Interest areas of English and French Twitter users who are interested in US politics

Omid Setayeshfar, Mojtaba Sedigh Fazli, Saber Soleymani

1.Introduction

Today, Twitter plays an indispensable role not only on social media platforms but also in everyday life of the people. Twitter is still one the most popular social media which is growing rapidly. Based on some statistics as of March 2017, there were more than 328 million active users, creating more than 600 million tweets a day [1]. So, it is always a good source of information that encompasses a versatile range of people with a different scope of ideas, race, education level, etc. Twitter has some advantages compared to other social web application. It is simple, mostly text-based, real-time and information can reach a large number of users in a very short time. Moreover, some other information which is available in profile information is accessible through a Twitter API which is designed exactly for this purpose. This information is including name, age, geographical location and summary of interests. Furthermore, other relevant attributes, such as explicit and implicit interests or political preferences are usually omitted but still, there could be some extra information which could be useful for analysis. As a conclusion, it is an apposite source of data for statistical analysis, and there has been an increasing trend to analyze Twitter data in past years.

One of the significant applications of the Twitter is in the analysis of political issues using machine learning techniques. Recently, thousands of researchers are applying machine learning tools for different data analysis, especially in political science. For instance, opinion mining which has done in 2015[2]. In another political science application, the phenomenon of political disaffection defined as the "lack of confidence in the political process, has been analyzed recently by Corrado Monti and his colleagues [3]. Also, there is a very solid paper from Simon Noah team at University of Washington - Seattle in 2010 which connects measures of public opinion measured by polls with sentiment measured from text [7]. Thus, here we aim to use the machine learning techniques for a political application. To make this we have collected and analyzed 22 million tweets of more than 19000 users. Results show interesting facts about attitudes and areas of concern from people interested in US politics both from English and French speaking twitter accounts. In this research we made the following contributions:

1) we propose a framework to gather information about the people interested in a particular area.

2) we evaluate our framework

3) we answer the research questions described in the following section.

4) we propose using language instead of the location in categorizing the location in a country level accuracy.

5) we publish a clean dataset of 14 million tweets related to US politics for further analysis.

1.1 Research question

In this project, our research question is, *"Can we find the interest areas of the people with different culture and language which are interested in US politics"?*

We make this assumption that when people talk about the US politics in different languages, they have different concerns and perspectives. The goal of this project is to find out what are the main topics people in different languages are concerned about. Through their tweets; To find out whether twitters are interested in US politics or not; and what are their viewpoints and categorize them into their languages to see the correlation between the language and concerns about the US politics. First, we planned to use the posts with more languages like Arabic, Farsi, Turkish, English and French. Because of the time limitation and language expertise requirements we did not possess, we decided to limit the languages to two languages, English and French and try to identify the differences between domestic political interested views and some other people from different culture views.

2. Methodology

In this section, we will discuss our data and the methodology which we used to solve the problem. We also describe modules developed through this research in detail.

For this project, we have mainly used a Dell Machine equipped with one Intel Corei7 8 core, 7th Gen processor, 32 GB of memory and a high-end SSD storage device. Some very few experiments –due to limited time on this machine- have also been conducted on a server machine with 24 Xenon Cores and 128 GB memory.

2.1 Data

As we discussed before, we gathered the data from the twitter. To do that we have developed 3 programs. One is for extracting the data from the twitter using the search API, but because of the limitations which are imposed by the Twitter (The limitation is we could not extract either more than 3000 tweets or the tweets which are for more than 7 days ago). Thus, we designed another program to listen to the twitter's stream API and extract the data periodically. Then we developed another program to collect users and then read their profiles using the profile API; this latter approach is what we describe in the following sections.

2.2 pipeline

The pipeline is illustrated in figure 1. In the first phase, we have extracted the data using Python and Tweepy which is a Tweeter API for Python (the details are explained in next sections). In next step, we will clean the data and prepare it for the topic modeling.

Then, we will apply our machine learning method on cleaned data, and finally, the topics will be analyzed, and extracted final results will be represented.

