

Final project in NLP

Guide: DR Nava Shaked

**Should there be culture-specific approaches to depression?**



**Daniel Sabba**

**Omer Pesah**

**Table of Content:**

1. Introduction……………………………………………………….…….… **3**
2. Experiment steps and design………………………………….…...**4**
   1. Gather data…………………………………………………….…...**4**
   2. Basic feature extraction……………………………….…….…**4**
   3. Basic Text Pre-processing……………………..………………**5**
   4. Advance Text Processing……………………………………...**5**
3. Advance Text Processing……………………………………………...**5**
4. Experiments results……………………………………………………...**6**
   1. Basic features…………………………………………………….…**6**
   2. Time differences……………………………………………….….**7**
   3. Bi-grams……………………………………………………………....**9**
   4. POS tagging…………………………………………………………**10**
   5. Most common words………………………………………....**11**
   6. sentiment analysis……………………………………………...**13**
   7. Topic modelling……………………………………………….…**15**
5. Experiment conclusions……………………………………….…….**17**
6. Bibliography……………………………………………………………….**18**
7. **Introduction**

The topic of our project is depression across cultures.

Major depression constitutes a serious challenge in personal and public health, millions of people each year suffer from depression and The World Health Organization now ranks major depression as one of the most burdensome diseases in the world.

As we see the serious damaging depression does to the society, we know we have the deal with it.

One of the ways to deal with depression is with adequate treatment, but first, we will need to find those people, with numbers growing as we mentioned above, this task is really hard.

One of the ways to find those people who suffer from depression is from social networks, Tweeter, Facebook, Reddit and more, as we advance in technology, there are more and more tools from the fields of AI, NLP, and ML, many of these fields have many tools that does a great job in text analysis, and we will even use some in our project, they can predict whether an individual is depressed or not.

1. **Experiment steps and design**

**2.1** The first stage of our project is to gather data

How we collected the data, and what kind of data we focused on getting.

The data we collected are posts from forums of depression.

Because we explore depression across cultures, we needed data in several languages, we chose English, Spanish, and French.

We built several crawlers to scarp those forums to get our posts, while focusing on getting long posts, with a lot of words and expressions.

**2.2** Basic feature extraction using text data

To gather insights from the cleaned data, we had to do some manipulations on the text, e.g. word count, char count, average word length, counting the stop words, the lexical diversity of the posts, the date that the post was written, etc.

**2.3** Basic Text Pre-processing of text data

First, we have raw text, that is filled with tags, numerical values, punctuations, misspellings, common non-sensical txt (/n), upper letters, spam posts, links, Frequent words removal, Rare words removal, and a lot of stop words that adds noise to the context, we will need to remove them.

**2.4** Advance Text Processing

We used several methods such as - Tokenization, stemming, lemmatization, n-grams, term frequency(tf), inverse document frequency(idf) and tf-idf, part of speech tagging, sentiment analysis, topic modeling, and principal component analysis.

All those methods helped us build a model, gather more insights on our data, and adjusting it the best, so our model wont overfit the data.

**3. Targets and measurements**

Our main target is to prove or contradict our null hypothesis, in order to do so, we will need to extract all the measurements we can from each language, and examine the difference and similarity between each measurement.

The measurements we look at, are detailed in sections 2.2 and 2.4.

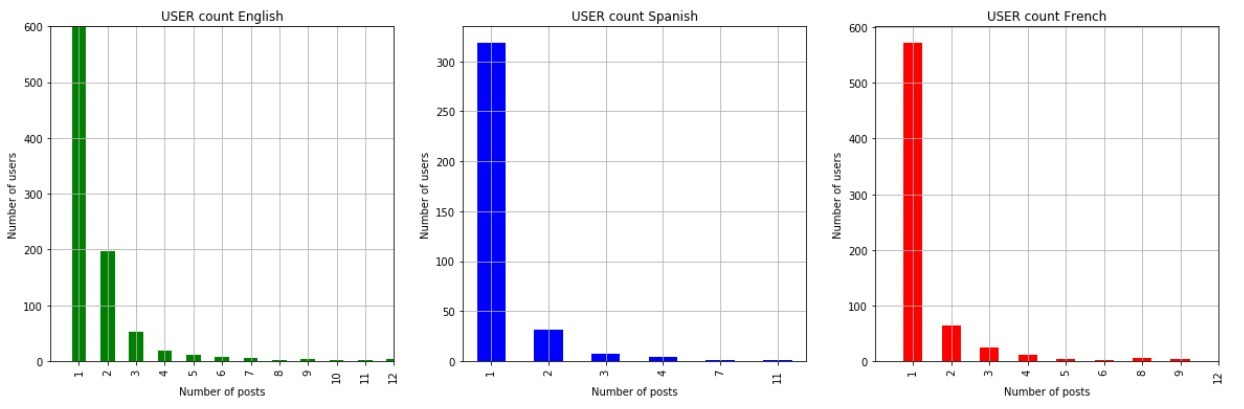
**4. Experiments results**

In this section, we will explain how we got our conclusion.

