# Automatic generation of comments on twitter based on news

Carlos Casar Morejon

Director: Javier Bejar
Specialization: Computation

# Contents

# 1 Context and scope of the project

## 1.1 Context and problem formulation

In our modern society, the information is playing an important role. We receive tons of information every day by many different devices. In addition to the classic media like television and radio, we have now social media [1] [2]. Nowadays an important part of the citizens get the news from social media and in this fast-paced society, it is important to have good resumes of the news.

As some studies show [3], the time spent reading news online is pretty low in average compared to print news. To keep the attention of the users, the news must be short enough to be fully read by the users, that is the reason that we need good resumes.

One of the most trendy fields in the past years in computation is Machine Learning, a branch of Artificial intelligence. The reason for his popularity is the power and scalability on it. On the last few years, Machine learning has been used in a wide range of applications[4][5][6].

Machine Learning allows the machine to learn and make decisions by itself. It takes large amounts of data to do that, that is why Machine Learning is the answer to the computation problems where the explicit programming is not possible. That is the reason why in this days, the information age, Machine Learning is a trend.

The objective of this project is to create an AI with machine learning techniques, to be able to comment news on twitter. To perform that, first, the application will use the classification method to categorize the news. Secondly, the machine, with some clustering method, will categorize tweets from people that comments news. With that training the machine will be able to categorize news and then comment on them by itself.

## 1.2 Actors

The development of the project involves several actors, which are listed and described in the following lines.

**Developer:** Is the person in charge of research, implement and document the whole project. In this project, there is only one developer, myself, and I am the person in charge to accomplish the deadlines.
**Director:** Is the main responsible for guiding and giving advice to the developer. In this project, Javier Bejar, from the Computer Science department, has acted as director.

**Beneficiaries:** The main beneficiary of this project is an online newspaper, that wants a bot that comments news by itself. Also, it can be the interest of different kind of researchers working in related topics or using similar techniques.

## 1.3 State of the art

To put it briefly, machine learning is a subject undergoing intense study, so there is a lot of research and improvements going on these days. Since machine learning has become a thing, almost every field has been affected, from healthcare to financial trading to smart cars. Drawing conclusions from texts has not been an exception [7][8], being capable of analyze large amounts of data has always being important.

One example of a real application drawing conclusions are the virtual assistants,
like the bot Alexa from amazon, that after each sentence draws a conclusion and answers.

## 1.4 Objectives

The main objective of our project is to develop a tool that is capable of creating new and original comments of news on Twitter. We aim to design the program in a way that the sentences of the comments make syntaxes sense.

Another objective for this project is to have a high accuracy with the machine learning algorithms, in order to avoid critical mistakes like classify a sports news into a political news. A high accuracy will reduce the possible errors and will make easy figure out where the mistakes are made.

An additional objective is to find several data sets for the training of the classification algorithm. This is an important part of the algorithm because a good data set will carry on to good results.

## 1.5 Scope

In order to be able to solve the problem that defines our project first, we will program a machine learning code to clasificate news, from sports to politics to science. First, we need to find a good amount of data, to train our machine. That data has to contain, at least, the text of the new and the data type. Also the data has to come from an official and reliable source, like a university[9]. If we don't find an appropriate set of data, we will have to train with the deep learning technique, so the machine will classify the data only with the text.

One more important algorithm is the comunication with Twitter, that will enable us to collect data and Tweet our comments. For this algorithm we will use the Tweepy library for accessing the Twitter api. As we will explain in the comment algorithm, this library will help to the creation of comments, collecting tweets that comment news.

When our machine is able to recognize different types of news, and we are able to communicate properly with twitter, we will proceed with the comments. First, we need to create another machine learning code, this time training with tweeters that usually comment news on twitter, using the Tweepy library that we mention before. These tweets will be classified in the same way as the news are, therefore, we will have for each type of new,

many different answer. With all of this data collected, we will try to generate a comment for a new, first classifying it, and then, generating the comment.

## 1.6 Methodology

Since the timetable is really tight, the project will be developed in an agile methodology. This fast and flexible methodology will help us develop the project.

