A generic framework to perform comprehensive analysis of tweets

Marie-Noelle Bessagnet

A generic framework to perform comprehensive analysis of tweets

Marie-Noelle BESSAGNET[1]

[1] Laboratoire LIUPPA
IAE Pau-Bayonne
Université de Pau et des Pays de l'Adour
64012 PAU
marie-noelle.bessagnet@univ-pau.fr

Abstract. Recently, there has been an increased interest in the use of social media data as important traffic information sources and as a data research in the field of Human and Social Sciences. Social media are used, for example, to identify Twitter user communities in the context of altmetrics. In this paper, we highlight the potential use of social media data analysis by non computer science researchers. We present an approach based on a multi-dimensional analysis which combines the thematic, temporal and spatial features of tweets. We detail an experimentation using different tools in order to create the "perfectible" toolbox for researchers in HSS (for example territorial marketing) or for local territory managers. This approach can be applied in the context of altmetrics.

Keywords: *Geographical information, Data analysis, Territorial Policy, Almetric.*

1 Introduction

Twitter, a popular microblogging service, has received much attention recently. This online social network is used by millions of people around the world to remain socially connected to their friends, family members, and coworkers through their computers and mobile phones [1]. Twitter asks one question, "What's happening?" Answers must be fewer than 140 characters

A tweet is often used as a message to friends, colleagues and other people in order to inform about situations, to share opinions, feelings, etc… about many subjects. We think we can use the information embedded in a tweet in order to identify correlations between more or less structured data.

Counting articles and citations, analyzing citations and co-authors graphs have become ways to assess researchers and institutions performance. But, before using these ways to assess researchers, the research data on which they are working are very important. For researchers in the field of Human and Social Sciences, a set of tweets can be seen as a research data and requires different steps known as the research data lifecycle: collecting, recording, processing, analysing and then results can be published. Tweet analysis is used in many domains such as Real-Time Event Detection, political opinion analysis, improvement of social media strategy, marketing strategy, environment, almetrics, etc. We can quote the research work of [2] who

demonstrates that publications from the social sciences, humanities and the medical and life sciences show the highest presence of altmetrics, indicating their potential value and interest for these fields. In another domain, we can quote the research work of [3] who works on meta data linked to the tweets in order to analyse the behavior of people within a metropolitan urban area In many organizations the job of "community manager" is created in support of the organization's strategy. Results from tweets analysis can be done as statistics figures like *Key Performance Indicator (KPI)*[1].

We know as computer scientists that it is possible to extract, explore, synthesize and visualize knowledge based on a large mass of information available on Twitter. And sometimes, we are competent to do so. But the question is whether we can propose a framework and some tools to non computer scientist researchers to enable them to achieve these actions in a simply way. What kind of comprehensive analysis can we perform using a collection of tweets? To what end?

The goal of this paper is twofold. It first presents a framework devised from a geographical approach based on different theories and tools. Secondly it illustrates the use of dedicated tools for tweet analysis in order to identify correlations between thematic areas (what?), location of events inside the tweets or location of authors tweets (where?), time of the event (when?), and identification of sentiment.

The next section presents the general framework.

2 Defining a generic framework to perform comprehensive analysis

In this section we will present our generic framework and a toolbox we can imagine to process with this kind of research data.

2.1 A generic framework

The flowchart in Figure 1 depicts how tweets can be analyzed to assess the 5 W dimensions (who, when, what, where, why), three of which (When, Where and What) emphasize the geographical information. We can know who was tweeting, how, what about, what is the opinion expressed in the tweet, what kind of extraction we can do concerning a territory or a publication. As we said, for researchers in the field of Human and Social Sciences (HSS), a set of tweets can be seen as a research data. Thanks to this research data (meta data linked to the tweet or content of the tweet), researchers in the field of HSS can work on non structured textual document. For them, tweets are like others non structured textual documents: poems, epistolary exchanges between artists, ancient documents…

The main objective of the method is to analyze a collection of tweets from a multidimensional perspective. The general approach describes a process made up of different steps of analysis conducted to build a dashboard adapted to user profile: (1) preparation and validation of tweets, (2) first analysis based on user profiles and information embedded in the different fields of the tweets, (3) the multidimensional information analysis of the content of the tweet which permits the exploration of the collection of tweets and (4) summarization step

[1] KPIs evaluate the success of an organization or of a particular activity (such as projects, programs, products and other initiatives) in which it engages.

thanks to dashboard, map, timeline, KPI,....

After collecting the tweets we propose two types of analysis:

1. The first one is based on information mentioned in user profiles where we can analyze, in a basic way, the who, when and where dimensions,
2. The second one deals with the content of the tweet where more sophisticated analyses can be made of the 5 W dimensions.