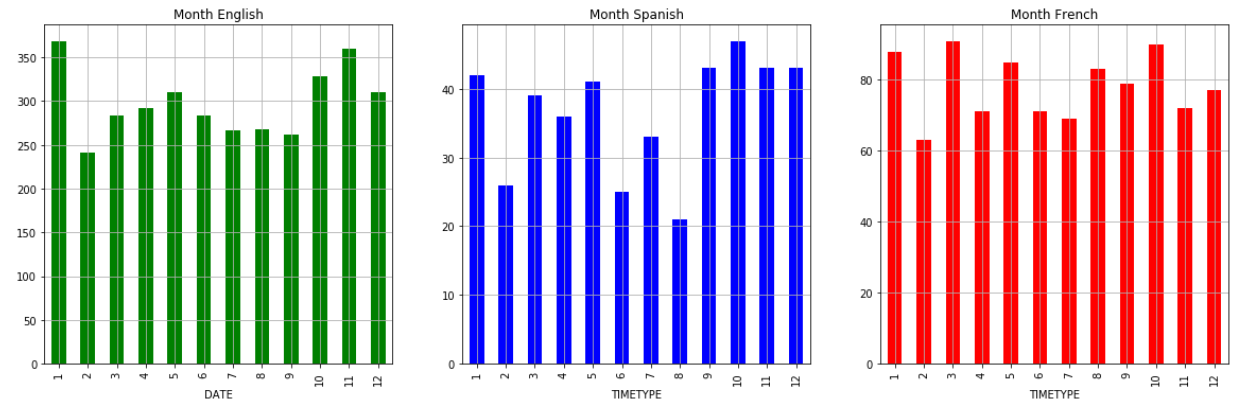
The first thing we could investigate was the basic feature extraction, most of them didn't really helped, but some had clear difference.

**4.1** Basic features

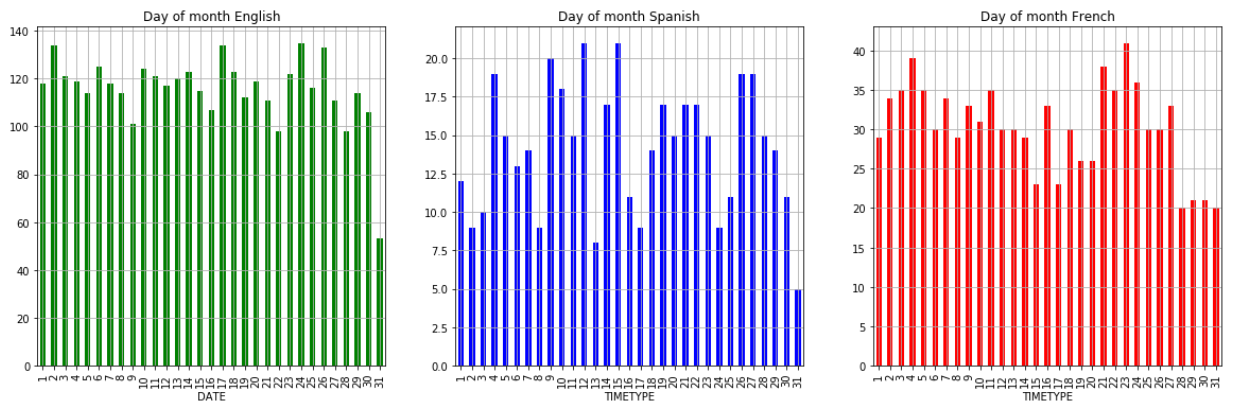
**The average word length-** We can see that in French and Spanish the highest values is between 4-5 characters, while English is between 5-6.

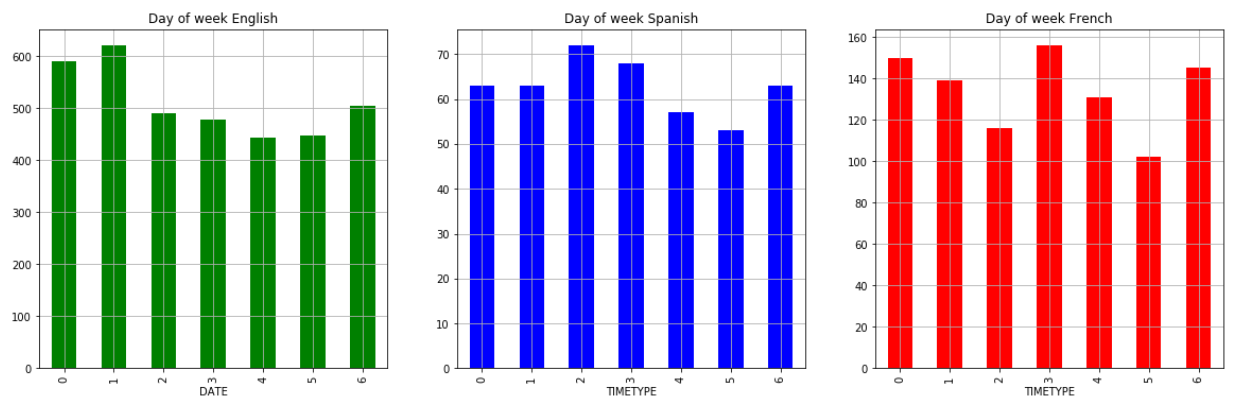
**Returning users** - we also examined the frequency of returning users, English people 28% for more than 1 post, Spanish with 27% and French with 40%

**4.2** Time differences

Month:

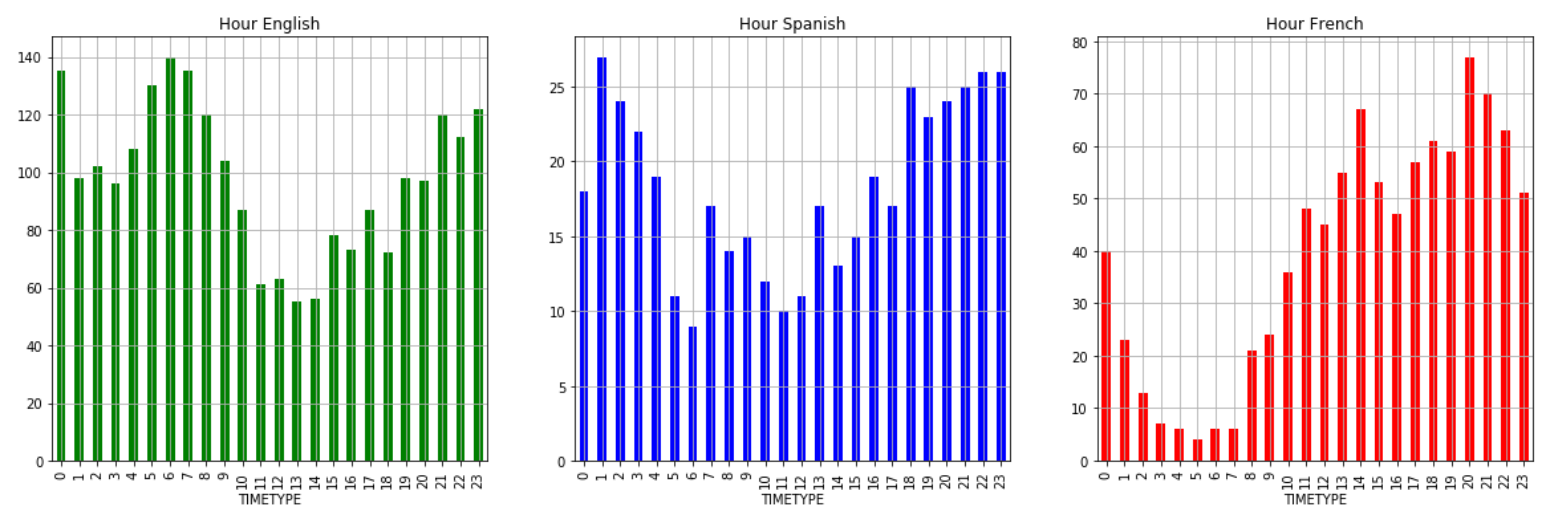
Day in month:

Only by month and day of month we couldn’t extract any information, due the "wav" pattern,

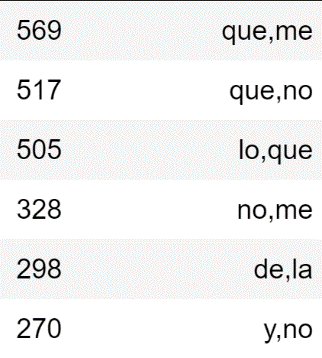
Day of week:

By looking on days of week (0 – Monday, 6 – Sunday), we can see that in Spanish and French the number of posts increases in the middle of the week, and in Sunday and Monday, unlike the Australians that increases only Monday – Tuesday

# I HATE MONDAYS

Hour:

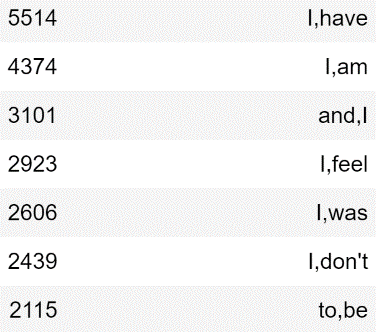
We can see that there is a difference in the activity hour, French people don't post at night times (1 am to 6 am) while English and Spanish people do, unlike the noon hours, where French people are more active.

**4.3** Bi-grams

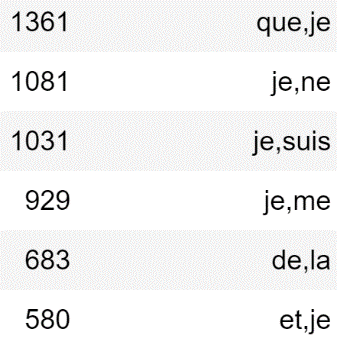
Spanish:





English:

French:



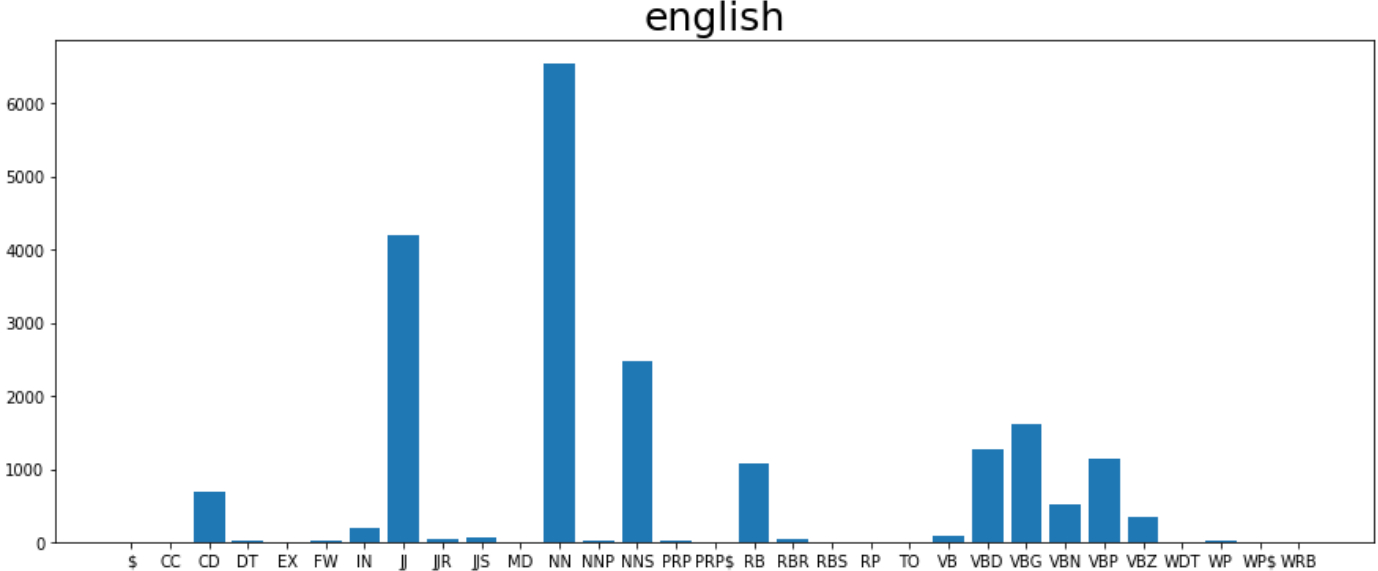


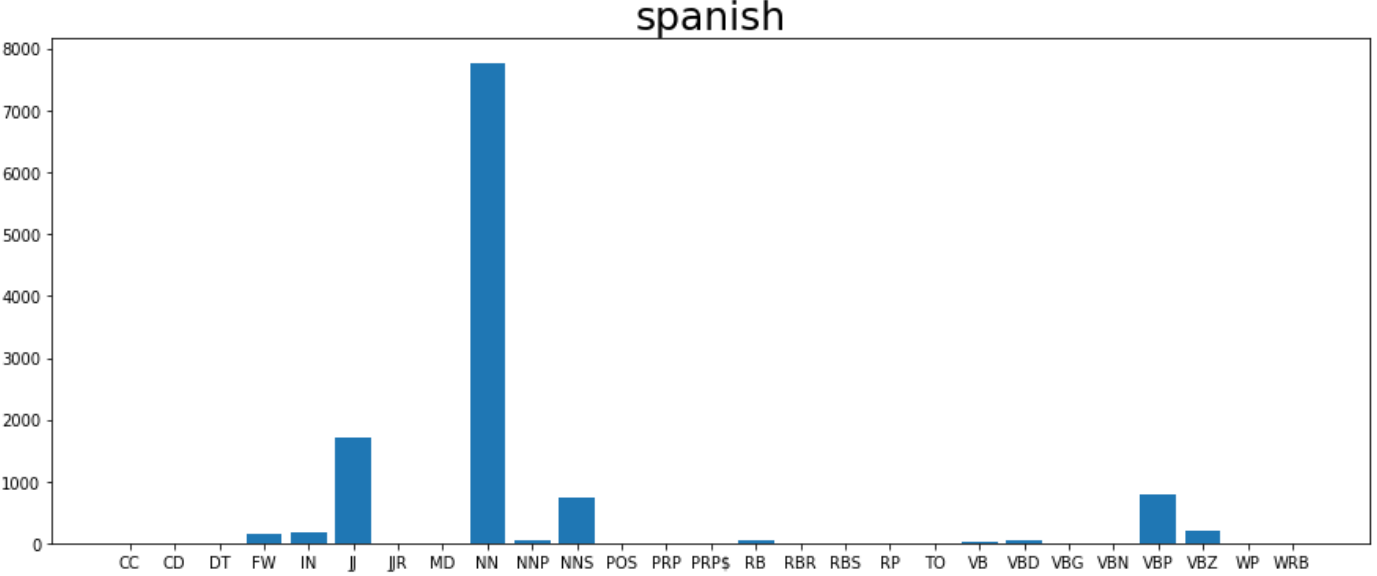
By analysing 2-grams, we could get a "**hunch** "on what all those posts commonly said, we noticed that French and Spanish talks a lot on their families, English people talks a lot more on their feelings, and the past,