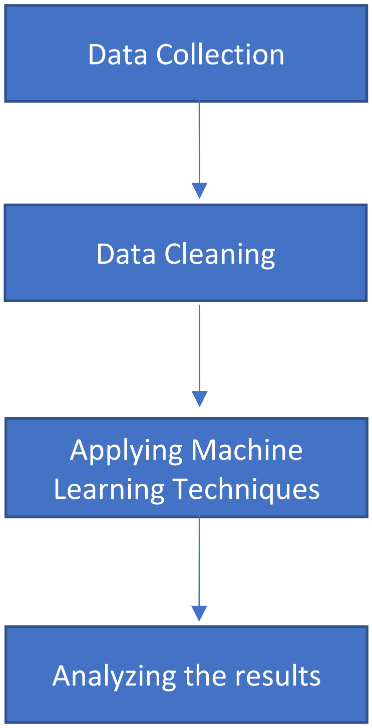


Fig. 1: The pipeline for our methodology for answering the research question

2.2.1 Data Collection

As mentioned earlier we have implemented three methods for this phase, but we only discuss the model in which we read the tweets using the user profile API. This section will describe its details.

First, we manually made a list of great politic-related tweeter accounts. This list includes 32 accounts of individuals, government agencies, political sections of news agencies of different parties, and independent reporters reporting political news from various parties. This list is a filtered list from the initial list of 60 accounts we found reported as most famous political twitter accounts to follow by [1], and other sources including the US government's open data page explaining the list of official tweeter accounts for government branches and agencies; this filter has been done based on the number of followers of each account.

Then we extracted the list of followers of each account, in total we collected 290,690 users in this part. The resulting list of users for each account has been cross-referenced to find the users satisfying our initial criteria –whom he follows at least two of the 32 accounts in the list on Twitter is interested in US politics-. A total of 38,278 users did satisfy our criteria.

Then using the user profile API from Twitter we downloaded 700 tweets from each user; amongst them, 19,764 accounts did have accessible profiles –the rest had private profiles, or their privacy settings blocked our access-. There is an interesting point here which is, when you try to extract the data of a user, the Tweeter API will not consider any restriction like 7 days or a limited number of tweets. So it gives us a golden chance to extract whatever we are seeking for easily. In total, we collected 14,858,400 tweets with an approximate size of 150 GB. To downsample, we filtered the number to 100 tweets per user. The filtered data contained the tweet along with all the meta information it has for 2,486,665 tweets with a size of 18GB. This data set is what we used mainly for the rest of this research. Overall these steps we have used tweepy[6] a free wrapper package on tweeter API for Python, this package offers JSON objects received from the tweeter API. Our implementations are fine-tuned to address the API limitations of the tweeter.

As an effort to make the data smaller to address our limited time and processing power, we then removed all the meta tag information from the tweets and only kept the text body and the language for each tweet. But the primary dataset with all the information would be published open source for further research by others.

2.2.2 Data Cleaning

In this section, we describe our cleaning procedure. The data cleaning part is also done by a python program which we called “CreatesCSV”. In this file, we have the data in JSON format. The reason for that is, we need to do some operations on the data which works better in JSON. This program outputs the data in the CSV format to be used in following modules.

The NLTK library –a python based library for Natural Language Processing- has been used. In this module, we start by converting all the text to lower case letters. Then we removed all the stop words using NLTK's word dataset as well as words without meaning. Another pass of cleaning was done before the NLTK package using beautiful soup a package for filtering and managing web content.

Then the texts have been converted to words using the "text_to_word" function. Then finally the cleaned data has been saved to a CSV file for further processing.

2.2.3 Applying Machine Learning Method

In this section we describe the machine learning methods applied to the data. In this research, we used LDA to extract topic models as outlined in this section.

Our topic modeling phase is implemented in a program "**LdaOnCSVs.py"** in which we use the cleaned data and then the LDA model is created using LDA library of python. then, we create the topic_table.csv file which contains the document number and topics probabilities.

2.2.3.1 LDA

LDA stands for Latent Dirichlet Allocation which is an unsupervised machine learning technique which identifies latent topic information in large document collections. It uses a "bag of words" approach, which treats each document as a vector of word counts. Each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a number of words. LDA defines the following generative process for each document in the collection:

1. *For each document, pick a topic from its distribution over topics.*
2. *Sample a word from the distribution over the words associated with the chosen topic.*
3. *The process is repeated for all the words in the document.*

More formally, each document in the collection is associated with a multinomial distribution over T topics, which is denoted as θ. Each topic is associated with a multinomial distribution over words, denoted as φ. Both θ and φ have Dirichlet prior with hyper- parameters α and β respectively. For each word in one document d, a topic z is sampled from the multinomial distribution θ associated with the document and a word w from the multinomial distribution φ associated with topic z is sampled consequently. This generative process is repeated Nd times where Nd is the total number of words in document d. [4][5]

Since LDA is very fit to our purpose, we used it as a machine learning tool to do the topic modeling. We created an LDA model using LDA library in python. The fitted that model on our data set. In model creation we started by 2 topics then moved to 5 and then finally, we tried 10 topics for both English and French language tweets. Also, the number of iteration is heuristically set between 200 to 1000 iteration, and finally we found that 500 iterations is enough for our purpose.