By using an agile method, we will be able to review regularly to have the project in good control. Furthermore, we will fix little goals each week to accomplish, in order to keep the project fresh and not digress from the goal

### 1.6.1 Development tools

For the development of the main program we will use Python, with his library tweepy, that allows us to work with the api of Twitter. Additionally, we will use many math and machine learning libraries, to make the coding and algorithms easier. The main reasons for using Python above other high level languages is his deep documentation about twitter bots and machine learning.

The monitoring tool to keep track the development of the project is going to be github. This will improve the communication about the code between director and developer.

### 1.6.2 Validation of results

To check if the program is working as we want, we will have periodical meetings with the director to check the results.

## 1.7 Possible obstacles of the project

### 1.7.1 Data sets

One of the main problem we can have with the project is not finding the appropriate data sets for training our machine. If this occurs, we would probably change to a deep learning approach, where the data sets we need are easier to find.

### 1.7.2 Classificatory Algorithm

The main obstacle that we can find while working on this algorithm is reaching a high accuracy classification the topics of the news.

### 1.7.3 Comment Algorithm

The comment generator algorithm is probably the hardest task of this project. It will need a lot of accuracy in order to make sentences that make sense, and that means a lot of work making an optimized code.
Another possible problem for the algorithm is to find a good amount of tweets that comment news. Additionally, these comments have to be of high quality, which makes this task even more difficult.
Also. we could also find some problems creating comments with only machine learning, more than likely we should help the construction of sentences with other algorithms.

### 1.7.4 Bugs

Considering that the machine learning framework is complex, and we will have long algorithms, we will make unit test to ensure that there is no bugs.

# 2 Temporal planning

This section will describe the tasks that are going to take place in the project and the time to accomplish them. An action plan will be provided in order to finish the project in the desired time. However, the initial planning could be revised and modified as a result of the evolution of the project.

The estimated project duration is of about 4 months and a half, starting in 1st September, 2017 and the deadline is on 21st January, 2017.

## 2.1 Task description

### 2.1.1 Learning machine learning

The first step toward this project is getting a wide background in machine learning, due to is the core of the algorithms in this project. I started by reading this online book[10], that gave me a general view and the basics to understand the topic. Afterwards, I continue with an online practice tutorial[11], that introduced me on the programming part of the machine learning.
This process of learning took a month for the reason that learning from 0 machine learning is not an easy task.
This task did not requiered any material resources, but it did requiered human resources to read and understand all the information about machine learning.

### 2.1.2 Project planning

This is the task covered by the GEP course, that defines what is going to be done and how is going to be done in this project. It can be divided in the following three stages:

- Context and scope of the project
- Project planning
- Budget and sustainability

This task did not requiered any material resource(except the material about the GEP course in atenea), but it did requiered human resources to research, write and understand all the information about machine learning.

### 2.1.3 Tweepy library

Tweepy is a Python library for accessing the Twitter API. This library is core for our project, because we need to retrieve information for the learning process of our machine and also Tweet our comments. Our main source for learning how to use this library is the official documentation for Tweepy[12].

As material resource, the Python idle is needed in order to practice.This task also requires human resources to read and understand all the information about the library.

### 2.1.4 Data sets

As we already explain in the process to create a machine learning algorithm, we need large data sets, with the proper structure, to train our program to be able to recognize the topic of a new. Searching this data sets is not a simple task, we need some official and reliable repositories from a good source.
This task did not requiered any material resources, but it did requiered human resources to read and research all the information about data sets.

### 2.1.5 Recognize the topics of news algorithm

For this first algorithm we will use the classification method, to separate the news by topics. There is many and different algorithms for the classification method, but for this specific project we think that k-nearest neighbors[13] algorithm fits better with our project. With the proper data set, this algorithm will work perfectly, because every new of the same topic has similarities, then this will ensure a high accuracy.
However, if the accuracy of the algorithm is not as high as we want, we will have to try other algorithms, looking for a higher accuracy.
For the data sets, we will try more than one, looking for the best performance, maybe training with more than one at the time.
Once the algorithm is working correctly, we will do some test to prove is working as we want. We will try with some news from newspapers to find if the algorithm decides his topic correctly.
As material resource, the Python idle is needed in order to code.This task also requires human resources to code and understand.