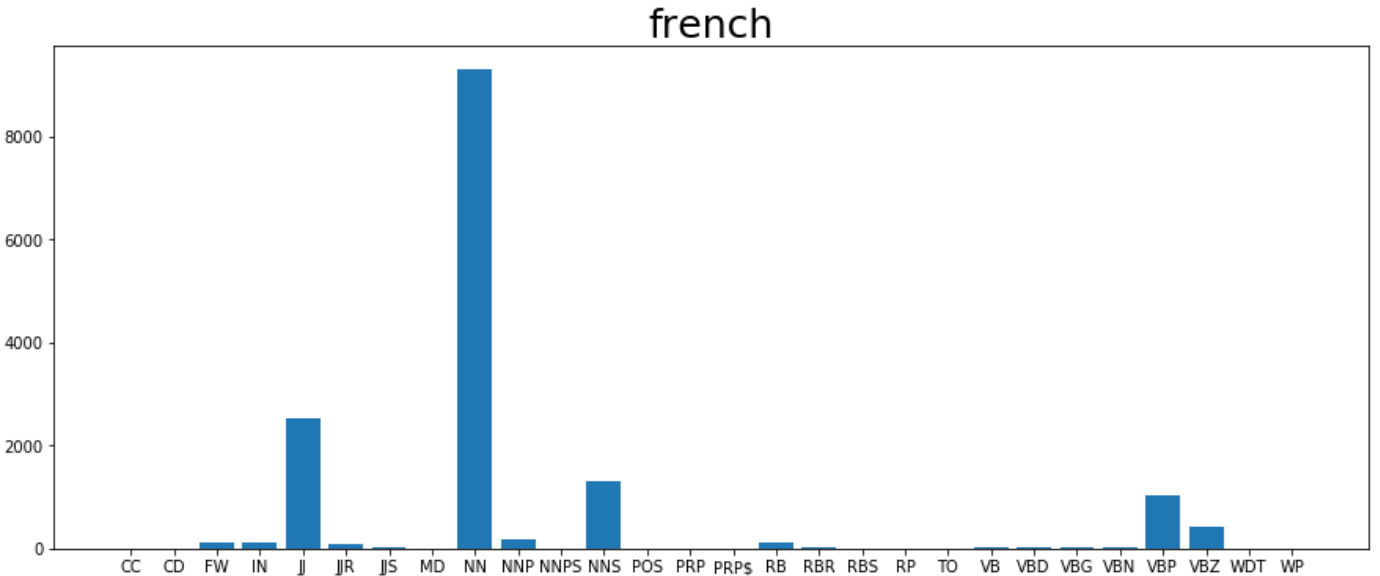
French talk about "trust" and the impression they get from things, While in Spanish people talks more about their personality and their life.

We could see as well that there's almost no difference between the top 2-grams that **means 2-grams syntax doesn’t really play a part in determine what culture\language these 2-gram belong.**

**4.4** POS tagging

**Part of speech** tagging is a measurement that helped us understand whether different culture people use different **morphologic and syntactic properties**.

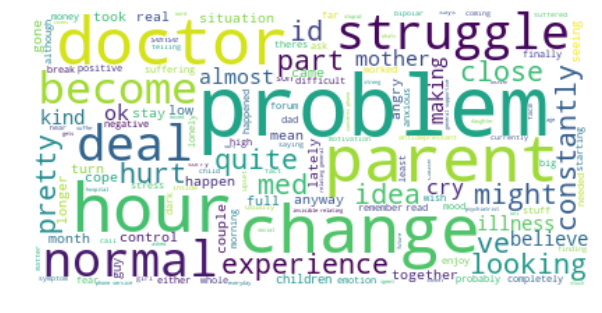
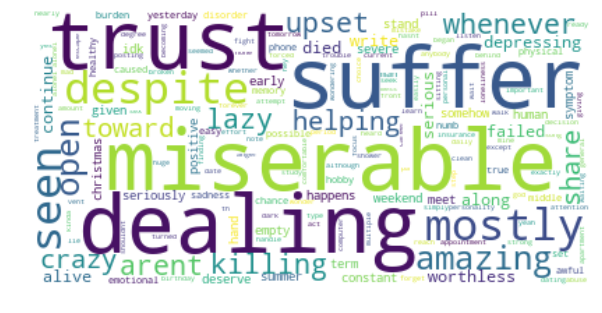


From the graphs above, you can see that all languages uses a lot of NN(nouns) and JJ(adjectives), furthermore, we noticed that English posts uses more verbs then Spanish and French, we concluded that **Australian people describes things more than Spanish and French**, while Spanish and French uses more nouns, same with RB(adverbs).

Another difference is CD (cardinal digit e.g. 'one', 'two'), English speakers tend to use more CD while others doesn’t.

**4.5** Most common words

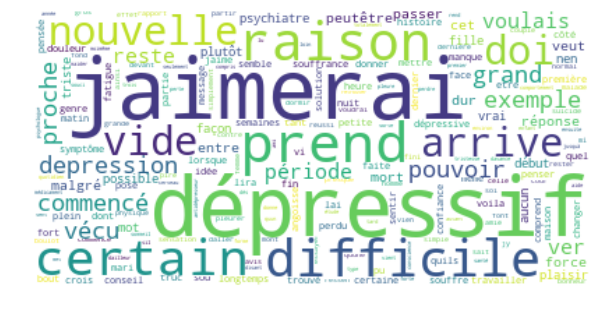
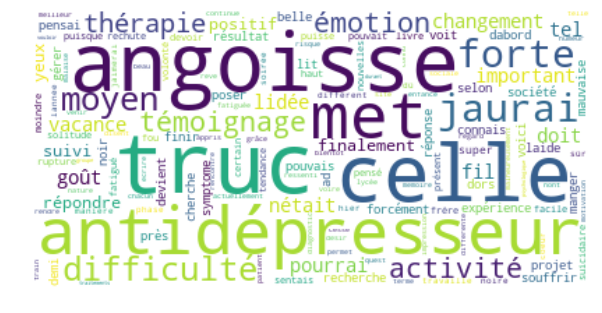
In this section we took the most common words and we tried to understand the general subject.

English:

We can see that there are different subjects, but the general topic is the struggle dealing with depression,

Spanish:

Same with Spanish, although we can see the need for help and everyday difficulties.

French:

Here we can see the need to tell about depression, anxiety along with the request for treatment.