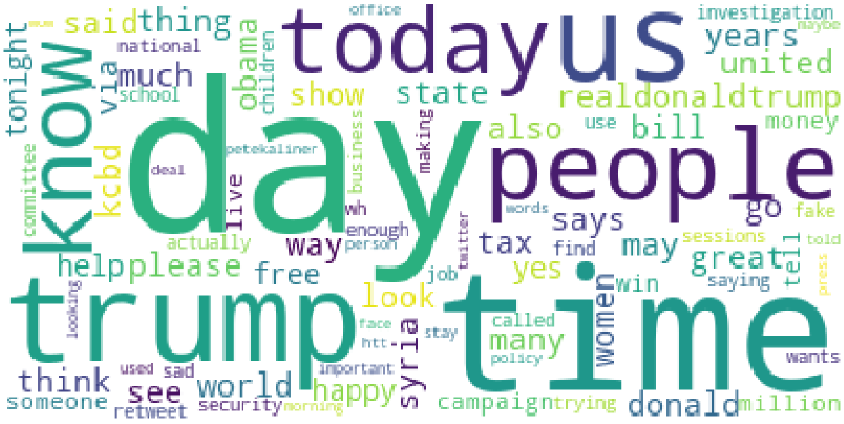
3. Results

In this section, we present our results of topic modeling on final English and French CSV file. First, we show the topic modeling of 5 topics in English, and French. Then show the result of topic modeling when k=10 for both languages.
Then, we display the word cloud of 200 most frequent terms in both k=5 and k=10 for both languages.

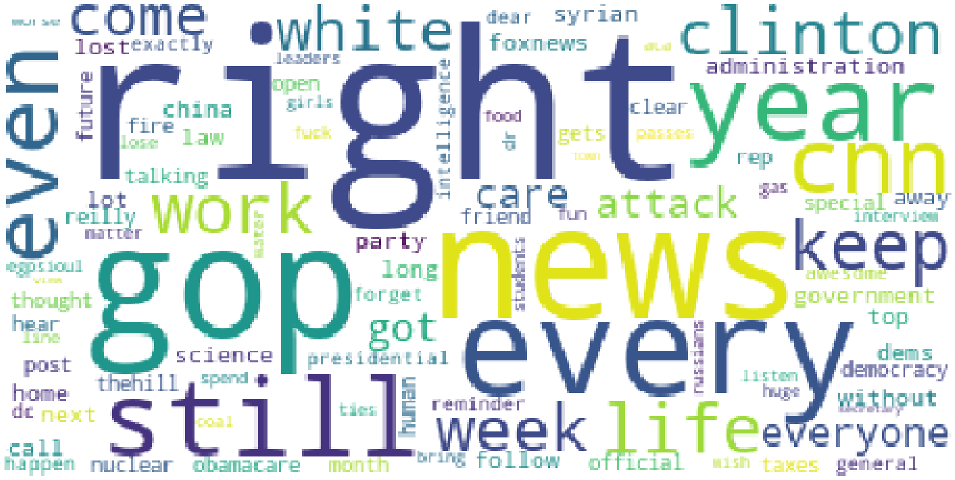
English Topics – k=5

Topic Number	**Possible Topic Title**	**Top words**
Topic 0	Election results	trump us like people today day time know realdonaldtrump would
Topic 1	Media	news gop p every still year cnn even g keep
Topic 2	Foreign affairs	one right potus russia back love b could election putin
Topic 3	Populist opinion	new w good need r want make let read say
Topic 4	Domestic support	trump u house president first man c n american America