Thanks to data in the user profiles, some statistics can be processed on the collection of tweets summarizing some of their features: number of tweets per day; users classified by location, by country; the relevant platforms, the most cited hashtags, the number of retweets,… This kind of analysis can be performed without cleaning the tweets, in a simple way. As we said, since the text is essentially informal, many challenges must be taken up in order to perform the different tweet analyses according the different dimensions.

2.2 The perfectible Toolbox

Based on our experiments we present a toolbox that although simple, can easily be extended with new and/or improved methods. The Twitter REST search API was used for collecting tweets using GET requests. We implemented a crawler to harvest tweets automatically, every day between January 2017 and July 2017, which could contain at least the following term: #Bearn.

First step of preprocessing

According to [4], the literature presents some proposals for dealing with such information, namely: a) Filtering: removal of URLs, Twitter user names (starting with @) and Twitter special words (“RT”, “via”, ...); b) Removal of stop words; c) Use of synonyms for the decomposed terms; d) Part of speech tagging usage (POS tagging); e) Recognition/Extraction of entities; f) Stemming: method for reducing a term to its radical, removing endings, affixes, and thematic vowels; and g) Treatment of the composite terms containing HashTags. The terms are normally separated according to the capitalization of letters. For example, “#VeryGood” becomes “Very Good” - a blank space is added between the words.

The classification and geocoding steps are responsible for classifying sentiment polarity and inferring geographical locations mentioned in the text, respectively.

IRAMUTEQ Environment

IRAMUTEQ is an environment dedicated to lexicometry. Thanks to this environment, a researcher can analyze much of the corpora. We can define lexicometry as the measurement of the frequency with which words occur in text. It provides statistical indicators and graphical representations that permit to investigate written text such as tweets. This kind of analysis completes content analysis [5].

ELIXA Environment

There are three levels of sentimental analysis –1) document level sentiment classification; 2) sentence level sentiment classification; and 3) aspect level sentiment analysis. This study has carried out a sentiment analysis by using EliXa which is an Aspect based sentiment analysis platform developed by the Elhuyar Foundation. Given that a sentence may contain multiple opinions, they define a window span around a given opinion target (5 words before and 5 words after). Both scores are calculated as the sum of every positive/negative score in the corresponding lexicon divided by the number of words in the sentence [6].

GATE Environment

We have developed a processing environment with the GATE platform ([7] ; [8]). This environment especially embeds the POS tagger Treetagger [9] and is also relevant for French language. Two steps are required: step 1 performs texts and ontology (GEONTO) to produce the same texts in which concepts and relationships are annotated. This process begins with the lemmatization of the texts (thanks to French Tokeniser and TreeTagger-FR-No-Tokenization modules) and continues with the annotation of terms suited to defined labels in the ontology (Flexible Gazetteer module). Each annotation includes the following details: the original term, the corresponding lemma, the identified label, name of the concept or the relation, object type (instance, class or relation). Step 2 performs rules (JAPE Transducer_00B7B module) derived from regular expressions combining domain concepts, codomain concepts and relations. It relies on annotations of step 1 and the ontology in order to validate the annotation of the semantic relations of the triplets (relation, domain, codomain).

Other processes allow us to annotate spatial entities and temporal ones. Details are given in [10].

2.3 In the context of altmetrics

This framework and the tools described are adapted to analyse tweets in the context of Almetrics. The two types of analysis we propose on the 5W dimensions are adapted to tweets linked to set of scientific papers. Instead of waiting months, we can rapidly know the sentiment expressed in a tweet linked to a scientific paper. For example, we can measure the e-reputation of researchers thanks to positive sentiment. A major limitation is done to altmetrics findings. We do not know their meaning in the process of scholarly communication but we think almetrics can change the way to recognize the scientific contributions of researchers.

.

