The #Idéo2017 Platform

Julien Longhi, Claudia Marinica, Nader Hassine, Abdulhafiz Alkhouli, Boris Borzic

The #Idéo2017 Platform

Julien Longhi[1], Claudia Marinica[2], Nader Hassine[1,2], Abdulhafiz Alkhouli[2], Boris Borzic[2]
[1] University of Cergy-Pontoise, AGORA
[2] ETIS Lab UMR 8051 University of Paris-Seine, University of Cergy-Pontoise, ENSEA, CNRS
33 Boulevard du Port, 95000 Cergy-Pontoise, France
julien.longhi@u-cergy.fr, claudia.marinica@ensea.fr, nader2hassine@gmail.com,
abdulhafiz.alkhouli@ensea.fr, boris.borzic@ensea.fr

Abstract

The #Idéo2017 platform allows citizens to analyze the tweets of the 11 candidates at the French 2017 Presidential Election. #Idéo2017 processes the messages of the candidates by creating a corpus in almost real time. By using techniques from linguistics supplied with tools, #Idéo2017 is able to provide the main characteristics of the corpus and of the employment of the political lexicon, and allows comparisons between the different candidates.

Keywords: NLP for social media, NLP applications, textometry, tweets mining

1. Introduction

Social networks are becoming an important source for citizen's information, concerning mainly their "consumption" of information (Mercier, 2014). Twitter, the most known micro-blogging platform (Kaplan and Haenlein, 2010), by allowing the publication of short messages (140 characters), gives to social networks a new dimension. Indeed, Twitter can be used to assess how users react to social (Longhi and Saigh, 2016), political (Longhi et al., 2014; Conover et al., 2011), or economic issues. Therefore, the textual data (messages) sent on Twitter can be used to extract emotions, feelings, opinions, etc., of the users (Johnson and Goldwasser, 2016).

The analysis of political tweets during the election campaigns, or specific events, is increasing and can be seen as a specific type of political discourse (Longhi, 2013). Among the studies on this subject, Roginsky and Cock (2015) propose a qualitative analysis of interactions on Twitter, but they are limited to "the discursive and communicational analysis of the types of expression on Twitter that we can observe, with a particular interest in the way studied actors present and put forward themselves". Johnson and Goldwasser (2016) propose a classification of positionings based on the most frequent words, while the analysis of Vidak and Jackiewicz (2016) focuses on emotions. Moreover, many studies have proposed approaches to predict the result of the presidential elections (or to explain why the prediction is not possible) by analyzing the tweets (Tumasjan et al., 2010; Gayo-Avello et al., 2011; Metaxas et al., 2011).

Thus, there is an extensive literature on the analysis of political tweets, but these works are difficult to gather because they come either from the computer sciences, either from the humanities and social sciences (communication sciences, linguistics). Moreover, despite the unquestionable interest in outlining political facts, these results are not accessible by citizens interested in this subject.

In this context, this article presents #Idéo2017, a new and innovative platform making analytic information available to citizens. #Idéo2017 proposes a tool for analyzing tweets and speeches (relayed on Twitter) of the 11 candidates at the presidential election in France in 2017. #Idéo2017 analyzes the messages of the candidates by creating a corpus in almost real time (updated every 24 hours) with the tweets published in candidates' official accounts (from September 1st 2016 to May 7th 2017). Using techniques and metrics derived from linguistic tools, the new platform provides the main characteristics of the corpus and allows comparisons between the different candidates.

The rest of the paper is structured as follows. In Section 2., we provide a general description of the tool, and in Section 3. we present the analyses that can be carried out. Then, in Section 4. we detail the tool's development, as well as the technological choices that we made. Section 5. concludes the papers and provides a set of perspectives.

2. Description of #Idéo2017

#Idéo2017 is a web platform available online allowing to analyze the messages, posted on Twitter, related to political news (meetings, debates, television broadcasts, etc.). Its objective is to make available on the web for average citizens a set of statistical analyses and data visualization tools applied on the Twitter messages. The choice of a web platform rather than software to be installed comes from the fact that we want citizens that are non-specialists of tools and software to be able to have access to the analyses' results without going through the phases of corpus formation, tagging, etc. Thus, the citizens can make their own queries (based on linguistic and textometric criteria, more precisely, the most used words by political personalities, analyses of similarities, ALCESTE algorithm, etc.) and obtain comprehensible result.