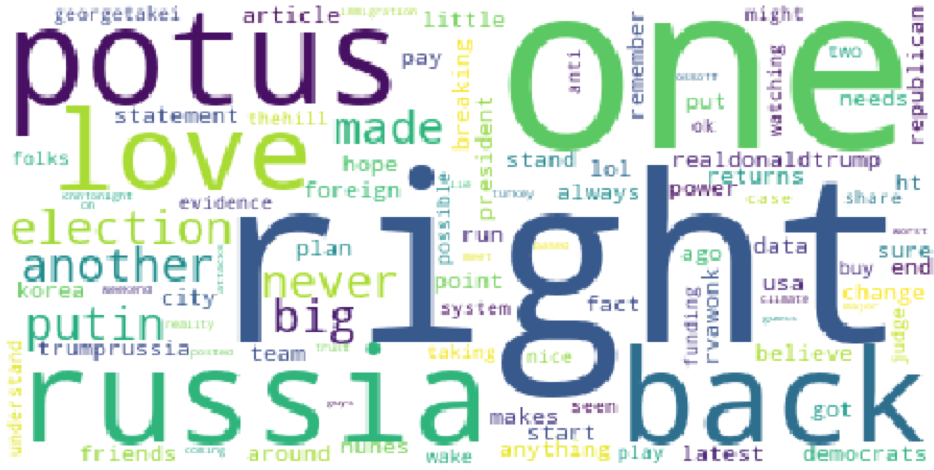
Topic #0



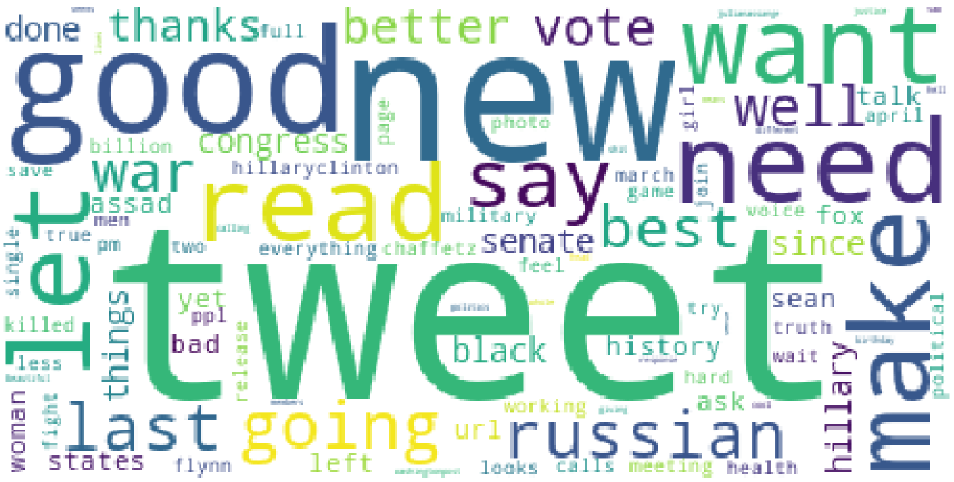
Topic #1



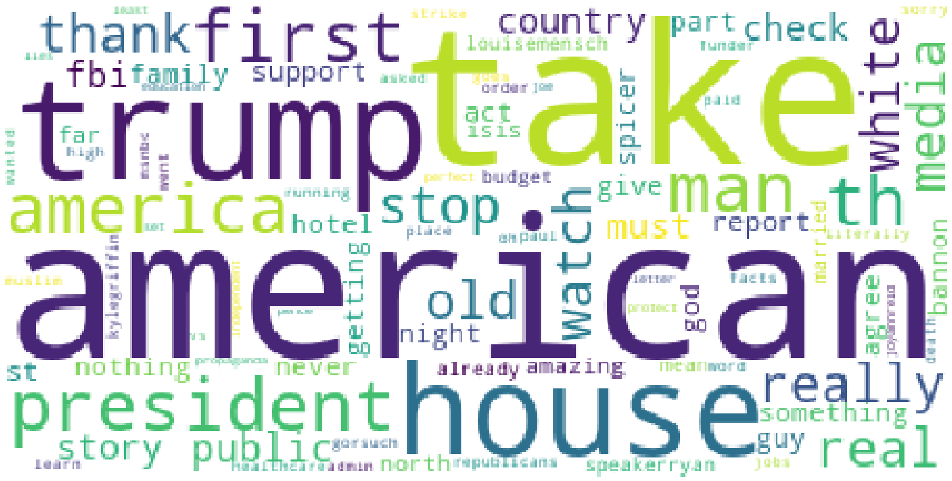
Topic #2



Topic #3



Topic #4



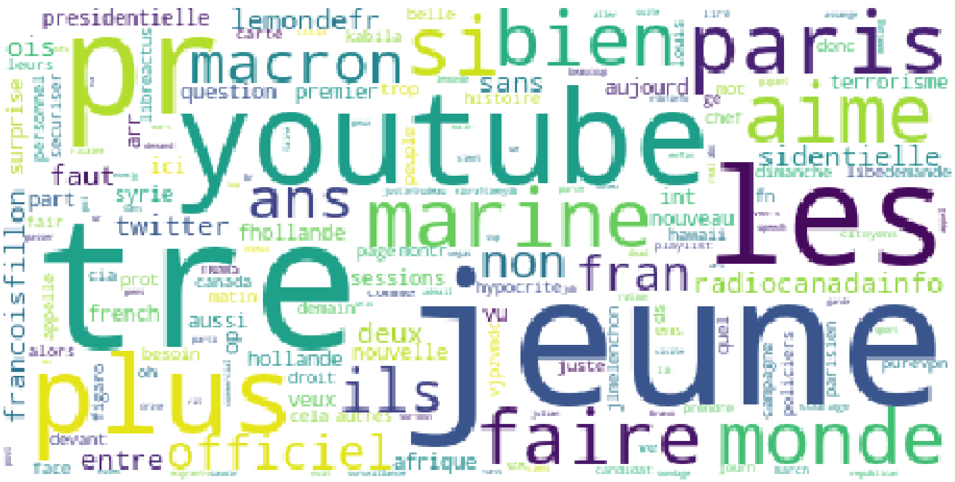
English Topics – k=10

Topic Number	Possible Topic Title	Top words
Topic 0	Election day	us like people one new w today day c know
Topic 1	Media bios	gop b cnn g work z life everyone never free
Topic 2	Middle east	trump time realdonaldtrump obama make syria house l every bill
Topic 3	Foreign policy	russian h keep thing law foxnews q still yet power
Topic 4	Populist rhetoric	back election first always trumprussia article take fact enough important
Topic 5	Email Scandal	russia potus could war fbi breaking team change, please tweet
Topic 6	Congress relations	read going love congress thanks, senate since done makes call