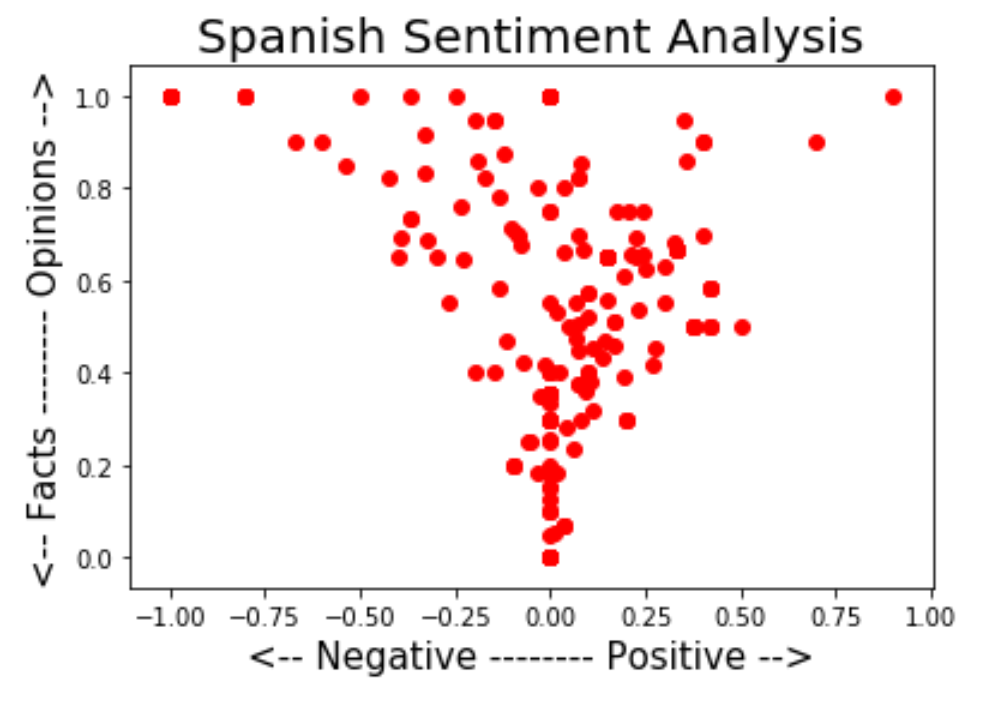
**4.6** Sentiment analysis

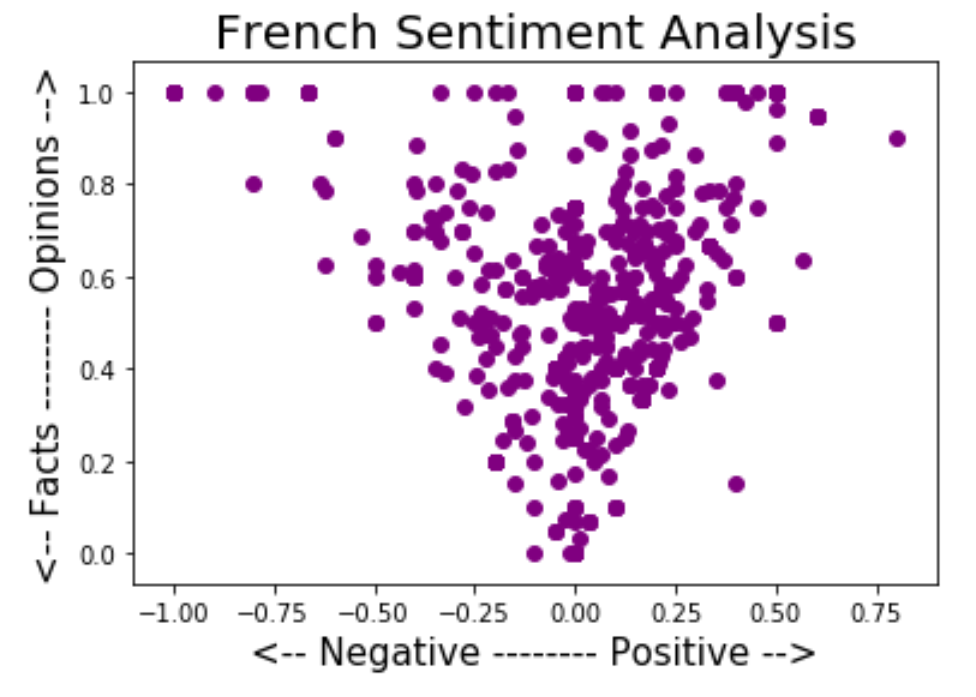
Sentiment Analysis also known as *Opinion Mining* is a field within NLP that builds systems that try to identify and extract opinions within text.