The #Idéo2017 platform follows the processing chain shown in Figure 1: (1) retrieving the set of tweets of the candidates, (2) setting up a backup of tweets, (3) indexing tweets to facilitate the search process, (4) applying a set of linguistic analyses on tweets, (5) setting up a search engine on the tweets, and (6) displaying the results on a web page. In this processing chain, we are firstly interested in extracting candidates' tweets: we want to extract the tweets daily and to propose to the users to analyze the current database; for example, on April 4th, 2017 the users are able to analyze

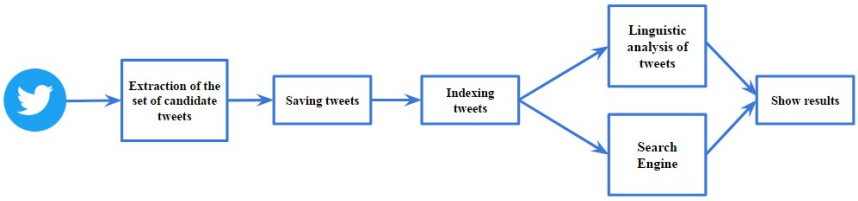

Figure 1: Processing chain in the #Idéo2017 platform.

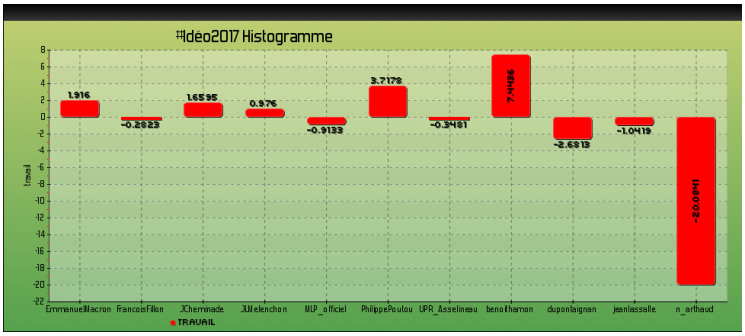

Figure 2: Analyses for the word *work*: under- / over- use.

the tweets sent by the candidates until April 3rd, 2017. To this end, we took advantage of the work that we carried out in the context of the extraction of the *Polititweets* corpus (Longhi et al., 2014), which is available and documented on the Ortolang platform[1].

In the tweets backup and indexing steps, we are interested in the issues of tweets storage, and, respectively, in the implementation of an indexing system. These two steps facilitate the access to the tweets and the development of an intelligent search engine.

In the linguistic analysis step, we propose to the user a set of analyses to be carried out on the set of tweets. These analyses, described in the next section, concern: the use of a specific word and its derivatives by the different candidates, the words associated with a specific word, the word cloud, themes, relations between words, and the specificities of the different candidates. In addition, we have developed an intelligent search engine based on the faceted search process that allows to the user to perform searches on tweets using complex filters.

3. Analyses and Search Engine Description

The #Idéo2017 platform, available at http://ideo2017.ensea.fr/plateforme/, proposes two types of analyses and the search engine. These three elements are described below.

3.1. The Analysis "I Analyze the Tweets that Contain the Word [Word]"

The analysis "I analyze the tweets that contain the word [word]" allows to the user to choose a word among the 13 words that are often used in political debates (Alduy, 2017).

[1]https://repository.ortolang.fr/api/content/comere/v3.3/cmr-polititweets.html

Our choice on limiting the user to 13 words is related to the computation time of analyses and graphics which would be too high if performed in real time. So, all the computations for the 13 words are performed at the same time (during the night) and the results are kept on the drive and displayed when requested. In a new version of our platform we plan to improve this aspect. On the other side, if the user wish to search for a word in the tweets he/she can use the search engine.

The list of 13 selected words is: *France*, *state*, *Republic*, *people*, *law*, *work*, *freedom*, *democracy*, *security*, *immigration*, *terrorism*, *Islam* and *secularism*.