Topic 7	Hillary's plans	good let last best vote well things better hillary hope
Topic 8	-	right president say u n trump v ever get stop
Topic 9	Populist rhetoric	trump u man need american th thank america real media

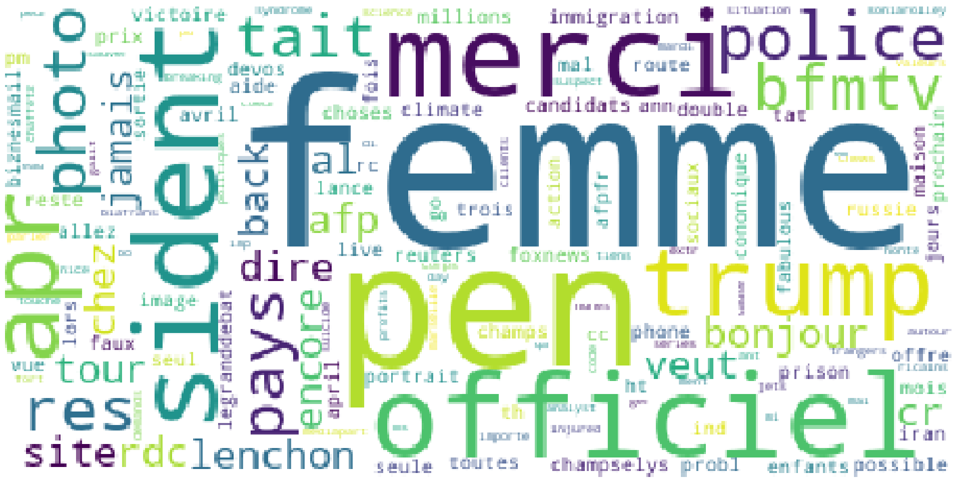
French Topics – k=5

Topic Title	Top words	English Translation
media	a e pr les youtube plus r f p paris	a e pr the youtube more r f p Paris
Politics in France	g pen merci u officiel sident trump apr and police	g pen thank you official u president trump apr and police
Politics in France	fr v b fillon excellent cette va w tr peut	fr v b Fillon excellent this goes w TR can
Science	france vid i q in x contre to fait ais	France vid i q in X against TB fact AIS
US Politics	les r via o h vote mlp fran trump the	R via o h MLP vote fran trump the

