Seminar Thesis for the course "Data Science in Techno-Socio-Economic Systems"

Understanding Public Opinion Dynamics Using Twitter Sentiment Analysis – Case Study: French Presidential Election 2017

Andreas Psimopoulos, Barthélémy Launet

ETH Zurich

Abstract

During last years, social networks have become an integral part of our everyday life. They are not just a communication means; they gradually compose a miniature of the real society for a variety of aspects. One of these aspects is the dynamics of public opinion formation. People use social media in many ways to express themselves for a plethora of topics. Due to its major societal importance, Politics is one of the most interesting areas for investigation through social networks. In this paper, we examine how someone can exploit data from Twitter in order to understand the public opinion dynamics during a pre-election period. Having as a case study the French presidential election 2017, we present a data science approach for i) predicting the final result of the elections and ii) optimizing a candidate's political strategy using text analytics tools. We found that it is very difficult to make a good approximation of the final results using sentiment analysis on Twitter posts. A very interesting part of this analysis is that one can find interesting facts regarding thousands of public opinions, such as relations between topics, just by looking at word infographics and word-network representations.

1. Introduction

Social networks provide us the opportunity to explore and analyse public opinion dynamics in real time. The goal of modeling opinion dynamics is to determine the opinion states in a population and the transitional processes between such opinion states (Castellano, et al., 2009). In this project, we take into consideration both opinion states and their transitional processes, regarding the recent French presidential election.

Our first goal is to provide an estimation of the election result, based on the opinion states. One of the first published papers that investigate the potential of predicting election results using Twitter[1], shows that the mere number of messages mentioning a party reflects the election result (Tumasjan, et al., 2010). For the case of the US presidential election 2016, in contrast to the forecasts of many polls, social media analysts could predict Donald Trump's win by using only sentiment analysis (Perez, 2016). Regarding the UK EU Referendum, while most of the polls for voter intention predicted a win of "Remain" even during the last weeks before the vote day, a sentiment analysis based on Twitter posts[2] correctly predicted that "Brexit" would be the final result (Porcaro & Müller, 2016). Therefore, there is some evidence that – by extracting data from social networks – predictions with regard to election results can be more precise than the conventional polls. Consequently, it is worth investigating for such methods and this is one of this project's objectives.

The second goal of our project is to provide some analytical tools that can be used by a candidate for political strategy optimization. Here are the transitional processes between opinion states that come to the front. By studying the reasons behind an increase or a decrease in a candidate's popularity, it is becoming possible to identify the determinant factors of such changes and, consequently, provide feedback regarding a candidate's campaign. The way we choose to explore these factors is by using word clouds and

[1] Case study: Federal election of the German national parliament, September 2009.

[2] Using a method which developed at the Dortmund Center for data-based Media Analysis (DoCMA).

networks based on the words which appear more frequently. For instance, assume that a decline in one candidate's popularity is observed during the period X. By searching for the most frequent words related to the candidate, in tweets with negative sentiment polarity, we can discover which topics are related with the aforementioned decline during that period. Hence, a real-time feedback is made available to the candidate and his/her consulting team.

2. The challenge and the desirable outcomes

The system we have studied is the Twittersphere of France. To be more specific, we studied posts provided by the Twitter API, which were written in French and their content was related to the candidates and the political parties which took part in the French presidential election 2017. Under the assumption that this network is a representative of the electorate, we try to estimate the election result and draw conclusions about the opinion dynamics of the French society. Of course, it is known that not all the voters are represented fairly in the Twittersphere and not all the people express their real opinion publicly. But, since we use data from social media, we compromise with the restrictions that accompany them.

In this paper, we examine several hypotheses. First of all, we examine the potential of predicting election results using data from Twitter. The examples mentioned in Introduction provide some evidence that it is possible to estimate election outcomes, but we need more real examples to verify this hypothesis.

Our second hypothesis is the fact that we can optimize a candidate's political strategy by getting feedback from sentiment analysis on tweets. In detail, the idea is that by showing the most frequent words that are related to positive or negative sentiments, the politician can adjust his/her strategy by discovering the sources of good and bad reputation. If big changes are observed in the reputation time-series after important events (like, a debate), the infographics can provide more information about what went right or wrong. Additionally, by constructing word networks for specific periods, we make it feasible to depict for each of these important words which other words coappear with. This means that not only the importance (i.e. reference frequency) of a topic can be found, but also the magnitude of its relation with other topics. With regard to this hypothesis, we had to compromise with the restriction that we could not test in practise the optimization of an actual candidate's strategy. However, the purpose of this analysis is to explore and show which potentialities arise from this approach for further research.

Regarding the desirable outcomes of this research, first of all we want to understand the public opinion dynamics which are relevant to political issues. The pre-election period is probably the most important period for the democratic societies, as their near future is determined during this. So, it is very important to understand which topics define people's opinions, the election result and the political stage in general. Besides that, we want to build a sentiment analysis algorithm that is able to predict real outcomes of social collective processes (like the elections) and provide feedback about the reasons which formulate these outcomes. Its potential applications can be extended to marketing purposes and other fields of interest as well.

3. Why following a data science approach?

Since we have to deal with data from social media, our research questions are inextricably linked with the so-called *big data analysis*. Inference based on social media analytics requires techniques which make it feasible to deal with such a big amount of data. Moreover, the fact that we have to do with real-time data does not leave room for other approaches than data science, since both their volume and velocity make them difficult (if not impossible) to be handled otherwise. As it is explained thoroughly below, in order to get even a summary table or a meaningful word infographic

from the analysis of millions of tweets, many intermediate processing stages are required. Thus, there are plenty of reasons which lead us to follow a data science approach.

Searching in literature for research relevant to public opinion dynamics, we find that for similar problems a data science approach is followed as well. For instance, in the paper of Xiong & Liu (2014) we find that sentiment analysis was applied to over a million of tweets in order to investigate the dynamics of public opinion (such as the opinion formation process, conditions that lead to an ordered state, etc.). Mohammad, et al. (2015), showed that electoral tweets are rich in emotions and they developed supervised automatic classifiers for detecting the emotional properties of unseen posts. Last but not least, Burnap, et al. (2016) predicted correctly the order of the top three parties for the UK 2015 General Election, by applying sentiment analysis on Twitter posts. However, they had less accuracy to the prediction of the final result in seat shares.

Regarding the limitations of a data science approach, it is worth mentioning that such methods are not a panacea for every kind of problem. Of course, we can process a very big amount of data by taking advantage of a computer cluster, but there are kinds of problems which cannot be solved just with more computing power. For example, a problem that we faced in this project is that of the sarcasm effect, i.e. the fact that we get a positive sentiment score for a tweet which is actually negative due to sarcasm. This is an important problem which requires more intelligent applications in order to tackle it with a data science approach. Actually, a Machine Learning-based Twitter sarcasm detector is a project currently under development (Majaski, 2016). Thus, data science approaches are not yet mature enough to solve problems for every case as efficiently as a human can do in a smaller scale.

One other limitation is relevant to the cost of the analysis. Even if we have a very good algorithm that does exactly what we want, it may be the case that it needs a lot of time (or computing capabilities) in order to complete an operation to the whole dataset. A serious limitation specifically of the sentiment analysis, comes to the front when researchers apply it in order to draw conclusions also for parts of the population which are not equally represented in the sphere of social media. While a conventional poll can be designed in such a way that all the groups of interest can be represented proportionally to their actual ratios, this cannot be done for sentiment analysis as each social network is restricted by its own specific demographics[3]. Therefore, this has to be taken into consideration before making statements about the whole population.

Especially for our project, what we get from Twitter API is a sub-sample of the posted tweets and not all of them. This means that we approximate the real results only if the sampling is random. But we don't know the exact mechanism behind the tweets selection and, if some bias is introduced, we are not able to account for that, which implies that there is one more limitation that arises from such a data science approach.

4. Data sources

To collect relevant tweets, we defined a list of keywords per candidate such as names, campaign slogans or political labels, and used an implementation of Apache Storm for Twitter kindly provided by Dr. Moise.

The collection spanned from March 20th to May 7th, producing a final output of 150 GB of disk space made of 4518 files (15 minutes slices) of raw text.

5. Data analytics pipeline

Formatting

The role of the formatting part is to convert each text file into a table of tweets containing useful information for our analysis. Each tweet is

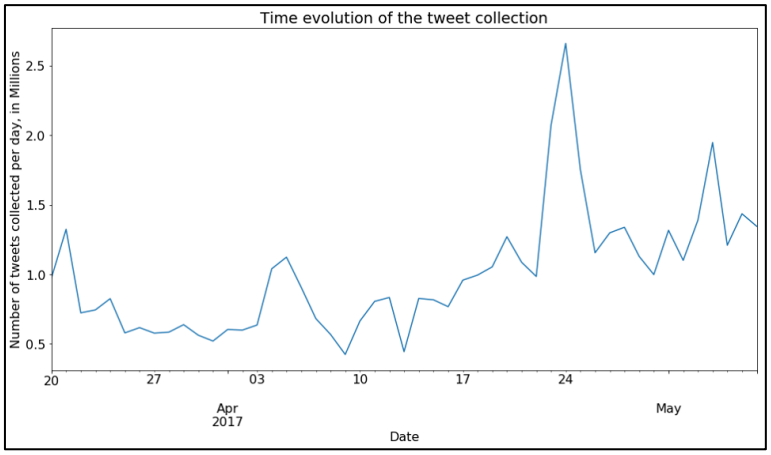
[3] A recent report about social media demographics can be found in York (2017).

uniquely identified by its *ID* and regroups many fields such as user information, text of the tweet, hashtags, date of post, etc. (Appendix, section A).

Pattern and keywords recognition were used to extract valuable data for each tweet. In case of a retweet, we saved the original tweet as a separate line in our table; this led to duplicated tweet *IDs* (multiple retweets of the same original), and we made sure to keep only the last occurrence of each, i.e. the one with the up-to-date number of favourite and retweet counts.

Only tweets in French language were kept, to avoid contamination by opinions from other observers who don't have voting power.

At the end of this formatting step, we end up with a total of 47 million tweets to analyse, and the disk space usage is reduced to 22 GB.



Processing

i. Candidate reference determination

Based on a list of keywords for each candidate, we look into *hashtags*, *text* and *userMentions* to determine which candidate the referring to. If none is found, we also look into *user_description*.

