

UNIVERSIDAD POLITÉCNICA DE MADRID

**ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN**



**GRADO EN INGENIERÍA DE TECNOLOGÍAS Y
SERVICIOS DE TELECOMUNICACIÓN**

TRABAJO FIN DE GRADO

SARCASM DETECTION ON TWITTER

JUAN ÁLVAREZ FERNÁNDEZ DEL VALLADO
2019

TRABAJO DE FIN DE GRADO

Título: Detección de Sarcasmo en Twitter
Título (inglés): SARCASM DETECTION ON TWITTER
Autor: Juan Álvarez Fernández del Vallado
Tutor: Carlos Angel 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



TRABAJO DE FIN DE GRADO

SARCASM DETECTION ON TWITTER

Enero 2019

Resumen

La aparición de la red social Twitter en el año 2006 llevó consigo un tremendo empujón a la Revolución Digital que ya se estaba viviendo desde finales del siglo 20. Como resultado, a día de hoy se twitteen unos 6000 tweets cada segundo. Se estima que para el año 2020 la red social Twitter contará con 275 millones de usuarios en todo el mundo. Tal es la popularidad de Twitter que casi el 74% de sus usuarios usan Twitter a diario para consultar las noticias del día a día. Se podría decir que la Revolución Digital ha convertido a Twitter en una parte de nuestras vidas. Es por estas razones que este proyecto tiene como objetivo desarrollar un clasificador con la habilidad de detectar el sarcasmo en textos cortos, como en tweets.

El propósito de este proyecto es conseguir la detección automatizada de sarcasmo. El principal desafío a tener en cuenta es la detección de emociones en frases que contienen sentimiento. Para abordar este enorme reto, se ha diseñado un clasificador que posee tecnologías de Machine Learning, el cuál podrá aprender a detectar sarcasmo. También han sido necesarias herramientas capaces de procesar lenguaje natural: Natural Language Processing. Con el fin de obtener unos parámetros adecuados en el clasificador se han llevado a cabo pruebas para ver que características aportan información relevante sobre el sarcasmo al clasificador. Las mas notables fueron los ngramas, las características léxicas y Latent Dirichlet Allocator (LDA). Además, se ha llevado a cabo un Grid Search con el propósito de buscar los parámetros óptimos del clasificador. Esto ha consistido en obtener el F1 score para cada permutación de una tabla de parámetros a probar. Una vez se obtuvieron los mejores parámetros, la plataforma senpy fue usada para dar una interfaz intuitiva en la que un usuario puede introducir texto y el clasificador indica si el texto tiene sarcasmo.

Para terminar, los resultados obtenidos han oscilado, presentando un mínimo F1 score de 90.22% y un máximo de 96.68%.

Palabras clave: Sarcasmo, Twitter, Aprendizaje Automático, Big Data, Python, NLP, Detección de Sarcasmo, Sentimientos, Emociones y Análisis.

Abstract

The emergence of the social network Twitter in 2006 resulted in a tremendous boost of the Digital Revolution taking place since the end of the 20th century. As a result, there are on average at least 6000 tweets being tweeted every second. It is estimated that by 2020 the number of users on Twitter will increase up to 275 million users worldwide. Twitter has become so popular that roughly 74% of all the Twitters' users claim to use the social network as a source to get the daily news. One could say that Twitter has become part of our lives and such a statement would not be far from the truth. For these reasons, this project is committed to developing a classifier with the ability to detect sarcasm in short texts, such as Twitter.

Automatic sarcasm detection is the task of recognizing sarcastic patterns in a text. Considering the challenges of detecting emotions in a sentiment-bearing text, the approach followed was the design of a Machine Learning classifier, with the ability to perform a classification between sarcastic and non-sarcastic tweets. Natural Language Processing tools played a big role in the task. Moreover, feature extraction facilitated the learning and generalization steps. To ensure optimum Machine Learning parameters and provide the best score, several optimization tests were conducted to detect those features which give more information to the classifier. The most noteworthy features to mention were: Latent Dirichlet Allocator (LDA), ngrams and lexical features. Furthermore, a cross validation Grid Search searched over specified parameters to optimize the classifier and get the best score. These consisted of evaluating the F1 score by looping over a set of parameters. Once the classifier was optimized, the senpy engine was used to provide a framework where a user could insert a short text and the classifier could detect whether or not the text is sarcastic.

To conclude, the scores obtained varied between a minimum F1 score of 90.22% and a maximum F1 score of 96.68%.

Keywords: Sarcasm, Twitter, Machine Learning, Big Data, Python, NLP, Sarcasm Detection, Sentiments, Emotions and Analysis.

Agradecimientos

A mi familia por apoyarme en mis decisiones, a la Universidad Politécnica de Madrid por ofrecerme la oportunidad de hacer un Trabajo de Fin de Grado, al Grupo de Sistemas Inteligentes de la UPM por asistirme todas aquellas veces que me dirigí a ellos y a mi tutor, Carlos Iglesias.

Contents

Resumen	VII
Abstract	IX
Agradecimientos	XI
Contents	XIII
List of Figures	XVII
1 Introduction	1
1.1 Context	1
1.2 Project goals	2
1.3 Structure of this document	2
2 State of the Art	5
2.1 Introduction	5
2.2 Sarcasm	5
2.2.1 Sarcasm Definition	5
2.3 Dataset	6
2.4 Approaches used for Sarcasm Detection	7
2.4.1 Rule-based Approaches	7
2.4.2 Feature sets	7
2.5 Main Issues in Sarcasm Detection	8
2.6 Figures of Merit	9
3 Enabling Technologies	13
3.1 Introduction	13
3.2 Natural Language Processing	13
3.2.1 Natural Language Toolkit	14
3.2.1.1 Definition	14
3.2.1.2 Application	15

3.3	Machine Learning	16
3.3.1	Sklearn	19
3.3.1.1	Definition	19
3.3.1.2	Application	19
3.4	Senpy	21
3.4.1	Architecture	21
3.5	Bitter	21
3.5.1	Architecture	22
3.6	Elasticsearch	22
3.6.1	Architecture	23
3.7	Pickle	24
4	Model	25
4.1	Introduction	25
4.1.1	Overview	25
4.2	Dataset	27
4.3	Preprocessing	28
4.4	Feature Extraction	29
4.4.1	Lexical Features	29
4.4.2	Syntactic Features	30
4.4.3	Ngram Features. Pipeline	31
4.4.4	Latent Dirichlet Allocator (LDA) Features	31
4.4.5	Frequency Inverse Document Frequency (TFIDF)	31
4.5	Classifiers	32
4.5.1	Naive Bayes	32
4.5.2	Support Vector Machine	33
4.5.3	K-Nearest Neighbors (k-NN)	34
4.5.4	Logistic Regression (LR)	35
4.6	Quality of a Classifier	35
4.6.1	Precision	36
4.6.2	Recall	36
4.6.3	F1 Score	36
4.6.4	Model Creation	36
4.7	Parameter Optimization	37
4.7.1	Grid Search	38
4.7.1.1	Multinomial Naive Bayes	38
4.7.1.2	Support Vector Machine	38

4.7.1.3	K-Nearest Neighbors	39
4.7.1.4	Logistic Regression	39
4.8	Senpy	40
4.8.1	Results	40
4.9	Elasticsearch	43
5	Conclusions and future work	45
5.1	Conclusions	45
5.2	Achieved goals	46
5.3	Future work	47
5.4	Problems encountered	47
A	Cost of System	I
A.1	Introduction	I
A.2	Physical requirements	I
A.3	Human resources	II
A.4	License	II
B	Impact of the Project	III
B.1	Introduction	III
	Bibliography	IV

List of Figures

2.1	Features used for Statistical Classifiers [23]	8
3.1	Image is taken from http://www.gartner.com/newsroom/id/2819918	14
3.2	Linear regression plotted [28]	17
3.3	Underfitting [32]	18
3.4	Overfitting [32]	18
3.5	A Pipeline. [6]	20
3.6	Senpy's Architecture [31]	21
3.7	Bitter Architecture	22
3.8	Elasticsearch [9]	23
3.9	Elasticsearch and a Database [10].	24
4.1	The Machine Learning [18] Model	26
4.2	Dataset	28
4.3	Classifier [36]	33
4.4	Model Construction [1]	37
4.5	Sarcastic example using Senpys' Playground [31]	41
4.6	Non-sarcastic example using Senpys' Playground [31]	42

Introduction

1.1 Context

The emergence of the social network Twitter in 2006 led to a tremendous boost of the Digital Revolution taking place since the end of the 20th century. As a result, there are on average at least 6000 tweets being tweeted every second. It is estimated that by 2020 the number of users on Twitter will increase up to 275 million users worldwide [4]. Twitter has become so popular that roughly 74% of all the Twitters' users claim to use the social network as a source to get the daily news [7]. One could say that Twitter has become part of our lives and he would not be far from the truth.

Combining the data allocated in the Twitter database with mathematical models can create insights that we are not capable of imagining. Moreover, the Twitter API offers the possibility of making topic-based queries, allowing the developer to get an enormous amount of data from almost any subject. Therefore, Twitter is like a gold mine of data. Additionally, spatial data can also be obtained by using the Twitter API. This can be very interesting if the developer wants to see what is the feeling of the people from an area towards a certain product or company.

According to [33], the world has already generated 1 zettabyte (ZB) of data and it is supposed to increase up to 44 ZB by 2020. The 44 ZB will contain at least half of lexical data, coming from social networks like Facebook, Twitter and instant messaging platforms.

Furthermore, it is expected that this humongous amount of data will continuously grow due to the influx of digital technologies that have already sprung up in the digital era.

Since an enormous amount of text data has arisen, there is a great need to give some insight to raw data. In that context, text mining uses Natural Language Processing (NLP) and Machine Learning (ML) techniques to process the data. Usually, text analysis is conducted for two purposes [33]. The first purpose is to analyze peoples' sentiment on an issue. The second purpose is to evaluate peoples' opinion and sentiment on a product or topic.

Finally, to give a glimpse of the results achieved, the senpy engine was implemented to provide a framework where a user could insert a short text and the classifier could detect whether or not the text is sarcastic.

1.2 Project goals

The main objective of this project is to define a classifier capable of detecting sarcasm in texts, particularly in tweets. To do so, the following challenges shall be overcome:

- *Tweet body retrieval*: The data provided will only contain tweets id and a sarcastic indicator. Hence, software capable of downloading the body of the tweets is required.
- *Dataset preprocessing*: To combine different datasets and encode categorical values, a preprocessing stage is mandatory.
- *Feature extracion*: If the mathematical classifier is to learn from the body of the tweets, a feature extraction process is required.
- *Model creation*: Create classification models using different mathematical learning techniques.
- *Optimization*: For the classifier to be optimal, the model will have the best parameters, most relevant features and greatest mathematical learning technique.
- *Senpy implementation*: Implement the optimum classifier in the senpy engine as a plugin to create a framework where a user can insert a short text and the classifier detects whether or not the text is sarcastic.

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:

Chapter 1 discusses the context wherein this project is located and gives an insight into the topics that this project will cover.

Chapter 2 describes what are the results already achieved in similar works. Serves as a benchmark.

Chapter 3 from a generic perspective explains the technologies chosen to create the model.

Chapter 4 analyzes extensively how are the technologies from chapter 3 implemented and shows the results achieved.

Chapter 5 finalizes the project showing the conclusions drawn and shows what were the problems faced.

State of the Art

This chapter will serve as a compilation of past work in automatic sarcasm detection. Firstly, a brief introduction to the main learning technologies used for text-based learning will be given. Moreover, an analysis of automatic sarcasm detection will be briefed. Finally, the main issues found in automatic sarcasm detection shall be overviewed.

2.1 Introduction

Sarcasm detection is an important component for NLP very relevant to natural language understanding, dialogue systems and text mining [25].

The perspective followed in this chapter is merely to give a glimpse to the reader of the state-of-the-art technologies used by other researchers in this field.

The chapter will begin by defining sarcasm and its context.