Once the word is chosen, the user has access to four analyses:

- The first analysis allows to identify the use of the chosen word by the different candidates, and the results are presented in the form of two graphs: one for the computation of specificities (the under- / over- use of the word by the candidates), and the other one for the frequency of use of the word by the candidates.
- The second analysis detects the words associated with the chosen word for all candidates. This analysis of co-occurrences is presented in the form of a graph of associated words.
- The third analysis consists in computing the use of the chosen word and its derivatives (on contrary to the first analysis) by the different candidates.
- The last analysis creates a word cloud that allows to display graphically the lexicon.

To exemplify these analyses, let us consider the word *work* (*travail* in French). We can see in Figure 2 the under- / over- use of the word *work* (computation of specificities),

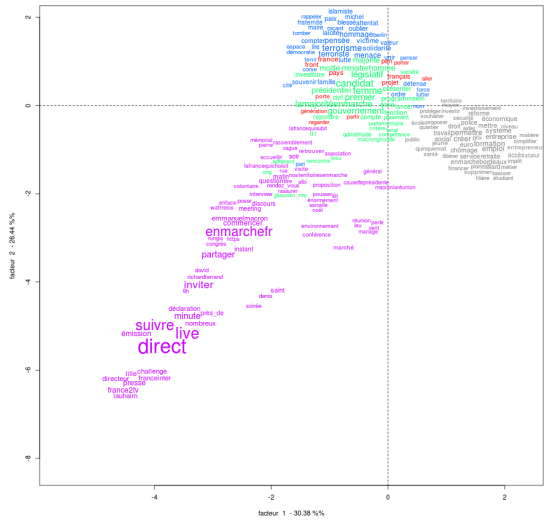

Figure 3: Analysis of the tweets of the candidate Emmanuel Macron: the themes (by lexical categories).

and in Figure 5 in Appendix Section the frequency of this word in all candidates' tweets.
The Figure 6 in Appendix Section allows us to see the words associated with the word *work* (analysis of similarities), and the Figure 10 in Appendix Section the word cloud.
It is important to outline that the analyses are performed after lemmatizing each word (for example, for the lemma liberty, we can have the forms liberty, liberties, etc.).

3.2. The Analysis "I Analyze the Tweets of [Candidate]"

The analysis "I analyze the tweets of [candidate]" allows to the user to perform a set of analyses on each of the 11 candidates: *N. Arthaud*, *F. Asselineau*, *J. Cheminade*, *N. Dupont-Aignan*, *F. Fillon*, *B. Hamon*, *J. Lassalle*, *M. Le Pen*, *E. Macron*, *J.-L. Mélenchon* and *P. Poutou*.
Once a candidate is chosen, the possible linguistic analyses are the following:

- The first analysis allows to detect and analyze the words the most used by the candidate;
- The second analysis, called "Themes", proposes, for the chosen candidate, to group together the words that are semantically close in order to outline the important themes discussed by the candidate;
- The third one analyzes the similarity between the words of each candidate in a graphical form;
- The forth analysis displays the lexicon of the tweets in the form of a word cloud;
- The last one is devoted to the analysis of the tweets of all the candidates. It allows to identify the specific words and categories of the different candidates and to compare them.

For example, let us consider the candidate Emmanuel Macron; the Figure 8 in Appendix Section presents the most frequent words used in his tweets, and the Figure 3 exposes the themes of his tweets.

3.3. The Search Engine

In the previous 2 sections we presented two type of analyses available in the #Idéo2017 platform; the third feature of #Idéo2017 is an intelligent search engine in real time (the graphical interface is shown in Figure 9 in Appendix Section) which offers a faceted search over the tweets: by candidate, by hashtag or by mention. Moreover, it provides complete search flexibility allowing to the user to compare the candidates using complex queries, and it also allows to sort the results by the date or by the commitment.
Twitter includes already a classical search engine in its interface; in order to propose a richer search experience, we built our search tool as a hybrid system bringing together the results of real time queries on the tweets and the synthesis of several tweets by aggregating the information via several facets (filters) and linguistic computations or word clouds. Thus, for a specific word or theme, our goal is to provide an access to the original tweets for each candidate, but also to compute the exact distribution of tweets per candidate and per theme.
The tweets' distribution information permits to contextualize each query, because our objective is at the same time to build a search engine, but also to propose a business intelligence system allowing to study the communication strategies of the candidates. It is important to outline that the two features (search engine and business intelligence system) have completely different goals: meanwhile the search engine struggles against the noise (all the answers should be the most pertinent), a business intelligence system, as a benchmark, aims to reduce the silence (all pertinent tweet should be shown to the user). However, when we try to reduce the silence, we increase the noise, and the more we fight against the noise, the more the silence becomes loud. In computer science, this complexity is assessed by two complementary metrics named the precision and the recall. To overcome this challenge, we took advantage of the applications in business intelligence (BI), the tools in reporting and the systems of knowledge management. Generally dedicated to dashboarding or back-office tools, we propose to the citizens to extend the queries results with a synthetic information provided by the linguistic analyses and integrated visually and progressively.