We chose to ignore tweets referring to multiple candidates. We identified two typical cases for multiple reference: very descriptive tweets only mentioning general information about candidates (TV appearance at a debate for example) or tweets opposing two candidates ideas. In the first case, the tweet doesn't vehicle any sentiment and is therefore not relevant for our analysis. In the second one, being able to separate the sentiment of the tweet into positive for one candidate and negative for the other is too error-prone and would introduce noise.

With this operation, we reduced our sample by ~30% in number of tweets to analyse.

To save computational time, this candidate reference determination is done only on original tweets; we then propagate the result to potential retweets using their tweet *id*.

ii. Sentiment determination

Because our dictionaries of scored words are lemmatized, we first needed to lemmatize the content of each tweet to make it comparable.

We achieved this operation with the TreeTagger program, a decision tree algorithm able to determine the lexical type of a word ('tag') and its lemma (Schmid, 1995) (Schmid, 1994) in combination with a French corpus of tagged words.

However, TreeTagger is very sensitive to spelling errors and typos, as it can lemmatize only existing words in its corpus, and a big effort had to be made to clean and correct each tweet *text*.

1. Filtering words only

To feed the lemmatizer with a correct input, we first had to filter out every 'non-word' entity: emojis, http links, hashtags, user mentions, numbers and dates are thus first removed from the *text*. We then split the text into lexical units ('tokens') using TreeTagger tokenizer, and removed all punctuation and quotation marks.

2. Correction

The next step is to make sure that this cleaned *text* is correctly spelled. For this purpose, we used Hunspell, an opensource spellchecker used by LibreOffice and Mozilla, combined with OpenOffice French dictionary. This tool can spot misspellings but unfortunately not correct them.

Our first attempt to correct is to try some minor modifications on the word, identified as classical typos in French: missing accents (‘*reellement’* → *‘réellement’*) and duplicated letters (‘*apeler’* → ‘*appeller’*).

If this operation is unsuccessful, we assume that the error is probably due to abbreviations or phonetic language, widely used on Twitter, and that no classical dictionary or method can correct.

The website http://www.traducteur-sms.com/ has proven to be a performant tool to correct this type of mistakes, and an HTTP request was made to this website for each misspelling; if none is find, it returns the initial word.

Every correction is kept in a dictionary to save computation time (the HTTP request is particularly slow). We ended up with a correction dictionary of 290 000 words, with 11% of the words being actually correctly spelled after correction. This apparently low rate of success can be explained by the fact that most of the words detected as ‘misspelled’ are actually names of politicians, companies, TV channels, that actually should not be corrected and that anyway don’t carry any sentiment. Other kind of uncorrected mistakes appear for truncated words, that appear when the tweet contains a http reference (ex: ‘*@MLP_officiel @1Frenchpatriot Après avoir vu ça je confirme, candidat de rien, à non candidat pour un poste à Bruxe… https://t.co/gAzdsMEgai’).* The last word of such tweet is often incomplete, and thus misspelled. Truncated tweets represent only 3% of our sample, and if they probably affect strongly our success rate in the correction, they don’t bias our scoring method.

3. Scoring

This corrected and cleaned text is then lemmatized, and compared to our scored dictionary, looking first for expression scores and then words scores. We compute separately the sum of positive and negative words and output them both for each tweet, to enable differentiate analysis afterwards.

iii. Relevant words determination

For each tweet, we keep for further analysis a selection of “relevant” words: from the list of lemmas, we select only words of length more than 2 and categorized by TreeTagger as “noun”, “name” or “adjective”.

iv. Final output

This processing step was run on Euler ETH supercluster for a total of approximately 3 days, using 4 cores and 3GB of RAM.

The final output is a table of 20 million unique tweets (2.5GB) containing all the relevant information for our analysis tasks (Appendix, section B). The data analytics pipeline can be found visualized in Appendix, section C.

6. Metrics and measurements

The main question of this project is to what extent we can predict the final result of French presidential elections 2017 using sentiment analysis on Twitter posts. The election had two rounds. In the first (23 April) eleven candidates took part, while in the second round (7 May) only the top two of the first round (Ministère de l'Intérieur, 2017). What we want to test is whether the tweets’ sentiment polarity can be used to predict the actual election outcome, for both rounds.

i) For the first round: We score each tweet according to the available word / phrase-to-sentiment score dictionaries (Appendix, section D). We define a neutral score zone (for the moment, is enough to say that this is a way to make the algorithm more robust to false ratings). For every user’s last tweet with total score[4] larger than the upper limit of the neutral zone, we consider

[4] I.e. the sum of each word’s score.

this tweet as a vote intention for the mentioned candidate. This means that we exclude from this part of our analysis tweets with score lower than the lower limit of neutral zone (negative sentiment), because their interpretation is not straightforward in favour of a specific candidate. Hence, we collect all these "vote intentions" and we consider the corresponding percentages as forecasts of the first-round results. These results will be compared to the real ones using a Chi-Square test and the Marascuillo procedure if it is necessary (details below). The null hypothesis in these tests is H_0*: percentage forecasts from sentiment analysis and election results are the same*. So, ideally, our method performs best if there is no rejection of H_0 from the statistical tests.

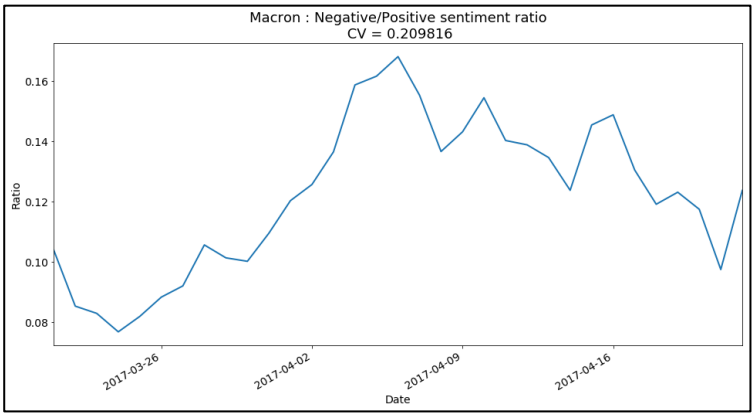
ii) For the second round (tweet collection during the period between the two election rounds): The procedure of evaluation for this case has only one difference. As we have only two candidates in this case, we can interpret negative tweets regarding candidate A, as positive in favour of B and vice versa. Therefore, we took into account both positive and negative tweets, by transforming the latter as positive for the other than the mentioned candidate.

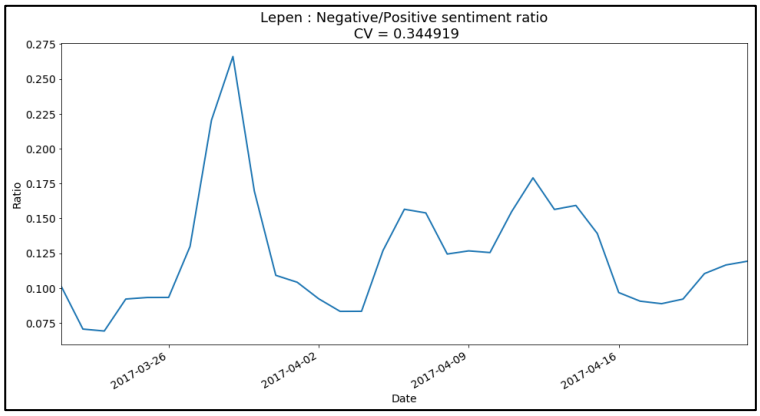
With regard to the second goal of our research (i.e. to provide a data science approach for political strategy optimization), the ideal method for evaluation would be to consult a candidate to change his/her rhetoric according to the findings, and afterwards study the changes in his/her population. For example, by restating opinions regarding a topic which is included in the list of negative words, it is likely that an increase in popularity probably will happen. But, as it is impossible to test that in practice, we can only present the findings of this procedure up to the point before political consulting.

7. Results and evaluation

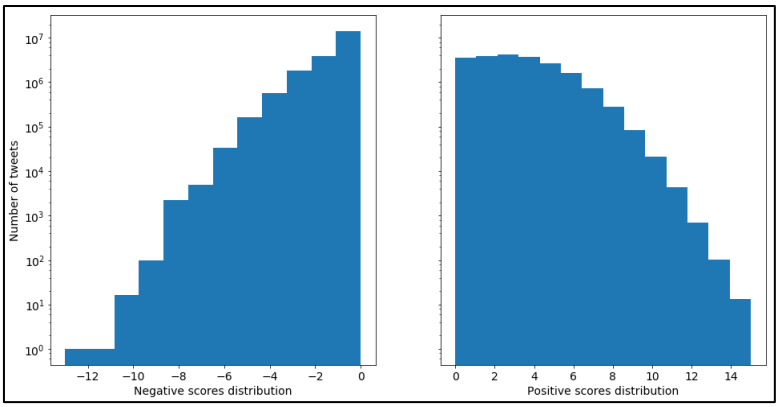
i) Definition of positiveness

The first step of our analysis was to define the notion of positiveness and negativeness that we would use. As already mentioned, one important problem of the sentiment analysis is the sarcasm effect. Sarcasm by definition is "*the use of remarks that clearly mean the opposite of what they say*" (The Cambridge Dictionary). It is easily conceivable why sarcasm is a big trouble for sentiment analysis algorithms: pieces of text are accounted as positive, while actually they bring a negative message. This means that for a tweet dataset where the sarcasm effect exists, there is an unknown fraction of positive tweets which actually are negative. What we observed for every candidate after the scoring process, was the very low ratio *negative to positive tweets*. Here we present this effect for Emmanuel Macron and Marine Le Pen:





We observe that, for the most of the days, this ratio is around 15%, which essentially means that positive tweets are over five times more than the negative ones, regardless the candidate. The problem is that this holds for the whole dataset and thus the variability around this low mean ratio is very low as well.



