand the samples. In addition, using image object detection and classification models have proven to be new ways to support the decision of the classifiers and add further information about the context of the tweets. Although this models have shown very promising results, further work is needed to fully exploit their potential. Moreover, we would like to highlight the results of the classifiers we have developed. This results have been obtained using a hand-

labeled dataset with not a large size, achieving good performance without the need of large amounts of data. The best classifying models are Random Forest, Logistic Regression and Extra Tress; and we have join all the three models in an ensembler to achieve more robust results. Although the theme classifier have not performed good enough due to the extra difficulty of a multi-class classifier, there is room for improvement using larger datasets and new methodologies.

However, this study is subject to several limitations. For a given province, tweets sent from that province were utilized to characterize fast-food. However, tweets could be sent by both residents and visitors. Additionally, users of social media tend to be younger than the general population; in 2015, 51% of individuals aged 18–34 years old used Twitter compared to 18% of individuals 45–54 years and 5% among those 55+years [61]. In addition, using social media data, like other data relies upon people's willingness to report. The content of tweets reflects information that people feel comfortable reporting and may not represent the true spectrum of their feelings or their experiences. Besides, some insights could be obtained looking at our dataset. Tweets from Madrid are the 26% of the total and tweets from Andalucía the 23%. This shows a biased dataset, which does not represents good enough the reality in Spain. To fully capitalize on emerging big data sources, further assessment of potential biases and continued development new methods are needed.

Overall, we think this is an innovative work that demonstrates the utility and cost-effectiveness of existing big data sources not produced for the purpose of this research. Despite we need a much larger dataset to better test our system capabilities, it is a very promising tool to evaluate health related issues in Spain and further work will be carried out in this line.

5.2 Future work

This section describes the possible new features or improvements that could be done in this project.

- **Increase the dataset**. Increase the size of the dataset to achieve more realistic results.
- **Real-time streaming analysis**. Development of a analysis system to characterize real-time fast-food related tweets and find patterns and trends in different time frames.
- **Food dictionary improvement**. Find new methodologies to create a food dictionary including all type of foods to be able to analyse whole nutrition patterns.

- **Geolocation of tweets improvement**. Use location field of users as the location of the tweet to increase the dataset of located tweets.
- **Custom food detection model**. Create a custom deep-learning model to detect the foods we are interested about.
- **City level analysis**. Perform the fast-food characterization at a city level analysis. For this purpose a huge amount of data is needed.
- **Test text classification deep-learning models**. Evaluate classification models based on neural networks and compare the results with the used machine learning models.
- **New social media sources**. Discover the capabilities of our system in new social media sources such as Instagram or food journaling social networks.

APPENDIX A

Project impact

This appendix reflects, quantitatively or qualitatively, on the possible social, economic and environmental impact jointly with ethical implications.

A.1 Social impact

This section describes the main social impacts our work might have. As a consequence of the raise of social media in recent years, a huge amount of public-available data is daily shared by people. Projects as this one aim to analyze and exploit this data for different purposes.

In our case, the study and the tool developed on fast-food will allow health researchers to carry out nutrition-related studies much more easily compared to traditional approaches, since it benefits from public available big data. Also, this system could be considered as a starting point for further work on the subject.

Studies using our system would be carried much more faster, giving useful insights to address different health issues related with fast-food consumption. This would be very useful to apply health policies that may benefit the overall of the population.

A.2 Economic impact

In this section we summarize the main economic impacts of our project in public institutions and, thus, all the population.

The possibility of carrying out health-related studies using social media mining, directly implies less duration and cost compared to traditional approaches. This studies are usually carried out by public institutions, so our system can potentially reduce the health costs in a country not only by reducing the time and cost of the studies, but by the insights obtained from our system than can reduce health issues and, thus, reduce the economic cost of health institutions.

A.3 Environmental impact

This section briefly describes the main environmental impacts of our project.

All systems based on machine learning and big data have an important ecological footprint due to the energy needed for, firstly, producing the huge amount of data these systems need and, secondly, training the machine learning models, which may last even months.

In particular, we have to mention the energy consumption of the computer and big data cluster needed for this project.

A.4 Ethical implications

Finally, this section depicts the main ethical considerations.

The main ethic issue that may concern us is related to privacy. Although the project has been developed using publicly available data, most of the population is not concerned about what really implies share their data publicly. However, people have agreed the use of their data for different purposes, including projects like this. Our conclusion in that more education about the value of privacy is needed to be able to not consider this issue.

APPENDIX B

Project budget

This appendix details an adequate budget to bring about the project. The project structure is described along with the activities undertaken to complete it and costs are evaluated including material and human resources, as well as taxes.