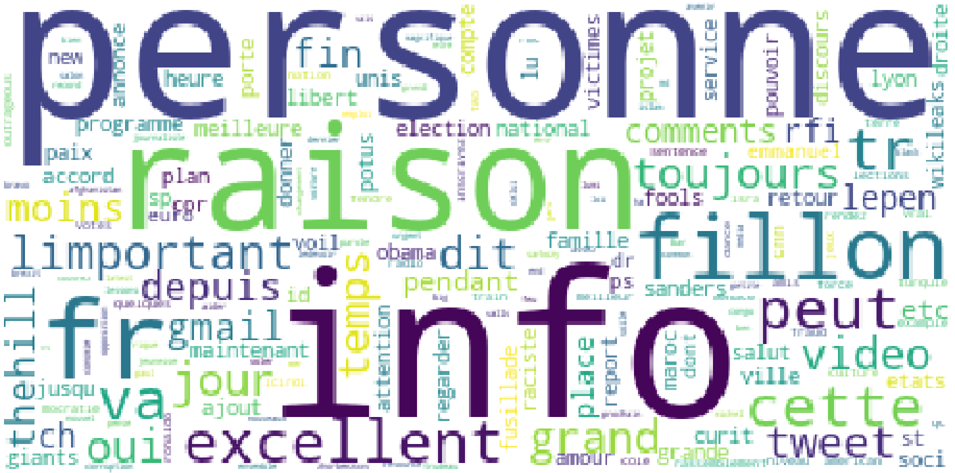
Topic #0



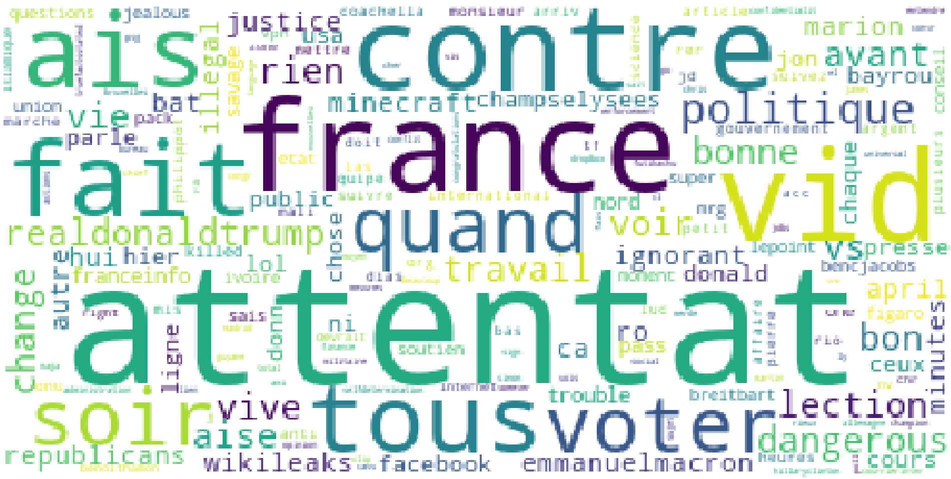
Topic #1



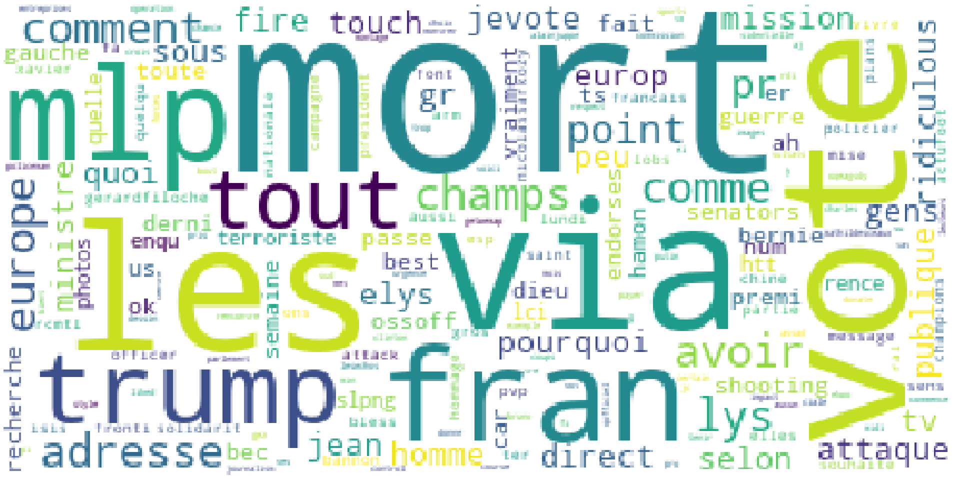
Topic #2



Topic #3



Topic #4



French Topics – k=10

Topic Title	Top words	English Translation
Media	a e via france pr o r youtube in plus	A E via France pr o r youtube in more
Politics France	paris les k pen and non police francois fillon faut is	Paris The K pen and not franco fillon police need is
Politics France	pen b jour rdc nouvelle afrique arr attentat fin gmail	Pen b day rdc New Africa arr attack end Gmail
International finance	merci apr change sidentielle r al fait lenchon lepen bon	Thank you exchange Apr sidentielle r al fact lenchon lepen good
Environment	q b cette peut dit marine cr depuis moins rien	q b This can marine said CR since less nothing

Canada relations	limportant res twitter int video comments canada pendant curit direct	important RES Twitter int video comments Canada during direct Safety
World Politics	contre x quand z pays toujours dangerous vie vs officiel	Against x When z countries always dangerous life vs Official
Politics	vid i to va tous soir politique are pourquoi grand	vid i to will all political evening are why large
Inter party connections	tout realdonaldtrump com selon republicans minecraft his april hui marion	Any realdonaldtrump com according to Republicans Minecraft his April Today Mariona
Radical Policies	les r h trump vote mlp excellent fran champs tre	r h trump MLP vote excellent fran fields much

4.Conclusion

Although we have just observed English, and French tweets, we have found the interest areas of the people with different culture and language who are interested in US politics are relatively different.

By observing topics and most frequent terms in both languages we have found some facts:

1- YouTube is a major source for French-speaking people to retrieve news of US politics.

2- Relation of France, and Canada with US politics is an important factor for French-speaking twitter users. Most probably, it's because of French speaking people who live in Canada and France. Among those the trade and international financial markets seems to be amongst the most important areas for French people.