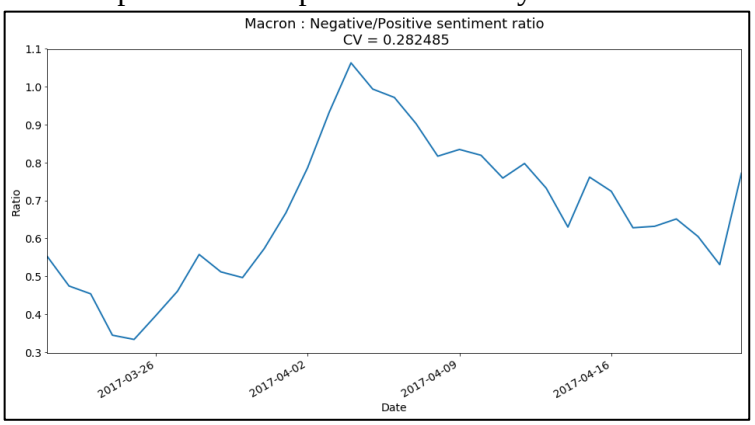
Scores distribution (full dataset)

What we decided to do in order to correct somehow this domination of positive tweets was to introduce a "neutral zone". The idea is that we don't consider as neutral only the tweets with total score=0, but also those with a score which lies in an interval of values close to zero. As sarcasm introduces a bias against the negative scores, we decided to treat the negative scores as we would do without any correction, but consider as neutral also the tweets with a low positive score.

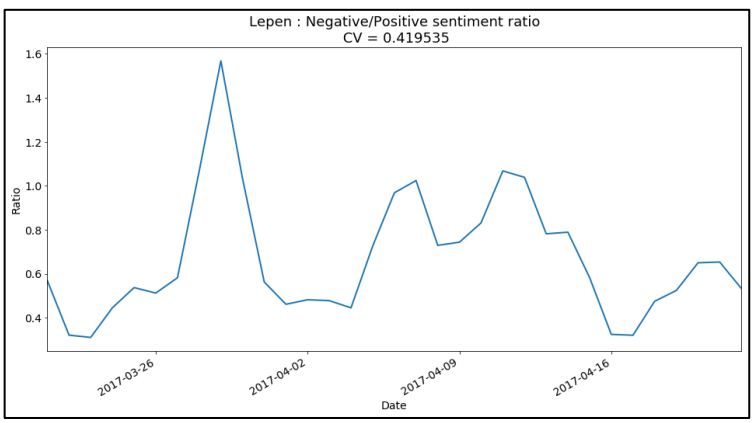
But now, the question is how to choose a good upper bound for such an interval. From the one side, we wanted to balance the ratio (i.e. values not very far from one) and from the other side, we wanted to choose an upper bound that introduces more variance in these ratio plots (because, even with a low mean ratio, if big variance exists, there is no strong bias against negative scores). An objective measure to use for such a decision is the Coefficient of Variation (CV). CV is simply the ratio of standard deviation to the mean, i.e. $CV = \frac{\sigma}{\mu}$ and its advantage is that it is a measure of variation independent of each tested group's mean size. Thus, we were able to compare the ratio plots of different neutral zones for the same candidate and select the one with higher mean CV. We tested three different neutral zones: Firstly, the case with no neutral zone, i.e. the interval is the zero point.

The second tested neutral zone was the interval [0,4] (because – among others – it produced ratio plots with properties close to the desired) and the third was an adaptive method. The idea behind an adaptive method is that the sarcasm effect differs day by day and among candidates, so we cannot use the same interval for the whole dataset, but rather a neutral zone which is adapted to each candidate's and each day's particularities. For the adaptive method, the transformation that we tested was the following: $Upper\ bound_{ij} = \frac{range_{ij}}{2} + \min(score_{ij})$, where *i* is a day, *j* is a candidate and score refers to the total score of each tweet. For example, if the range of scores for candidate A at day 1 is the [-2,5], the corresponding upper bound for the tweets of that day, for that candidate, is 7/2 + (-2) = 1.5. This means that, any tweet in day 1 for candidate A, with total score from 0 to 1.5 (included), will be considered as neutral. The idea behind this transformation is that we move the upper bound of neutral zone accordingly to the maximum observed score. Probably there are many transformations that one could implement in order to choose a good neutral zone, but in this project we tested only this.

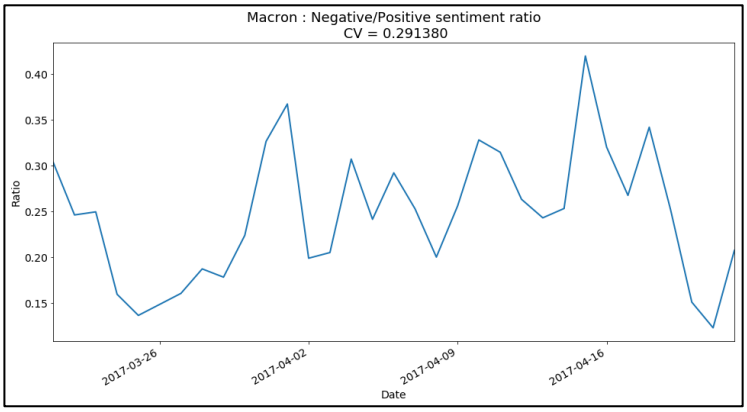
We concluded that, according to the CV criterion, the best choice for the neutral zone is the [0,4]. Here we provide some plots to show why:



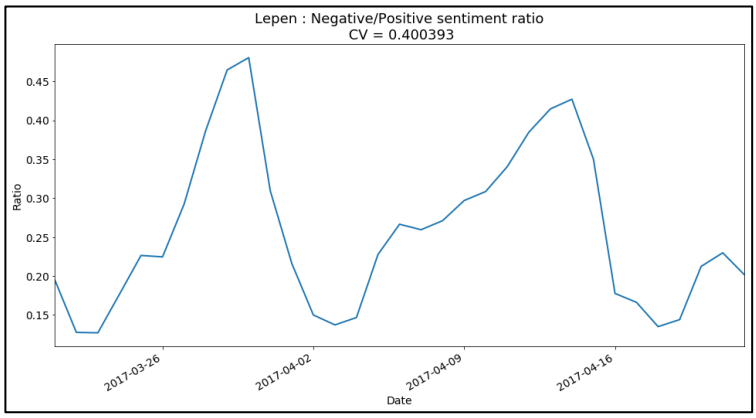
Neutral zone at [0,4].



Neutral zone at [0,4].



Adaptive neutral zone

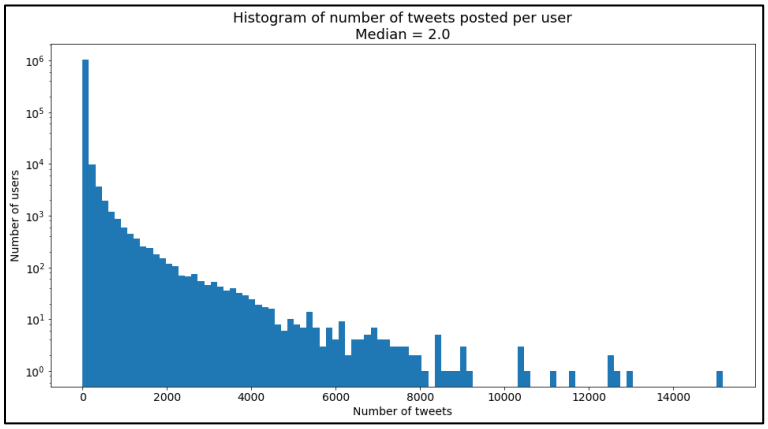


Adaptive neutral zone

For the two candidates of our example, we have the following mean CV values: No neutral zone (initial plots): 0.277, neutral zone at [0,4]: 0.351, adaptive neutral zone: 0.346. We observe that any neutral zone introduces more variance in the negative to positive ratio and, strictly speaking, in this example we should select the neutral zone at [0,4] due to the biggest mean CV. But there is one more important reason to do this selection: ratio balanceness. We observe that when having the neutral zone at [0,4], the negative to positive tweets ratio is much closer to one than in the other two cases. The same holds for the plots of the rest 9 candidates and that was the most important reason to set the neutral zone at [0,4] for the rest part of our analysis.

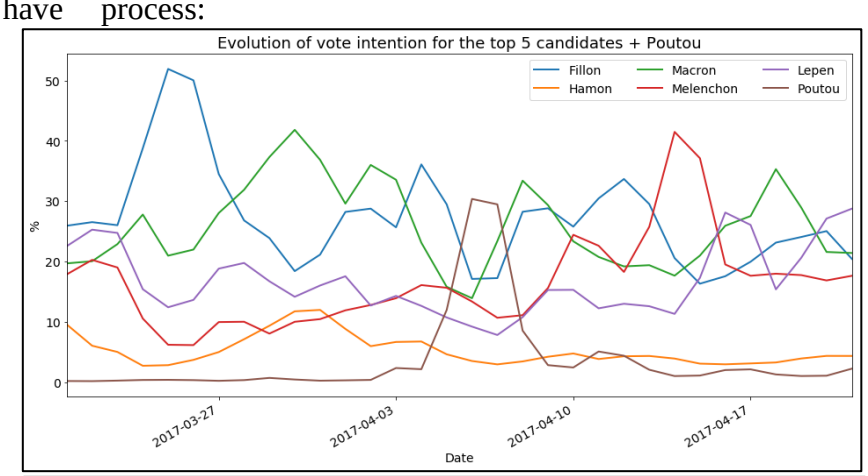
ii) Vote intention

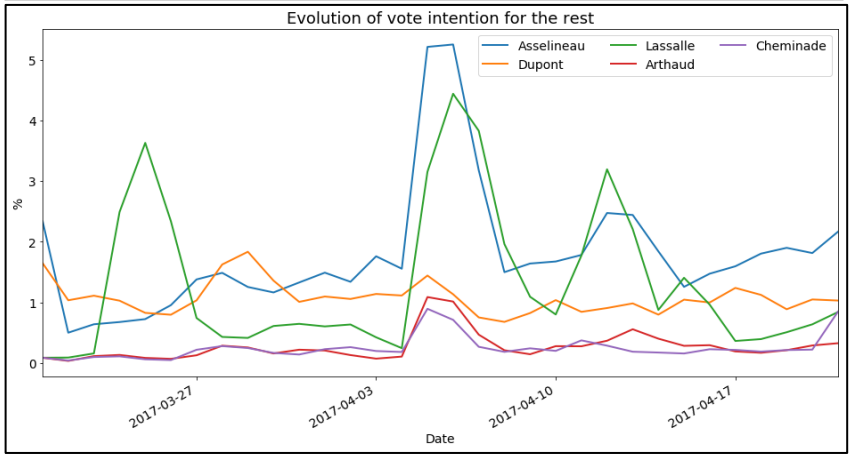
After defining the neutral zone, we were able to treat the positive tweets (i.e. score >4) as potential votes. In order to be as accurate as possible, it was important to take into account only the last tweet per unique user. Actually, in our dataset it was very common to have more than one tweets per user and here is the relevant distribution about this fact:



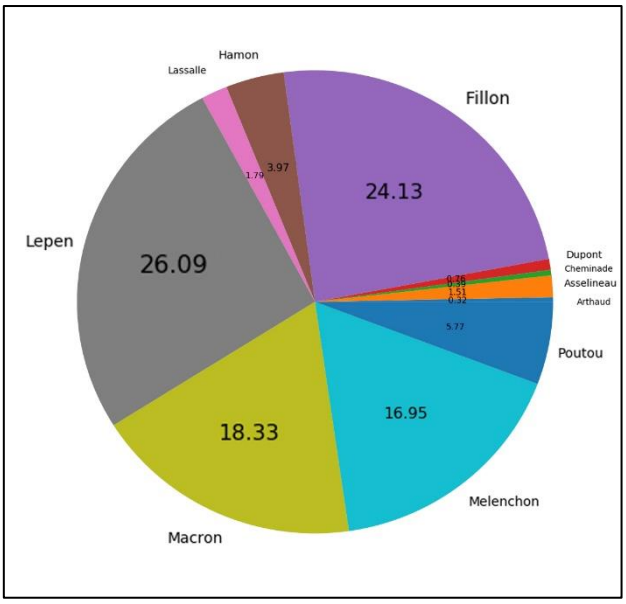
Regarding the long tail of this distribution, probably there are users like news agencies, who post very regularly, or even Twitter-bots which post hundreds of tweets per day. By using only the last tweet per user, we also mitigate this effect.

Firstly, we present some plots about the evolution of the vote intention during our tweet collection process:



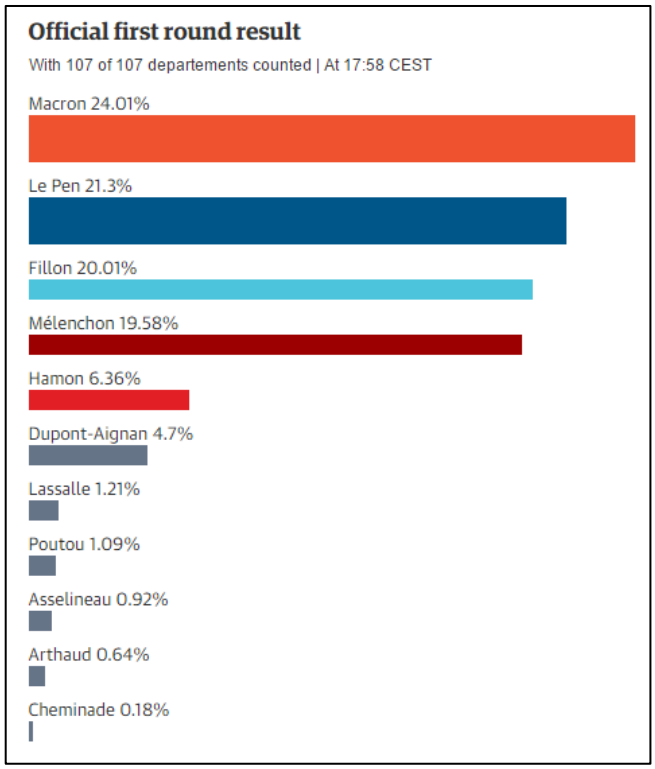


Each point in these plots is a moving average of a two-day window. For each point, we sum up all the positive tweets and we compute the percentages, up to the day before the election (i.e. the data for this analysis are up to 22/04/2017, 00:00). We observe that specific events are very likely to leave their mark on such plots. For example, we know that on 4th April a debate with all eleven candidates happened. In the plot above we can see that on this day many people talked positively more about the six candidates with the lower percentages, than those of the higher ones. The next graph is our forecast about the result of the first round:



Forecast for the first round (percentages).

The actual results of this round are the following (The Guardian, 2017):



It is obvious that there are differences between the two results. In order to see if the forecast as a whole approximated well the final results, we used the Chi-square test in R. The results are the following:

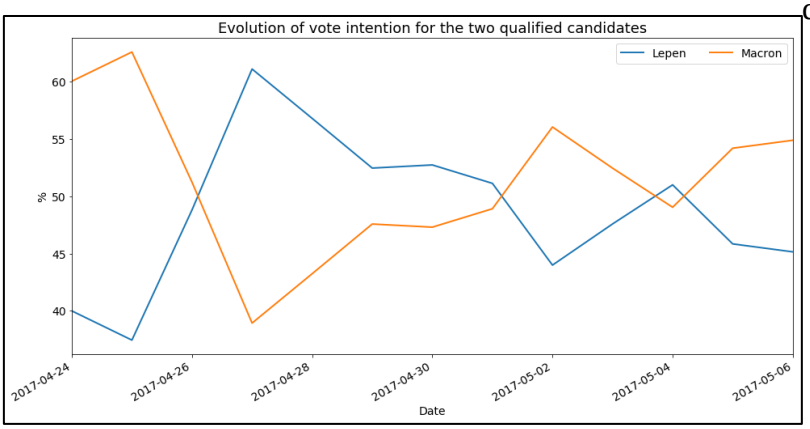
```
        11-sample test for given proportions without continuity correction

data:  est_vot out of rep(sum(est_vot), 11), null probabilities real
X-squared = 83462, df = 11, p-value < 2.2e-16
alternative hypothesis: two.sided
null values:
 prop 1  prop 2  prop 3  prop 4  prop 5  prop 6  prop 7  prop 8  prop 9 prop 10 prop 11
 0.2401  0.2130  0.2001  0.1958  0.0636  0.0470  0.0121  0.0109  0.0092  0.0064  0.0018
sample estimates:
     prop 1      prop 2      prop 3      prop 4      prop 5      prop 6      prop 7      prop 8
0.183330174 0.260857043 0.241283012 0.169527365 0.039683986 0.007590451 0.017864174 0.057679407
     prop 9     prop 10     prop 11
0.015111633 0.003168157 0.003904598
```

This almost zero p-value means that in general our sentiment analysis couldn't predict the election result. However, it seems that there are some estimated percentages which approximate well the real ones, while others not (and these are the reason why the Chi-square test suggested the rejection of our null hypothesis: "*the percentages do not differ significantly*"). In order to test for each percentage separately, we used the so-called Marascuillo procedure (NIST/SEMATECH, 2013). The results of this test were the following:

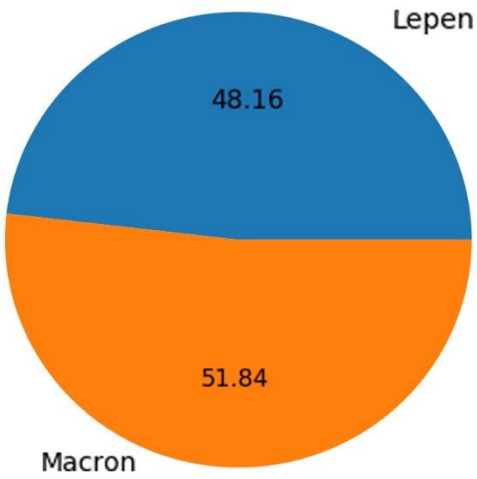
```
      Candidate     Abs Dif.  Crit. Val.  test
[1,]  Macron      "0.0568"  "0.0032"    "Reject"
[2,]  Le Pen      "0.0479"  "0.0036"    "Reject"
[3,]  Fillon      "0.0412"  "0.0035"    "Reject"
[4,]  Melenchon   "0.0263"  "0.0031"    "Reject"
[5,]  Hamon       "0.0239"  "0.0016"    "Reject"
[6,]  Dupont      "0.0394"  "7e-04"     "Reject"
[7,]  Lassalle    "0.0058"  "0.0011"    "Reject"
[8,]  Poutou      "0.0468"  "0.0019"    "Reject"
[9,]  Asselineau  "0.0059"  "0.001"     "Reject"
[10,] Arthaud     "0.0032"  "5e-04"     "Reject"
[11,] Cheminade   "0.0021"  "5e-04"     "Reject"
```

This procedure takes into account both the sample size and the number of people who voted, and it computes some critical values for the absolute differences between our estimations and the actual results. If an absolute difference is larger than the critical value, then this difference is considered statistically significant, which means, for our analysis, that the deviation from the real result was statistically significant. The conclusion from both tests is that our sentiment analysis failed to predict the final result of the first round.

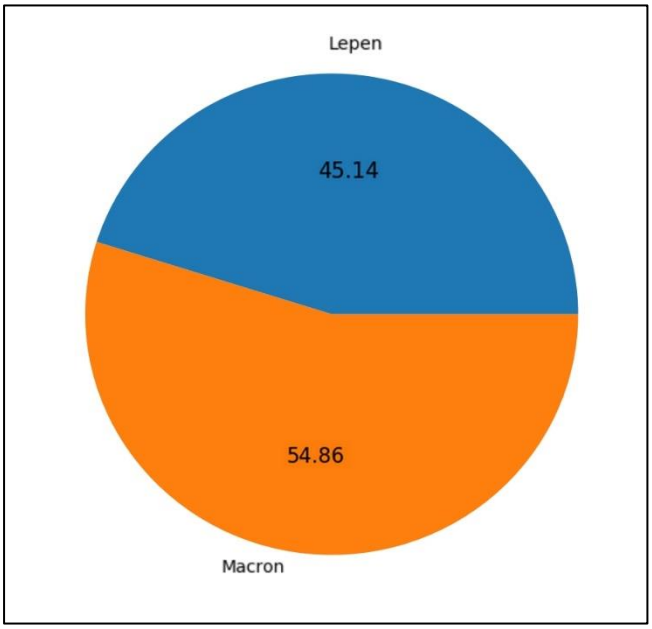
For the second round, the evolution of vote intention can be seen in the following plot:



In this plot also, a two-day moving average is used, in order to make the plot smoother. The actual results of this round were the following: Emmanuel Macron: 66.1%, Marine Le Pen: 33.9% (Ministère de l'Intérieur, 2017). By taking into account the sample from 24/04/2017 at midnight, up to 06/05/2017 at midnight, we obtained the following results:



So, according to our forecasts, Macron would win with a very small difference. But in reality, Macron had a clear win, as he got the 2/3 of the votes. By looking at the plot of vote intention revolution, we observe that such a big difference is observed only at the first time-point (i.e. the two days after the first round) and after that we observe many alterations until the end. In fact, if we take into account only the last two days (i.e. just the last time-point), our predictions come closer to the real result, but it is obvious that for the second round it was impossible to predict the actual percentage difference of the two candidates at any moment.