2.2 Sarcasm

2.2.1 Sarcasm Definition

The Free Dictionary¹ defines sarcasm as a form of verbal irony that is intended to express contempt or ridicule. The subjective characteristic of sarcasm makes sarcasm very difficult

¹www.thefreedictionary.com

to detect since a word may have a literal meaning and a sarcastic meaning. As a result, sarcasm automatic detection presents a big challenge [23].

Typically, sarcasm detection has been formulated as a classification problem. In that sense, given text, the goal is to predict whether or not the text is sarcastic. However, the big challenge is to label the text, since the classifier is supposed to distinguish between irony, humour and other sentimental meanings.

In linguistics, sarcasm is a form of figurative language where the literal meaning of words does not hold, and instead, the opposite interpretation is intended. Sarcasm is a type of irony.

2.3 Dataset

As a matter of a fact, the studies already done in automatic sarcasm detection depend entirely on the dataset. That is why it is convenient to divide them into three kinds [23]:

- *Short texts*: The social network Twitter is extremely popular due to its popularity among users. It can serve as an example of short text since every tweet is restricted to not exceed a certain number of words. One approach to obtaining labels for the tweets is to manually mark them as sarcastic or non-sarcastic. Another approach is hashtag-based supervision. Hashtags can reveal when a tweet is labelled as sarcastic. By introducing the *sarcasm* label, the author makes clear his intention to use sarcasm. This allows the dataset to be bigger. Many works using this variation have been reported since its popularity has constantly grown. One of the problems with Twitter is the text length restriction. If the tweet length exceeds the tweet restriction limit, the author is going to be forced to introduce abbreviations to the words used. As a consequence, sarcasm detection is going to be harder since the words are not going to be correctly spelt.
- *Long texts*: The information usually comes from reviews, posts, news, articles. Some of the popular platforms for sentiment analysis in long texts are imdb (for film critiques) and reddit.
- *Other data-sets*

It is noteworthy to mention the emphasis made on the Twitter social network since this project will be focusing on a Spanish set of tweets to detect sarcasm. Please note that this project will be developed for short texts. Hence, all the conclusions made are only valid for short texts.

2.4 Approaches used for Sarcasm Detection

In this section, several approaches used for detecting sarcasm will be discussed. In general, approaches for sarcasm detection can be classified into rule-based, statistical and deep learning-based approaches [23].

2.4.1 Rule-based Approaches

Rule-based approaches work by identifying sarcasm through a set of rules based upon evidence [23]. The evidence is captured by using rules relying upon sarcasm. For instance, one rule can rely on Google to determine how likely that smile (emoticon) is while another rule can come from hashtag analysis. By identifying the hashtags present in a tweet, a sarcasm pattern could be found. If the sentiment detected in the tweet does not match with the hashtag, then that tweet is labelled as sarcastic.

Furthermore, one could find sarcasm if a positive verb happens in a negative situation phrase in a sentence. The problem here is extracting the set of negative situation phrases. This is also known as a rule-based approach.

2.4.2 Feature sets

This kind of approach used some of the technologies described below [23]. In this project, there was a combination of features and learning algorithms but, in contrast, no deep learning technologies were implemented.

- *Features Used:* Most of these approaches use bag-of-words as features that have been found by statistical sarcasm detection. Apart from finding the features, some other works focus on designing pattern-based features, that are able to indicate sarcastic patterns on the corpus. The majority of the features used can be found in Figure 2.1. To allow the classifier to spot sarcastic patterns, these pattern-based features take three real values: exact match, partial match and no match. Furthermore, pragmatic features, like emoticons, can also be considered. For the sake of simplicity, the contextual sarcastic features (i.e. features that depend on previous messages) are not considered in this project.
- *Learning Algorithms:* A variety of classifiers have already experimented for sarcasm detection. Support Vector Machine (SVM) is quite often present in sarcasm detection classification techniques. Moreover, many other works have used Logistic Regression (LR), Multinomial Naive Bayes (MNB) and Random Forests.

Salient Features
<p>Sarcastic patterns, Punctuations</p> <p>User mentions, emoticons, unigrams, sentiment-lexicon-based features</p> <p>Ambiguity-based, semantic relatedness</p> <p>N-grams, POS N-grams</p> <p>N-grams, emotion marks, intensifiers</p> <p>Sarcastic patterns (Positive verbs, negative phrases)</p> <p>Skip-grams, Polarity skip-grams</p> <p>Freq. of rarest words, max/min/avg # synsets, max/min/avg # synonyms</p> <p>Synonyms, Ambiguity, Written-spoken gap</p> <p>Interjection, ellipsis, hyperbole, imbalance-based</p> <p>POS sequences, Semantic imbalance. Chinese-specific features such as homophones, use of honorifics</p> <p>Word shape, Pointedness, etc.</p> <p>Length, capitalization, semantic similarity</p> <p>Unigrams, Implicit incongruity-based, Explicit incongruity-based</p> <p>Readability, sentiment flips, etc.</p> <p>Pattern-based features along with word-based, syntactic, punctuation-based and sentiment-related features</p> <p>Affect-based features derived from multiple emotion lexicons</p> <p>Features based on word embedding similarity</p> <p>Cognitive features derived from eye-tracking experiments</p>

Figure 2.1: Features used for Statistical Classifiers [23]

- *Deep Learning-based Approaches:* Although nowadays deep learning has raised a lot of awareness, there are still few approaches reported for automated sarcasm detection. A similarity between word embeddings as features for sarcasm detection has been reported. This augments features based on similarity of word embeddings related to other word pairs and report an improvement on performance. Convolutional neural networks have also been implemented, reporting an improvement of 2% in performance.

To put it in a nutshell, different techniques for automatic sarcasm detection can be performed. Despite these numerous techniques, sarcasm automatic detection is difficult to detect since a word can have a literal meaning but also a sarcastic meaning that is not so easily labelled.

2.5 Main Issues in Sarcasm Detection

There are three main issues present in sarcasm automatic detection techniques [23]:

- *Data:* Even though hashtag-based labelling can provide large-scale supervision, the quality of the dataset can be doubtful. For example, let us take the hashtag *#not* very often present in tweets. Is this supposed to express sarcasm in the sentence or is it simply used to express a negation? In most of the works, this problem is tackled by removing the *#not* in the pre-processing step and analyzing the sentence. However, this may as well not be the optimum solution. Another solution is to use as test set some manually labelled tweets and as train set a hashtag labelled set.
- *Features as Sentiment:* The question is how can sentiment be detected in a sentence? In the case of sarcasm, some answers have already been given. If a negative phrase occurs in a positive sentence then that sentence has most likely sarcasm. In a statistical classifier, surface polarity can be used as a feature of the tweet. To capture surface polarity one has to analyze two emotions: activation and pleasantness. This can lead to a 4% improvement in the accuracy [23].
- *Skewed Data-sets:* Sarcasm is hard to find in an expression. For that reason, skew is reflected in data-sets.

2.6 Figures of Merit

This section is committed to showing other past works in Sarcasm Detection for tweets. The decisive parameters shown in this section will be the F-measure.

- Sarcasm Detection on Czech and English Twitter [29]:
 - *Description:* This paper presents a ML approach to sarcasm detection on Twitter in two languages - English and Czech. The classification was carried out by using SVM classifiers and Maximum Entropy. Since only SVM has been implemented in this project, the latter will be discarded.
 - *Results:* F-measure (balanced dataset): **0.947**. F-measure (unbalanced dataset): **0.924**. Algorithm: **SVM**
- Irony detection in short texts [2]:
 - *Description:* This document summarizes the result of analyzing ironic tweets. The language is spanish.
 - *Results:* F-measure (balanced dataset): **0.86**. F-measure (unbalanced dataset): **0.82**. Algorithm: **SVM**

Moreover, in [23] a table can be found with a collection of many other past works in sarcasm detection. However, the values on the table are not directly comparable, because different datasets and techniques have been used by the authors.

In the table 2.1 a compilation of performance values of sarcasm detection is shown.

The left column represents the citations used in the article [23]. Therefore, they do not correspond to this projects' citations.

The middle column is expressing whether the dataset was composed of short texts, such as tweets, or other long texts, like reddit.

Finally, the third column shows the score achieved. To simplify, the F1 Score (in the table is represented as F) will be considered the most relevant score. The average F1 Score is 76,771%. To compute the average, only tweets displaying a F1 Score in the table have been considered.

For more information on F1 Score see section 4.6.

	Details	Reported Performance
[67]	Conversation transcripts	F: 70, Acc: 87
[13]	Tweets	F: 54.5 Acc: 89.6
[68]	Reviews	F: 78.8
[69]	Similes	F: 88
[26]	Tweets	A: 75.89
[58]	Irony vs general	A: 70.12, F: 65
[56]	Reviews	F: 89.1, P: 88.3, R: 89.9
[41]	Tweets	AUC: 0.76
[45]	Discussion forum posts	F: 69, P: 75, R: 62
[55]	Speech data	Acc: 81.57
[59]	Irony vs humor	F: 76
[60]	Tweets	F: 51, P: 44, R: 62
[5]	Tweets	F: 62
[9]	Reviews	F: 71.3
[47]	Tweets	F: 91.03
[2]	Tweets	Acc: 85.1
[18]	Tweets	F: 83.59, Acc: 94.17
[20]	Tweets	Cosine: 0.758, MSE: 2.117
[22]	Tweets	F: 97.5
[28]	Irony vs politics	F: 81
[32]	Tweets/Disc. Posts	F: 88.76/64
[36]	Tweets	F: 88.2
[54]	Tweets	Acc: 83.46, AUC: 0.83
[73]	Reddits	P: 0.141, F: 0.377
[75]	Tweets	Macro-F: 69.13
[1]	Tweets	AUC: 0.6
[17]	Tweets	F: 82
[33]	TV transcripts	F: 84.4
[34]	Book snippets	F: 80.47
[48]	Tweets, Quotes and Reviews	F: 75.7
[50]	Reviews	F: 75.7
[63]	Tweets	Acc: 87.2

Table 2.1: Performance Values of Sarcasm Detection; Precision/Recall/F-measures and Accuracy Values indicated in Percentages [23]

Enabling Technologies

In this chapter, the technologies used to achieve this projects' objective shall be discussed.

3.1 Introduction

In this project two main technologies have been used: Natural Language Processing (NLP) and Machine Learning (ML). The implementation of these tools has been performed in a combined fashion. Furthermore, other technologies have also been used but they are less relevant than NLP and ML. On the one hand, with NLP it was possible to transform words into tokens, find syntactical connections between tokens in the same tweet and, finally, construct a numerical vector. On the other hand, ML was very useful to construct a classifier, define a model and feed the vectors from the previous step into the model to make predictions.

In short, this is how the project was carried out. Both technologies will be extensively explained in the next sections.