### 2.1.6 Communication with twitter

As we said in the 2.1.2, we will use the Tweepy library to use the Twitter API. First, we will program a function to Tweet things, for the comment section of the project.
Afterwards, we will program the information retrieval which will be used to relate news and tweets. With this information, we will be able to construct the second algorithm, the comment generator.
As material resource, the Python idle is needed in order to code.This task also requires human resources to code and understand.

### 2.1.7 Comment generator

This is the main task of this project, it will be the visible part of the project that shows that all the previous algorithms works. It will be in charge of comment news from a newspaper and Tweet it.
This algorithm has to be very precise, because as we said, it will be the one who generates the comments and any grammatical error or nonsense sentence will be terrible. For this algorithm, we will use again a classification algorithm.

First, it will select news from an online newspaper. Next, with the 2.1.5 algorithm, it will decide the topic of the news.
Secondly, we will choose some recognised tweeter, who usually comments news on Twitter, to train our machine learning algorithm. Once we have trained our machine with several tweeters, the machine will have a proper base to create comments and it will be ready for testing and see how the algorithm behave. If the algorithm is not working properly, we have three ways to continue:

- Continue the machine learning with more tweeters.
- Combine machine learning with normal code for sentence structure.
- Change the classification method to another one.

We will repeat this process until the algorithm works as we want to.
As material resource, the Python idle is needed in order to code.This task also requires human resources to code and understand.

### 2.1.8 Final task

In this task we are going to look over the whole project works as expected and we are going to prepare the delivery of the project, including the documentation and preparing the final presentation.

## 2.2 Time table

The table 1 is an estimation of the time spent in each task described in the previous section.

| Task | Estimated duration(Hours) |
|---|---|
| Learning machine learning | 80 |
| Project planning | 80 |
| Tweepy library | 20 |
| Data sets | 25 |
| Recognize the topics of news algorithm | 75 |
| Communication with twitter | 25 |
| Comment generator | 115 |
| Final task | 40 |
| Total | 460 |

## 2.3 Gantt chart

Figure 1 shows a Gantt chart of the different task of the project.

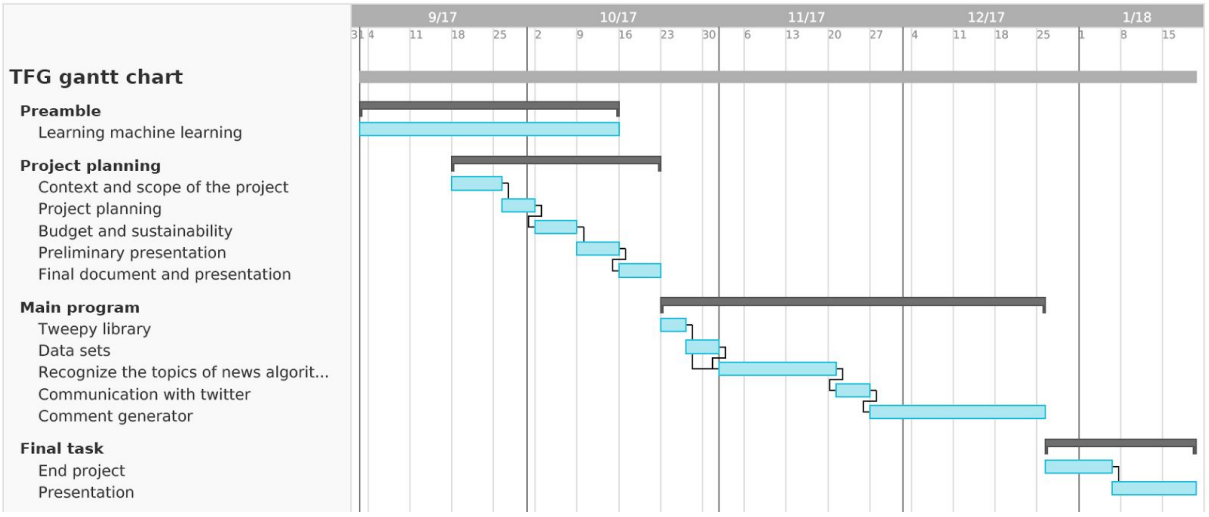


Figure 1: Gantt chart of the project

## 2.4 Alternatives and Action plan

In this section we are going to describe how are we going to execute the planning that we made in this section.