B.1 Project structure

In order of achieving the proposed goals, the project have been divided in the activities shown in Table B.1, where details about its duration and dependencies are provided. All activities require a person with good programming and machine learning background, a telecommunication engineer for example. As the project has been carried by only one person, effort of each task is not considered since it is directly related to the duration of the activity. **The total duration of the project has been 524 hours**, taking into account that one day corresponds to 4 working hours.

Activity	**Description**	**Dependencies with other tasks**	**Duration (days*)**
1.- Python and Machine Learning course overview	The fundaments of Python programming language and the main Machine Learning concepts are refreshed in order to have a basis good enough to carry out the project	-	10
2.- State of art research	Research about works related to the goal of our project	-	15
3.- Twitter API	Get to know how Twitter API works and create the module to capture tweets. Start capturing.	1	5
4.- Data munging and cleaning	Prepare the data to be able to use it easily.	1,3	10
5.- Data preprocessing	Prepare the data to feed the classifiers with it	4	14
6.- Dataset annotation	Label dataset by hand. Prepare annotated datasets.	4	2
7.- Text classification	Prepare the machine learning algorithm to perform text classification. Models evaluation. Ensemblers.	5,6	25
8.- Image object detection	Research about image object detection algorithms. Implementation of Object Detection using ImageAI library.	4	7
9.- Image classification	Research about image classification algorithms. Implementation of selected model based on Transfer learning.	4	5
10.- Sentiment analysis	Research and implementation of Senpy	4	3
11.- Case study	Use of the developed system on a specific case	7,8,9,10	5
12.- Report writing	Writing of the TFG	2,11	30

Table B.1: Project structure division by activity.

1

B.2 Costs evaluation

This section summarizes the material resources needed to develop the project, divided in software and hardware resources, as well as the human resources and taxes associated to it.

B.2.1 Material resources

B.2.1.1 Software

All the project has been developed using open-source software that is available on Internet for free. For this reason, there is no cost associated to software.

B.2.1.2 Hardware

To carry out this project a computer and a big data cluster, both provided by the Intelligent Systems Group, has been used.

The specifications of the computer are:

- Intel Core i5 CPU of 3.2GHzx4
- 8 GB of RAM
- Hard disk of 500 GB

On the other hand, the big data cluster specifications are:

- DELL PowerEdge R320
- Intel® Xeon® E5-2430 v2
- 4x32GB RDIMM
- 3x3TB, SATA

A computer with this specifications costs 700€ approximately, while the approximate cost of the big data cluster is 2600€. Thus, **the hardware cost is 3300€**. Although it should have been taken into account, amortization has not been computed so the final hardware cost should be lower or even 0€ if we consider the hardware have already been amortized.

B.2.2 Human resources

We have needed one telecommunication engineer to accomplish this project. The salary have been based on the GSI Research scholarship, through which this project has been carried out, stipulated in 450€/month for 80 working hours (20 working hours per week). Thus, the fee per hour would be 5,625€. Since the project duration has been estimated in 524 hours, the total fee for the project ascend to 2947,5€. We have multiplied this number by 1.3 to approximately include Social Security, thus, **the total cost of human resources is around 3850€**.

B.2.3 Taxes

In case the final product is sold to an interested company, taxes related to a software engineering project must be taken into account. The fees paid by the company would be the corresponding VAT established in the local country

B.3 Conclusion

The project had a **duration of 524** hours and the **total cost ascends to 7150€**.

Bibliography

[1] Obesity and overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. (Accessed on 01/3/2020).

[2] Aviva Must, Jennifer Spadano, Eugenie H. Coakley, Alison E. Field, Graham Colditz, and William H. Dietz. The Disease Burden Associated With Overweight and Obesity. *JAMA*, 282(16):1523–1529, October 1999.

[3] P. Kopelman. Health risks associated with overweight and obesity. *Obesity Reviews*, 8(s1):13–17, 2007.

[4] Juan Oliva, Laura González, José M. Labeaga, and Carlos Álvarez Dardet. Salud pública, economía y obesidad: el bueno, el feo y el malo. *Gaceta Sanitaria*, 22:507–510, December 2008.

[5] Martha L. Daviglus, Kiang Liu, Lijing L. Yan, Amber Pirzada, Larry Manheim, Willard Manning, Daniel B. Garside, Renwei Wang, Alan R. Dyer, Philip Greenland, and Jeremiah Stamler. Relation of body mass index in young adulthood and middle age to Medicare expenditures in older age. *JAMA*, 292(22):2743–2749, December 2004.

[6] Robert W. Jeffery, Judy Baxter, Maureen McGuire, and Jennifer Linde. Are fast food restaurants an environmental risk factor for obesity? *International Journal of Behavioral Nutrition and Physical Activity*, 3(1):2, January 2006.

[7] S. A. French, M. Story, D. Neumark-Sztainer, J. A. Fulkerson, and P. Hannan. Fast food restaurant use among adolescents: associations with nutrient intake, food choices and behavioral and psychosocial variables. *International Journal of Obesity*, 25(12):1823–1833, December 2001.