3.2 Natural Language Processing

NLP has a crucial role in this project since texts (tweets) is what this project is trying to deal with. Hence, the information that used to train the ML algorithm will come in the form of text. According to [14], at one extreme NLP could be as simple as counting

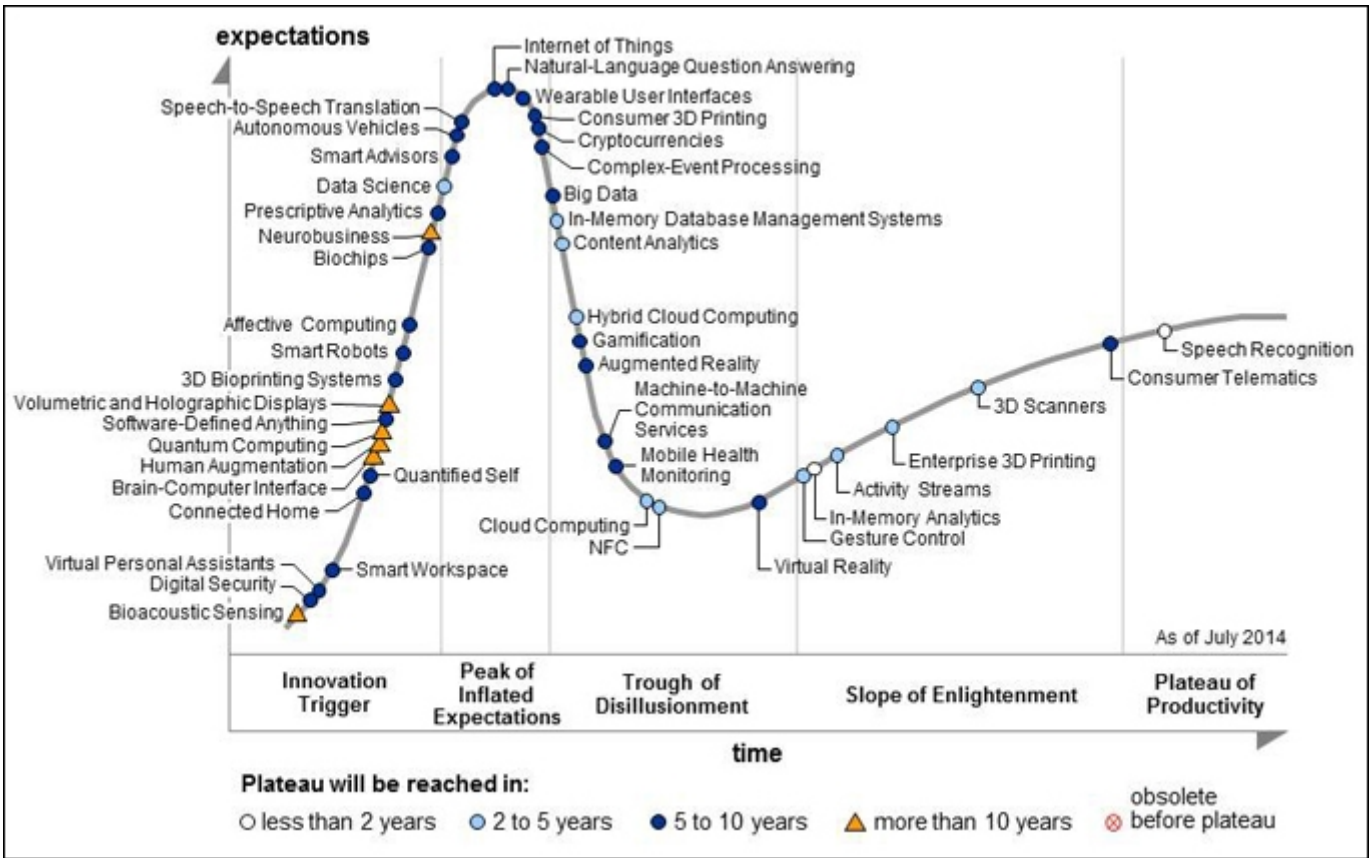


Figure 3.1: Image is taken from <http://www.gartner.com/newsroom/id/2819918>

word frequencies to compare different writing styles. At the other extreme, NLP involves "understanding" complete human utterances, at least to the extent of being able to give useful responses to them.

When the task is handling text, NLP is a very powerful tool. Such is the importance that many other applications that can be found everyday already implement it. Figure 3.1 shows many emerging technologies that are using NLP to provide their services. Many other examples of NLP applications are machine translation, optical character recognition, language detection, topic modelling and many more [20].

3.2.1 Natural Language Toolkit

3.2.1.1 Definition

Natural Language Toolkit (NLTK) is a platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-

strength NLP libraries, and an active discussion forum ¹.

The following applications explained beneath belong to the NLTK class.

3.2.1.2 Application

NLTK has provided the following tools to this project:

- *Tokenizer*: A word (Token) is the minimal unit that a machine can understand and process. So any text string cannot be further processed without going through tokenization. Tokenization is the process of splitting the raw string into meaningful tokens [21]. In this project, tokenization was required to get the words containing a tweet. Furthermore, some of the tokens were removed since they do not add meaningful information (e.g. yo, a). These words belong to a stoplist provided by NLTK for all supported languages [21].
- *Stemmer*: A stemmer is needed to remove affixes from a word, ending up with a stem. Stemming is frequently used for indexing words. Instead of storing all forms of a word, the algorithm stores only the stems, greatly reducing the size of the index while increasing retrieval accuracy [21]. Particularly, the SnowballStemmer was chosen because it supports stemming in 13 non-English languages, especially Spanish.
- *Part-of-Speech tagging*: A-part-of-Speech tagger (pos_tag) indicates how a word is functioning within the context of a sentence [13]. This part is crucial since a word can function in multiple ways and we would like to distinguish those cases. POS_tagging has been achieved by collecting each word feature and building up a dictionary with the word features' stats. The word features offered by NLTK are:
 1. *VERB*: Verbs (all tenses and modes)
 2. *NOUN*: Nouns (common and proper)
 3. *PRON*: Pronouns
 4. *ADJ*: Adjectives
 5. *ADV*: Adverbs
 6. *ADP*: Adpositions (prepositions and postpositions)
 7. *CONJ*: Conjunctions
 8. *DET*: Determiners
 9. *NUM*: Cardinal numbers
 10. *PRT*: Particles or other function words

¹<https://www.nltk.org/>

11. *X-other*: Foreign words, typos, abbreviations
12. *.*: Punctuation

The applications at the top of this section are more commonly known as the preprocessing part previously needed to any ML application. Since there are plenty of tasks, they could be easily coordinated by using pipelines.

More information on how were these technologies implemented in the project can be found in section 4.3.

3.3 Machine Learning

When it comes to defining models capable of learning from a text-based source, the two dominant technologies are Natural Language Processing and Machine Learning. These two main technologies are concerned about how does the machine process human text [28].

Given data and the proper statistical techniques, ML automates analytical model building capable of recognizing patterns with the aim of making predictions and learning from the model. To be able to do so, a learning process is required. ML algorithms can be classified into three categories [8]:

- *Regression*: Regression learning is predicting a continuous varying variable. Regression methods can be used to address classification and prediction models. To visualize a simple example of what regression is we can see the picture Figure 3.2.
- *Classification*: Classification is modeling based on delineating classes of output based on some set of input features. If regression gives us an outcome of "how much," classification gives us an outcome of "what kind."
- *Clustering*: Clustering is an unsupervised learning technique that involves using a distance measure and iteratively moving similar items more closely together. At the end of the process, the items clustered most densely around n centroids are considered to be classified in that group. *K – means* clustering is one of the more famous variations of clustering in machine learning.

Furthermore, depending on the dataset the learning process can be carried out in two different ways:

- *Supervised learning*: The classifier is trained using a percentage of the total dataset while feeding the rest of the dataset to make predictions and to evaluate how good the classifier is. Particularly, this project will be using supervised learning.

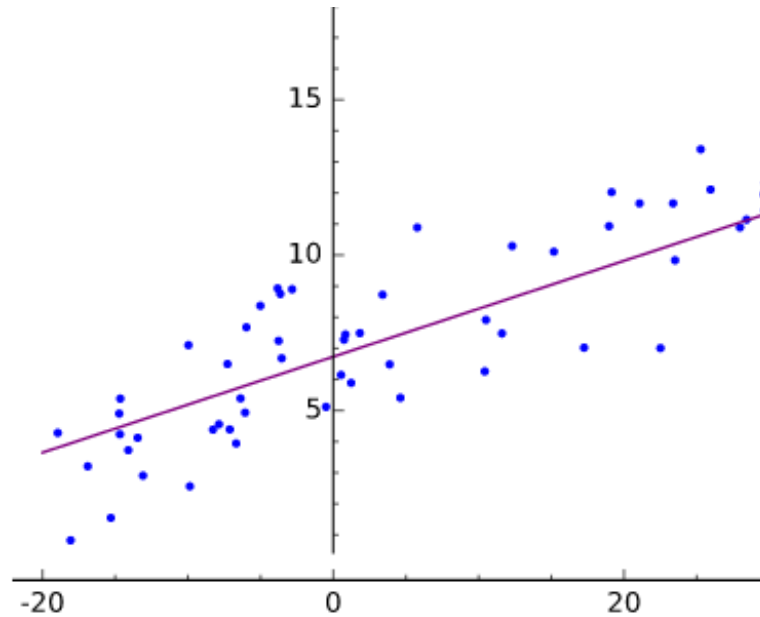


Figure 3.2: Linear regression plotted [28]

- *Unsupervised learning:* These algorithms base their learning properties on learning patterns present on the dataset. Unlike in supervised learning, there is no training and test dataset partition

Two major problems encountered in a ML application are underfitting and overfitting. These two problems appear due to the bias generated when training the model.

It turns out that depending on *how much trained* the classifier is, it can make more assumptions than desired. A learned function that can perform well enough on unseen data points as well as on the training data is termed a generalizable function [32]. A generalizable function is achieved through a generalizing fit. To get a proper fit, the function must not be biased, it must be perfectly balanced.

- *Underfitting:* Figure 3.3 shows that for any point in the diagram, the function always assigns the same predicted value. Solving the problem of underfitting means making the line approximate the data better.
- *Overfitting:* Figure 3.4 shows that for any point in the diagram, the function performs very good. However, the function has become very inflexible and it will not perform good enough on new points. According to [28], overfitting can also be understood as a probable distribution of data. The training set of data that the line is being drawn through is just a sample of a larger unknown set, and the line which will be drawn needs to fit the larger set equally well if it is to have any predictive power. Therefore, it must be assumed that our sample is loosely representative of a larger set.

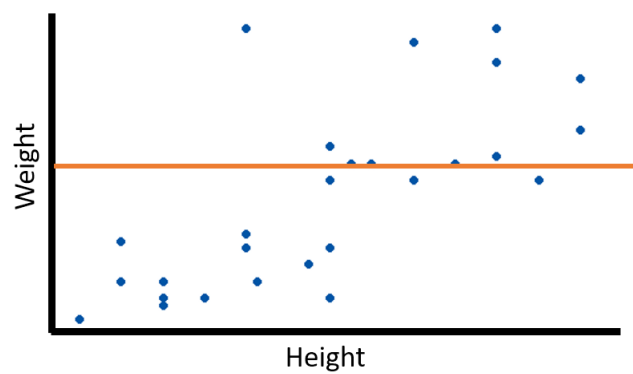


Figure 3.3: Underfitting [32]

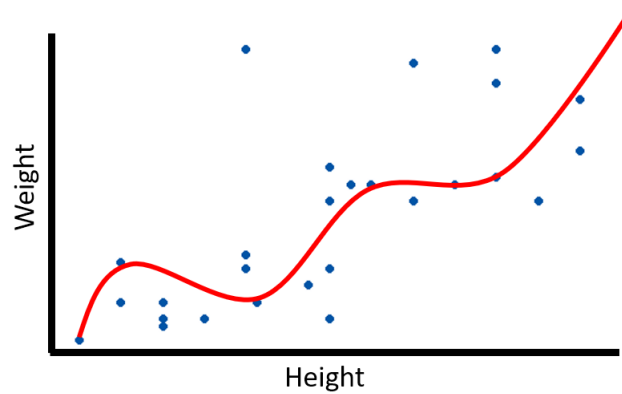


Figure 3.4: Overfitting [32]

Another field of Artificial Intelligence is NLP, which provides tools for analyzing and deriving meaning in human language. To do so, NLP provides a means to process lexical properties as well as syntactic recognition of the human language. However, human language can be rarely precise and ambiguous. Even though NLP has a very wide range of applications, such as automatic text summarizing, relationship extracting, stemming, sentiment analysis and more, it is a difficult problem to tackle in computer science.

3.3.1 Sklearn

3.3.1.1 Definition

Sklearn (a.k.a scikit-learn) is defined ² as a simple and efficient set of tools for data mining and data analysis. It is built on NumPy and SciPy. Sklearn will provide not only the classifiers needed for the pipeline but also meaningful tools.

3.3.1.2 Application

Sklearn has provided the following tools for this project:

- *Train and Test splitting*: This method is responsible for splitting the main dataset into two parts, one for training the model and the other one for testing the model. It is a very common ML technique. In addition, the testing set will later be used to compute parameters such as accuracy and f1 score. More information on accuracy and f1 score can be found in section 4.6.
- *Pipeline*: A pipeline is a set of procedures connected in series, one after the other where the output of one process is the input to the next [24]. Image 3.5³ represents the idea of a pipeline, though it does not resemble exactly this projects' pipeline.
- *Frequency Inverse Document Frequency*: The Frequency Inverse Document Frequency (TFIDF) is used for every word. It contains two parts; the *tf* part, which represents the word frequency and the *idf* part, which represents inverse document frequency [27]. This term represents a words' weight in a text.
- *Count Vectorizer*: The algorithm consists of representing a token by considering how many times it appears in a text [15]. It basically counts the vocabulary and generates a dictionary which can later be fed into a pipeline, like in this project.