The idea is to work as we have planned, doing each task at the time we point out, but we know that every project has delays and obstacles that may make difficult to stick to the plan. If the problems occur and we run out of time we are going to try to get only a basic version only. Certainly, as every project we have priorities. We must have a working algorithm in most cases, that can comment any new properly. We will try to have the best accuracy possible in the machine learning algorithms, but having a working program is the goal.

We are going to have meetings with the director every time that an important stage of the project is finished.

Next, some examples of potential sources of delays and alternatives are mentioned.

### 2.4.1 Sources of information on Twitter

One of the biggest problems that we can have in this project is to find reliable comments on Twitter to build our machine learning algorithm for the comments. This could cause delays until we find a proper source of information or an alternative aproach.

### 2.4.2 Optimization problems

A bad selection of the optimizations to be done could result in a waste of time without a significant gain. This could cause delays, because optimizing the code is fundamental in the development of machine learning codes.

# 3 Budget and sustainability

## 3.1 Budget estimation

In this section, we are going to do a budget estimation for our project. In this document, we will take into account three resources: Hardware, software and human resources. At the end of the section we are going to show the total budget of the three resources.

To calculate the amortization of each resource, we are taking into account two factors, the first one being the useful life and the second one being the duration of our project, that will go on for 4 months. Also, the usual life duration will be, according to the tax office, 3-4 years for the hardware and 2-3 for the software.

### 3.1.1 Hardware resources

Table 1 contains the cost and the amortization of the hardware that we are going to use in the project.

| Product | Price (€) | Useful life | Amortisation (€) |
|---|---|---|---|
| PC (included all the needed devices) | 800.00 | 4 years | 43.83 |

Table 1: Amortisation and prince for the hardware products.

### 3.1.2 Software resources

Table 2 contains the cost and the amortization of the software that we are going to use in the project.

| Product | Price (€) | Useful life | Amortisation (€) |
|---|---|---|---|
| Windows 10 Pro | 260.00 | 3 years | 18.99 |
| Python IDLE | 0.00 | 3 years | 0.00 |
| Github | 0.00 | 3 years | 0.00 |
| Google Docs | 0.00 | 3 years | 0.00 |
| TeamGantt | 0.00 | 3 years | 0.00 |
| **Total** | 260.00 | - | 18.99 |

Table 2: Price and amortization for the software products

### 3.1.3 Human resources

This project is going to be developed by one person. For that reason, this person will be the Project manager, Software developer and tester. The 460 hours of the project will be distributed between the 3 roles. In table 3 and estimation of human resources is showed.

| Role | Price per hour (€/h) | Time (h) | Cost (€) |
|---|---|---|---|
| Project manager | 30 | 90 | 2700 |
| Software developer | 20 | 310 | 6200 |
| Tester | 15 | 60 | 900 |
| Total | - | 460 | 9800 |

Table 3: Cost estimation by role.

Following, table 4 provides the exact time that each role spends in the different task of the project that we previously have defined.

| Task | Duration (hours) | Dedication (hours) | | |
|---|---|---|---|---|
| | | Project manager | Software developer | Tester |
| Learning machine learning | 80 | 0 | 80 | 0 |
| Project planning | 80 | 80 | 0 | 0 |
| Tweepy library | 20 | 0 | 20 | 0 |
| Data sets | 25 | 0 | 25 | 0 |
| Recognize the topics of news algorithm | 75 | 0 | 55 | 20 |
| Communication with twitter | 25 | 0 | 20 | 5 |
| Comment generator | 115 | 0 | 85 | 30 |
| Final task | 40 | 10 | 30 | 5 |
| Total | 460 | 90 | 310 | 60 |

Table 4: Time estimation by role and task.