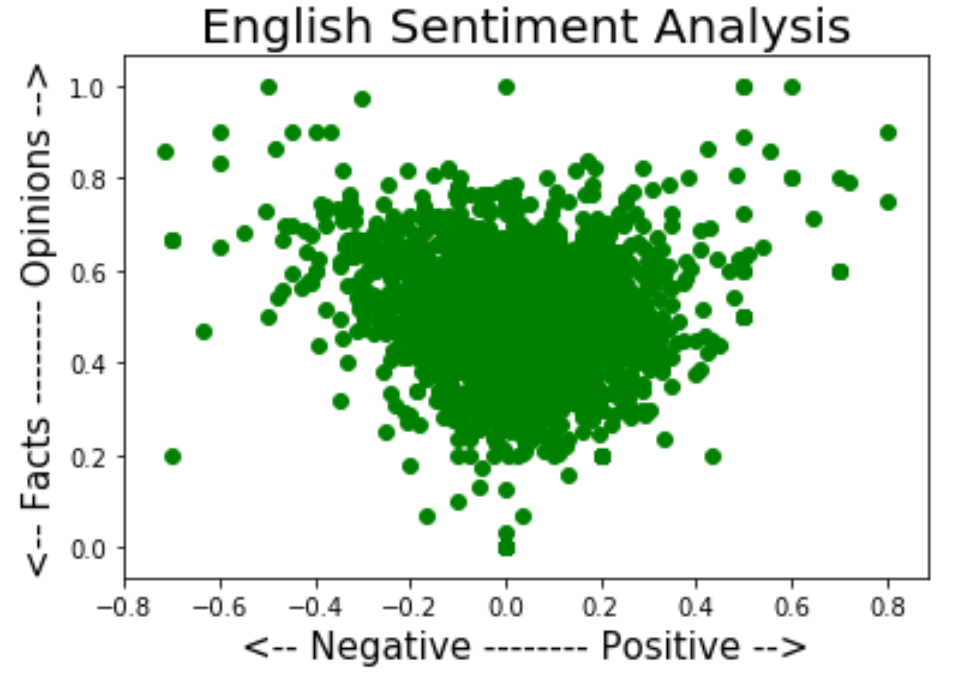
These systems extract attributes of the expression -

Polarity: if the speaker expresses a positive or negative opinion.

Subject: the thing that is being talked about.

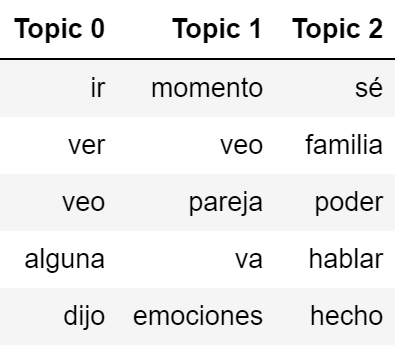




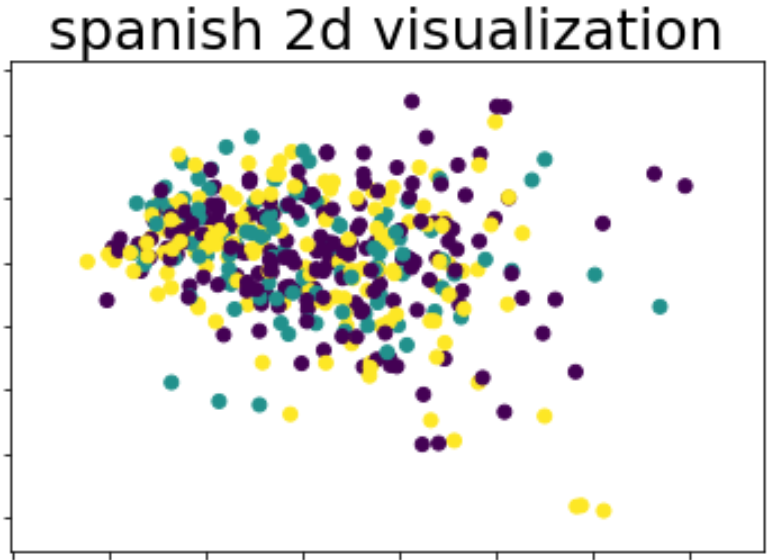


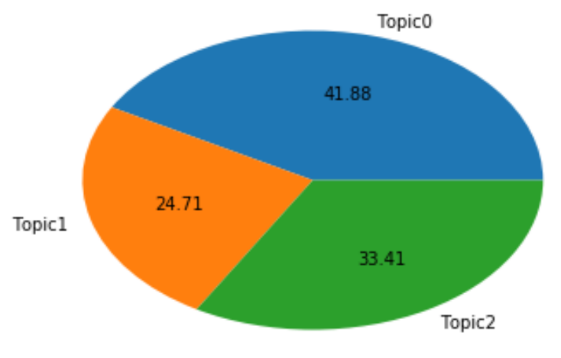
The scatter plots above, shows us the polarity and subjectivity of the people in that language, we see that Spanish and French.

**4.7** Topic modelling and principal component analysis

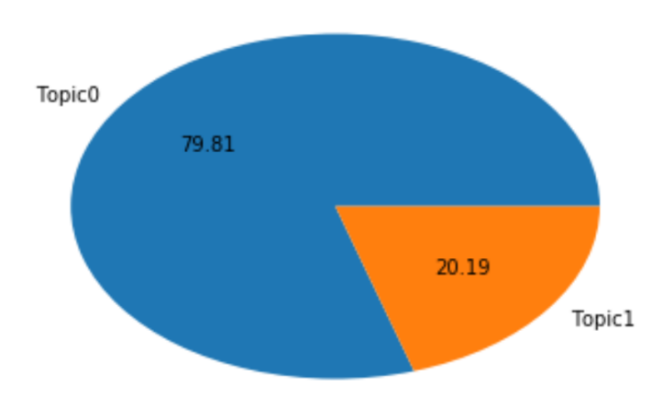
In order to cluster the data by topics, we used several methods of feature extraction (count vector), decomposition (Latent Dirichlet Allocation), and dimension reduction (PCA).