4. Tool Development

For the development of the tool, we had to tackle different technological problems. We will present the solutions, shown in Figure 11 in Appendix Section, that we have chosen for each problem. First, we used the Twitter API to retrieve the tweets directly from the official accounts. Then, we stored these tweets in the MongoDB[2] NoSql database; its advantage consists in a flexible, document-oriented structure that does not require complex queries to access the data. Then, we decided to use Elasticsearch to

[2]www.mongodb.com

store the tweets; Elasticsearch (Kononenko et al., 2014) improves the response time of our tool especially when using the search engine.
Given that Elasticsearch's standard method performs a classical search without dealing with the derivatives of a word, before sending the data from MongoDB to Elasticsearch, we prepare a data index that takes into account the derivatives of the words. All the communication between these tools uses the Java language. For the analysis part, we compared several software packages for the linguistic analysis of corpus (Hyperbase, Lexico3, Trameur, TXM and Iramuteq). We studied their analyses but also their availability in open source and/or API. After our study, we decided to use several features of Iramuteq[3] that are implemented in PHP and available in open source. To this end, several modifications were needed in the implementation of Iramuteq. We also used PHP Word Cloud[4] for word clouds and pChart[5] for graphics.
After the election, a second version of the platform was released with new features: the first one, for data visualizations (as shown in Figure 7 in Appendix Section), and the second one, for sub-corpus extraction from the complete corpus by the candidate and the period (as shown in Figure 4). The latter is very important and useful for the humanities and social sciences community. Indeed, researchers in this area are interested in specific political issues, but they do not have access to tools/platforms allowing them to extract and structure their corpus before usage.

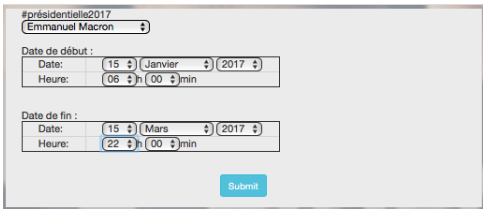

Figure 4: Corpus generator feature.

5. Conclusion and Perspectives

#Idéo2017 combines different technologies and inputs, which give to citizens the opportunity to grasp a part of the discursive issues of the election. This development, which can be enriched, allows to easily use a set of features usually accessible by software requiring different transformations of the data.

5.1. Creation of the #Idéo2017 Corpus

For the period from September 1st 2016 to May 7th 2017, 42290 tweets were extracted for the 11 candidates. These tweets were gathered in a collection that will be published in a TEI corpus in the standards of the Ortolang platform. The publication of this corpus under the requested standard is founded with the support of the CORLI consortium[6]. This process will follow the guidelines listed in the acquisition report (Longhi, 2014) written when we released the Polititweets corpus. Concerning the juridical issues, the creation of this corpus is legal. The position of Twitter is the following:

- "Please review the Twitter Rules (which are part of these Terms) to better understand what is prohibited on the Service. We reserve the right at all times (but will not have an obligation) to remove or refuse to distribute any Content on the Services, to suspend or terminate users, and to reclaim usernames without liability to you. We also reserve the right to access, read, preserve, and disclose any information as we reasonably believe is necessary to (i) satisfy any applicable law, regulation, legal process or governmental request, (ii) enforce the Terms, including investigation of potential violations hereof, (iii) detect, prevent, or otherwise address fraud, security or technical issues, (iv) respond to user support requests, or (v) protect the rights, property or safety of Twitter, its users and the public."