Taking into account only the last two days (4th and 5th May 2017).

The Chi-square test in this case gave the following results:

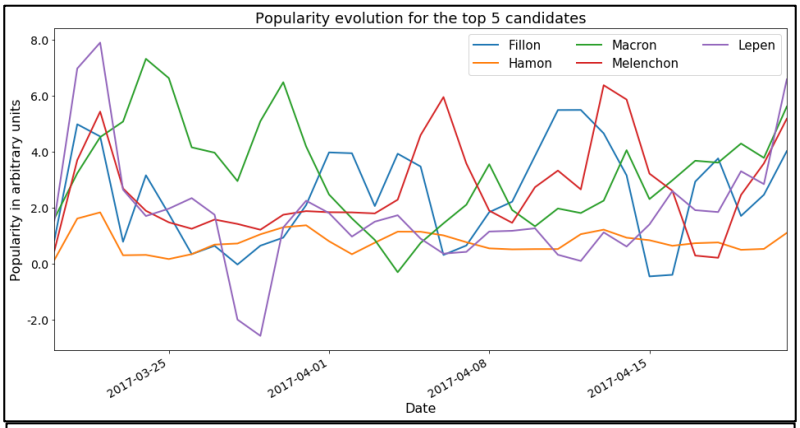
```
        2-sample test for given proportions with continuity correction

data:  c(181188, 195010) out of c(376198, 376198), null probabilities c(0.339, 0.661)
X-squared = 68305, df = 2, p-value < 2.2e-16
alternative hypothesis: two.sided
null values:
prop 1 prop 2
 0.339  0.661
sample estimates:
   prop 1    prop 2
0.4816294 0.5183706
```

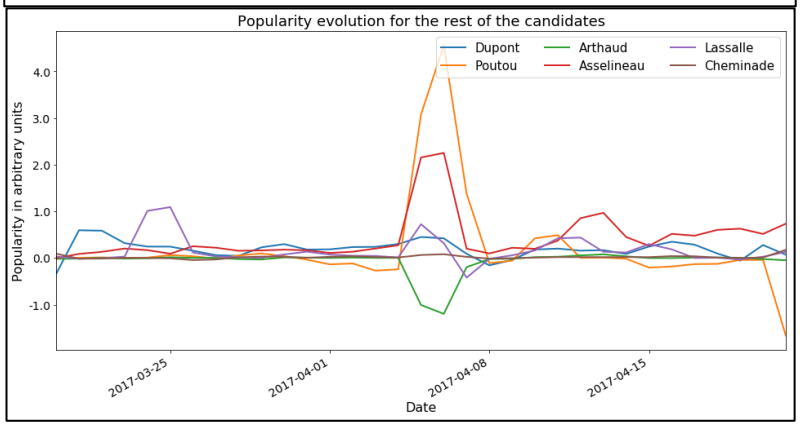
As we can see in this output, it may be the case that we predicted Macron's win, but the difference with the actual percentages was very significant (again a p-value very close to zero).

iii) Progress of popularity

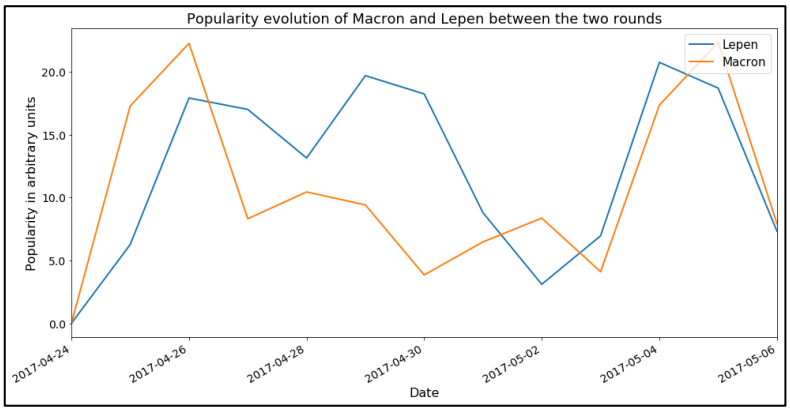
Apart from the time series regarding the vote intention, we also used a different way to plot the popularity of each candidate. $Popularity = sign(score) * (\#\ likes + \#\ retweets + 1)$.
Here the idea is the following: When a tweet has some likes and some retweets, this means that there are people who agree at least with its polarity and that's why they like/share it. Therefore, from each tweet we can extract information about the

number of people who interacted with it in a positive way and use this information to represent a candidate's popularity. We used the formula above and we obtained the following plots:





These plots are referred to the first round. We observe that while they produce different results from those of vote intention, they either cannot provide us a clear image about the real election result. The same holds for the second round:



There is no evidence of the upcoming Macron's emphatic victory by looking at this plot. One possible reason why these plots fail even to approach the reality, is that – in contrast to the previous ones – they allow for users to express their opinion multiple times.

The conclusion from this part of our analysis is that, by implementing sentiment analysis to Twitter posts like the way we did, it is very difficult to draw safe conclusions regarding an election result. Probably further effort is needed for the selection of a representative sample, but, at least for Twitter, such a proper sampling seems rather impossible.

iv) Political strategy optimization

One other interesting outcome of our work is the way it can be used to identify topics or trends about which a candidate receives positive or negative mentions, and thus provide some feedback to the candidates about how their ideas are perceived.

Following our definition of the polarity, we split our dataset per candidate and per polarity and determine the most frequent words in each subset.

We display here the top 8% of the words appearing in positive and negative tweets about candidates Macron and Le Pen, during the 9 days preceding the first round (Appendix, Section E).

Even if some words appear in common, probably because they carry themselves some negativity (terrorism), we can see some interesting differences: 'Mohammedsaou', for Macron, is a member of En Marche, Macron's political party, who had polemical words about January 2015 terrorist attacks in France.

In Section F of Appendix, we provide a translation for a part of the words from both candidates' positive and negative tweets.

Again, we can spot some interesting differences here with, for example, positive mentions for Macron about his program for higher education.

As this analysis can be performed for any period during the campaign (and any time scale; hour, day, week, etc.), this tool can be very beneficial to observe the evolutions of word clouds and, for example, the impact of a speech or a meeting.

Furthermore, we can build weighted graphs with these most frequent words: we weight the edge between two words by the number of their co-appearances in tweets.

The display uses the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991), where the distance between nodes is function of the weight of the edges; this way, we can spot clusters of words appearing together.

In Section G of Appendix, we present the computed word networks for Macron and Le Pen.

We can clearly spot clusters in Le Pen mentions, for example around terrorism, fear, attacks, meaning that some expressions always come together in mentions about Marine le Pen.

On the other hand, clustering is less obvious in Macron graphs, as if the number of expressions about him was broader. A parallel conclusion that can probably be inferred with his program, is that it does not focus only on immigration and terrorism like Le Pen; we leave further analysis of this to sociologists.

Again, these graphs can be built for any time period and any candidate, allowing very fine analysis of processes regarding the opinion dynamics.

8. Conclusions and future work

To summarize this work, we had two main research questions to answer. The first one was if it is possible to predict election results using sentiment analysis in Twitter posts. We had the chance to collect data during the pre-election period of French presidential election 2017 and try to predict the final result one day before the election days (first and second round). The results of our analysis show that an accurate prediction is extremely difficult to achieve. Despite the fact that we had in our disposal a very big sample size to analyse, it seems that there are factors which introduce biases and noise in our dataset, making our forecasts inaccurate.

A very important issue is the role of the dictionaries. We noticed that in many words which we considered them as neutral, our dictionaries provide a small but positive score for them. One more problem regarding the scoring process, is the sarcasm effect. We saw that by taking the raw polarity of the tweets (i.e. without introducing a neutral zone, but considering as a neutral score only the zero value), positive tweets are over five times more than the negative ones. One reason for that may be the sarcasm effect as, by definition, tweets with sarcasm are counted as positive while they are actually negative.

Regarding our second question about political strategy optimization, despite the fact that we were not able to test in practice the feedback from our analysis to the candidates' popularity, it seems that interesting facts may come to the front. We can see which topics are related to positive and negative sentiments and the magnitude of their importance. Moreover, by using network visualization, we can show not only which topics are related with positive or negative feelings, but also their relations with other topics in clusters of words.

Taking into account the problems that arose from our analysis, we can mention some topics that could be investigated in future. A very interesting topic to investigate is constructing more "intelligent" scoring processes. For example, there are cases where a word has a depending-to-context polarity and therefore its score is not the same for all the texts. But in our analysis, we had a static dictionary and every available word was scored with the same value regardless the context. One other topic is the solution to the sarcasm effect. As we already mentioned, there is ongoing research about how to find sarcasm in text and the development of such a mechanism could improve the scoring procedure a lot. Last but not least, a topic we suggest for further research is the development of algorithms that can "understand" not only text but also other kind of information. For example, we observed that many tweets were posts of images with a clearly positive or negative message. Many people choose this way to express themselves, but due to our text-only analysis, we

omitted this kind of data. Developments in deep learning technologies could provide efficient tools to cope with this kind of problems.

References

Abdaoui, A., Azé, J., Bringay, S. & Poncelet, P., 2016. FEEL: French Expanded Emotion Lexicon. *Language Resources and Evaluation*.

Burnap, P. et al., 2016. 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies*, Volume 41, pp. 230-233.

Castellano, C., Fortunato, S. & Loreto, V., 2009. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81(2), p. 591–646.

Fruchterman, T. M. & Reingold, E. M., 1991. Graph Drawing by Force-directed Placement. *Software - Practice and Experience*, 21(11), pp. 1129-1164.

Majaski, C., 2016. *Researchers develop sarcasm detector for Twitter, and that's no joke.* [Online]
Available at: https://www.digitaltrends.com/social-media/sarcasm-detector-twitter/
[Accessed 13 May 2017].

Ministère de l'Intérieur, 2017. *Election présidentielle 2017.* [Online]
Available at: http://elections.interieur.gouv.fr/presidentielle-2017/
[Accessed 13 May 2017].