²<http://scikit-learn.org/stable/>

³<https://www.slideshare.net/databricks/dataframes-and-pipelines>

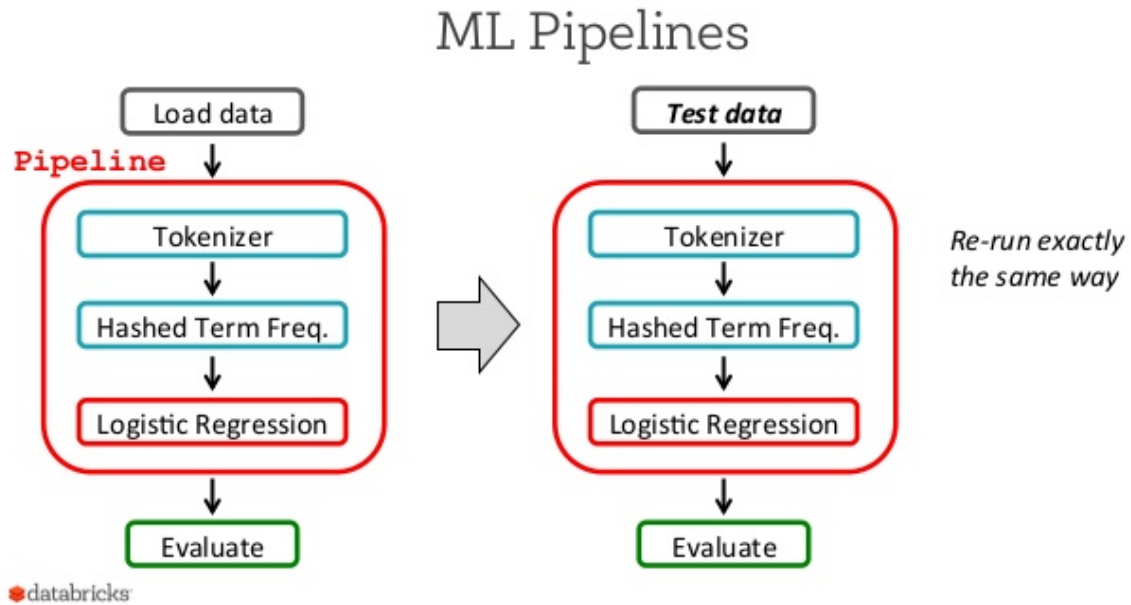


Figure 3.5: A Pipeline. [6]

- *Multinomial Naïve Bayes*: A multinomial distribution useful to model feature vectors where each value represents a number of occurrences of a term (or its relative frequency) [15].
- *Support Vector Machine*: Given a set of training examples, each marked as belonging to one or the other of two categories, a Support Vector Machine (SVM) training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier⁴.
- *K-Nearest Neighbors*: Is a non-generalized ML algorithm that computes distances from one point to all the training dataset points and chooses the k nearest points [15]. The KNeighbors classifier has been fed into the pipeline as input.
- *Logistic Regression classifier*: Is analogous to multiple linear regression, except the outcome is binary. Various transformations are employed to convert the problem to one in which a linear model can be fit [16].

All the methods illustrated above are widely explained altogether with an implementation clarification in section 4.5.

⁴https://en.wikipedia.org/wiki/Support_vector_machine

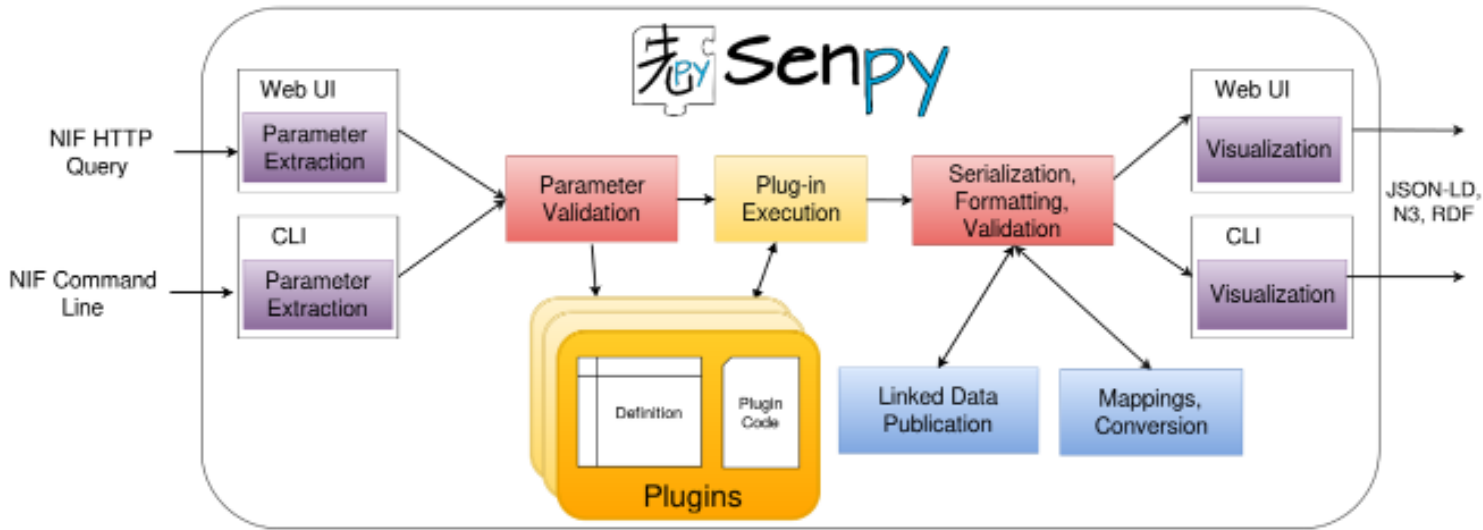


Figure 3.6: Senpy's Architecture [31]

3.4 Senpy

Senpy [31] is a framework for text analysis using linked data. It aims at providing a framework where analysis modules can be easily integrated as plugins and, at the same time, provides the core functionalities (data validation, user interaction, logging, etc).

3.4.1 Architecture

Senpy introduces a modular and dynamic architecture which allows implementing different algorithms in an extensible way while offering a common interface and offering common services that facilitate development, so developers can focus on implementing new and better algorithms.

To do so, the architecture consists of two main modules:

- Senpy core: Building block of the service.
- Senpy plugins: Analysis algorithms.

Figure 3.6 depicts a simplified version of the processes involved in Senpy analysis framework.

3.5 Bitter

In order to download the Tweets which integrate partially the dataset, the tool bitter was used. Bitter is able to automate several actions (e.g. downloading Tweets) by using a

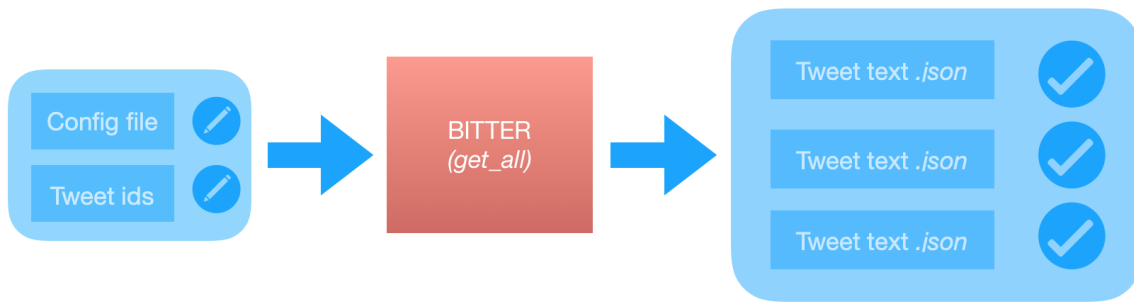


Figure 3.7: Bitter Architecture

Python wrapper over Twitter which adds support for several Twitter API actions ⁵. Bitter could be used after installing it and running commands on the UNIX terminal.

3.5.1 Architecture

Figure 3.7 explains how was bitter used in this project. On the left, the reader can see the information provided to the bitter tool. The configuration file is needed to download the tweets, otherwise, nothing will be downloaded. The file contains the user credentials for accessing the Twitter database.

The tweet id file is a *.csv* database containing in one column the desired tweets. Finally, bitter will download in *.json* format all the tweets in the indicated folder (indicated on the command). For those tweets that, by any reason, could not be downloaded, a folder named *error* will be created containing the tweet ids that failed.

3.6 Elasticsearch

Elasticsearch is a highly scalable open source search engine with a very powerful analytical engine. The data stored in Elasticsearch is Javascript Object Notation (JSON) formatted. The primary way of interacting with Elasticsearch is via **REST API** [10]. The most powerful tool that Elasticsearch provides is indexing. Indexing becomes a very important task when dealing with lots of amount of data, in this case, with many tweets.

⁵<https://github.com/balkian/bitter>

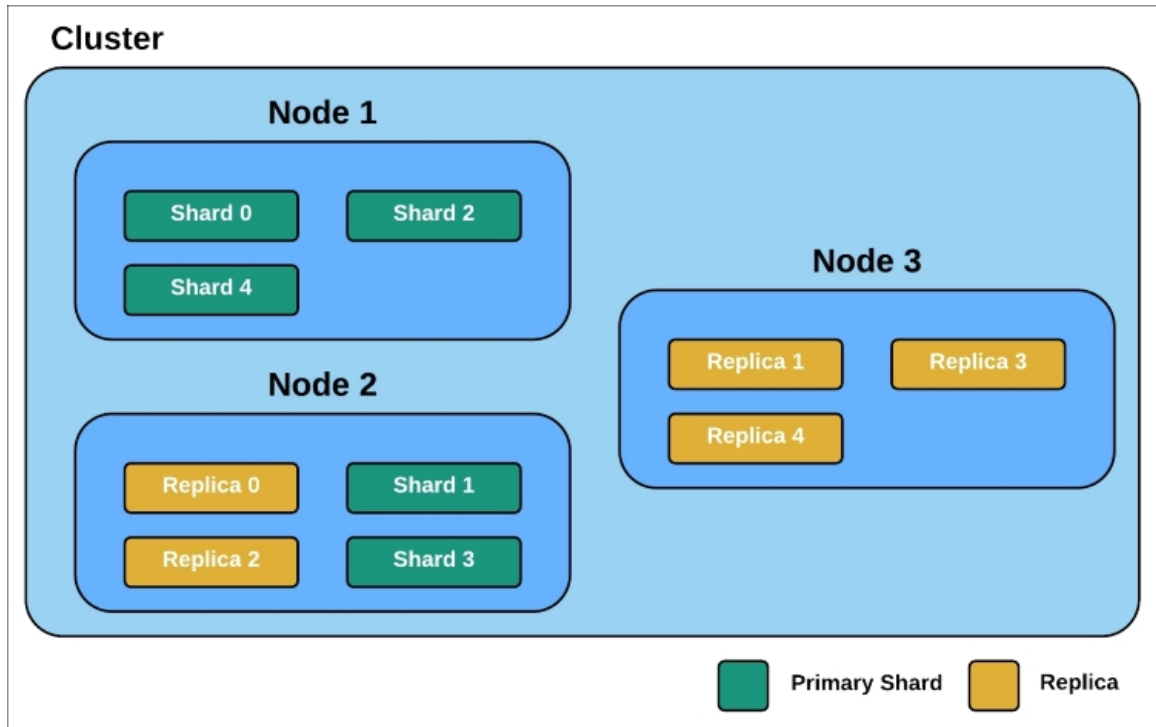


Figure 3.8: Elasticsearch [9]

3.6.1 Architecture

All data stored in Apache Lucene's data is stored in *inverted index*. This means that the data is being transformed. The process of transforming data is called analysis, which relies in two basic pillars: **tokenizing** and **normalizing**. The normalizing step is mandatory to make the tokens readable. Inverted index processes are performed by analyzers. An analyzer is composed of a tokenizer and one or more token filters. During the indexing time, when Elasticsearch processes a field that must be indexed, it checks whether an analyzer is defined in several levels. [10]

When it comes to indices, Elasticsearch uses shards and replicas in order to distribute data around the cluster.