- "Except as permitted through the Services, these Terms, or the terms provided on dev.twitter.com, you have to use the Twitter API if you want to reproduce, modify, create derivative works, distribute, sell, transfer, publicly display, publicly perform, transmit, or otherwise use the Content or Services."

Thus, Twitter does not disclose personally identifying information to third parties except in accordance with their Privacy Policy. Moreover, Twitter encourages and allows broad re-use of content. The Twitter API exists to enable this.

5.2. #Idéo2017 as a Prototype: the Reproducibility of the Platform

#Idéo2017 allowed to the French electors to analyze the discourse of the candidates by means of their tweets. But, the utility of this platform is not limited to the French election, because it can be modified to different needs and usages, and adapted to other contexts. Thus, after the presidential election, we released two other versions of the platform: #législatives2017 (http://ideo2017.ensea.fr/legislatives2017/) and #quinquennat (http://ideo2017.ensea.fr/quinquennat/) which allow, in the first case to analyze the tweets of the main political parties during the election of deputies to the French National Assembly, and in the second case to daily analyze the beginning of Emmanuel Macron's presidential mandate via the tweets of the current political protagonists.
A set of new features were also proposed such as statistical analyses of the hashtags and mentions with Kibana. As a perspective, the #Idéo2017 prototype could be adapted to be used as a competitive intelligence, a measurement and a visualization tool analyzing people's opinion on Twitter. We can imagine dealing with political, social or cultural subjects through the integration of influential accounts, but also individual accounts of the users.

[3] www.iramuteq.org
[4] github.com/sixty-nine/PHP_Word_Cloud
[5] www.pchart.net
[6] https://corli.huma-num.fr/

6. Acknowledgements

UCP Foundation, Vo tv/Vo News, IUT of Cergy-Pontoise, University of Cergy-Pontoise, ETIS Lab, AGORA Lab.

7. References

Alduy, C. (2017). *Ce Qu'Ils Disent Vraiment. Decoder Le Discours Des Presidentiables*. SEUIL.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Flammini, A., and Menczer, F. (2011). Political polarization on twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Gayo-Avello, D., Metaxas, P. T., and Mustafaraj, E. (2011). Limits of electoral predictions using twitter. In Lada A. Adamic, et al., editors, *ICWSM*. The AAAI Press.

Johnson, K. and Goldwasser, D. (2016). "all I know about politics is what I read in twitter": Weakly supervised models for extracting politicians' stances from twitter. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2966–2977.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68.

Kononenko, O., Baysal, O., Holmes, R., and Godfrey, M. W. (2014). Mining modern repositories with elasticsearch. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR 2014, pages 328–331, New York, NY, USA. ACM.

Longhi, J. and Saigh, D. (2016). A textometrical analysis of french arts workers "fr.intermittents" on twitter. In *4th conference CMC and Social Media Corpora for the Humanities*, Ljubljana, Slovenia, September.

Longhi, J., Marinica, C., Borzic, B., and Alkhouli, A. (2014). Polititweets : corpus de tweets provenant de comptes politiques influents. In Chanier T., editor, *Banque de corpus CoMeRe*. Ortolang.fr.

Longhi, J. (2013). Essai de caractérisation du tweet politique. *L'information grammaticale*.

Longhi, J. (2014). Proposition pour l'acquisition d'un corpus de tweets (cmr-polititweets-teiv1-manuel.pdf). In J. Longhi, et al., editors, *Polititweets : corpus de tweets provenant de comptes politiques influents*. Banque de corpus CoMeRe. Ortolang.fr : Nancy. [cmr-polititweets-tei-v1].

Mercier, A. (2014). Twitter l'actualité : usages et réseautage chez les journalistes français. recherches en communication. *Recherches en communication 40*, pages 111–132.

Metaxas, P. T., Mustafaraj, E., and Gayo-Avello, D. (2011). How (not) to predict elections. In *SocialCom/PASSAT*, pages 165–171. IEEE.

Roginsky, S. and Cock, B. D. (2015). Faire campagne sur twitter. modalités d'énonciation et mises en récit des candidats à l'élection européenne. *Les Cahiers du numérique*, 11(4):119–144.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

Vidak, M. and Jackiewicz, A. (2016). Les outils multimodaux de twitter comme moyens d'expression des émotions et des prises de position. *Cahiers de praxématique*, (66).