Ministère de l'Intérieur, 2017. *Résultats globaux du second tour de l'élection du Président de la République 2017.* [Online]
Available at: http://www.interieur.gouv.fr/fr/Elections/Election-presidentielle-2017/Resultats-globaux-du-second-tour-de-l-election-du-President-de-la-Republique-2017
[Accessed 13 May 2017].

Mohammad, S. M., Zhu, X., Kiritchenko, S. & Martin, J., 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing and Management*, Volume 51, pp. 480-499.

NIST/SEMATECH, 2013. *Comparing multiple proportions: The Marascuillo procedure.* [Online]
Available at: http://www.itl.nist.gov/div898/handbook/index.htm
[Accessed 14 May 2017].

Perez, S., 2016. *Analysis of social media did a better job at predicting Trump's win than the polls.* [Online]
Available at: https://techcrunch.com/2016/11/10/social-media-did-a-better-job-at-predicting-trumps-win-than-the-polls/
[Accessed 11 May 2017].

Piolat, A. & Bannour, R., 2009. An example of text analysis software (EMOTAIX-Tropes) use: The influence of anxiety on expressive writing. *Current psychology letters [En ligne]*, 25(2).

Porcaro, G. & Müller, H., 2016. *Tweeting Brexit: Narrative building and sentiment analysis.* [Online]
Available at: http://bruegel.org/2016/11/tweeting-brexit-narrative-building-and-sentiment-analysis/
[Accessed 11 May 2017].

Schmid, H., 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees.* s.l., Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Schmid, H., 1995. *Improvements in Part-of-Speech Tagging with an Application to German.* s.l., Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

SentiStrength, 2011. *SentiStrength versions for other languages.* [Online]
Available at: http://sentistrength.wlv.ac.uk/
[Accessed 26 March 2017].

The Cambridge Dictionary, n.d. *Definition of "sarcasm" from the Cambridge Advanced Learner's Dictionary & Thesaurus © Cambridge University Press.* [Online]
Available at: http://dictionary.cambridge.org/dictionary/english/sarcasm
[Accessed 13 May 2017].

The Guardian, 2017. *French presidential election: first round results in charts and maps.* [Online]
Available at: https://www.theguardian.com/world/ng-interactive/2017/apr/23/french-presidential-election-results-2017-latest
[Accessed 13 May 2017].

Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M., 2010. *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.* s.l., Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 178-185.

Xiong, F. & Liu, Y., 2014. Opinion formation on social media: An empirical approach. *Chaos*, March, 24(1), p. 013130.

York, A., 2017. *Social Media Demographics to Inform a Better Segmentation Strategy.* [Online]
Available at: http://sproutsocial.com/insights/new-social-media-demographics/
[Accessed 13 May 2017].

APPENDIX

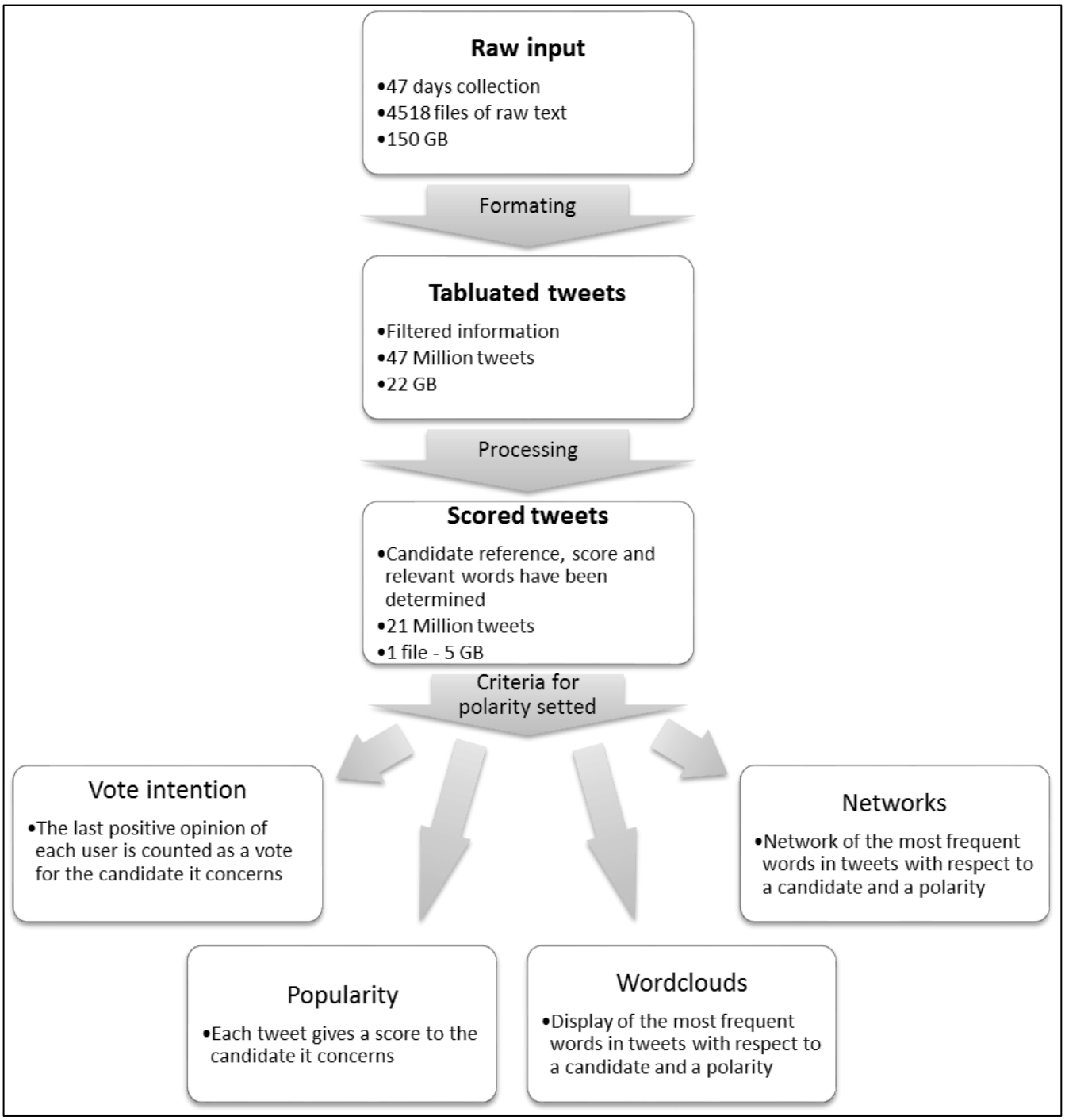
A. : Tableau 1: Valuable keywords that were considered for each tweet

Name of the field	Description
createdAt	Date at which the tweet was posted (microsecond precision)
id	Unique number identifying the tweet
text	Text content of the tweet
favoriteCount	Number of time the tweet has been set to 'favorited'
retweetCount	Number of time the tweet has been retweeted
lang	Language in which the tweet is written
retweetedStatus	If the tweet is a retweet, this field contains all the information of the original tweet. We kept only the original id, and saved the rest as a separate tweet.
userMentionEntities	Mentions of users contained in the text
hashtagEntities	Hashtags contained in the text
user	Field containing many information about the user; we keep only the unique *user_id* identifier and its *user_description*

B. Tableau 2: Description of quantities available for each tweet after the processing step

Name of the field	Description
createdAt	Date at which the twet was posted
id	Unique number identifying the tweet
user_id	Unique number identifying the user
favoriteCount	Number of time the tweet has been set to 'favorited'
retweetCount	Number of time the tweet has been retweeted
candidate_ref	Name of the candidate the tweet is referring to
Relevant words	list of words contained in the original text identified as relevant
score_neg	sum of the scores of negative words
score_pos	sum of the scores of positive words

C. Diagram 1: The data analytics pipeline



D. Dictionaries

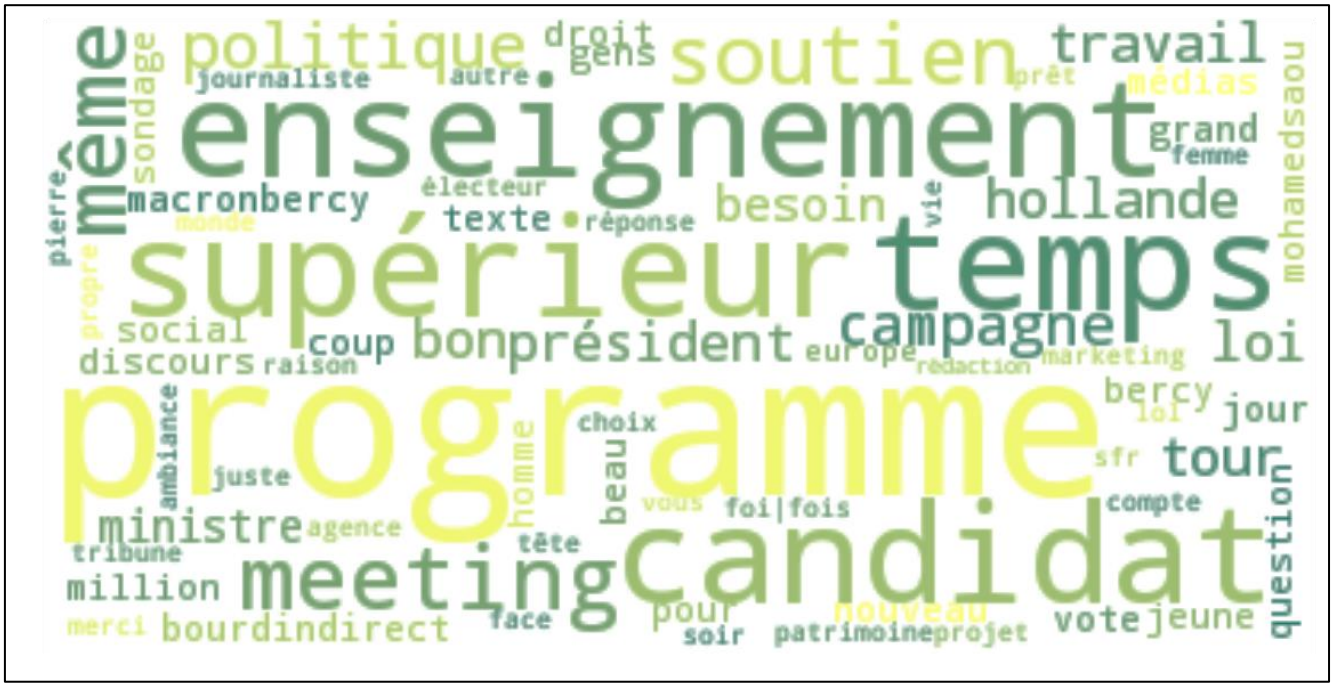
To determine the polarity of a tweet, our approach was to sum the polarity of each word or expression found in the tweet.