Spanish topics:

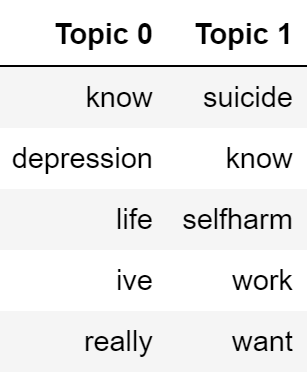


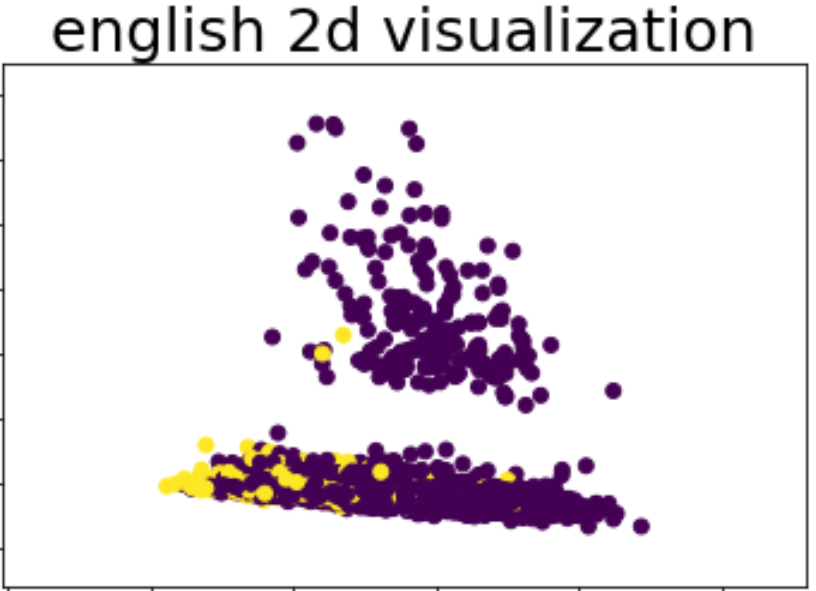


We can see that the first topic talk about searching for help (personal or medical help), the second topic talks about feeling and the need of the writer to "unload ", the third topic relates to the other two in the facet of family.

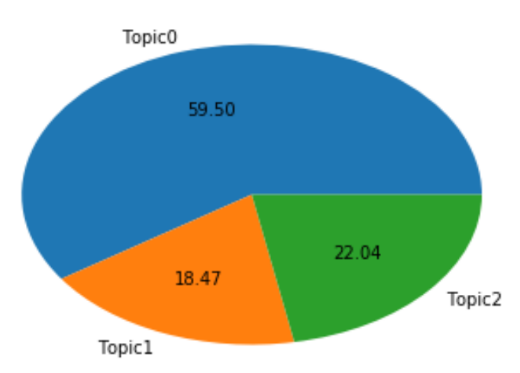


English topics:

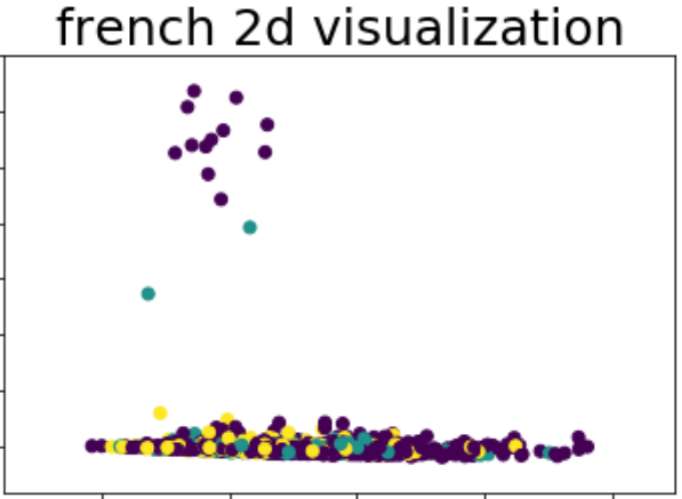




We can see that the two subjects are really overlapping, yet, the first topic talk about depression and the dealing with it, while the other talk about self-harming that comes from depression.

French topics:





We can see that the first topic talk about family, the second topic talk about the features of depression, while the third one talk about the features of depression as well and loneliness,

By analysing the topics of each language, we found that they are really similar, and there isn’t a clear difference between them, we noticed that English speakers are more polite despite the situation, and Spanish speakers have more day to day struggle.

**5. Conclusions**

To sum up the process above, we can see that we answered our question, there is a clear difference of expressing depression across cultures, along with that, during the process we have seen the resemblance between each of the languages, by examining the all those measurements, we concluded that by selecting different methods for different languages (cultures) by their properties, will drastically improve the research.

We think the next stage to improve our study, is to examine real life habits, add more languages, and sorting by areas,

We hope our project will help others understand the properties of languages and cultures, and help the world with fighting depression.

1. **Bibliography**

**Articles:**

# Nairan Ramirez-Esparza(1), Cindy K. Chung(2), Ewa Kacewicz(2), James W. Pennebaker(2) ; University of Washington (January 2008) ; The Psychology of Word Use in Depression Forums in English and in Spanish: Testing Two Text Analytic Approaches (study 1). 102-108

<https://www.aaai.org/Papers/ICWSM/2008/ICWSM08-020.pdf>

# Munmun De Choudhury, Michael Gamon, Scott Counts, Eric Horvitz; Microsoft Research, Redmond WA 98052 (June 2013); Predicting Depression via Social Media. 129-137

<https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6124/6351>

**Internet sources:**

* Topic Modeling

<https://www.machinelearningplus.com/nlp/topic-modeling-python-sklearn-examples/>

* Ultimate guide to deal with Text Data

<https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/>