### 3.1.4 Total budget

Table 5 shows the total cost of the project, using the data shown in tables 4 3 and 2.

| Concept | Cost |
|---|---|
| Hardware resources | 800 € |
| Software resources | 260 € |
| Human resources | 9800 € |
| Total | 10860 € |

Table 5: Total budget cost

## 3.2 Budget control

As previously mentioned, our budget will need modifications if we can't follow the established plan.

We could have difficulties in our project, but is improbable that we need more hardware resources aside for the ones already mentioned. We might need more software resources for the development of the project, but there is plenty of free options for the software development.

The most difficult tasks in this project are the main machine learning algorithms, which could take longer to develop. This task involves the software developer and the tester, so we have to take into account that the money spent on them could grow if the problem becomes harder than we expected.

## 3.3 Sustainability

In this section we are going to measure the sustainability of our project in this three dimensions: Economic dimension, social dimension and environmental dimension. This measure will be based on the application of the sustainability matrix, as shown on table 6.

| | **PPP** | **Useful life** | **Risks** |
|---|---|---|---|
| **Envoirimental** | Design consumption | Ecological footprint | Environmental risks |
| | 9/10 | 16/20 | - |
| **Economical** | Bill | Viability plan | Economical risks |
| | 7/10 | 15/20 | - |
| **Social** | Personal impact | Social impact | Social risks |
| | 9/10 | 14/20 | - |
| **Sustainability range** | 25/30 | 45/60 | - |
| | 70/90 | | |

Table 6: Sustainability matrix

### 3.3.2 Environmental dimension

The development of this project uses the minimum amount of resources possible, only the electricity required for the PC to work. Therefore, searching alternatives to reduce the consumption is pretty much impossible. Also, the reuse of resources in this project is difficult too.

The project will make automatic comments on news by Twitter, which has a very little consumption of electricity.

### 3.3. Economic dimension

A detailed budget has been done for this project, including material and human resources as shown in previous sections of this document.

The proposed solution will be less expensive than current solutions from an economic point of view, because of the efficiency of the algorithms and his performance it will be cheaper in time and energy for the final user.

### 3.3.3 Social dimension

The execution of this project has teached me how hard is to develop a big project, and to put some perspective to the problems. Obviously, the project entails the knowledge and improvements in programming machine learning techniques.

As mentioned in the introduction section, machine learning is a subject undergoing intense study. For that reason, all the researchers that work on this field could benefit from this project.

# 4 References


[1] “News Use Across Social Media Platforms 2016”
http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

[2] “Social Media Use for News and Individuals' Social Capital”
http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2012.01574.x/full

[3]”National press online readers average 30 seconds per day versus 40 minutes for print”
http://www.pressgazette.co.uk/study-national-press-online-readers-average-30-seconds-per-day-versus-40-minutes-for-print/

[4] "Machine learning in genetics and genomics - NCBI - NIH."
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5204302/.

[5] "Machine learning applications in cancer ... - ScienceDirect.com."
http://www.sciencedirect.com/science/article/pii/S2001037014000464.

[6] "End to End Learning for Self-Driving Cars." 25 abr.. 2016,
https://arxiv.org/abs/1604.07316.

[7] "Thumbs up? - ACM Digital Library - Association for Computing ...."
http://dl.acm.org/citation.cfm?id=1118704.

[8] "Text categorization with Support Vector Machines... - ResearchGate."
https://link.springer.com/chapter/10.1007%2FBFb0026683?LI=true.

[9] “UC Irvine Machine Learning Repository”
https://archive.ics.uci.edu/ml/datasets.html

[10] "Introduction to Machine Learning - Alex Smola."
http://alex.smola.org/drafts/thebook.pdf

[11] "Machine Learning with Python tutorial series - PythonProgramming.net."
https://pythonprogramming.net/machine-learning-tutorial-python-introduction/.

[12] "Tweepy Documentation"
http://docs.tweepy.org/.

[13]"A Branch and Bound Algorithm for Computing k-Nearest Neighbors."
http://ieeexplore.ieee.org/abstract/document/1672890/