We used for this purpose the combination of three dictionaries:

1. FEEL: a French Expanded Emotion Lexicon (Abdaoui, et al., 2016). This dictionary gathers 14000 entries associated with a polarity (and other emotions we didn't use). These entries are, as in a dictionary, *lemmas*: base words without any mark of feminine, plural or conjugation, to avoid duplicated entries.
2. EMOTAIX (Piolat & Bannour, 2009). This dictionary classified many words and expressions between positive and negative emotions. To make it comparable with the FEEL dictionary, and more robust for our analysis, we lemmatized each entry thanks to TreeTagger.
3. SentiStrength (SentiStrength, 2011). SentiStrength is a sentiment analysis tool which provide dictionaries for 17 languages, were each word has either a positive or a negative value, depending on its meaning.

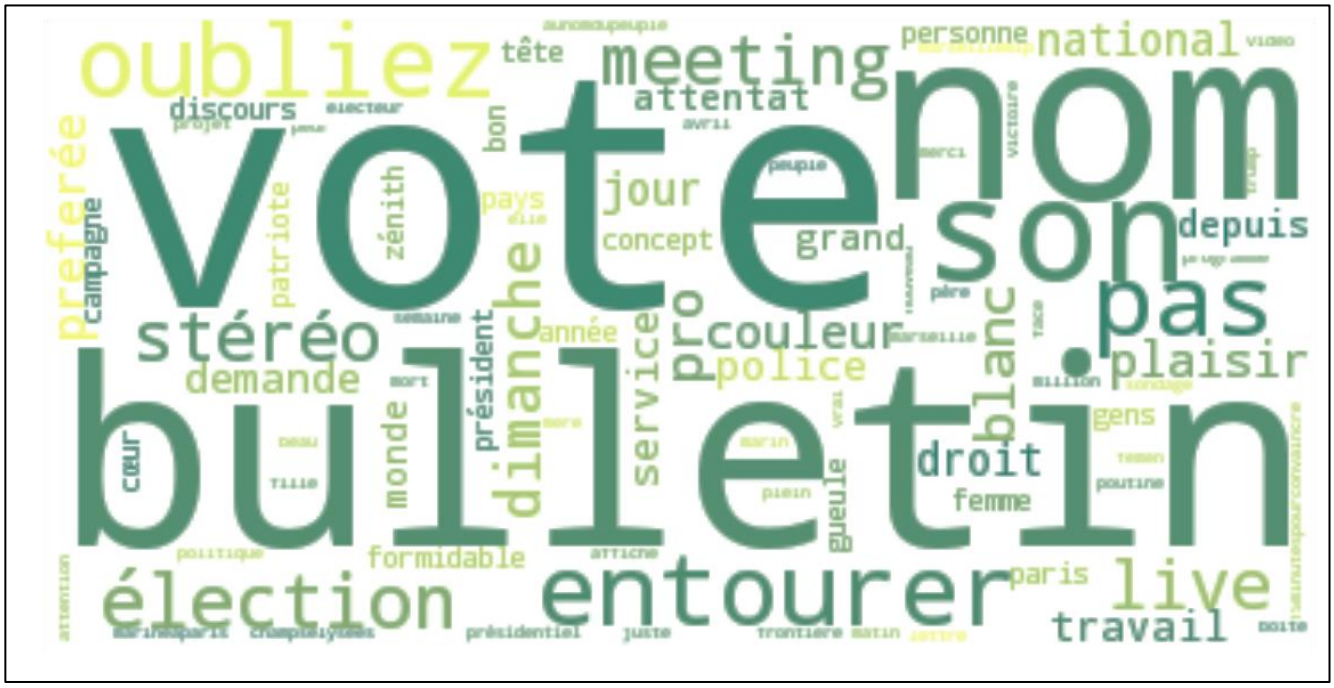
These three resources have been merged to build a unique dictionary of words/expressions vs polarity, with lemmatized entries.

E. Word clouds, first round, Macron and Le Pen



Macron – positive tweets

Macron – negative tweets



Le Pen – positive tweets

Le Pen – negative tweets

F. Part of words translated in English

Translation Macron+

enseignement	teaching
campagne	campaign
temps	time
supérieur	higher
même	same
soutien	support
patrimoine	inheritance
droit	rights
sondage	poll

Translation Le Pen+

bulletin	ballot
oubliez	forget
nom	name
entourer	surround
préférée	favorite
monde	world
travail	work
monde	world

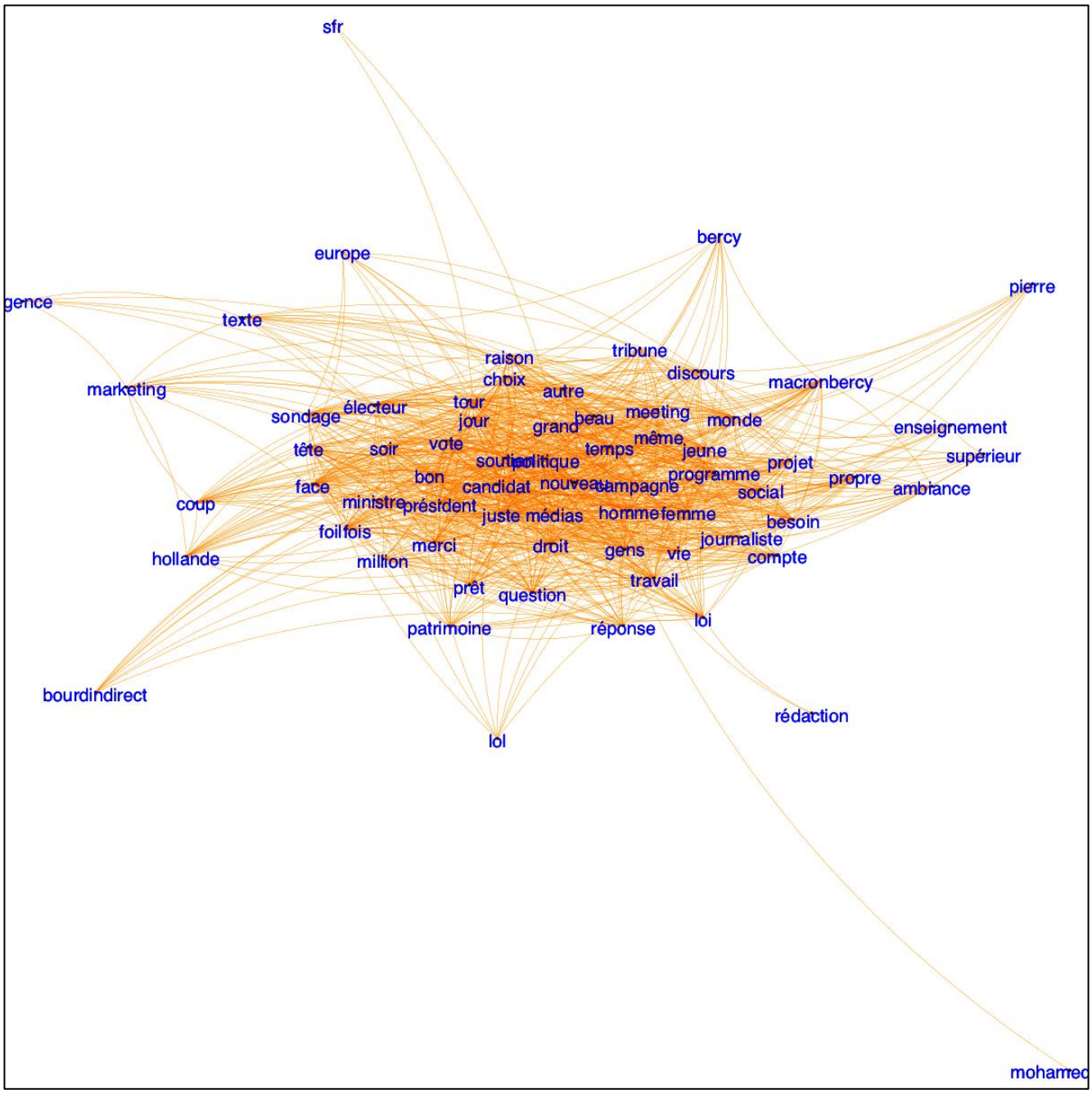
Translation Macron-

petit	small
communitariste	communitarianist
lutte	fight
radical	extremist
porte-parole	spokeperson
seul	alone
hollande	name of the previous predisent
censure	censorship
histoire	history
gauche	left wing
bercy	Ministry of Economy