Appendix

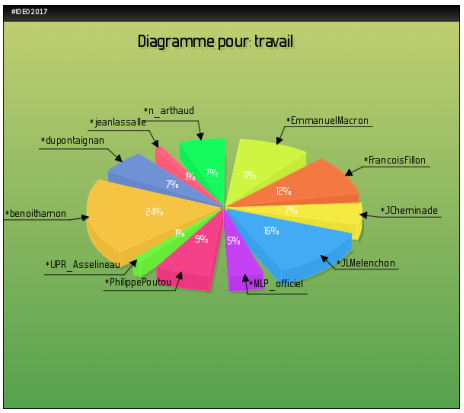

Figure 5: Analyses for the word *work*: frequency.

Figure 6: Analyses for the word *work*: associated words.

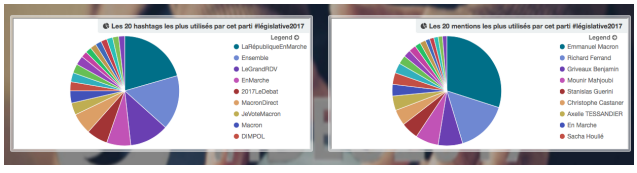

Figure 7: Data visualization.

Les mots les plus utilisés de @EmmanuelMacron

Forme	Fréquence	Type
france	223	Nom
français	184	Adjectif
projet	155	Nom
politique	153	Nom
aller	136	Verbe
pays	132	Nom
europe	125	Non reconnue
donner	89	Verbe
ensemble	85	Adverbe

Figure 8: Analysis of the tweets of the candidate Emmanuel Macron: the most frequent words.

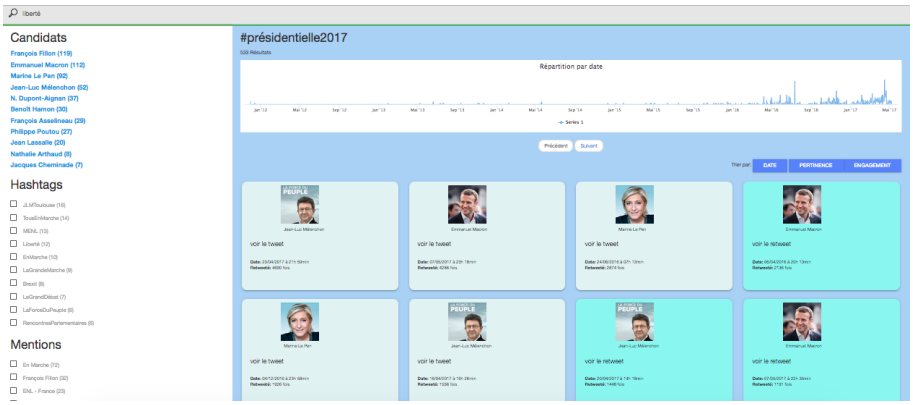

Figure 9: The user interface of the search engine.

Figure 10: Analyses for the word *work*: word cloud.

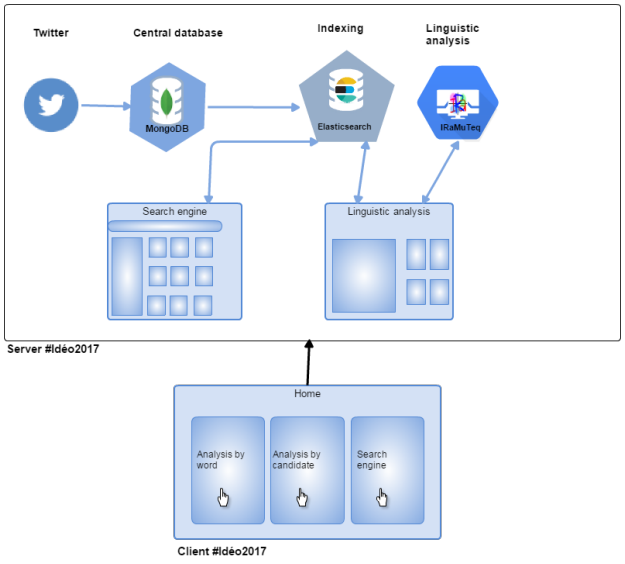

Figure 11: Schema of the tool.