3- Middle east, and foreign policy are among important areas of interest of English speaking twitter. Whereas, French speaking users tend to have more interest to trade and policy.

4- French speaking people have more tendency toward populism rhetoric of American capaigns.

Based on the facts we observed, we can conclude that our research question is satisfied by our methodology. In next step, we can analyze more languages and explore more facts about their interest toward US politics.

5. References

[1] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[2] Hasnat, Md, et al. "Opinion mining from twitter data using evolutionary multinomial mixture models." *arXiv preprint arXiv:1509.07344* (2015).

[3] Corrado Monti, Alessandro Rozza, Giovanni Zappella, Matteo Zignani, Adam Arvidsson, and Elanor Colleoni. 2013. Modelling political disaffection from Twitter data. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining* (WISDOM '13). ACM, New York, NY, USA, Article 3 , 9 pages. DOI=http://dx.doi.org/10.1145/2502069.2502072

[4] Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics* (SOMA '10). ACM, New York, NY, USA, 80-88. DOI=http://dx.doi.org/10.1145/1964858.1964870

[5] Shuang-Hong Yang, Alek Kolcz, Andy Schlaikjer, and Pankaj Gupta. 2014. Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '14). ACM, New York, NY, USA, 1907-1916. DOI: http://dx.doi.org/10.1145/2623330.2623336

[6] http://docs.tweepy.org/en/v3.5.0/api.html

[7] O'Connor, Brendan T., Ramnath Balasubramanyan, Bryan R. Routledge and Noah A. Smith. “From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.” *ICWSM* (2010).

Appendix 1, most frequent languages in tweets

en : English
es : Espanola
fr : french
ar : arabic
tl : tlingit
tr : turkish
da : danish
ht : Haitian Creole
nl : Dutch
cy : welsh
de : delaware
sl : slovenian
no : Norwegian
lt : Lithuanian
fa : Persian
it : Italian
fi : finish
ja : Japanese
pt : Portuguese