[8] Jay Maddock. The Relationship between Obesity and the Prevalence of Fast Food Restaurants: State-Level Analysis. *American Journal of Health Promotion*, 19(2):137–143, November 2004.

[9] Janet Currie, Stefano DellaVigna, Enrico Moretti, and Vikram Pathania. The Effect of Fast Food Restaurants on Obesity and Weight Gain. *American Economic Journal: Economic Policy*, 2(3):32–63, August 2010.

[10] R. Rosenheck. Fast food consumption and increased caloric intake: a systematic review of a trajectory towards weight gain and obesity risk. *Obesity Reviews*, 9(6):535–547, 2008.

[11] Gunther Eysenbach. Infodemiology and Infoveillance: Tracking Online Health Information and Cyberbehavior for Public Health. *American Journal of Preventive Medicine*, 40(5, Supplement 2):S154–S158, May 2011.

[12] Twitter for Business | Twitter tips, tools, and best practices. https://business.twitter.com/en.html. (Accessed on 01/3/2020).

[13] Food4health | EIT Food. https://www.eitfood.eu/crosskic/projects/food4health. (Accessed on 01/10/2020).

[14] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1568–1576, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. event-place: Edinburgh, United Kingdom.

[15] Aron Culotta. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language Resources and Evaluation*, 47(1):217–238, March 2013.

[16] Daniel Suarez, Oscar Araque, and Carlos A. Iglesias. How Well Do Spaniards Sleep? Analysis of Sleep Disorders Based on Twitter Mining. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 11–18, October 2018.

[17] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. Recognizing Depression from Twitter Activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3187–3196, New York, NY, USA, 2015. ACM. event-place: Seoul, Republic of Korea.

[18] Palakorn Achananuparp, Ee-Peng Lim, and Vibhanshu Abhishek. Does Journaling Encourage Healthier Choices?: Analyzing Healthy Eating Behaviors of Food Journalers. In *Proceedings of the 2018 International Conference on Digital Health*, DH '18, pages 35–44, New York, NY, USA, 2018. ACM. event-place: Lyon, France.

[19] Predicting individual well-being through the language of social media | Biocomputing 2016. https://www.worldscientific.com/doi/abs/10.1142/9789814749411_0047. (Accessed on 01/3/2020).

[20] Yelena Mejova, Ingmar Weber, and Michael W. Macy. *Twitter: A Digital Socioscope*. Cambridge University Press, May 2015.

[21] Chi Y. Bahk, Melissa Cumming, Louisa Paushter, Lawrence C. Madoff, Angus Thomson, and John S. Brownstein. Publicly Available Online Tool Facilitates Real-Time Monitoring Of Vaccine Conversations And Sentiments. *Health Affairs (Project Hope)*, 35(2):341–347, February 2016.

[22] Amir Karami, Alicia A. Dahl, Gabrielle Turner-McGrievy, Hadi Kharrazi, and George Shaw. Characterizing diabetes, diet, exercise, and obesity comments on Twitter. *International Journal of Information Management*, 38(1):1–6, February 2018.

[23] Aron Culotta. Estimating County Health Statistics with Twitter. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1335–1344, New York, NY, USA, 2014. ACM. event-place: Toronto, Ontario, Canada.

[24] Keiji Yanai and Yoshiyuki Kawano. Twitter Food Photo Mining and Analysis for One Hundred Kinds of Foods. In Wei Tsang Ooi, Cees G. M. Snoek, Hung Khoon Tan, Chin-Kuan Ho, Benoit Huet, and Chong-Wah Ngo, editors, *Advances in Multimedia Information Processing – PCM 2014*, Lecture Notes in Computer Science, pages 22–32. Springer International Publishing, 2014.

[25] Vijaya Kumari Yeruva, Sidrah Junaid, and Yugyung Lee. Exploring social contextual influences on healthy eating using big data analytics. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1507–1514, November 2017.

[26] 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, August 2016.

[27] 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, and Ken R. Smith. Twitter-derived neighborhood characteristics associated with obesity and diabetes. *Scientific Reports*, 7, November 2017.

[28] 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, October 2014.

[29] 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*, CHI '15, pages 3197–3206, New York, NY, USA, 2015. ACM. event-place: Seoul, Republic of Korea.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] Moses and John Olafenwa. Imageai, an open source python library built to empower developers to build applications and systems with self-contained computer vision capabilities, mar 2018–.

[32] Opencv. `https://opencv.org/`. (Accessed on 12/29/2019).

[33] COCO - Common Objects in Context. `http://cocodataset.org/#home`. (Accessed on 01/3/2020).

[34] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[35] Home - keras documentation. `https://keras.io/#you-have-just-found-keras`. (Accessed on 12/29/2019).