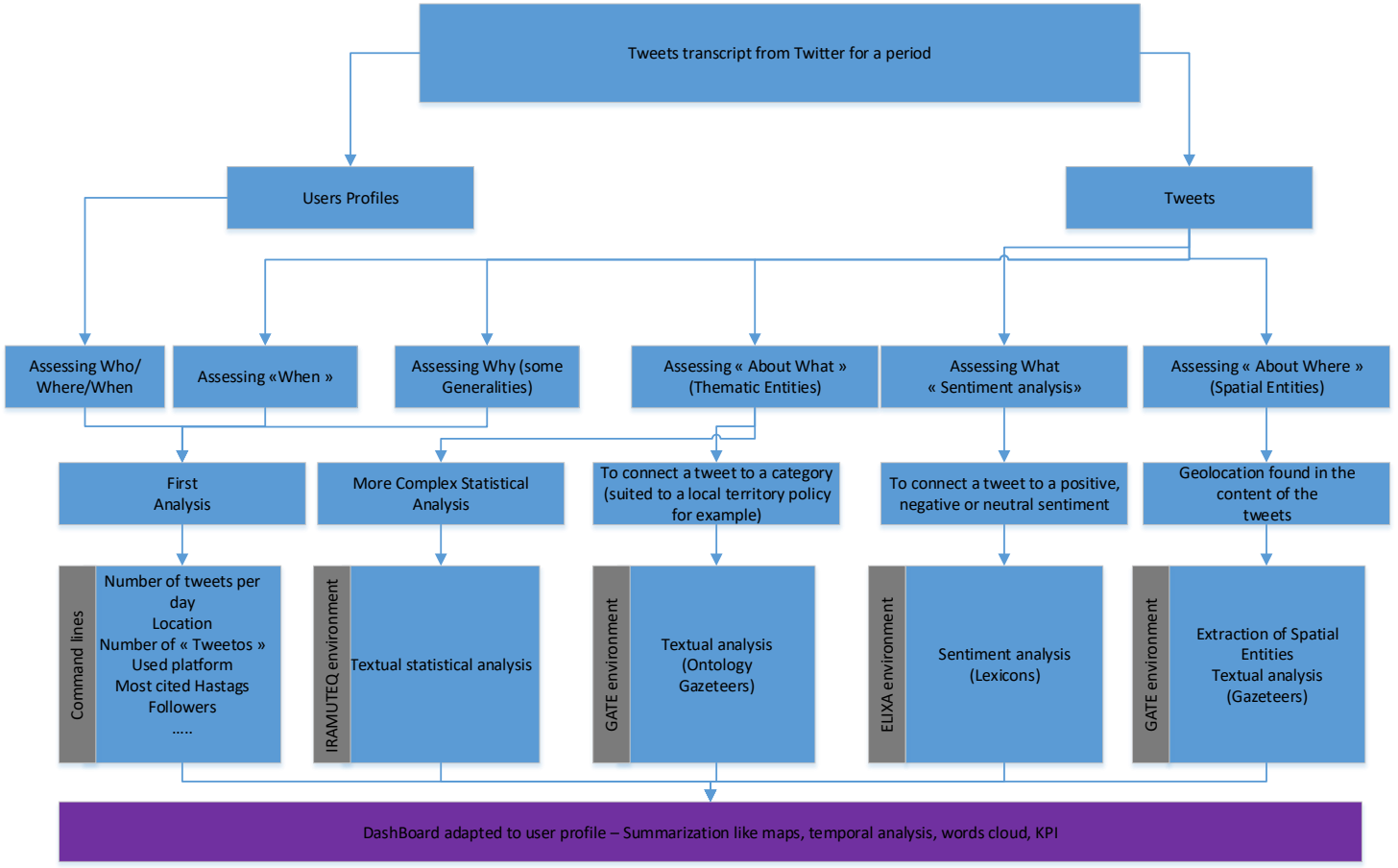

Fig. 1. A general framework (about tweets corpora analysis).

References

[1] Panteras, G., Wise, S., Lu, X., Croitoru, A., Crooks, A., & Stefanidis, A. (2015). Triangulating social multimedia content for event localization using Flickr and Twitter. *Transactions in GIS*, *19*(5), 694-715.

[2] Rodrigo Costas, Zohreh Zahedi, Paul Wouters, Do altmetrics correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective, https://arxiv.org/abs/1401.4321, last access 3/2/2018.

[3] Françoise Lucchini, Bernard Elissalde, Leny Grassot et Julien Baudry, « Paris tweets, données numériques géolocalisées et évènements urbains », *Netcom*, 30-3/4 | 2016, 207-230.

[4] Alves, André Luiz Firmino ; de Souza Baptista, Cláudio ; Firmino, Anderson Almeida ; de Oliveira, Maxwell Guimarães ; de Paiva, Anselmo Cardoso: A Spatial and Temporal Sentiment Analysis Approach Applied to Twitter Microtexts.. In: JIDM, 6 (2015), Nr. 2, S. 118-129

[5] Daniel Pélissier, "Comment préparer l'analyse de textes de sites Web grâce à la lexicométrie et au logiciel Iramuteq ?," dans *Présence numérique des organisations*, 14/04/2016, https://presnumorg.hypotheses.org/187

[6] I. San Vicente, X. Saralegi, y R. Agerri, «EliXa: A modular and flexible ABSA platform», in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, 2015, pp. 748-752.

[7] Cunningham H., Gaizauskas R., Wilks Y. (1995). A General Architecture for Text Engineering (GATE) – a new approach to Language Engineering R&D. Rapport technique no CS – 95 – 21. Department of Computer Science, University of Sheffield.

[8] Bontcheva K., Tablan V., Maynard D., Cunningham H. (2004). Evolving GATE to Meet New Challenges in Language Engineering. Natural Language Engineering, vol. 10, no 3/4, p. 349–373.

[9] Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings of international conference on new methods in language processing, p. 44-49.

[10] Buscaldi D., Bessagnet M.-N., Royer A., Sallaberry C. (2013). Using the semantics of texts for information retrieval: A concept- and domain relation-based approach. In B. Catania et al.(Eds.), Adbis (2), vol. 241, p. 257-266. Springer.