Elasticsearch distributes the shards by performing a process called *sharding*. The indices (or index) are called shards and Elasticsearch automatically manages the shards, like a low-level working unit. The most common terms used in Elasticsearch are:

- *Node*: An instance of Elasticsearch running on a machine.
- *Cluster*: The name under which one or more nodes are connected.
- *Document*: A JSON object containing the actual data in key-value pairs.
- *Index*: Logical namespace in which Elasticsearch stores data.

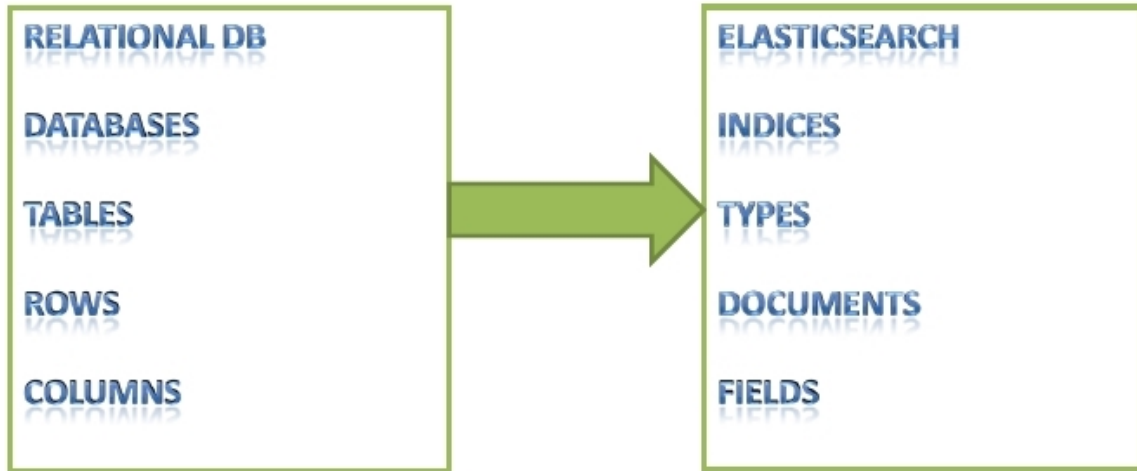


Figure 3.9: Elasticsearch and a Database [10].

- *Doc types*: A class of similar documents. A type consists of a name and a mapping.
- *Shard*: Containers which can be stored on a single node or multiple nodes. A shard can be either primary or secondary. A primary shard is the one where all the operations that change the index are directed. A secondary shard is the one that contains duplicate data of the primary shard and helps in quickly searching the data as well as for high availability; in a case where the machine that holds the primary shard goes down, then the secondary shard becomes the primary automatically.
- *Replica*: Duplicate copy of the data living in a shard for high availability.

Finally, as mentioned before, Elasticsearch can be related to a traditional database managing system. This is depicted in Figure 3.9.

3.7 Pickle

The pickle module makes it possible to store an object in a file in a standard format for later use by the same or a different program. The stored can be any type of object as long as it does not require an operating system resource such as a file handle or network socket [34]. Pickle is very useful for using an already trained classifier in different modules.

4.1 Introduction

The social network Twitter has become extremely popular worldwide. It has developed into a tool that allows anyone to interact directly with people of his interest. Even world-wide leaders, like the president of the United States of America, announce important decisions using Twitter. In other words, Twitters' simplicity and easy-to-use interface has completely transformed the way citizens interact with the society. Altogether with the globalised world that we are living in and the limitless scope that the internet presents makes Twitter a resourceful platform of data.

In this chapter, a technical description of the implemented model will be given.

Firstly, an overview of the model shall be presented, followed by the dataset details. Furthermore, the data munging preprocessing will be made clear. Later on, the model building alongside with the classification options shall be extensively explained, making emphasis on the best classification results. Finally, the parameter optimization will be briefed altogether with the best-founded parameters results.

4.1.1 Overview

In this subsection, a wide view of the system will be given. The first part of this project consists of elaborating the ML model. Figure 4.1 shows the steps followed during the ML

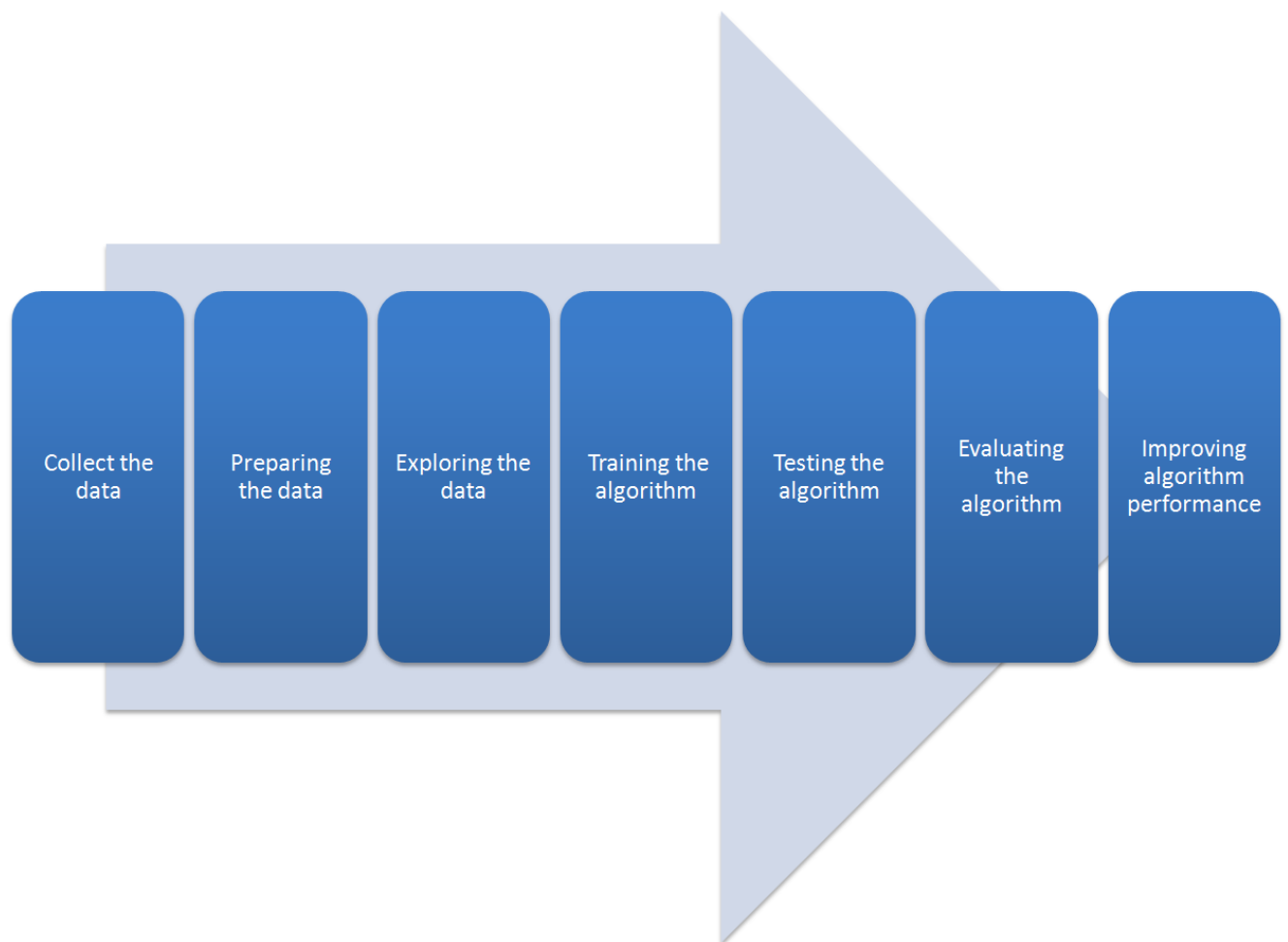


Figure 4.1: The Machine Learning [18] Model

stage.

The first two steps are needed to collect the data from the Twitter social network, combine it with more data sources, generate a final *.csv* file and prepare the data frame (see section 4.2). Once the data frame is generated, data exploration and cleaning will be performed (i.e. delete empty tweets, encode categorical variables and so on). It is then when an algorithm is built and trained. The algorithm is also known as a classifier and it will be responsible for predicting sarcasm in new tweets (see section 4.5). Finally, an accuracy report of the classifier will be carried out to evaluate the quality of the classifier (see section 4.6).

The second part of the project consists of defining a framework, where the ML process can be added as an analysis module. This can be accomplished thanks to *senpy* [31].

The way *senpy* is implemented is by defining just one plugin, which will contain the complete ML process previously designed. In other words, the plugin will contain the classifier. The plugin will be run every time the user types in some text (see Figure 4.5) and, finally, the results will be visualized in the web.

4.2 Dataset

This section explains the dataset choice, as well as the main label present in the dataset.

This project is formerly designed to detect sarcasm in Spanish written texts, particularly tweets. Therefore, all the datasets used in this project are written in Spanish.

The dataset body will contain two columns, the tweet body, i.e. the text, and a binary value expressing the sarcasm nature of that tweet.

The final dataset will be composed of three different datasets.

The initial database [2] consists of 4529 sarcastic tweets and 335 non-sarcastic tweets. At first, the dataset contained only the tweet identifier and the sarcastic label, meaning there was no text present on the dataset. In order to obtain the tweet text, it was necessary to use the *bitter* tool (see section 3.5). While doing so, many tweet bodies were no longer available to retrieve since the Twitter platform had removed them or the owners had privatized them. As a consequence, new sarcastic tweets were provided by a colleague from the Intelligent Systems Group in the UPM.

Even though the numerous amount of sarcastic tweets was a good start, the dataset was completely unbalanced. Consequently, new tweets had to be searched to obtain a properly balanced dataset. Considering the fact that the majority of the dataset is sarcastic, finding new non-sarcastic tweets was not a difficult duty.

In the end, the result was 5638 sarcastic tweets and 5444 non-sarcastic. That can be rewritten as 49.12% of non-sarcastic and 49.5% of sarcastic tweets.

	tweet	ironic
0	Algunas personas sufren en las discos mientras...	1
1	@jacevedoaraya es para sostener el marcador.....	1
2	Alguna de estas imágenes te sacara una sonrisa...	1
3	@_Eurovision2014 en 2013 falta esdm jajajajaja...	1
4	Hooo que buen padre...#Sarcasmo #GH2015	1
5	@JhoynerV ja ja ja ja así o más claro, cas...	1
6	@patronbermudez con todo respeto lo principios...	0
7	Gran rapidez todo en la UPO y no iban a ser me...	1
8	¿Que humilde es Simeone no? #ironía #llorón #f...	1
9	¿Alguien se viene a la playa conmigo? #resfria...	1
10	Tantos enamorados que no son novios 🙄💔💔💔💔 \ny tan...	0

Figure 4.2: Dataset

The Figure 4.2 displays the dataset structure. As explained before, there is one column showing the tweet text and another column showing the sarcastic value of the tweet. The '1' refers to sarcastic, while '0' is non-sarcastic.

4.3 Preprocessing

The preprocessing stage is conceived to dispose of any information that will not be needed during the ML step [35]. The preprocessing was applied individually to each tweet. The changes in the dataset that have been made are:

- *False items*: Since almost half the dataset tweets come from Twitter, it was required to download those tweets from the social network. Some of the captured tweets were false.
- *Categorical values*: Originally the labels of the dataset were either 'True' or 'False'. These two values were encoded into 'True'= 1 and 'False' = 0.
- *Randomization*: This step was used to provide a randomized dataset and have the tweets mixed up. This was important because initially almost every sarcastic tweet

was at the beginning and that is not desirable.

At the beginning of the preprocessing stage, the dataset is split into two parts, a training part and a testing part. The purpose of this split is to train a classifier using the train sequence and evaluate the accuracy of the classifier using the test sequence. The split is taken randomly. The size of each sequence is 75% for the train sequence and 25% for the test sequence (8311 and 2771). More information on can be found on section 4.6.4.