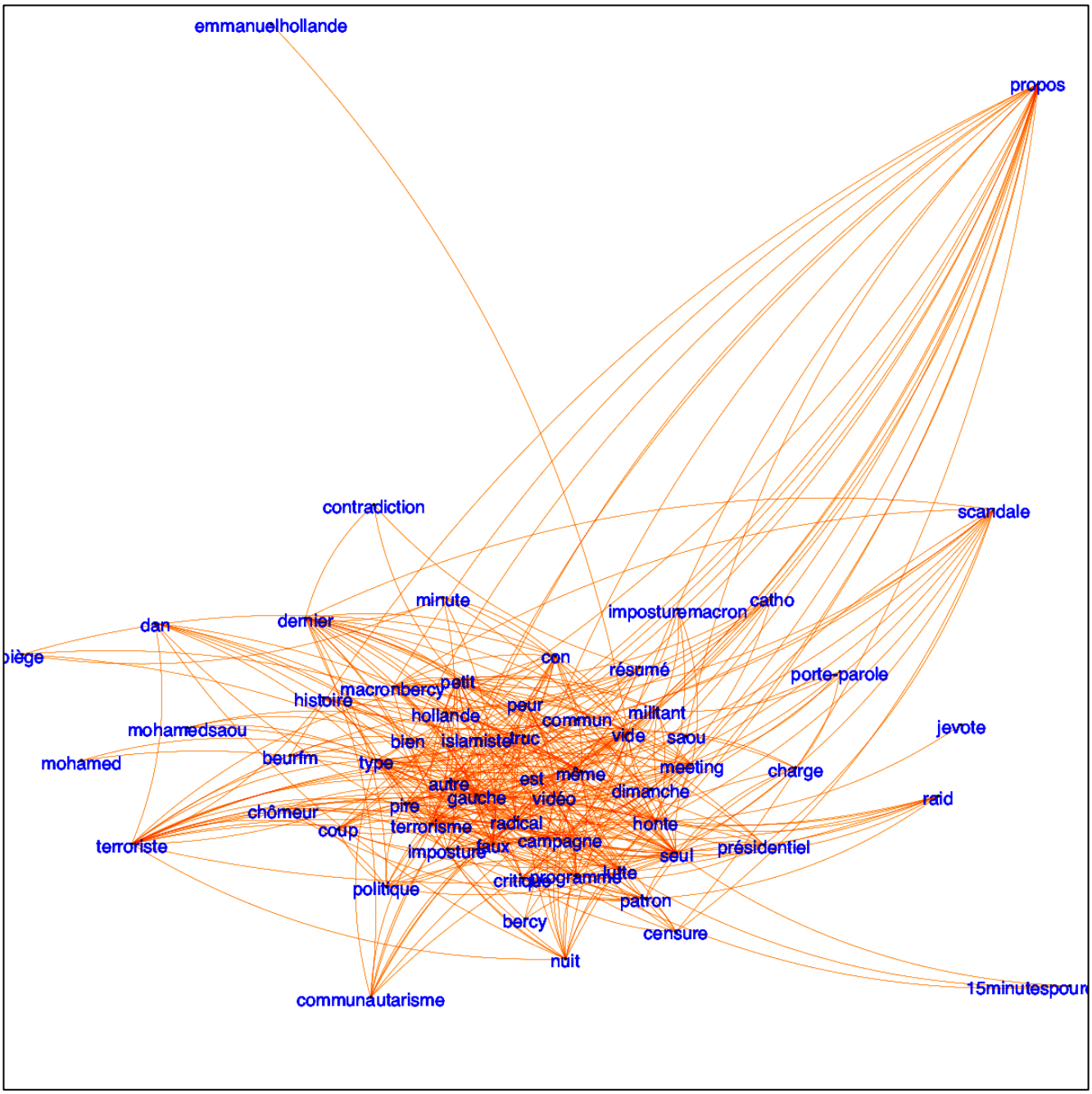
Translation Le Pen-

cause	reason
étranger	foreigner
vif	lively
attentat	terrorist attack
frontier	border
braquage	robbery
merde	shit
fusillade	shooting
haine	hate
honte	shame

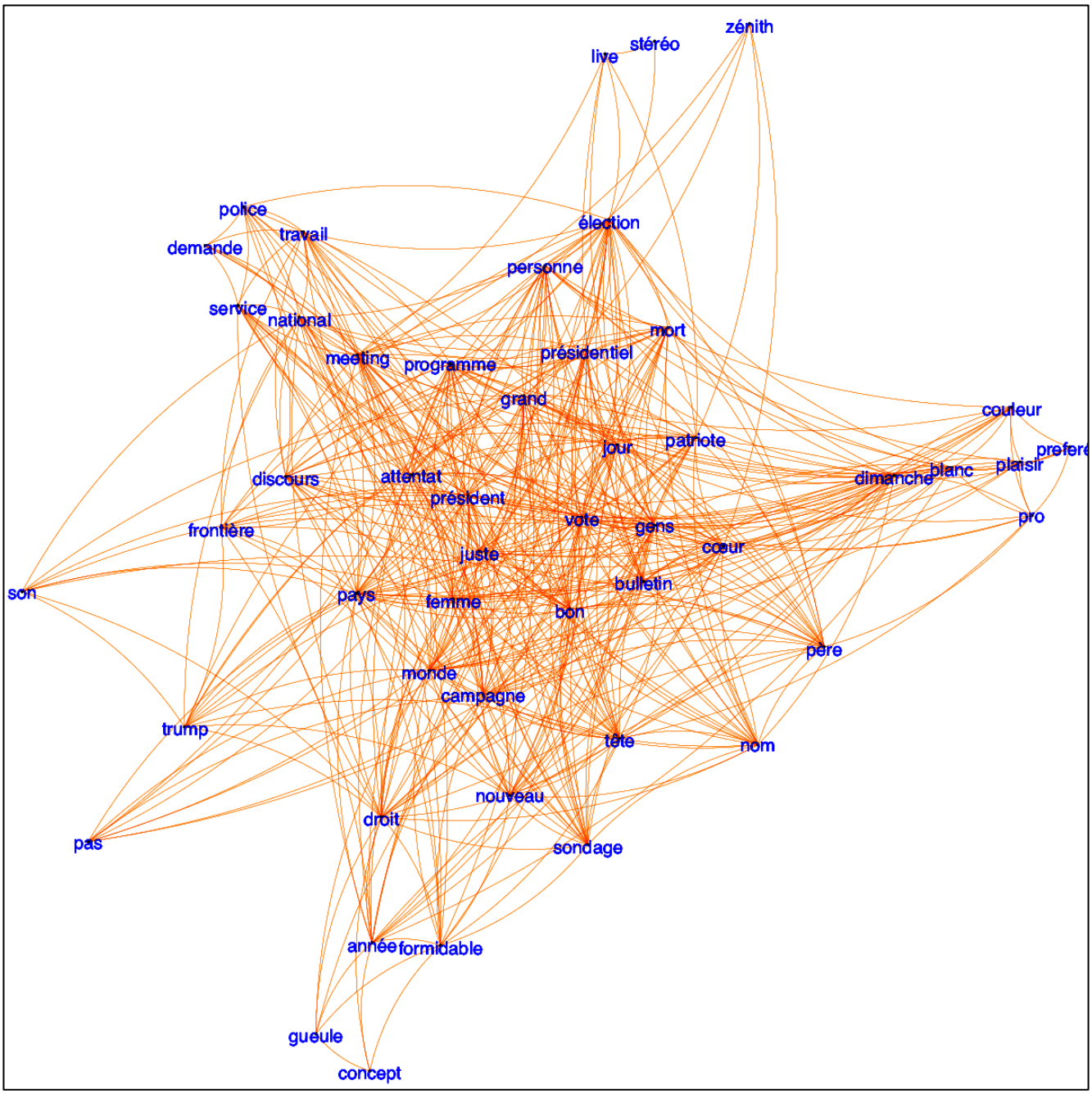
G. Word networks



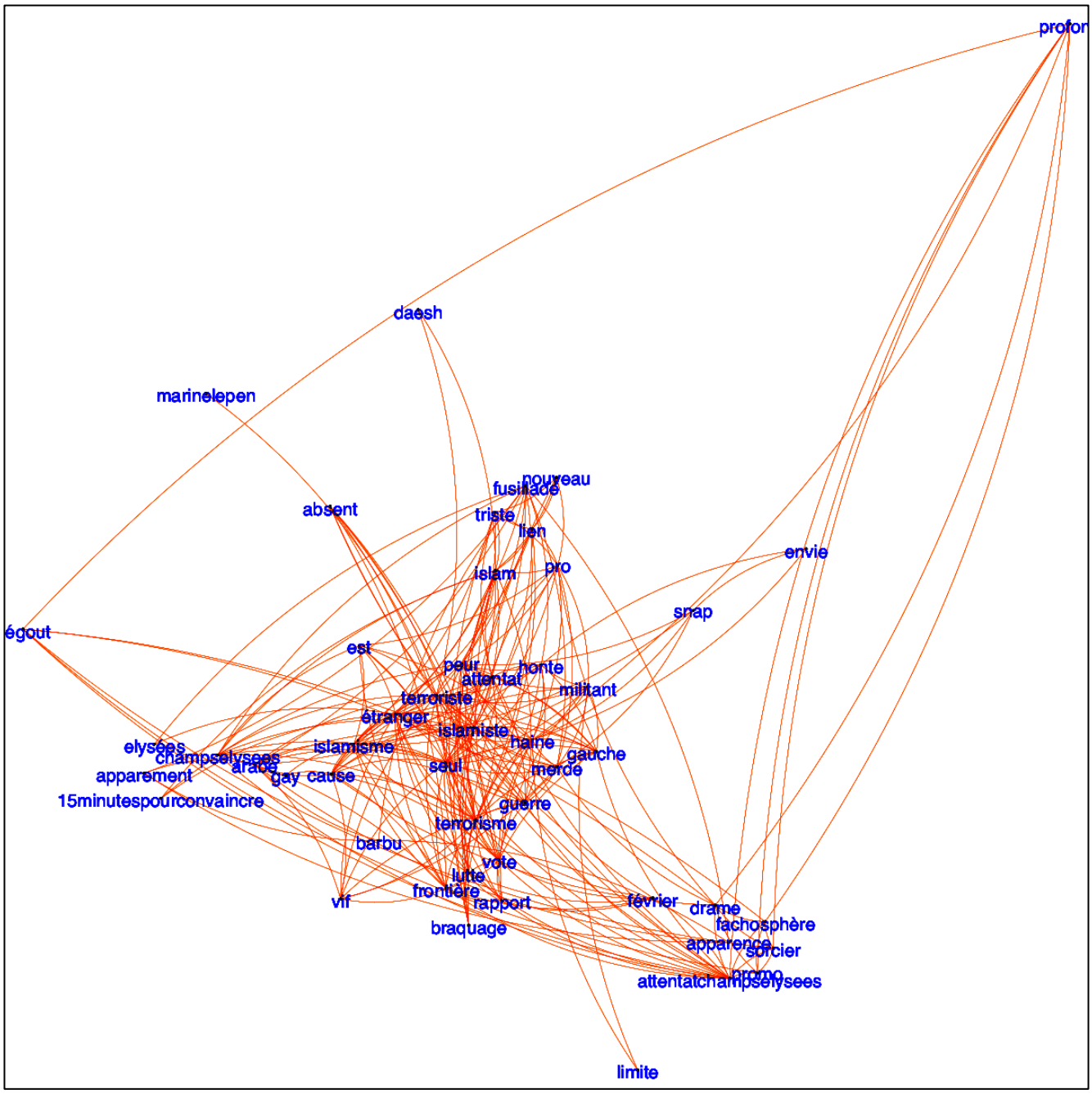
Macron, positive tweets



Macron, negative tweets



Le Pen, positive tweets.



Le Pen, negative tweets