[36] Tensorflow. `https://www.tensorflow.org/`. (Accessed on 12/29/2019).

[37] Python data analysis library — pandas: Python data analysis library. https://pandas.pydata.org/. (Accessed on 12/29/2019).

[38] Geopandas 0.6.0 — geopandas 0.6.0 documentation. http://geopandas.org/. (Accessed on 12/29/2019).

[39] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[40] Github - residentmario/geoplot: High-level geospatial data visualization library for python. https://github.com/ResidentMario/geoplot. (Accessed on 12/29/2019).

[41] Natural language toolkit — nltk 3.4.5 documentation. https://www.nltk.org/. (Accessed on 12/29/2019).

[42] Github - gsi-upm/gsitk. https://github.com/gsi-upm/gsitk. (Accessed on 12/29/2019).

[43] Stanford topic modeling toolbox. https://nlp.stanford.edu/software/tmt/tmt-0.4/. (Accessed on 12/29/2019).

[44] Github - clips/topbox: Python 2 & 3 wrapper around the stanford topic modeling toolbox. intended to be used for hassle-free supervised topic classification with labeled latent dirichlet allocation (l-lda, llda, slda). https://github.com/clips/topbox. (Accessed on 12/29/2019).

[45] Developer. https://developer.twitter.com/. (Accessed on 12/29/2019).

[46] Tweepy. https://www.tweepy.org/. (Accessed on 12/29/2019).

[47] Iglesias C. A. Corcuera I. & Araque Ó. Sánchez-Rada, J. F. Senpy: A pragmatic linked sentiment analysis framework. https://senpy.readthedocs.io/en/latest/index.html, October 2016. (Accessed on 12/30/2019).

[48] Estudios Fintonic: Restauración en España 2018. https://www.fintonic.com/blog/estudios-fintonic-restauracion-en-espana-2018/, April 2018. (Accessed on 01/3/2020).

[49] Federación Española de Bebidas Espirituosas. http://www.febe.es/El-sector-en-cifras/consumo-bebidas-espirituosas/. (Accessed on 01/3/2020).

[50] Ashutosh Singla, Lin Yuan, and Touradj Ebrahimi. Food/non-food image classification and food categorization using pre-trained googlenet model. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, MADiMa '16, page 3–11, New York, NY, USA, 2016. Association for Computing Machinery.

[51] Github - sanxlop/tfm_etsit: Development of a food image classification system based on transfer learning with convolutional neural networks. https://github.com/sanxlop/tfm_etsit. (Accessed on 01/02/2020).

[52] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

[53] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[55] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[56] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[57] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *Processing*, 150, 01 2009.

[58] Oscar Araque, Lorenzo Gatti, Jacopo Staiano, and Marco Guerini. Depechemood++: a bilingual emotion lexicon built through simple yet powerful techniques. *arXiv preprint arXiv:1810.03660*, 2018.

[59] V Estudio sobre los usuarios de Facebook, Twitter e Instagram en España. Technical Report V, 2019.

[60] Natalie Dixon, Bruno Jakic, Roderick Lagerweij, Mark Mooij, and Ekaterina Yudin. FoodMood: Measuring Global Food Sentiment One Tweet at a Time. In *Sixth International AAAI Conference on Weblogs and Social Media*, May 2012.

[61] Estadísticas de los usuarios de twitter ¿cómo son y se comportan? `https://www.iebschool.com/blog/estadisticas-usuarios-twitter-como-son-redes-sociales/`. (Accessed on 01/10/2020).

[62] The disease burden associated with overweight and obesity | obesity | jama | jama network. `https://jamanetwork.com/journals/jama/fullarticle/192030`. (Accessed on 11/05/2019).

[63] Sahasporn Paeratakul, Daphne P. Ferdinand, Catherine M. Champagne, Donna H. Ryan, and George A. Bray. Fast-food consumption among US adults and children: Dietary and nutrient intake profile. *Journal of the American Dietetic Association*, 103(10):1332–1338, October 2003.

[64] Joanne L. Slavin and Beate Lloyd. Health Benefits of Fruits and Vegetables. *Advances in Nutrition*, 3(4):506–516, July 2012.

[65] WHO | Promoting fruit and vegetable consumption around the world. `https://www.who.int/dietphysicalactivity/fruit/en/`. (Accessed on 01/3/2020).

[66] Shapely — shapely 1.7a2 documentation. `https://shapely.readthedocs.io/en/latest/`. (Accessed on 12/29/2019).

[67] Genderize.io | determine the gender of a name. `https://genderize.io/`. (Accessed on 12/30/2019).

[68] Inebase / sociedad /salud /encuesta nacional de salud / resultados. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176783&menu=resultados&idp=1254735573175. (Accessed on 01/03/2020).