4.4 Feature Extraction

The feature extraction process consists of reading the text and extracting certain features. Furthermore, those extracted features will be fed to the classifier and finally the model will be completed. As a result, the accuracy of the model will be directly related to the choice of features. Therefore, it is very important to choose wisely the features to extract. This section will be committed to extensively explaining the selected features.

The extraction will be performed inside a Pipeline, which will execute each extraction sequentially. In this project, two pipelines are defined, the first aims at extracting the ngrams. With the ngrams extracted from the first pipeline, a second one will extract other features and unify the ngrams with the other extracted features. Even though there are two pipelines defined, the results of the ngrams pipeline are added to the other pipeline. This is done thanks to the Feature Union class provided by Scikit-Learn.

4.4.1 Lexical Features

Usually extracting lexical features involves getting the number of sentences and the number of words in each sentence. However, since this projects' dataset is made out of tweets and the number of characters in a tweet is limited, I have considered that each tweet contains only one sentence. As a consequence, the lexical feature extraction becomes very simple. The extraction consists of tokenizing each sentence and stemming each token (or word) to get the root. At the same time, the words belonging to the Spanish stop list are removed (see section 4.3 for more details). In section 4.4.5 is explained what happens with the extracted words in this stage. The list of operations made for the lexical feature extraction are as follows:

1. *Stopwords*: In ML it is a very common practice to delete words which barely provide any information to the text (e.g. 'a', 'I', 'you'). In this project, the list of stopwords was chosen for the Spanish language due to the language of the tweets. The list of words can be consulted in [20].

2. *Tokenization*: This process consists of splitting a string (tweet) into tokens. Usually, a token is the same as a word.
3. *Stemming*: This process consists of reducing the tokens into its root. Basically, the words are transformed into its root. Stemming was accomplished using a Snowball stemmer.

After performing these operations, the remaining words were passed to a Frequency Inverse Document Frequency transformer (see section 4.4.5).

4.4.2 Syntactic Features

The syntactic feature extraction consists of counting the number of nouns, adjectives, verbs, adverbs, conjunctions, pronouns and numbers present in each tweet.

The table 4.1 details the syntactic features extracted in this project. The approach to do so consisted of labelling every token present in a tweet, assigning to the token the grammar tag (see table 4.1) and generating a dictionary with the number of verbs, number of nouns, and so on.

Finally, the dictionary will be processed by the TFIDF (see section 4.4.5 for more details on this stage).

Tag	Meaning	Example
NOUN	noun	year, home, costs
ADJ	adjective	new, good, high
VERB	verb	is, say, told
ADV	adverb	really, already, still
CONJ	conjunction	and, or, but
ADP	adposition	on, of, at
PRON	pronoun	he, their, her
NUM	numeral	twenty-four, fourth, 1991

Table 4.1: POS Used Tagset [13]

4.4.3 Ngram Features. Pipeline

Ngram extractions consist of grouping the words into bag of words. If we group the words individually then a sentence with five words will be considered as a five words sentence. If we group two words together then a sentence of six words will be considered as a three word sentence. So each pair of words are understood by the estimator as one.

Since it is very difficult to guess what is the optimum bag of words (i.e. number of words) CountVectorizer method provided by the Scikit-Learn class allows us to define a range. Finally, to select the optimum range, a Grid Search (GS) will be performed (see section 4.7.1).

4.4.4 Latent Dirichlet Allocator (LDA) Features

LDA ¹ is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

In other words, LDA extraction can be used to extract how many topics are being spoken about in the dataset. This can be observed by analyzing each words' presence.

The main parameter of this statistical model is the number of topics. Similarly as with the ngrams features, a GS will be performed to find the optimum number of topics (see section 4.7.1).

4.4.5 Frequency Inverse Document Frequency (TFIDF)

TFIDF ² is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. Particularly, the TFIDF recognizes those words that are rare in the corpus but may be of great importance. The TFIDF uses a word as an argument and outputs the inverse frequency of that word.

The equation used is [28]:

$$tf_{t,d} = \frac{count(t)}{count(alltermsindocument)} \quad (4.1)$$

with

$$idf_t = \log\left(\frac{N}{df_t}\right) \quad (4.2)$$

and

$$tfidf_{t,d} = tf_{t,d} \times idf_t \quad (4.3)$$

The number of occurrences of a word in a complete document is computed with eq. (4.1). Equation (4.2) represents the *idf* of a word. That amount serves as how much information a word provides. As explained before, terms with less frequency are considered to provide more

¹https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

²<https://en.wikipedia.org/wiki/Tf\OT1\textendashidf>

information to the whole document as more common terms. Finally, eq. (4.3) computes the weight of a word.

4.5 Classifiers

This section will explain extensively the classifiers used for the learning process.

Classification [36] is the task of predicting the class to which an object, known as a pattern, belongs. The pattern is assumed to belong to one and only one among a number of a priori known classes. Each pattern is uniquely represented by a set of measurements, known as features. One of the early stages in designing a classification system is to select an appropriate set of feature variables. These should “encode” as much class-discriminatory information, so that, by measuring their value for a given pattern, to be able to predict, with high enough probability, the class of the pattern. Figure 4.3 is an example of two linearly separable classes, where a straight line can separate the two classes.

In this project, there will be three classifiers.

4.5.1 Naive Bayes

The MNB classifier is a typical and popular example of a suboptimal classifier. The basic assumption is that the components (features) in the feature vector are statistically independent; hence, the joint Probability Density Function (PDF) can be written as a product of l marginals [36]:

$$p(x|\omega_i) = \prod_{k=1}^l p(x_k|\omega_i) \quad (4.4)$$

Considering that both ω_i and x are Gaussian variables, it is only necessary to compute the mean and the variance for each pair of variables. That leads to a total of $2 \times l$ unknown variables for a subclass. Computationally speaking, $2 \times l$ complexity is achievable in a reasonable amount of time for a large l . This is the great advantage of the MNB.

The MNB classifier was defined setting the parameter $\alpha = 0.01$. Nevertheless, it is not very important what parameters to set when defining the classifier since a GS will be performed with the purpose of discovering the best classifier parameters. (See section 4.7.1) In this project, the results obtained for the MNB with no parameter optimization can be observed in table 4.2. The quality of a classifier will be established using the *F1-Score* field from the classification report. For more details about this precedent see section 4.6.

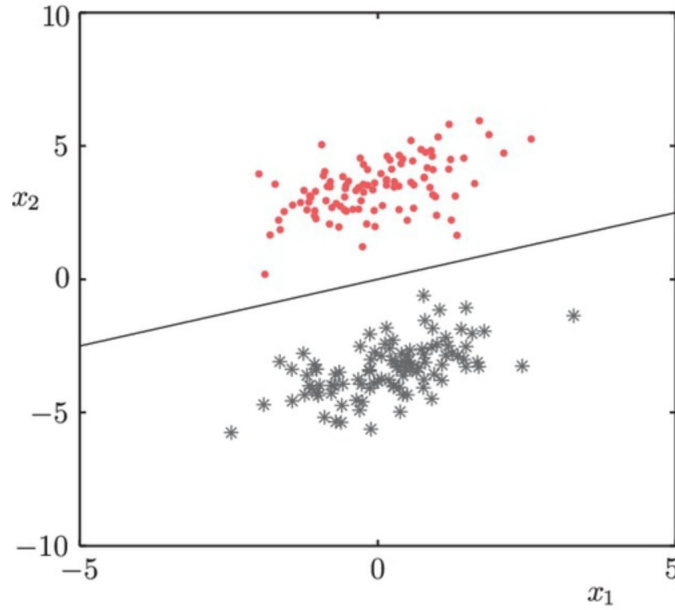


Figure 4.3: Classifier [36]

Variable	Precision	Recall	F1-Score	Support
0	0.81733	0.78781	0.80230	1329
1	0.81074	0.83773	0.82401	1442
avg/total	0.81390	0.81379	0.81360	2771

Table 4.2: Multinomial Naive Bayes Classification Report

4.5.2 Support Vector Machine

Support vector machine [22] is a supervised learning algorithm that can be used for classification and regression. It is able to classify data linearly and nonlinearly using kernel methods. Each data point in the training dataset is labelled, as it is supervised learning, and mapped to the input feature space, and the aim is to classify every point of new data to one of the classes. A data point is an N dimension number, as N is the number of features, and the problem is to separate this data using $N-1$ dimensional hyperplane and this is considered to be a linear classifier.

The SVM was defined setting as parameters $C = 1$ and $kernel = linear$. In other words, the initial SVM model was the linear SVM model.

With such parameters, the results can be observed in table 4.3. The quality of the classifier

was established following the *F1-Score* field from the classification report. For more details see section 4.6.

Variable	Precision	Recall	F1-Score	Support
0	0.92590	0.92387	0.92488	1366
1	0.92614	0.92811	0.92712	1405
avg/total	0.92602	0.92602	0.92602	2771

Table 4.3: Support Vector Machine Classification Report

4.5.3 K-Nearest Neighbors (k-NN)

k-NN is based on the idea that samples that are close with respect to a predefined distance metric are also similar, so they can share their peculiar features. To measure that distance, a distance function is defined. Usually, that distance function is the Minkowski metric [15]. For each data point in the out-sample data, we calculate its distance from all data points in the in-sample data. Each data point has a vector of distances and the K distance which is close enough will be selected and the final decision about the class of the data point is based on a weighted combination of all k neighborhoods [22].

The k-NN was defined setting as parameters *neighbors* = 7.

The results can be observed in table 4.4. The quality of the classifier was established following the *F1-Score* field from the classification report. For more details see section 4.6.

Variable	Precision	Recall	F1-Score	Support
0	0.79561	0.76940	0.78229	1366
1	0.78276	0.80783	0.79510	1405
avg/total	0.78909	0.78888	0.78878	2771

Table 4.4: K-Nearest Neighbors Classification Report

4.5.4 Logistic Regression (LR)

Logistic Regression is a classification method that is based on the probability of a sample belonging to a class. It consists of maximum-entropy classification. Mathematically, the aim is to compute the minimum of eq. (4.5) [21].

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^n \log(\exp(-y_i(X_i^T w + c)) + 1) \quad (4.5)$$

The LR classifier was defined setting the parameter $C = 1$.

For such a parameter, the results obtained can be observed in table 4.5. The quality of the classifier was established following the *F1-Score* field from the classification report. For more details see section 4.6.

Variable	Precision	Recall	F1-Score	Support
0	0.90007	0.90074	0.90040	1360
1	0.90426	0.90361	0.90393	1411
avg/total	0.90220	0.90220	0.90220	2771

Table 4.5: Logistic Regression Classification Report

4.6 Quality of a Classifier

In order to ensure a sufficient classifier accuracy, precision is not always the best indicator. Let us consider a binary class dataset where 99% of the data belongs to one class and only 1% of the data belongs to the other class. Now, if a classifier were to always predict the majority class for every data point, it would have 99% accuracy. But that would not mean that the classifier is performing well. To have a better idea of the classifiers' efficiency we first must define some parameters [12]:

- *True Positive (TP)*: Cases where the actual and the predicted classes are both positive.
- *True Negative (TN)*: Cases where the actual and the predicted classes are both negative.
- *False Positive (FP)*: Cases where the actual class is negative but the predicted class is positive.

- *False Negative (FN)*: Cases where the actual class is positive but the predicted class is negative.

To better understand these definitions let us imagine a classifier that predicts a sarcastic when in fact that tweet is not sarcastic, that would be a false positive. On the other hand, if the classifier predicts that a tweet is not sarcastic when in reality is sarcastic, that would be a false negative.

4.6.1 Precision

Precision is defined as the ratio of the number of positive cases that were correct to all the cases that were identified as positive [12]. Mathematically:

$$Precision = \frac{TP}{TP + FP} \quad (4.6)$$

4.6.2 Recall

Recall is defined as the ratio of the number of positive cases that were identified to the all positive cases present in the dataset [12]. Mathematically:

$$Recall = \frac{TP}{TP + FN} \quad (4.7)$$

4.6.3 F1 Score

is the harmonic average of the precision and recall. It is a metric that conveys the balance between precision and recall [12]. Mathematically:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.8)$$

Since the F1 score takes into account both the precision and the recall (see eq. (4.8)), it is the most suitable for determining a binary classifiers' quality. In this project, a tweet can either be sarcastic or non-sarcastic. Therefore, F1 score report is the appropriate quality measure.

4.6.4 Model Creation

The Figure 4.4 depicts the methodology followed for training and evaluating the classifier. Firstly, the dataset is split into two sets, the training and the test. (In the Figure 4.4, there is another subdivision called *val* but that will not be necessary for this project). Once the split is done, a model is constructed. Finally, the quality of that model is evaluated using the test set and applying the F1 score.

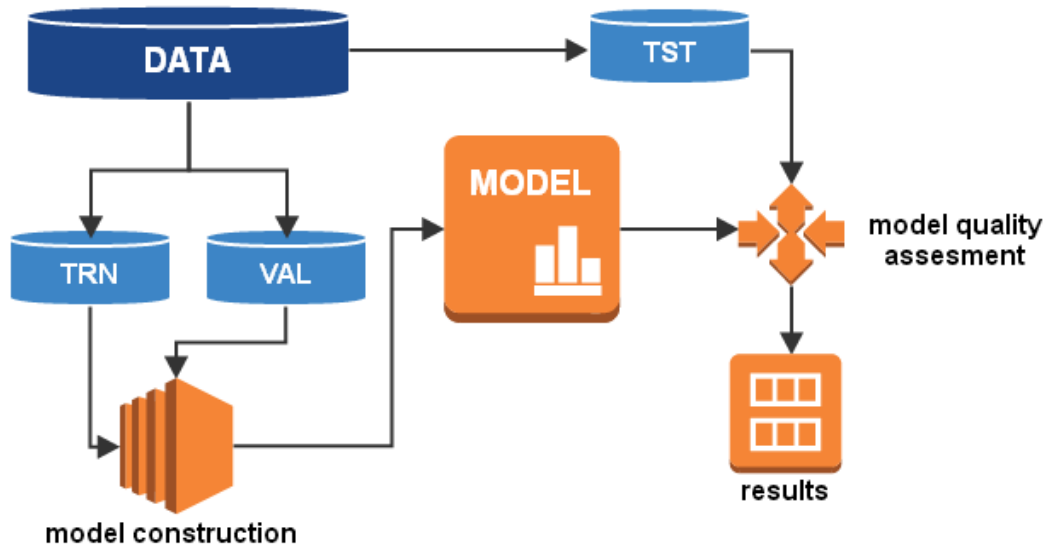


Figure 4.4: Model Construction [1]

4.7 Parameter Optimization

This section will cover how was the parameter optimization performed and what were the results obtained for each classifier.

To understand the parameter optimization, first, we must comprehend how is the classifier built in this project.

The simplest definition of text classification is that it is a classification of text based on the content of that text. Now, in general, all the machine learning methods and algorithms are written for numeric features/variables. One of the most important problems with text corpus is how to represent text as numeric features [21]. That stage is accomplished by making use of the pipeline described in section 4.4. Depending on how the feature extraction pipeline was created, the parameters of the final classifier change. Therefore, the classifier is assembled by both the pipeline described in section 4.4 and one of the classifiers chosen in section 4.5.

In order to find the right classifier parameters, it must be carried out a parameter search for the pipeline parameters and the classifier parameters.

From section 4.4, the pipeline extracts three main features; ngrams, words (custom tokenizer) and topics (LDA). From these features, the only parameters that will be searched are the number of topics from LDA and the number of ngrams.

From section 4.5, the parameters searched will depend on the type of classifier. On the following subsections from this section, the topics will be covered. The parameter search

will be done using the GridSearchCV from the Scikit Learn class. The principle of the GridSearchCV is:

1. Define a dictionary of parameters to change.
2. Fit the grid search with the training dataset (this process can take a very long time).
3. Print best score alongside with the best parameters.
4. Print the metrics accuracy score by testing the classifier with the best parameters found. For this, it is required a test dataset.

4.7.1 Grid Search

A grid search is an exhaustive search through a manually-specified subset of the hyperparameter space. The grid search algorithm requires a performance metric, such as cross-validation error or validation-set error to evaluate the best possible parameter [32]. The grid search method comes from the Scikit Learn class.

A grid search was optimized for each classifier to get the best performance. During the process of optimization not only the parameters of the classifiers were optimized, but also the type of features to extract.

The results found extracting ngrams, lexical features and the number of topics shall be presented alongside with the scores found extracting ngrams, lexical features, syntactic features and number of topics. The optimized parameters are the same either for the four feature extraction than for the three feature extraction.

4.7.1.1 Multinomial Naive Bayes

- **Parameters to optimize:**

1. *Alpha*. Belongs to classifier
2. *Number of topics*. Belongs to pipeline

- **Results**

4.7.1.2 Support Vector Machine

- **Parameters to optimize:**

1. *C*. Belongs to classifier
2. *Number of topics*. Belongs to pipeline
3. *Number of ngrams*. Belongs to pipeline

- **Results**

Variable	Precision	Recall	F1-Score	Support
0	0.96481/0.87704	0.96765/0.86740	0.96623/0.87219	1360/1365
1	0.96873/0.87262	0.96598/0.88193	0.96735/0.87726	1411/1406
avg/total	0.96680/0.87480	0.96680/0.87477	0.92602/0.87476	2771/2771

Table 4.6: Multinomial Naive Bayes Classification Report Optimized. Three features extracted/Four features extracted

Variable	Precision	Recall	F1-Score	Support
0	0.91131/0.89993	0.90662/0.89597	0.90896/0.89794	1360/1365
1	0.91044/0.89943	0.91495/0.90327	0.91269/0.90135	1411/1406
avg/total	0.91086/0.89968	0.91086/0.89968	0.91086/0.89967	2771/2771

Table 4.7: Support Vector Machine Classification Report Optimized. Three features extracted/Four features extracted

4.7.1.3 K-Nearest Neighbors

Since the results of the k-NN were very low (see table 4.4) there was no grid search performed to predict the best classifier parameters. Furthermore, k-NN it is not a recommended classifier for the purpose of the project.

4.7.1.4 Logistic Regression

- **Parameters to optimize:**
 1. *Number of topics.* Belongs to pipeline
 2. *Number of ngrams.* Belongs to pipeline
- **Results**

Variable	Precision	Recall	F1-Score	Support
0	0.92819/0.90767	0.90361/0.89304	0.91573/0.90030	1359/1365
1	0.90953/0.89776	0.93272/0.91181	0.92098/0.90473	1412/1406
avg/total	0.91868/0.90264	0.91844/0.90256	0.91841/0.90254	2771/2771

Table 4.8: Logistic Regression Classification Report Optimized Three features extracted/-
Four features extracted

4.8 Senpy

Senpy was used in this project as a framework for wrapping the classifier work in a plugin, providing an interactive interface where the user only needs to type in text and senpy detects the sarcastic property of the text.


To make senpy work, the following was required:

- Install senpy using command line³
- Integrate the classifier as a plugin

4.8.1 Results

The following pictures illustrate the result of using senpy using as examples a sarcastic sentence and a non-sarcastic one.

³<https://github.com/gsi-upm/senpy>

 **Senpy** Playground

v0.1.4

About Test it

Hola me llamo ironia

Select the plugin.

sarcasm-twitter +

Plugin to detect sarcasm in Tweets

> Basic API parameters

▼ Plugin extra parameters

language es

Analyse!


/api?algo=sarcasm-twitter&i=Hola%20me%20llamo%20ironia&conversion=full&expanded-jsonld=false&informat=text&intype=direct&outformat=json-ld&urischeme=RFC5147Stri

Viewer Raw

```

▼ object {5}
  @context : http://localhost:5000/api/contexts/Results.jsonld
  @id : _:Results_1544546619.8002653
  @type : results
  ▼ analysis [1]
    0 : endpoint:plugins/sarcasm-twitter_1.0
  ▼ entries [1]
    ▼ 0 {3}
      @id : Entity0
      is_ironic : Ironic
      text : Hola me llamo ironia

```



This development has been partially funded by the European Union through the MIXEDEMOTIONS PROJECT (project number H2020 655632), as part of the RIA ICT 15 Big data and Open Data Innovation and take-up programme.


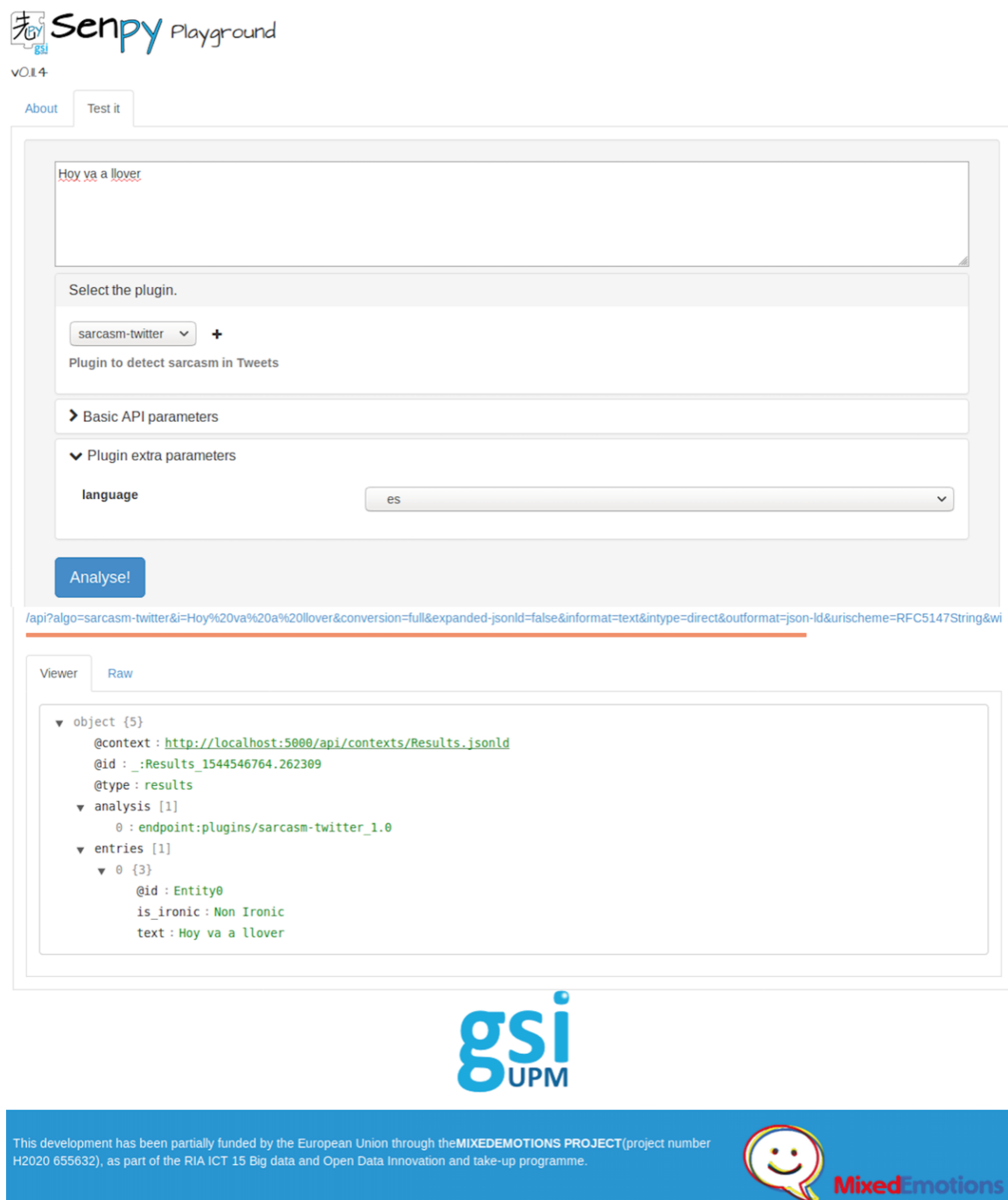
 **MixedEmotions**

Figure 4.5: Sarcastic example using Senpys' Playground [31]



The screenshot displays the Senpys Playground web interface. At the top, the logo "Senpy Playground" is visible, along with "v0.1.4" and tabs for "About" and "Test it". The main input area contains the text "Hoy va a llover". Below this, a section titled "Select the plugin." shows a dropdown menu with "sarcasm-twitter" selected and a "+" button. A description below reads "Plugin to detect sarcasm in Tweets".

Underneath, there are two expandable sections: "Basic API parameters" and "Plugin extra parameters". The "Plugin extra parameters" section is expanded, showing a "language" dropdown menu set to "es". A blue "Analyse!" button is positioned below these settings.

The output area, labeled "Viewer" and "Raw", displays a JSON response. The URL bar shows the API endpoint: `/api?algo=sarcasm-twitter&i=Hoy%20va%20a%20llover&conversion=full&expanded-jsonId=false&informat=text&intype=direct&outformat=json-ld&urischeme=RFC5147String&wi`.

```

{
  "@context": "http://localhost:5000/api/contexts/Results.jsonld",
  "@id": "_:Results_1544546764.262309",
  "@type": "results",
  "analysis": [
    {
      "endpoint": "plugins/sarcasm-twitter_1.0"
    }
  ],
  "entries": [
    {
      "@id": "Entity0",
      "is_ironic": "Non Ironic",
      "text": "Hoy va a llover"
    }
  ]
}

```

At the bottom of the interface, the "gsi UPM" logo is displayed. A blue footer bar contains text about funding from the European Union through the MIXEDEMOTIONS PROJECT (project number H2020 655632), as part of the RIA ICT 15 Big data and Open Data Innovation and take-up programme. To the right of the text is the "MixedEmotions" logo, which features a stylized face with a rainbow smile.

Figure 4.6: Non-sarcastic example using Senpys' Playground [31]

4.9 Elasticsearch

Elasticsearch was implemented in this project as a place to store the tweets. The advantage of using Elasticsearch is its versatility to introduce new improvements to the project, such as implementing a dashboard using Sefarad [5].

Conclusions and future work

The conclusions reached in the previous chapter will be brought together in this chapter. Moreover, a description of the achieved goals and the lines of future work will be explained. Finally, the problems faced will be presented.

5.1 Conclusions

This project resembles an example of automatic sarcasm detection. The approach followed was training a classifier using a dataset consisting of two columns, one for the text body and the last column to feature the sarcastic property of the text body. This goal was fulfilled by making use of ML techniques and NLP techniques, with the ability to detect sarcasm in short text, such as Twitter.

To compile the final dataset, three datasets containing tweets were assembled and put together generating a final dataset with the tweet body and the feature.

Later on, the model could be built by introducing a preprocessing stage designed to encode categorical variables and delete tweets lacking a text body. To ensure a non-dependency between the tweets, the rows of the dataset were randomized.

Once the dataset was ready a pipeline was designed, which would be responsible for the feature extraction process. The features that gave more information to the classifier were:

1. **Lexical features:** The word repetition frequency had a relationship with the sarcastic

property of a short text.

2. **Ngram features:** This feature shows that depending on how many words uses the algorithm to consider a noun, the model can be more precise.
3. **Number of topics feature (LDA):** Taking into account the possible number of topics that a text can be talking about is a relevant characteristic of sarcasm.

As an interesting fact, the model predicted with more accuracy having only these three features extracted rather than counting on the syntactic features.

The next step was to feed the features extracted to a classifier. According to [26], a classifier is a kind of rule-based system with general mechanisms for processing rules in parallel, for adaptive generation of new rules, and for testing the effectiveness of new rules. The classifiers which had an outstanding performance were:

1. **Multinomial Naive Bayes:** This classifier was by far the best one, reaching an F1 score of 0.92.
2. **Logistic Regression:** This classifier had an F1 score of 0.91.
3. **Support Vector Machine:** This classifier had an F1 score of 0.91.

The three classifiers were trained using a training set and the F1 score was extracted by using a testing set. Additionally, the K-Nearest Neighbors classifier was also implemented but it had a very bad performance.

Finally, the senpy engine was implemented to provide a user-friendly interface allowing the user to insert text and get a message saying if the tweet is sarcastic. An example can be found in Figure 4.5.

5.2 Achieved goals

The main goals achieved in this project were:

Construction of a ML classifier capable of detecting sarcasm in short texts, such as tweets.

This was the main objective of this project.

Implementation of the classifier in a web engine which allows any user to detect sarcasm in a user-inserted text.

This goal could be achieved because of the Senpy engine.

5.3 Future work

In regard to the future development of the model, the following improvements could be done:

1. **More feature detection.** In sum, only four features have been extracted in this project, however, more features could give more information concerning sarcasm. However, this classifier has shown a better performance when analyzing three features rather than four. Therefore, if many more features are considered, the performance of the classifier may be degraded due to the overfitting phenomenon.
2. **Take into account previous tweets.** It is very common that tweets are used to answer other tweets. Sometimes the sarcasm is hidden in an answer to a message. In this project, that possibility was never considered, deeming a sarcastic tweet only if the tweets' text field is sarcastic. It could be a good improvement to consider the text field of a tweet and also the predecesing tweets' text field.
3. **Implement a dashboard.** In section 4.9 it was discussed the possibility of using the elastic search engine to generate a dashboard and offer a better interface to the user. To do so, the Sefarad engine offered by the Intelligent Systems Group at the UPM (GSI) is very useful. In [5] more information regarding the dashboard can be found.

5.4 Problems encountered

The problems faced in this project shall be explained

1. **Technologies used:** At the beginning of the project, it was required that I learned all the technologies used as well as the libraries implemented. This required a considerable amount of time.
2. **Retrieval of tweets:** The provided datasets contained only a column indicating the tweet id and sarcastic value ('True' or 'False'). To retrieve the text bodies that would later be fed to the classifier, those text bodies were obtained by using the Bitter(section 3.5) software.

Cost of System

A.1 Introduction

This appendix is committed to giving an overview of the economic requirements that this project presents.

A.2 Physical requirements

The only physical device needed for the development of the project is composed of a computer capable of handling the ML tasks and the deployment of the system. Since the requirements are not demanding there are many other possibilities than the one given in this appendix.

The technical characteristics capable of providing enough computational power for the needs presented in this project could be:

- *RAM*: 8 GigaByte
- *Hard Drive Disk*: 500 GigaByte
- *Processor*: Intel Core I7 with 4 cores

Considering the market of today and fulfilling the technical characteristics explained above, a computer of €800 would be eligible.

A.3 Human resources

This section will explain the cost of having a programmer to develop the system and maintain it.

Considering that this project is made out of twelve ECTS and each ECTS accounts for 27 hours, the total amount of time invested in this project are at least 324 hours. Considering the average salary in Madrid (€8.12/hour), the project would cost around €2630.

Moreover, a person dedicated to the maintenance should also be considered. To that end, someone with knowledge about ML and NLP, like a Telecommunication Engineer, could be hired. In that case, a salary of €24000 could suffice.

A.4 License

This project has been done without any private license. Therefore, there is no need to spend money on a software license. This is the main benefit of using open source software.

Impact of the Project

B.1 Introduction

Bibliography

- [1] Classification models. <http://algolytics.com/tutorial-how-to-determine-the-quality-and-correctness-of-classification-models-part-1-introduction/>.
- [2] Hacia la detección de ironía en textos cortos. http://ru.ffyl.unam.mx/bitstream/handle/10391/4787/VII_CoLiCo_G_Jasso_Mesa_3_2015.pdf?sequence=2&isAllowed=y.
- [3] Luigi documentation. <https://media.readthedocs.org/pdf/luigi/latest/luigi.pdf>.
- [4] Number of twitter users worldwide from 2014 to 2020 (in millions). <https://www.statista.com/statistics/303681/twitter-users-worldwide/>.
- [5] Sefarad architecture. <https://sefarad.readthedocs.io/en/latest/sefarad.html>.
- [6] Spark dataframes and ml pipelines. <https://www.slideshare.net/databricks/dataframes-and-pipelines>.
- [7] Twitter by the numbers: Stats, demographics and fun facts. <https://www.omnicoreagency.com/twitter-statistics/>.
- [8] Charu C Aggarwal and Chandan K Reddy. *Data clustering: algorithms and applications*. CRC press, 2013.
- [9] Huseyin Akdogan. *Elasticsearch Indexing*. Packt Publishing Ltd, 2015.
- [10] Abhishek Andhavarapu. *Learning Elasticsearch*. Packt Publishing Ltd, 2017.
- [11] Oscar Araque. Design and Implementation of an Event Rules Web Editor. Trabajo fin de grado, Universidad Politécnica de Madrid, ETSI Telecomunicación, July 2014.
- [12] Rounak Banik. *Hands-On Recommendation Systems with Python: Start building powerful and personalized, recommendation engines with Python*. Packt Publishing Ltd, 2018.
- [13] Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda. Applied text analysis with python: Enabling language-aware data products with machine learning. 2018.
- [14] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.
- [15] Giuseppe Bonaccorso. *Machine Learning Algorithms*. Packt Publishing Ltd, 2017.
- [16] Peter Bruce and Andrew Bruce. *Practical Statistics for Data Scientists: 50 Essential Concepts*. " O'Reilly Media, Inc.", 2017.
- [17] Giuseppe Ciaburro. *MATLAB for Machine Learning*. Packt Publishing Ltd, 2017.

- [18] Giuseppe Ciaburro. *Regression Analysis with R*. 2018.
- [19] Bharvi Dixit. *Elasticsearch Essentials*. Packt Publishing Ltd, 2016.
- [20] Nitin Hardeniya. *NLTK essentials*. Packt Publishing Ltd, 2015.
- [21] Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur. *Natural Language Processing: Python and NLTK*. Packt Publishing Ltd, 2016.
- [22] Param Jeet and Prashant Vats. *Learning Quantitative Finance with R*. Packt Publishing Ltd, 2017.
- [23] Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):73, 2017.
- [24] Russell Journey. *Agile Data Science 2.0, 1st Edition*. " O'Reilly Media, Inc. ", 2017.
- [25] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579*, 2017.
- [26] Pascal Bugnion; Patrick R. Nicolas; Alex Kozlov. *Scala:Applied Machine Learning*. 2017.
- [27] Sinan Ozdemir and Divya Susarla. *Feature Engineering Made Easy: Identify unique features from your dataset in order to build powerful machine learning systems*. Packt Publishing, 2018.
- [28] Josh Patterson and Adam Gibson. *Deep Learning: A Practitioner's Approach*. " O'Reilly Media, Inc." , 2017.
- [29] Tomáš Ptáček, Ivan Habernal, and Jun Hong. Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223, 2014.
- [30] J. Fernando Sánchez-Rada. Design and Implementation of an Agent Architecture Based on Web Hooks. Master's thesis, ETSIT-UPM, 2012.
- [31] J Fernando Sánchez-Rada, Carlos A Iglesias, Ignacio Corcuera, and Oscar Araque. Senpy: A pragmatic linked sentiment analysis framework. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pages 735–742. IEEE, 2016.
- [32] Dipanjan Sarkar, Raghav Bali, and Tamoghna Ghosh. *Hands-On Transfer Learning with Python: Implement advanced deep learning and neural network models using TensorFlow and Keras*. Packt Publishing Ltd, 2018.
- [33] Shahid Shayaa, Noor Ismawati Jaafar, Shamshul Bahri, Ainin Sulaiman, Phoong Seuk Wai, Yeong Wai Chung, Arsalan Zahid Piprani, and Mohammed Ali Al-Garadi. Sentiment analysis of big data: Methods, applications, and open challenges. *IEEE Access*, 6:37807–37827, 2018.
- [34] Mark G Sobell and Matthew Helmke. *A practical guide to Linux commands, editors, and shell programming*. Prentice Hall Professional Technical Reference, 2005.
- [35] Bhargav Srinivasa-Desikan. *Natural Language Processing and Computational Linguistics*. Packt Publishing Ltd, 2018.

- [36] Sergios Theodoridis. *Machine learning: a Bayesian and optimization perspective*. Academic Press, 2015.